Educational Psychology Review (2023) 35:43
https://doi.org/10.1007/s10648-023-09761-1

INTERVENTION STUDY

How do Students Regulate Their Use of Multiple Choice Practice Tests?

Sabrina Badali[1] · Katherine A. Rawson[1] · John Dunlosky[1]

Accepted: 8 March 2023

Abstract

Multiple-choice practice tests are beneficial for learning, and students encounter multiple-choice questions regularly. How do students regulate their use of multiple-choice practice testing? And, how effective is students' use of multiple-choice practice testing? In the current experiments, undergraduate participants practiced German-English word pairs. Students started with an initial study trial for each pair. Then, they had the options to restudy an item, take a practice test, or remove it from further practice. For comparison to students' use of multiple-choice practice questions, we included a second self-regulated group that had access to cued-recall practice questions. Participants chose to complete multiple-choice questions until they correctly answered each item about one time during practice, similar to students' use of cued-recall questions. We also included experimenter-controlled groups in which participants completed practice tests until they reached a higher number of correct answers during practice. As compared to the experimenter-controlled groups, participants who regulated their use of multiple-choice questions scored lower on final tests but also spent less time practicing items. Thus, when considering final test performance in relation to time spent practicing, students' choices to use multiple-choice practice questions to about one correct answer per item was comparatively effective.

Keywords Self-regulated learning · Practice testing · Multiple-choice · Criterion learning

Students encounter tests regularly in their academic lives, and these tests can be used for a variety of purposes. Tests are often used as summative assessments to evaluate student learning, but they might also be available as a learning resource to

✉ Sabrina Badali
sbadali@kent.edu

[1] Department of Psychological Sciences, Kent State University, P.O. Box 5190, Kent, OH 44242-0001, USA

Published online: 30 March 2023

help students prepare for upcoming assessments. When practice tests are available to students, they are frequently offered as an optional resource in classes, meaning that use of these questions is self-regulated. Using these optional practice tests would be in students' best interest, as retrieval practice is one of the top strategies available for improving long-term retention of material (e.g., Bjork et al., 2013; Dunlosky et al., 2013; Roediger & Butler, 2011). However, few studies have examined students' use of self-testing, particularly when it comes to students' use of different formats of practice questions. Multiple-choice questions are present in students' everyday academic lives, yet only one experiment has investigated how students choose to use multiple-choice practice questions (Pollock & Sullivan, 1990). Accordingly, the main questions of the present research are: How do students regulate their use of multiple-choice practice testing? And, how effective is students' use of multiple-choice practice testing for achieving long-term retention of material?

How will Students use Multiple-Choice Practice Testing?

To best understand predictions for how students will use multiple-choice practice tests, we first need to describe a few key aspects of the current method. In particular, undergraduate participants practiced foreign language word-pair translations. After an initial study trial, participants in self-regulated learning (SRL) groups made item-by-item choices for what they wanted to do next with each word pair. They could choose to restudy a word pair, take a practice test, or remove it from further practice. Students could continue restudying items and/or taking practice tests until they chose to remove all items from practice. Of primary interest, one SRL group received multiple-choice practice questions when they chose to test an item. The primary outcome for assessing use of practice testing was the number of times participants correctly answered a practice question about an item before choosing to remove that item from learning. This number of correct answers per item is referred to as the criterion reached during practice (e.g., if students correctly answered each item two times during practice, they reached a criterion of two). What criterion will students reach when using multiple-choice practice testing? Specifically, will they stop practicing (a) after they have correctly responded to a question a single time or (b) after they have correctly responded more than once? The answer to this question is open, and prior theory and research suggests that either outcome could plausibly occur.

Pertaining to theory, the *test-to-monitor* hypothesis states that students use practice testing to monitor their learning instead of using testing to enhance their learning. The idea is that students restudy an item until they judge they can answer a question correctly, and then they test themselves to evaluate whether they learned the item well enough to answer it correctly. In this case, students interpret a successful practice test attempt as an indicator that they have sufficiently learned the item and/or that they cannot improve their retention of the item with further practice during that session. Accordingly, students are expected to stop practicing an item after about one correct answer. Although this hypothesis has not been evaluated using multiple-choice practice tests, a few investigations involving students' use

of cued-recall practice tests provided initial support. For example, in Dunlosky and Rawson (2015), undergraduate participants learned key concept definitions and had the options to take a cued-recall practice test for an item, restudy an item, or remove that item from further practice. Consistent with predictions from the test-to-monitor hypothesis, participants chose several practice tests per item, but typically stopped engaging with items after about one correct response. These results are representative of findings from other self-testing research in which participants were given a similar amount of control over their learning (Ariel & Karpicke, 2018; Badali et al., 2022; Janes et al., 2018; Karpicke, 2009; Kornell & Bjork, 2008; for a review of the accumulating evidence consistent with the test-to-monitor hypothesis, see Rivers, 2021). In summary, the test-to-monitor hypothesis motivates the expectation that students will reach the same criterion (one correct response per item) when they are using multiple-choice practice tests or cued-recall practice tests. To evaluate this possibility and to replicate prior research, we also included a self-regulated learning group that used cued-recall tests during practice.

An alternative answer to our key question is motivated by the only prior investigation of students' use of multiple-choice practice questions. Pollock and Sullivan (1990) had seventh-grade students complete a science lesson with four sets of optional practice questions spaced throughout the lesson. When students reached the end of a section in the lesson, they encountered a set of practice questions and could choose to either complete these questions or bypass them. The format of these practice questions varied between participants; half of the students received multiple-choice questions in which they were asked to select the correct answer from four alternatives, and half received short-answer recall questions in which they were asked to type the correct answer. Regardless of practice question format, if a student's answer was incorrect they were told their answer was wrong and the correct answer was displayed. Students who received multiple-choice questions chose to complete more of the practice question sets than students who received short-answer questions. By extension, if students in the present experiments are willing to engage in more practice testing when using multiple-choice tests, students who have access to multiple-choice practice questions may reach a higher criterion than students who have access to cued-recall practice questions.

Other classroom studies also suggest that students may use optional multiple-choice practice questions to a criterion that is beyond one correct answer. For example, McDaniel et al. (2012) conducted a classroom experiment to investigate whether different formats of online quizzes would enhance performance on course assessments. Student participants were encouraged to take the quizzes multiple times and received course credit for up to four attempts. They found that students completed both multiple-choice and cued-recall quizzes several times (Experiment 1 multiple-choice $M = 3.2$ attempts, cued-recall $M = 3.9$ attempts). Performance on both multiple-choice and cued-recall quizzes was high even by the second quiz attempt. This suggests that students were reaching an overall criterion of more than one correct answer per item and were reaching this criterion for both multiple choice and cued recall formats. Additionally, Riggs et al. (2020) found that when students had access to student-authored multiple-choice practice questions, they often used them beyond what was required for course credit. Specifically, students used the multiple-choice

questions to earn their course points, but typically completed additional practice questions shortly before unit exams. However, performance on practice questions was not reported, so the criterion that students may have reached is unknown.

The two possible patterns discussed above concern the self-testing outcome of primary interest—the criterion students reach when using multiple-choice questions. Importantly, the test-to-monitor hypothesis also supports expectations for secondary outcomes in the self-regulated groups. As noted above, this hypothesis states that students will use practice tests to evaluate whether they know an item and motivates the prediction that students will drop items from further practice after about one correct response. According to this hypothesis, students will wait to test themselves until they judge that they will correctly answer the type of practice test they are completing (multiple-choice or cued-recall). Students believe multiple-choice questions are easier than other formats of questions (Parmenter, 2009; Scouller, 1998; Zeidner, 1987), so a prediction is that participants will start testing earlier (i.e., engage in less restudying before attempting a practice test) when they are using multiple-choice versus cued-recall practice questions. Additionally, on trials in which a student fails to answer the question correctly, they may choose to test again later to re-assess their learning. Incorrect responses are typically less likely with multiple-choice than with cued-recall questions, meaning that students using multiple-choice questions may require fewer test trials to reach their first correct response.

Effective Regulation for Multiple-Choice Practice Testing

In addition to examining *how* students regulate their use of multiple-choice practice testing, we also aimed to assess how *effective* students' use of multiple-choice questions is for achieving long-term retention of material. We view effectiveness as a balance between durability (which we measure using final test performance) and time costs associated with practice (which we measure using total number of practice trials and practice time per item). As described above, predictions for students' use of multiple-choice practice testing are that they either will practice items to a criterion of about one correct response per item or will reach a higher criterion when using multiple choice questions versus cued recall questions. But would either of these outcomes be an effective way to use multiple-choice practice testing? The answer is unknown because no research has examined the effects of criterion learning on durability of learning or time costs associated with practice when multiple-choice practice testing is used. Thus, we turn again to results from experiments that have used cued-recall practice questions to inform predictions about multiple-choice practice questions.

One way to reach high levels of retention on a memory test is to use cued-recall practice testing until items are correctly answered multiple times versus only one time (Karpicke & Smith, 2012; Pyc & Rawson, 2009); higher criterion levels are typically associated with higher final retention (Vaughn & Rawson, 2011; Vaughn et al., 2013). Reaching a higher criterion during practice may be particularly important when using multiple-choice practice questions, given that performance on final memory tests is usually lower when participants use multiple-choice practice tests,

compared to cued-recall practice tests (e.g., Clariana, 2003; Butler & Roediger, 2007; Kang et al., 2007, but see Bjork et al., 2015 and McDaniel & Little, 2019 for an explanation of exceptions).

Performance on final tests is a common way to measure effectiveness, as retention provides information about the durability of learning. However, we view durability as only one of two dimensions of effectiveness. To make conclusions about the overall effectiveness of students' use of multiple-choice practice questions, we will consider the durability of learning in relation to the time costs associated with students' choices. To help assess the effectiveness of participants' self-regulated choices, we included two experimenter-controlled criterion learning groups. Participants in these groups continued practicing items until they reached a predetermined criterion level (1, 3, or 5 correct responses). If one pattern of learning choices results in higher retention but also imposes a significantly greater time cost, it may not represent effective learning at an overall level. Alternatively, if two patterns result in the same retention but one takes less time, then the more efficient option would be most effective.

Overview of Current Experiments

The aim of the current set of experiments was to investigate students' use of multiple-choice practice testing and assess the effectiveness of these self-regulated choices. After an initial study trial, participants in the self-regulated learning (SRL) groups made item-by-item choices for what they wanted to do next with the word pair. Of primary interest, one SRL group received multiple-choice practice questions when they chose to test an item. We also included a second SRL group that received cued-recall practice questions, to allow for a comparison of multiple choice versus cued recall self-testing choices. All participants completed final multiple-choice and cued recall tests, two days later in Experiment 1 and seven days later in Experiment 2. Because retention interval was the only difference between the two experiments and did not have a substantive impact on final test performance, we adopted an internal meta-analytic approach and report Experiments 1 and 2 together.

In summary, the main questions of the present research are: How do students regulate their use of multiple-choice practice testing? And, how effective is students' use of multiple-choice practice testing for achieving long-term retention of material? We conducted two experiments to answer these questions by comparing self-testing choices, final test performance, and time costs associated with practice among the SRL groups and experimenter-controlled criterion learning groups.

Experiments 1 and 2

Methods

Design and participants Both experiments used a 2 (practice group: SRL vs. criterion) × 2 (practice test format: multiple-choice vs. cued-recall) between-participant design. Participants were randomly assigned to one of four groups: SRL multiple-choice (SRL-MC), SRL cued-recall (SRL-CR), criterion multiple-choice (criterion-MC), or criterion cued-recall (criterion-CR). In the criterion groups, items were practiced until they were correctly answered either 1, 3, or 5 times, manipulated within-participant. Finally, all participants took both a multiple-choice final test and a cued-recall final test.

For Experiment 1, an a priori power analysis was conducted using G*Power 3.1.9.6 (Faul et al., 2009) for an independent samples *t*-test to detect a medium effect ($d \geq 0.50$) for differences in the mean number of correct responses during practice for the SRL-MC and SRL-CR groups. We powered for a medium-sized effect because of the lack of prior research on which to base a more precise estimate and because a medium effect is often considered the smallest effect of practical interest. With power = 0.80 and two-tailed $\alpha = 0.05$, this analysis suggested 64 participants per group (total sample size of 256). We oversampled to allow for attrition. A total of 311 undergraduates participated in exchange for partial course credit (81% female; 80% white, 12% black, 6% Asian, 5% Hispanic or Latino, 4% First Nations; 38% were in their first year of college ($M = 2.1$, $SD = 1.2$); Age ($M = 19.9$, $SD = 3.8$); 34% were psychology majors).

For Experiment 2, an a priori power analysis was conducted using G*Power 3.1.9.6 (Faul et al., 2009) to detect a main effect of learning group for performance on the delayed cued-recall test. An analysis with power = 0.80 and an effect size $f = 0.16$ (based on the effect size from this analysis in Experiment 1) suggested a target sample size of 309 participants. We oversampled to allow for attrition. A total of 484 undergraduates participated in exchange for partial course credit (81% female; 83% white, 13% black, 6% Asian, 4% Hispanic or Latino, 4% First Nations, < 1% Pacific Islander; 41% were in their first year of college ($M = 2.2$, $SD = 1.2$); Age ($M = 19.9$, $SD = 2.6$); 30% were psychology majors).

Materials Materials included 48 German-English word-pair translations split into two blocks. Item difficulty data were available for 24 of these items and were used to distribute the items into two comparable blocks.[1] We then developed 24 additional

[1] In brief, these 24 items were used in a previous experiment involving cued recall practice questions after an initial study trial. We examined performance on the first cued recall practice trial in that experiment as an indicator of normative item difficulty. Item difficulty was similar for the 12 items assigned to the first block and the 12 items assigned to the second block in the current study (Block 1: $M = 28\%$, $SD = 15$; Block 2: $M = 26\%$, $SD = 11$).

word pairs, with 12 of these randomly assigned to each block. Within each block, the items were divided into sets of four items (including two word pairs that had item difficulty data and two new pairs). These sets were used to construct the answer choices for multiple-choice questions. Each question included four English words as the answer choices (the correct answer, plus the other three English words from that item's set). In both criterion groups, each block contained two sets (eight items) assigned to each of the three criterion levels (1, 3, and 5).

Procedure All instructions and tasks were administered via computer. In Experiment 1, some participants completed the experiment in an in-person lab setting (September 2019 – March 2020), and some completed it online remotely due to the COVID-19 pandemic (September 2020 – October 2020). The only difference in the procedure of in-lab participants versus online participants was that in-lab participants learned the items in a randomized order and online participants were randomly assigned to learn the items in one of three predetermined randomized orders. In Experiment 2 (January 2021 – September 2021), all participants completed the experiment online remotely. Participants in all groups started by completing a cued-recall pretest for all 48 German-English word pairs to assess prior knowledge. After completing this pretest, all participants were instructed that they will be learning 48 German-English word pair translations and the format of the final test was described. Specifically, they saw the instructions, "Your goal is to learn all 48 of these translations for a memory test two days from now. On this final test, for some items you will see the German word and be asked to type in the English translation. For other items, you will see the German word and be asked to select the English translation from a list of possible answers." Participants were also given instructions tailored to their randomly assigned group.

Self-Regulated Learning Groups In the two SRL groups, participants were told they would start with an initial study trial for each item and that after the initial study trial they would choose what they wanted to do next with that item – study, test, or drop. Participants were instructed that choosing "study it again later" meant the item would be placed at the end of the item list for them to restudy once the remaining items had been presented. Participants were instructed that choosing "take a practice test" meant the item would be placed at the end of the item list for them to take a practice test once the remaining items had been presented; participants were also informed that they would see the correct answer immediately after the practice test trial. Participants in the SRL-MC group were told that on a practice test trial, they would see the German word and be asked to select the correct English translation from a list of possible answers. Participants in the SRL-CR group were told that on a practice test trial, they would see the German word and be asked to type in the English translation. In both groups, participants were instructed that choosing "drop it from the list" meant the item would be removed from the list and not presented again for any additional restudy trials or practice tests. Finally, participants in both groups were told they could choose to study or test items as many times as they wanted before dropping them and would make a new choice for an item each time it was presented.

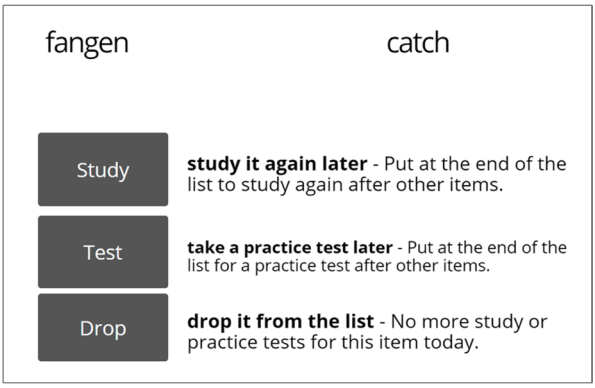

Fig. 1 Screenshot from the self-regulated learning groups, showing participants' three options for each item

Participants then started the learning task. Items were presented one at a time in either a random order (in-lab participants) or one of three different orders (online participants, randomly assigned to one of the item orders). The first time an item was presented, participants started by making an ease-of-learning (EOL) judgment. The German and English words were visible simultaneously. In Experiment 1, participants were asked, "How easy do you think it will be to learn this item well enough so that on the memory test two days from now, you'll be able to recall the English translation when shown the German word?" In Experiment 2, the prompt was the same except it referred to "the memory test one week from now." Participants responded using a slider scale ranging from 0 to 100 with the left side labeled "very difficult," and the right side labeled "very easy." Immediately after making this EOL, the item was presented for a self-paced study trial. After participants clicked a button to indicate they were done with the study trial, they saw a screen with buttons pertaining to the three choices for the item and descriptions of the options (see Fig. 1). When participants chose to study, the item was placed at the end of the item list for a self-paced restudy trial the next time it was presented. When participants chose to test, the item was placed at the end of the list for a self-paced practice test trial. On test trials in the SRL-MC group, the German word was presented as a cue along with the four English words from that item's set. On test trials in the SRL-CR group, the German word was presented as a cue along with a prompt to type in the English translation. For both types of practice tests, after participants were done trying to answer, the correct English translation was displayed for self-paced restudy. When participants chose to drop an item, the item was removed from the learning list and was not presented again. Immediately after choosing to drop an item, participants made a judgment of learning (JOL) for the item. In Experiment 1, participants were asked, "On the memory test two days from now, how likely is it that you will be able to recall the English translation when shown the German word?" In Experiment 2, the prompt was the same except it referred to "the memory test one week

from now." They answered using a slider scale ranging from 0 to 100 with the left side labeled "0% likely," and the right side labeled "100% likely."

Participants in the SRL groups continued engaging in restudy or test trials and continued making new choices each time an item was presented until they had chosen to drop all items in the first block. The number of restudy or test choices a participant could make was not limited. When all items were dropped, participants completed the same procedure for the second block of 24 items.

Criterion Groups In the two criterion groups, participants were told that they would start with an initial study trial for each item and then would complete practice tests for the items. Participants in the criterion-MC group were told that on each trial, they would see the German word and would be asked to select the correct English translation from a list of possible answers. Participants in the criterion-CR group were told that on each trial, they would see the German word and would be asked to type in the English translation. They were told that each item would continue to be presented for practice tests until they had correctly answered the item a predetermined number of times (1, 3, or 5), at which point it would be removed from the list and not presented again.

Participants in the criterion groups started with an EOL judgment and initial study trial for each item, involving the same procedure as in the SRL groups. Items were presented one at a time in either a random order (in-lab participants) or one of three orders (online participants). After EOLs and initial study trials were complete for all 24 items in the first block, participants started practice test trials for the items. Participants in the criterion-MC group saw the German word, four English words as answer choices, and after answering the question they were shown the correct answer for self-paced restudy. Participants in the criterion-CR group saw the German word, were asked to type the English translation, and after answering the question they were shown the correct answer for self-paced restudy. Regardless of the type of practice test, if an item was answered incorrectly, it was placed at the end of the item list and presented again after the remaining items had been tested. If an item was answered correctly but had not yet reached its assigned criterion level, the item was placed at the end of the list to be tested again. If an item was answered correctly and had reached its predetermined criterion level, the item was removed from the list and was not presented again. After the last test trial for an item, participants made a JOL using the same prompt and slider scale as the SRL groups. Participants continued cycling through the items until all 24 items in that block reached their criterion level and had been removed from the list. Participants then completed the same procedure for the second block of 24 items.

Session 2 Memory Tests In all groups, Session 1 ended after participants finished learning the second block of items. In Experiment 1, participants returned two days later to complete final memory tests, and in Experiment 2, participants returned seven days later to complete final memory tests. All participants took a cued-recall

test for a randomized half of the items and a multiple-choice test for the remaining half of the items. Test order was counterbalanced across participants, such that in each group half of the participants started with the cued-recall test and half started with the multiple-choice test. After completing these first two tests, participants then completed a multiple-choice test over the 24 items that had been assigned to their cued-recall test. Thus, all 48 items had a multiple-choice test trial, but half of these were completed after taking a cued-recall test for those items. For all tests, items were presented one at a time, in a random order, and trials were self-paced. After completing all memory tests, participants were asked questions regarding their thoughts about the effectiveness of multiple choice questions versus cued recall questions. In Experiment 2, participants were then asked to indicate whether they used any external memory aids during either session. The hypotheses, procedure, and data analysis plan for Experiment 2 were preregistered on the Open Science Framework (https://osf.io/vugkw).

Results and Discussion

Data exclusions and preliminary analyses Analyses excluded data for 58 participants who scored higher than 50% on the German-English pretest (Experiment 1 $n = 14$, Experiment 2 $n = 44$) and 22 participants who completed less than 25% of the learning task in Session 1 (Experiment 1 $n = 10$, Experiment 2 $n = 12$). Thirteen participants completed less than 50% of the second block of items (Experiment 1 $n = 3$, Experiment 2 $n = 10$). For these 13 participants, analyses excluded data for items from the second block. Session 2 analyses excluded data for one participant who had a mean response time of less than one second per trial for all final tests. Due to a computer error, 10 participants from Experiment 2 had multiple-choice final test data for only 24 of the 48 items. Finally, 108 participants did not return for Session 2 (Experiment 1 $n = 18$, Experiment 2 $n = 90$), but these 108 participants were included in Session 1 analyses. The final sample for Experiment 1 included 287 participants (SRL-MC $n = 68$, SRL-CR $n = 68$, criterion-MC $n = 76$, criterion-CR $n = 75$), and the final sample for Experiment 2 included 424 participants (SRL-MC $n = 104$, SRL-CR $n = 109$, criterion-MC $n = 108$, criterion-CR $n = 103$).

To check whether Experiment 1 outcomes could be attributed to our shift to online data collection, ANOVAs were conducted with location (in-person vs. online) as an additional independent variable. The results indicated that our main conclusions did not differ across location (all interaction $ps > 0.25$, except for one secondary outcome[2]). Accordingly, we collapsed across in-person and online participants for the analyses presented below. Although conclusions did not differ by location, some numerical differences motivated the inclusion of an additional question in

[2] In-person participants completed fewer restudy trials before their first test trial in the SRL-MC group ($M = .5$) compared to the SRL-CR group ($M = 1.4$) but online participants completed similar amounts of restudy trials before first test trial in both groups (SRL-MC $M = .6$, SRL-CR $M = .6$), $F = 7.38$, $p < .001$.

Experiment 2, in which participants were asked to report whether they used any external memory aids. Final test performance in Experiment 2 was higher than in Experiment 1, likely because with a fully online sample some participants used external memory aids. We checked to see if any conclusions changed based on self-reported use of external memory aid. First, ANOVAs were conducted with answers to this question (yes or no) as an additional independent variable. No interactions were significant ($ps > 0.33$), and thus conclusions did not differ based on whether a participant indicated they used an external memory aid. Additionally, we re-analyzed data from the Experiment 2 sample excluding participants who answered "yes" to this question, and conclusions based on statistical significance remained the same. Accordingly, we have used an intent-to-treat approach and analyzed data from the full Experiment 2 sample in the analyses below. Moreover, in Appendix 1, interested readers can find descriptive statistics separated by answer to external memory aid question, inferential statistics excluding participants who answered "yes" to this question, and more information about responses to this question.

Finally, two multiple-choice tests were used in case taking a cued-recall test before a multiple-choice test would impact scores. However, no interactions between group and score on these two multiple-choice tests were significant (interaction $ps > 0.18$, $\eta_p^2 < 0.02$), meaning that the qualitative pattern remained the same across both multiple-choice tests. Accordingly, we collapsed scores on these two multiple-choice tests to include more items and allow for more precise estimates; we now just refer to the multiple-choice final test.

To revisit, the primary research questions were: How do students regulate their use of multiple-choice practice testing? And, how effective is students' use of multiple-choice practice testing for achieving long-term retention of material? The design and methods of Experiments 1 and 2 yielded a rich data set with many outcome variables and possible comparisons. In the discussion below, we focus on the comparisons most relevant to our primary research questions. Specifically, comparisons between the SRL-MC and SRL-CR groups inform answers to both research questions, and comparisons between the SRL-MC group and the higher criterion conditions in the criterion-MC group inform answers about effectiveness. Other comparisons can be found in Appendix 2. Given the highly similar methods of the two experiments, we followed current recommendations for multi-study papers to focus on meta-analytic findings and pooled estimates instead of each experiment's individual significance tests (Braver et al., 2014).

How do Students Regulate Their Use of Multiple-Choice Practice Testing? The primary outcome of interest was the criterion that students achieved when self-regulating their use of multiple-choice practice questions. To revisit, two plausible predictions regarding criterion reached were outlined in the introduction. First, students may use multiple-choice practice tests similarly to how they use cued-recall practice tests and reach a criterion of about one correct response per item. Second, students may be willing to engage in more practice testing when using multiple-choice tests, meaning that students in the SRL-MC group would reach a higher criterion than students in the SRL-CR group.

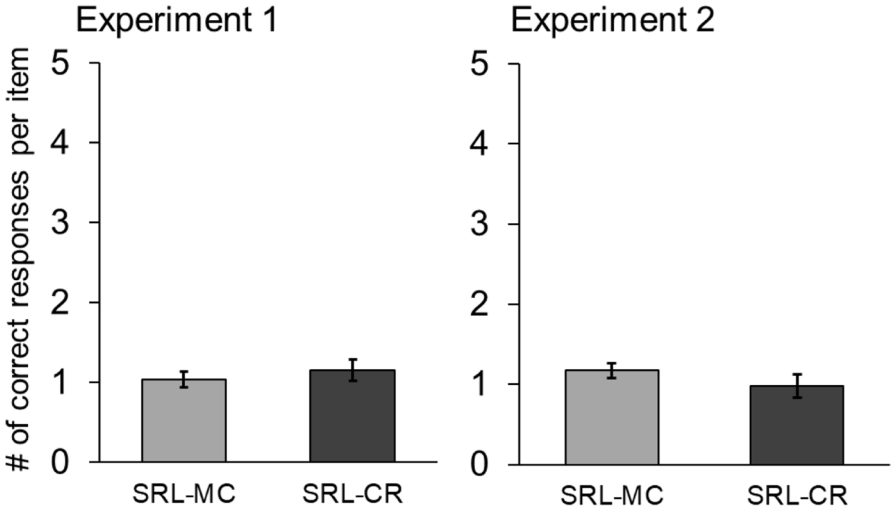

Fig. 2 Mean number of correct responses per item (i.e., criterion reached) for participants in the self-regulated learning group that had access to multiple-choice practice questions (SRL-MC) and the self-regulated learning group that had access to cued-recall practice questions (SRL-CR). Error bars report standard error of the mean

To evaluate these predictions, we compared students' use of multiple-choice practice questions to students' use of cued-recall practice questions in the two SRL groups. In both SRL groups, participants stopped testing items after about one correct response per item (see Fig. 2). Participants in the SRL-MC group reached a mean criterion of just over one correct response per item (pooled $M = 1.1$), which was similar to the criterion reached by participants in the SRL-CR group (pooled $M = 1.0$), pooled $d = 0.05$, 95% CI = -0.16, 0.26.

When examining criterion reached, students dropped multiple-choice and cued-recall questions after about one correct response on average. For items that reached at least one correct answer, participants' choice immediately after that first correct answer was to drop the item 70% of the time in the SRL-MC group and 74% of the time in the SRL-CR group. However, mean criterion achieved across items may hide some differences in number of correct responses. Examining criterion at an item level revealed that participants reached a criterion either higher or lower than one correct answer for some items and revealed some differences between the two SRL groups. Specifically, participants in the SRL-MC group reached a criterion of one correct answer for more items than participants in the SRL-CR group [(SRL-MC pooled $M = 21$ items, SRL-CR pooled $M = 17$ items), pooled $d = 0.24$, 95% CI = 0.03, 0.45]. Additionally, participants in the SRL-MC group reached a criterion of more than one correct answer for more items than participants in the SRL-CR group [(SRL-MC pooled $M = 12$ items, SRL-CR pooled $M = 9$ items), pooled $d = 0.22$, 95% CI = 0.01, 0.43], and participants in the SRL-MC group reached a criterion of zero correct answers for fewer items than participants in the SRL-CR group [(SRL-MC pooled $M = 14$ items, SRL-CR pooled $M = 21$ items), pooled $d = 0.40$,

Table 1 Descriptive statistics for additional Session 1 outcomes in the two SRL groups

	SRL-MC			SRL-CR		
	M	*SE*	Med	*M*	*SE*	Med
Experiment 1						
Ease of learning judgment magnitude	44	1		43	2	
Number of test trials per item	1.4	0.2	1.3	1.6	0.2	1.8
Number of restudy trials per item before first test choice for that item	0.4	0.1	0.2	0.8	0.1	0.5
Number of restudy trials per item	0.5	0.1	0.3	1.1	0.2	0.7
Judgment of learning magnitude	54	2		50	2	
Experiment 2						
Ease of learning judgment magnitude	44	1		44	2	
Number of test trials per item	1.5	0.1	1.4	1.6	0.2	1.0
Number of restudy trials per item before first test choice for that item	0.4	0.1	0.3	0.5	0.1	0.3
Number of restudy trials per item	0.5	0.1	0.2	0.8	0.1	0.3
Judgment of learning magnitude	56	2		51	2	

Ease of learning judgment magnitude could range from 0 (very difficult) to 100 (very easy). Judgment of learning magnitude could range from 0 (0% likely to be able to answer correctly) to 100 (100% likely to be able to answer correctly). *M* = mean. *SE* = standard error of the mean. Med. = median, which are reported for Session 1 learning choices that were not normally distributed

95% CI = 0.19, 0.61]. Thus, participants reached a mean criterion of one correct answer when averaged across all items (which is consistent with the approach taken in past research, see Dunlosky & Rawson, 2015; Janes et al., 2018). However, item-level analyses showed that students' choices are not the same for all items, and for at least some items, participants persist even after reaching a criterion of one correct answer. Future research will be needed to reveal why participants persist more for some items than others.

Did other aspects of students' testing choices differ? Secondary ways to measure use of practice testing were the mean number of restudy trials before the first test trial, which is an indicator of how early participants started taking practice tests, and the mean number of test trials per item. Table 1 reports descriptive statistics for these variables. Participants in the SRL-MC group chose fewer restudy trials before their first practice test trial (pooled $M = 0.4$) compared to participants in the SRL-CR group (pooled $M = 0.7$), pooled $d = 0.29$, 95% CI = 0.08, 0.50. Additionally, participants in the SRL-MC group completed fewer test trials per item (pooled $M = 1.5$) compared to participants in the SRL-CR group (pooled $M = 1.8$), pooled $d = 0.22$, 95% CI = 0.01, 0.43. Thus, participants who used multiple-choice questions started taking practice tests earlier in their learning and completed fewer test trials, compared to participants who used cued-recall questions.

How Effective was Students' Use of Multiple-Choice Practice Testing? Participants in the SRL-MC group completed practice test trials until they correctly answered an item about one time. But were these choices effective? We view effectiveness as the

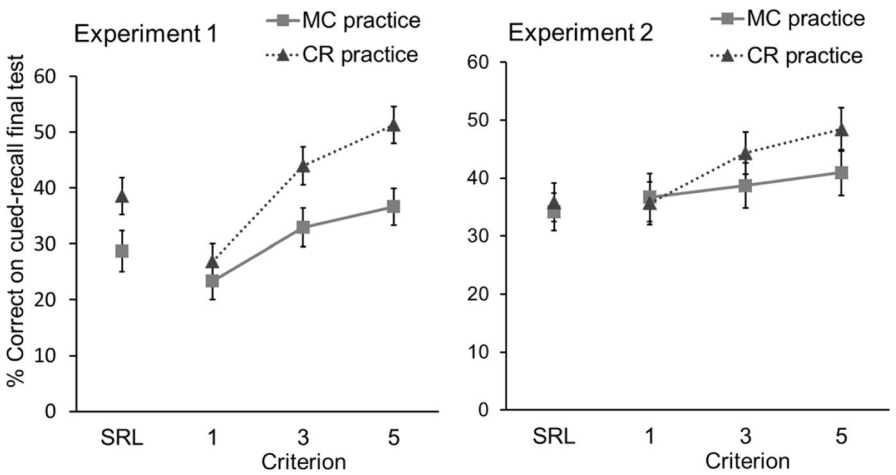

Fig. 3 Mean performance on the cued-recall final test for the self-regulated learning (SRL) groups and for each of the three criterion levels in the criterion groups, separated by Experiment 1 (two-day delay) and Experiment 2 (seven-day delay). MC practice and CR practice refer to the type of practice test questions that a participant completed in Session 1. Error bars report standard error of the mean

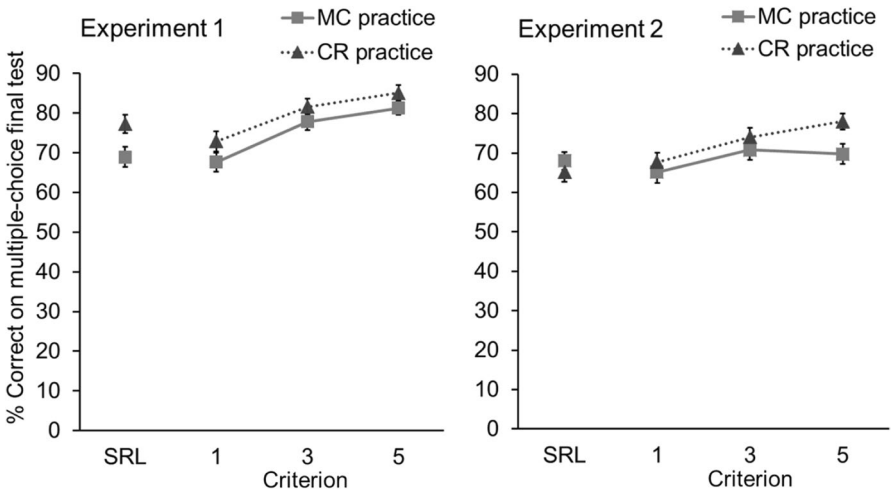

Fig. 4 Mean performance on the multiple-choice final test for the self-regulated learning (SRL) groups and for each of the three criterion levels in the criterion groups, separated by Experiment 1 (two-day delay) and Experiment 2 (seven-day delay). MC practice and CR practice refer to the type of practice test questions that a participant completed in Session 1. Error bars report standard error of the mean

combination of durability (which we measure using final test performance) and time costs associated with practice (which we measure using total number of practice trials and practice time per item). If one pattern of learning choices leads to higher final test performance and does not require a substantial increase in practice time, then it would be a more effective option. Alternatively, if two patterns of learning

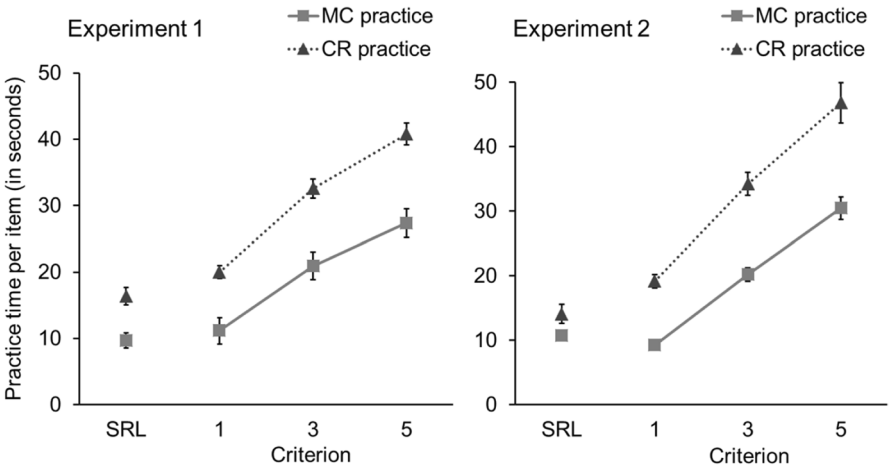

Fig. 5 Mean Session 1 practice time per item (in seconds) for the self-regulated learning (SRL) groups and for each of the three criterion levels in the criterion groups. MC practice and CR practice refer to the type of practice test questions that a participant completed in Session 1. Error bars report standard error of the mean

choices lead to similar final test performance, but one pattern requires less time to achieve that performance, then the more efficient option would be more effective.

Starting with durability of learning, Fig. 3 reports performance on the cued-recall final test, and Fig. 4 reports performance on the multiple-choice final test. Participants in the SRL-MC group performed similarly to participants in the SRL-CR group on the cued-recall final test [(SRL-MC pooled $M = 32\%$, SRL-CR pooled $M = 37\%$), pooled $d = 0.18$, 95% CI = -0.05, 0.40] and on the multiple-choice final test [(SRL-MC pooled $M = 68\%$, SRL-CR pooled $M = 70\%$), pooled $d = 0.10$, 95% CI = -0.13, 0.33]. When considering durability, which is our first dimension of effectiveness, the SRL-MC group is currently on equal footing with the SRL-CR group.

Nevertheless, participants in the SRL-MC group could have performed better if they had continued practicing items beyond one correct response. In particular, performance on the cued-recall final test was lower for participants in the SRL-MC group (pooled M = 32%) compared to items that participants in the criterion-MC group practiced to criterion 5 (pooled $M = 40\%$), pooled $d = 0.28$, 95% CI = 0.05, 0.51. Performance on the multiple-choice final test was lower for participants in the SRL-MC group (pooled M = 68%) compared to items that participants in the criterion-MC group practiced to criterion 3 [(pooled $M = 74\%$), pooled d = 0.28, 95% CI = 0.05, 0.51] and items practiced to criterion 5 [(pooled $M = 75\%$), pooled $d = 0.35$, 95% CI = 0.12, 0.58]. These outcomes suggest that participants in the SRL-MC group could have performed better on the final tests if they had continued practicing items to a higher criterion.

Based on durability alone, the current outcomes suggest that participants' choices in the SRL-MC group were equally as effective as participants' choices in the SRL-CR group but less effective than the higher criterion conditions in the criterion-MC

group, particularly for the criterion 5 items. However, time cost outcomes (our second dimension of effectiveness) support more nuanced conclusions concerning overall effectiveness.

Our two indicators of time cost were the mean number of Session 1 practice trials that a participant completed for each item and mean Session 1 practice time per item. Mean number of practice trials included initial study trials, restudy trials, and test trials. Similarly, practice time per item included time spent on initial study trials, restudy trials, and test trials. Figure 5 reports practice time per item in each of the four groups.

Participants in the SRL-MC group completed fewer practice trials (pooled $M = 3.0$) than participants in the SRL-CR group (pooled $M = 3.7$), pooled $d = 0.37$, 95% CI = 0.15, 0.58. Similarly, participants in the SRL-MC group also had a lower practice time per item (pooled $M = 10$ s) than participants in the SRL-CR group (pooled $M = 15$ s), pooled $d = 0.42$, 95% CI = 0.21, 0.63. Given that participants in these two groups achieved similar final test performance but using multiple-choice practice tests allowed participants to reach that performance more efficiently, the SRL-MC group was more effective overall.

Now we turn to the effectiveness of the SRL-MC group relative to the criterion-MC group. To revisit, participants' performance in the SRL-MC group was lower than for criterion 5 items for both types of final tests. However, participants in the SRL-MC group completed fewer practice trials than participants in the criterion-MC group needed to reach criterion 5 [pooled *M*s = 3.0 versus 7.4, respectively; pooled $d = 2.46$, 95% CI = 2.19, 2.74], and participants in the SRL-MC group spent less time per item than was needed to reach criterion 5 [pooled *M*s = 10 s versus 29 s, respectively; pooled $d = 1.30$, 95% CI = 1.07, 1.53]. Thus, participants in the SRL-MC group were more efficient in this comparison, although they reached lower levels of final performance.

To assess the balance between durability and time cost, we created a derived measure that combined both indicators. We divided each participant's total percent correct on each final memory test by the total practice time for all 48 items, which represents gains per minute. In the SRL groups, total practice time was the observed time spent practicing all 48 items. In the criterion groups, total practice time was an estimated total for how long it would have taken participants to practice all 48 items to criterion 5 (observed total practice time for all 16 criterion 5 items, multiplied by 3 to get estimated time for all 48 items). Gains per minute were numerically greater in the SRL-MC group versus in the SRL-CR group for the multiple-choice final test [(SRL-MC pooled $M = 17\%$, SRL-CR pooled $M = 13\%$), pooled $d = 0.19$, 95% CI = -0.03, 0.42], although not for the cued-recall final test [(SRL-MC pooled $M = 8\%$, SRL-CR pooled $M = 8\%$), pooled $d = 0.01$, 95% CI = -0.21, 0.24]. The advantage on the multiple-choice final test further supports the conclusion that the SRL-MC group was more effective overall than the SRL-CR group. Additionally, participants in the SRL-MC group had greater gains per minute compared to items practiced to criterion 5 in the criterion-MC group for both the cued-recall final test [(pooled *Ms* = 8% and 3%, respectively), pooled $d = 0.40$, 95% CI = 0.17, 0.63] and the multiple-choice final test [(pooled *Ms* = 17% and 7%, respectively), pooled $d = 0.72$, 95% CI = 0.48, 0.95]. This

pattern suggests that the SRL-MC group was more effective overall as compared to reaching criterion 5 in the criterion-MC group.

We will return to this trade-off between durability and time cost in the General Discussion. We focused on comparisons between the SRL-MC and criterion-MC groups here, but additional analyses specific to the criterion groups can be found in supplemental materials (https://osf.io/vugkw). The full CCMA results for each comparison between the SRL-MC vs. SRL-CR groups and SRL-MC vs. each criterion level in the criterion-MC group can be found in Appendix 2.

General Discussion

Multiple-choice questions are frequently part of students' academic lives. However, despite the widespread availability of multiple-choice practice tests, only one prior experiment had examined how students choose to use this format of practice question while learning. The current project was the first to examine students' self-testing choices for multiple-choice questions when students could test or restudy items as many times as they liked. Across the present experiments, we investigated two key research questions: How do students regulate their use of multiple-choice practice testing? And, how effective is students' use of multiple-choice practice testing for achieving long-term retention of material? To help answer these questions, we compared students' self-regulated use of multiple-choice questions to students' self-regulated use of cued-recall questions. To assess the effectiveness of self-testing choices, we included an experimenter-controlled criterion group in which participants continued completing multiple-choice practice tests until each item was correctly answered either 1, 3, or 5 times.

How did Students Regulate Their Use of Multiple-Choice Practice Testing?

To revisit, the test-to-monitor hypothesis states that students use practice testing to monitor their learning and not to enhance their learning per se. This hypothesis motivated the prediction that students will choose to reach an average criterion of one correct answer per item, and students' choices in the SRL-MC group were consistent with this prediction. Students typically completed multiple-choice test trials until they achieved one correct answer for a word pair. After this first correct answer, students tended to drop items from practice. Students in the SRL-CR group made similar choices. These results suggest that students were using practice tests to monitor their learning, and this did not change based on which format of practice test they used.

Students in both SRL groups reached a similar criterion during learning, but other aspects of self-testing choices differed between students who had access to multiple-choice questions versus those who had access to cued-recall questions. In particular, students in the SRL-MC group started testing themselves earlier (i.e., they completed fewer restudy trials before their first test trial) and completed fewer test trials overall, compared to students in the SRL-CR group. The test-to-monitor hypothesis can explain these findings, based on the auxiliary assumption that students perceive multiple-choice questions as being easier than cued-recall questions. Accordingly,

because this hypothesis predicts that students test themselves when they believe they will answer the practice question correctly, it follows that students will judge themselves as ready to check their knowledge earlier when using multiple-choice than cued-recall questions. On trials in which they fail to answer the question correctly, they may choose to test again later to re-assess their learning after more studying. Incorrect responses are less likely with multiple-choice than with cued-recall questions, meaning that multiple-choice questions would take fewer test trials to reach their first correct answer. Thus, even though these two aspects of self-testing choices differed between students who had access to multiple-choice versus cued-recall questions, these differences can be explained by the test-to-monitor hypothesis.

How Effective was Students' Use of Multiple-Choice Practice Testing?

In the current experiments, we viewed effectiveness as a balance between durability (measured using final test performance) and time cost associated with practice (measured using number of practice trials and practice time per item). Students in the SRL-MC and SRL-CR groups had similar durability, yet students' use of multiple-choice practice tests was more efficient. Because of this difference in time cost, students' use of multiple-choice practice testing was more effective than students' use of cued-recall practice testing. This conclusion differs from most lab experiments that compared multiple-choice and cued-recall practice testing, in which cued-recall practice tests are typically more effective (Butler & Roediger, 2007; Clariana, 2003; Kang et al., 2007). Methodological differences may explain these different conclusions. First, the current experiments examined self-regulated use of these two formats and evaluated both durability and time cost, whereas other experiments used a fixed number of practice test trials and did not include time cost outcomes. Differences in target material between the current experiments (German-English word-pair translations) and the prior research (facts or concepts from scholarly journal articles in Kang et al., 2007, facts from a class lecture in Butler & Roediger, 2007, and terms from a textbook in Clariana, 2003) may also contribute to these differences.

Even without considering time cost outcomes, prior research has shown that multiple-choice practice questions are effective at boosting performance when they involve competitive lures (i.e., alternatives related to other material studied). When multiple-choice questions have competitive incorrect alternatives, learners may retrieve information regarding the other alternatives in order to narrow down the correct answer. Because of this additional processing, multiple-choice questions with competitive lures can boost performance as much as cued-recall questions (Bjork et al., 2015; Little & Bjork, 2015; Little et al., 2012). In the current experiments, the alternatives for multiple-choice questions were other German words from the learning materials. These competitive lures meant that participants could not just rely on target familiarity and instead had to think about the associations between the German words and their English translations.

Conclusions about the overall effectiveness of participants' choices in the SRL-MC group are more nuanced when compared to items practiced to criterion 5 in the criterion-MC group. Compared to SRL-MC, items practiced to criterion 5 had

higher performance on both final tests, but students also needed more practice trials and more time per item to reach this higher criterion. Based on participants' average choices in the SRL-MC group, they would have needed to complete at least four more test trials per item to reach criterion 5, which would have been at least 192 additional test trials. Now consider the time cost associated with completing those additional trials. Participants in the SRL-MC group spent an average of 8 min completing practice test trials and restudy trials for all 48 items, and we estimated that they would have needed to spend an additional 15 min to complete the extra trials required to reach criterion 5 for all items. Would this additional time have been well spent? Reaching criterion 5 was associated with an 8-percentage point increase in performance on the cued-recall final test and a 7-percentage point increase on the multiple-choice final test, compared to performance in the SRL-MC group. Thus, stopping after one correct response was more efficient, but continuing practice to a higher criterion would have led to better durability.

So which practice schedule is better overall? The answer depends on a student's performance goal and how much time they have available. First, consider a student who has a high-performance goal and is willing to devote extra time to reach that goal. In this case, practicing material to criterion 5 would be best. However, students typically need to balance their time among several different classes and commitments. If a student is trying to balance their time, they may emphasize having low time costs associated with their practice, and then the choices made by participants in the SRL-MC group would be the most effective. The 15-min time cost discussed above would have occurred when students were learning 48 word-pair translations. Now consider scaling this up to the number of items students would be expected to learn for a typical course exam. If a student was taking a foreign language course, they would likely be expected to learn many more than 48 individual words for a given exam. Thus, the additional time cost associated with reaching a criterion of 5 correct answers per item could grow substantially, particularly when considering that most students take several classes per semester.

Future Directions

The only prior research that has examined students' self-regulated use of multiple-choice practice tests gave students limited control options – specifically, students only had the option to skip practice tests (Pollock & Sullivan, 1990). The current experiments were the first to examine students' use of multiple-choice practice questions when they had more control over their learning. This leaves a substantial amount of room for future experiments. Below we discuss two directions for further research that are arguably important.

First, the conclusions in the current project are based on students' use of multiple-choice practice testing during a single learning session. Students often use single-session learning to prepare for exams (i.e., they cram the night before an exam), but this is not the most effective learning schedule for long-term retention. Instead, students should spread their studying across multiple sessions (Rawson et al., 2013). Accordingly, one important future direction is investigating how

students use multiple-choice practice testing over several learning sessions. For example, Janes et al. (2018) investigated students' use of cued-recall practice questions across four learning sessions and found that students reached a criterion of about one correct response in each learning session. If students use multiple-choice questions similarly, this pattern would be important to consider when making conclusions about effectiveness. In the current experiments, a tradeoff occurred between the durability of learning and time costs associated with practice in the SRL-MC group compared to reaching criterion 5. Consider a slightly modified comparison group, in which participants learn items to criterion 5 during the first learning session, and then practice items to criterion one during three relearning sessions. The higher criterion during the first learning session would result in an initial boost in performance compared to the SRL-MC group (as was observed in the current two experiments), but this initial boost in performance would likely be attenuated by the benefits of the relearning sessions (this is known as *relearning override;* see Rawson et al., 2018). If participants in the SRL-MC group practiced items to criterion 1 during each learning session, then the SRL-MC group could reach the same performance levels as this modified criterion 5 group while still being the most efficient option.

Second, in the current experiments, students had access to *either* multiple-choice or cued-recall practice testing. However, we do not know what students would choose to do if they had access to both formats. Students might see a unique utility in each type of question and choose to use both multiple-choice and cued-recall practice questions. If students choose to use both formats, they may reach a criterion of one correct answer for each format of practice test, which would result in reaching a higher overall criterion and could impact the durability of learning. Alternatively, students may believe that using just one format of practice test is sufficient. Even with access to both formats of practice tests, students may reach an overall criterion of one correct answer per item, as in the current experiments.

Closing Remarks

Taking practice tests is an effective learning strategy that can help students succeed in their courses. Investigating how students use optional practice tests is important for understanding what choices students make and how effective those choices are. The current experiments focused on students' use of multiple-choice practice questions. Students chose to use multiple-choice practice questions until they correctly answered each item about one time, and this finding is consistent with the test-to-monitor hypothesis. Students could have achieved higher performance on the final tests if they had continued practicing to a higher number of correct answers, but that would also impose a significant time cost. Thus, students' use of multiple-choice questions was comparatively effective because of the lower time costs associated with practice. These experiments provide an important foundation for future research to further investigate how students regulate their use of multiple-choice practice questions.

Appendix 1

Table 2 Descriptive statistics for Experiment 2 separated by self-reported use of external memory aids

	"No", did not use external aid		"Yes", used external aid	
	M	*SE*	*M*	*SE*
SRL-MC group				
Criterion reached per item	1.2	0.1	0.9	0.2
Test trials per item	1.6	0.1	1.2	0.2
Restudy trials before first test trial	0.4	0.1	0.3	0.1
Restudy trials per item	0.5	0.1	0.4	0.1
Cued-recall final test performance (%)	25	3	59	5
Multiple-choice final test performance (%)	63	3	82	2
Practice trials per item	3.1	0.2	2.5	0.2
Learning time per item (in seconds)	11	1	11	2
SRL-CR group				
Criterion reached per item	1.0	0.2	0.8	0.2
Test trials per item	1.7	0.2	1.2	0.3
Restudy trials before first test trial	0.6	0.1	0.3	0.1
Restudy trials per item	0.9	0.2	0.4	0.1
Cued-recall final test performance (%)	23	3	66	6
Multiple-choice final test performance (%)	59	3	80	4
Practice trials per item	3.6	0.3	2.6	0.3
Learning time per item (in seconds)	15	2	11	2
Criterion-MC group				
Test trials per item				
Criterion 1	1.5	0.1	1.3	0.1
Criterion 3	4.0	0.2	3.5	0.2
Criterion 5	6.3	0.3	5.8	0.3
Cued-recall final test performance (%)				
Criterion 1	29	4	66	9
Criterion 3	33	4	60	9
Criterion 5	35	5	65	8
Multiple-choice final test performance (%)				
Criterion 1	61	3	80	6
Criterion 3	68	3	82	5
Criterion 5	67	3	79	6
Learning time per item (in seconds)				
Criterion 1	9	1	11	2
Criterion 3	20	1	23	2
Criterion 5	29	2	38	5

M = mean. *SE* = standard error of the mean

Table 3 Inferential statistics for Experiment 2, excluding participants who answered "yes" to using an external memory aid

	p	*Cohen's d*
SRL-MC group vs. SRL-CR group		
Criterion reached per item	0.31	0.16
Restudy trials before first test trial	0.18	0.21
Test trials per item	0.54	0.10
Final multiple-choice test performance	0.32	0.17
Final cued-recall test performance	0.70	0.07
Practice time per item	0.03	0.34
Number of trials per item	0.12	0.24
Final multiple-choice test gains per minute	0.12	0.28
Final cued-recall test gains per minute	0.96	0.01
SRL-MC group vs. Criterion 3 in Criterion MC group		
Final multiple-choice test performance	0.18	0.58
Final cued-recall test performance	0.11	0.63
Practice time per item	< 0.001	1.25
Number of practice trials	< 0.001	1.70
SRL-MC group vs. Criterion 5 in Criterion MC group		
Final multiple-choice test performance	0.23	0.56
Final cued-recall test performance	0.06	0.68
Practice time per item	< 0.001	1.65
Number of practice trials	< 0.001	2.56
Final multiple-choice test gains per minute	< 0.001	0.78
Final cued-recall test gains per minute	0.06	0.34

If a participant selected "yes" to the close-ended external memory aid question but elaborated in the open-ended question that they used a mnemonic, they were still included in these analyses

Additional details about responses to external memory aid question:

Participants were asked a yes or no question regarding whether they used an external memory aid during either Session 1 or Session 2. After answering this yes or no question, participants saw a textbox to type in the type of external memory aid they used. Many participants who answered "yes" to this question indicated they took notes and/or used a web source to help them with the answers. For example, one participant said, "I googled some of the answers, as well as wrote down some to study". Participants also commonly indicated they used notes or web sources to help them learn the items but did not use them during the final tests (e.g., "I wrote down some of the word pairings while learning them. But I didn't use my notes during the recall test"). Interestingly, some participants reported that they used an external memory aid, but their responses described a mnemonic or other memory strategy. For example, one participant said, "I would use other words that would rhyme the with german word to remember the word in english." Although some participants also indicated they wrote down the mnemonic they created (e.g., "On ones that I knew I couldn't remember I wrote down like "MM" on a sheet of paper. For example, Messen is measure so they both start with M thats how I remember them").

Appendix 2 CCMA outcomes

Table 4 Continuously Cumulating Meta-Analysis (CCMA) outcomes for all Session 1 and Session 2 variables of interest

	Experiment 1					Experiment 2					CCMA		
	s_{pooled}	*t*	*p*	*d*	*Z*	s_{pooled}	*t*	*p*	*d*	*Z*	*p*	*d*	*Z*
Criterion reached per item													
SRL-MC group versus:													
SRL-CR	0.98	0.66	0.51	0.11	0.65	1.25	1.15	0.25	0.16	1.14	0.63	0.05	1.27
Number of test trials per item													
SRL-MC group versus:													
SRL-CR	1.39	2.62	0.01	0.45	2.58	1.55	0.52	0.60	0.07	0.52	0.04	0.22	2.19
Number of restudy trials per item before first test trial													
SRL-MC group versus:													
SRL-CR	0.88	2.80	0.01	0.48	2.76	0.73	1.27	0.21	0.17	1.26	0.01	0.29	2.84
Cued-recall final test performance													
SRL-MC group versus:													
SRL-CR	26.61	1.99	0.05	0.37	1.97	31.11	0.36	0.72	0.05	0.36	0.13	0.18	1.65
Criterion-MC 1	27.64	1.11	0.27	0.20	1.10	33.97	0.49	0.63	0.07	0.49	0.72	0.04	1.12
Criterion-MC 3	28.49	1.29	0.20	0.23	1.28	32.56	0.92	0.36	0.14	0.92	0.12	0.18	1.56
Criterion-MC 5	27.32	2.16	0.03	0.39	2.13	32.84	1.35	0.18	0.21	1.35	0.02	0.28	2.46

Table 4 (continued)

	Experiment 1					Experiment 2					CCMA		
	s_{pooled}	*t*	*p*	*d*	*Z*	s_{pooled}	*t*	*p*	*d*	*Z*	*p*	*d*	*Z*
Multiple-choice final test performance													
SRL-MC group versus:													
SRL-CR	18.66	2.42	0.02	0.45	2.38	22.64	0.82	0.41	0.12	0.82	0.39	0.10	2.26
Criterion-MC 1	19.93	0.37	0.71	0.07	0.37	22.62	0.85	0.40	0.13	0.85	0.37	0.10	0.86
Criterion-MC 3	18.32	2.71	0.01	0.49	2.67	21.71	0.83	0.41	0.13	0.83	0.02	0.28	2.29
Criterion-MC 5	16.64	4.14	< 0.001	0.74	4.00	22.16	0.52	0.60	0.08	0.52	0.003	0.35	3.20
Number of practice trials per item													
SRL-MC group versus:													
SRL-CR	1.88	3.76	< 0.001	0.64	3.66	2.11	1.42	0.16	0.19	1.41	< 0.001	0.37	3.58
Criterion-MC 1	1.12	2.24	0.03	0.38	2.22	1.12	2.95	0.003	0.41	2.92	< 0.001	0.39	3.64
Criterion-MC 3	1.32	9.71	< 0.001	1.62	8.49	1.44	10.61	< 0.001	1.46	9.48	< 0.001	1.52	12.71
Criterion-MC 5	1.60	16.01	< 0.001	2.67	12.08	1.94	16.97	< 0.001	2.33	13.45	< 0.001	2.46	18.05
Practice time per item													
SRL-MC group versus:													
SRL-CR	10.10	3.85	< 0.001	0.66	3.74	12.46	1.97	< 0.001	0.66	0.74	< 0.001	0.42	4.03
Criterion-MC 1	14.01	0.62	0.53	0.10	0.62	7.10	1.48	0.14	0.20	1.48	0.46	0.08	1.48
Criterion-MC 3	14.49	4.63	< 0.001	0.78	4.47	9.65	7.15	< 0.001	0.98	6.76	< 0.001	0.90	7.94
Criterion-MC 5	14.97	7.01	< 0.001	1.18	6.52	14.20	10.11	< 0.001	1.39	9.12	< 0.001	1.30	11.07

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10648-023-09761-1.

Data Availability Data available upon request.

References

Ariel, R., & Karpicke, J. D. (2018). Improving self-regulated learning with a retrieval practice intervention. *Journal of Experimental Psychology: Applied, 24*, 43–56.

Badali, S., Rawson, K. A., & Dunlosky, J. (2022). Do Students Effectively Regulate Their Use of Self-Testing as a Function of Item Difficulty? *Educational Psychology Review, 34*, 1651–1677.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology, 64*, 417–444.

Bjork, E. L., Soderstrom, N. C., & Little, J. L. (2015). Can multiple-choice testing induce desirable difficulties? Evidence from the laboratory and the classroom. *The American Journal of Psychology, 128*, 229–239.

Braver, S. L., Thoemmes, F. J., & Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science, 9*, 333–342.

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.

Clariana, R. B. (2003). The effectiveness of constructed-response and multiple-choice study tasks in computer aided learning. *Journal of Educational Computing Research, 28*, 395–406.

Dunlosky, J., & Rawson, K. A. (2015). Do students use testing and feedback while learning? A focus on key concept definitions and learning to criterion. *Learning and Instruction, 39*, 32–44.

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.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149–1160.

Janes, J. L., Dunlosky, J., & Rawson, K. A. (2018). How do students use self-testing across multiple study sessions when preparing for a high-stakes exam? *Journal of Applied Research in Memory and Cognition, 7*, 230–240.

Kang, S. H., McDermott, K. B., & Roediger, H. L., III. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology, 19*, 528–558.

Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General, 138*, 469–496.

Karpicke, J. D., & Smith, M. A. (2012). Separate mnemonic effects of retrieval practice and elaborative encoding. *Journal of Memory and Language, 67*, 17–29.

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

Little, J. L., & Bjork, E. L. (2015). Optimizing multiple-choice tests as tools for learning. *Memory & Cognition, 43*, 14–26.

Little, J. L., Bjork, E. L., Bjork, R. A., & Angello, G. (2012). Multiple-choice tests exonerated, at least of some charges: Fostering test-induced learning and avoiding test-induced forgetting. *Psychological Science, 23*, 1337–1344.

McDaniel, M. A., & Little, J. L. (2019). Multiple-choice and short-answer quizzing on equal footing in the classroom: Potential indirect effects of testing. In J. Dunlosky & K. Rawson (Eds.), *Handbook of cognition and education* (pp. 480–499). Cambridge University Press.

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.

Parmenter, D. A. (2009). Essay versus multiple-choice: Student preferences and the underlying rationale with implications for test construction. *Academy of Educational Leadership Journal, 13*, 57–71.

Pollock, J. C., & Sullivan, H. J. (1990). Practice mode and learner control in computer-based instruction. *Contemporary Educational Psychology, 15*, 251–260.

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.

Rawson, K. A., Dunlosky, J., & Sciartelli, S. M. (2013). The power of successive relearning: Improving performance on course exams and long-term retention. *Educational Psychology Review, 25*, 523–548.

Rawson, K. A., Vaughn, K. E., Walsh, M., & Dunlosky, J. (2018). Investigating and explaining the effects of successive relearning on long-term retention. *Journal of Experimental Psychology: Applied, 24*, 57–71.

Riggs, C. D., Kang, S., & Rennie, O. (2020). Positive impact of multiple-choice question authoring and regular quiz participation on student learning. *CBE—Life Sciences Education, 19*, 1–9.

Rivers, M. L. (2021). Metacognition about practice testing: A review of learners' beliefs, monitoring, and control of test-enhanced learning. *Educational Psychology Review, 33*, 823–862.

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.

Scouller, K. (1998). The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay. *Higher Education, 35*, 453–472.

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., & Pyc, M. A. (2013). Repeated retrieval practice and item difficulty: Does criterion learning eliminate item difficulty effects? *Psychonomic Bulletin & Review, 20*, 1239–1245.

Zeidner, M. (1987). Essay versus multiple-choice type classroom exams: The student's perspective. *The Journal of Educational Research, 80*, 352–358.