Do Students Effectively Regulate Their Use of Self-Testing as a Function of Item Difficulty?

Sabrina Badali¹ · Katherine A. Rawson¹ · John Dunlosky¹

Accepted: 7 February 2022 / Published online: 5 March 2022

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract

Retrieval practice is beneficial for both easy-to-learn and difficult-to-learn materials, but scant research has examined students' use of self-testing for items of varying difficulty. In two experiments, we investigated whether students differentially regulate their use of self-testing for easy and difficult items and assessed the effectiveness of students' self-regulated choices. Undergraduate participants learned normatively easy and normatively difficult Lithuanian-English word pair translations. After an initial study trial, participants in the self-regulated learning groups chose whether they wanted to restudy an item, take a practice test, or remove an item from further practice. Participants chose to test items repeatedly while learning but dropped both easy and difficult items after reaching a criterion of about one correct recall per item. Consequently, final test performance 2 days later was lower for difficult items versus easy items, and performance was lower in the self-regulated learning group than in an experimenter-controlled comparison group (in Experiment 1). In Experiment 2, we tested hypotheses for *why* participants reached a similar number of correct recalls for both easy and difficult items. Three new groups included different scaffolds aimed at minimizing potential barriers to effective regulation. These scaffolds did not change participants' learning choices, and as a result, performance on difficult items was still lower than on easy items. Importantly, participants planned to continue practicing items beyond one correct recall and believed that an optimal student should practice difficult items more than easy items, but they did not execute this plan during the learning task.

Keywords Self-regulated learning · Retrieval practice · Practice testing · Criterion learning

✉ Sabrina Badali
sbadali@kent.edu

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

When students are studying for a class, they are typically engaging in self-regulated learning and making their own decisions about what material to study and what learning strategies to use. In almost any class, the learning materials will vary in difficulty, meaning that students will need to learn some material that is easier than others and some material that is more difficult than others. How students regulate their learning for items of varying difficulty will have important implications for their subsequent retention of that information; in particular, difficult material will require more practice during learning to reach a similar level of retention, even if the same effective strategy is used. For example, retrieval practice is a powerful memory strategy in which the act of retrieving information from long-term memory enhances retention of that information. Retrieval practice is beneficial for both easy and difficult items (de Jonge & Tabbers, 2013; Karpicke, 2009), but difficult items require more retrieval practice during learning to obtain the same retention as easier-to-learn items (Vaughn et al., 2013). When students are making their own self-regulated choices about what material to study and how much to study it, they may not differentially regulate their use of self-testing for easy and difficult items, which could have a detrimental impact on retention of difficult items.

The main questions of the present research are: Will students differentially regulate their use of self-testing for easy and difficult items? And how effective is students' use of self-testing for achieving long-term retention of both easy and difficult items? Below, we first briefly discuss research indicating what normatively effective use of self-testing would involve when learning easy and difficult items. This provides an example of what students could do to effectively regulate their learning of easy and difficult items. Next, we consider what students may do when regulating their learning of easy and difficult items through the lens of the agenda-based regulation framework (Dunlosky & Ariel, 2011). This illustrates different ways that students may dysregulate and achieve suboptimal performance for difficult items, compared to easy items. We then review relevant research on self-regulated learning. We start this review by discussing the two previous investigations of students' use of self-testing for items of varying difficulty. However, these two studies alone are not enough to inform predictions for the current set of experiments, so we also broaden this review by discussing the most relevant aspects of the self-testing literature and the literature on students' restudy choices for easy and difficult items. Together, these studies are then used to inform our predictions about how students may choose to use self-testing in the current set of experiments.

Effective Regulation for Easy and Difficult Items

One of the most effective ways to use retrieval practice is to continue practicing items until they are correctly recalled multiple times during learning (Pyc & Rawson, 2009). This method of practicing items to a certain number of correct recalls is known as *criterion learning*, in which items are presented for practice tests until they have reached their predetermined *criterion level* (i.e., the number of times an item needs to be correctly recalled during practice before it is removed from the list of learning materials). However, if easy and difficult items are practiced to the same

criterion level, performance on a memory test will be lower for the difficult items than for the easy items. For example, Vaughn et al. (2013) had students learn Lithuanian-English word pair translations that were normatively easy and normatively difficult. These easy and difficult items were assigned to one of six criterion levels (1, 3, 5, 7, 9, or 11 correct recalls during practice). Equating the criterion (number of correct recalls) for easy and difficult items was not enough to equate performance on a delayed cued recall test (e.g., performance in the Criterion 1 condition was 33% for easy items versus 18% for difficult items). Instead, the difficult items needed to be practiced to a higher criterion before final memory performance was similar to easy items (e.g., 58% for difficult items at Criterion 5, 61% for easy items at Criterion 3). This experiment provides an example of what students could do to effectively regulate their learning for items of varying difficulty: if students reach a higher criterion for difficult items compared to easy items, then they could perform equally well on both difficult and easy items on a final test.

Agenda-Based Regulation Framework of Self-Regulated Learning

When students are regulating their own learning, they have many choices to make, and they may not choose to implement a learning plan like the one outlined above. A theoretical framework of self-regulated learning is useful to consider the regulatory choices that learners may make. According to the agenda-based regulation framework (Dunlosky & Ariel, 2011), self-regulated learning is goal oriented and involves reducing a discrepancy between a perceived state and a goal state (this assumption is key to other frameworks as well, e.g., Winne & Hadwin, 1998). Learners set goals and then develop agendas (or plans) to work toward this goal state. These agendas guide various decisions that the learner makes, including which items to select for learning, which learning strategies to use, and how much time to allocate. Agendas play an important role in students' learning and can explain when regulation is optimal (e.g., when a learner develops an effective agenda and sticks to this agenda throughout learning) and when regulation is suboptimal (e.g., when the agenda constructed was suboptimal, the execution of an agenda was suboptimal, or the learner had a low learning goal). When executing agendas, learners can monitor and evaluate whether their goals have been met. If their goal has not been met, they can decide to continue engaging with an item, change the learning strategy they are using, or change their learning goal.

Consider a task in which students are learning easy and difficult items and can choose to restudy an item, take a practice test, or remove an item from further practice. This task gives students several options, and there are many ways students could dysregulate and achieve suboptimal memory performance. For example, students may enter this task with a learning goal (e.g., goal state) that involves learning only the easy items and tolerating lower performance for the difficult items. Alternatively, students may have the goal to perform well on all items but may not realize that difficult items will need more practice than easy items to achieve this goal. Thus, they may plan to make similar learning choices for all items, which would result in lower retention of difficult items compared to easy items. Alternatively, students could

plan to practice the difficult items more but may execute this agenda ineffectively and treat items similarly during learning, which would also result in lower retention of difficult items compared to easy items. As illustrated by the agenda-based regulation framework, several possible patterns of self-regulated behavior may emerge when engaging in self-testing for easy and difficult items. We turn now to reviewing the relevant self-regulated learning literature to help inform predictions about how students may choose to use self-testing for easy and difficult items in the current set of experiments.

How Do Students Use Self-Testing for Easy and Difficult Items?

When reviewing previous literature, it is important to consider that the primary purpose of the current experiments contains two key components: (1) students' use of self-testing and (2) for items of varying difficulty. For prior research to be directly relevant, it must include both components. The literature on self-regulated learning contains two relevant types of research, survey research and behavioral research.

Concerning survey research, only one study has examined students' use of self-testing for easy and difficult items. Wissman et al. (2012) found that students reported using flashcards differently depending on the difficulty of material. College students were asked how they decide to remove items from a stack of flashcards they are learning. Most participants indicated that they rarely remove items that seem too difficult, and most indicated that they often remove items that seem too easy. Other survey studies have asked students about their use of self-testing, but these remaining studies do not provide information about item difficulty (e.g., Blasiman et al., 2017; Hartwig & Dunlosky, 2012).

Concerning behavioral research, despite the vast literatures on retrieval practice and on self-regulated learning more generally, few studies have examined self-regulated use of retrieval practice. Of greatest interest for present purposes, only two prior studies have examined whether students' use of retrieval practice depends on item difficulty (Toppino et al., 2018; Tullis et al., 2018). In both of these studies, college students learned word pairs that were normatively easy or difficult. After an initial study trial, participants were given only one additional learning trial per item. For each item, participants could either choose to restudy the word pair or to take a cued recall practice test. In both studies, participants chose to test the majority of easy items and to restudy the majority of difficult items. These two studies provide information about students' self-testing choices when given one additional learning trial per item; however, students' choices for easy and difficult items may change if they are allowed to make multiple learning choices per item.

Although the studies discussed above are the only ones that contain both key components relevant to the primary aim of the current research (students' self-regulated use of self-testing and for items of varying difficulty), other behavioral studies contain one of the two components and provide information about what students choose to do in related learning conditions. For example, the few other behavioral studies that examined self-regulated use of retrieval practice did not include item difficulty, but the outcomes still shed light on what students choose to do when they

can make more than one learning choice per item. When allowed to make multiple learning choices, participants typically choose to test themselves on most items and engage in several test trials per item, but they stop after about one correct retrieval (Ariel & Karpicke, 2018; Dunlosky & Rawson, 2015; Karpicke, 2009; Kornell & Bjork, 2008). These results come from college students' choices for items of similar difficulty, but this pattern may hold for both easy and difficult items. Accordingly, when students are learning items of varying difficulty, they may choose to use practice testing for all items and choose to stop learning an item after about one correct recall, regardless of item difficulty.

Concerning the second key component of the current research (i.e., students' choices for items of varying difficulty), a large literature exists regarding students' restudy decisions for easy and difficult items. Students typically choose to restudy difficult items more often and spend more time restudying difficult items than easy items (Son & Metcalfe, 2000). These results come from students' restudy choices for easy items and difficult items, but this pattern may hold for test choices. Accordingly, when students can choose to restudy items and/or to test items, they may engage in more restudying and more testing for difficult items. This possibility may seem contradictory to the expectation that students will reach a criterion of one for both easy and difficult items, but these two expectations are complementary. Difficult items may require more restudy trials and more test trials before being correctly recalled for the first time. Thus, a student could engage in a higher number of restudy trials and test trials for difficult items and still reach the same criterion as easy items.

Overview of Current Experiments

The main questions of the present research are: Will students differentially regulate their use of self-testing for easy and difficult items? And how effective is students' use of self-testing for achieving long-term retention of both easy and difficult items? After an initial study trial, participants in a self-regulated learning group chose whether they wanted to restudy an item, take a practice test, or remove an item from further practice. Participants could make as many learning choices as they wanted for each item. Our initial hypothesis was that participants in the self-regulated learning group will reach a similar criterion for both the easy and difficult items, resulting in lower performance on a delayed memory test for the difficult items compared to the easy items. To foreshadow, the results of Experiment 1 confirmed these predictions.

The aims of Experiment 2 were to replicate the novel outcomes of Experiment 1 and to extend to investigate *why* participants did not adjust the criterion they reached for easy and difficult items. Agenda-based regulation framework motivated four potential hypotheses for why participants did not achieve a higher criterion for the difficult items. First, on an item level, students may not have been aware of which items would be easier to learn and which would be more difficult to learn. Second, students may not have had the same learning goal for easy items and for difficult items. Third, even if students wanted to reach similar performance for easy and

difficult items, their learning agenda may not include additional practice for difficult items. Fourth, even if participants had an effective agenda that included a plan to practice difficult items more than easy items, they may not have executed this agenda effectively.

Experiment 1

The primary purpose of this experiment was to examine whether students differentially regulate their use of self-testing for easy and difficult items. We also examined how effectively students regulate their use of self-testing, with respect to the level of retention they achieve for easy and difficult items. Participants studied Lithuanian-English word pair translations that were normatively easy and normatively difficult. Participants were randomly assigned to one of two groups: self-regulated learning (SRL) or criterion. Participants in the SRL group chose whether they wanted to take a practice test for an item, restudy an item, or drop an item, and could make as many choices as they wanted for each item. They continued until they chose to drop all items. The criterion group was an experimenter-controlled group that utilized a normatively effective learning schedule (i.e., criterion learning). Participants in the criterion group completed practice tests until each item was correctly recalled to its predetermined criterion level of 1, 3, or 5 correct recalls. This criterion group served as a comparison group to investigate the effectiveness of SRL participants' choices.

Methods

Design and Participants The experiment used a 2 (practice group: SRL vs. criterion) $\times$ 2 (item difficulty: easy vs. difficult) mixed design. Participants were randomly assigned to either the SRL group or the criterion group, and item difficulty was manipulated within participant. In the criterion group, items were practiced until they were correctly recalled either 1, 3, or 5 times, manipulated within participant.

The target sample size was 98 participants, based on an a priori power analysis conducted using G*Power 3.1.9.6 (Faul et al., 2009) to detect a main effect of practice group in a mixed ANOVA with power set at .80, $\alpha = .05$, and $f = .25$. Powering for this between-participants effect also ensured that our sample size was sufficient (power = .93) to detect a medium effect in all within-participant analyses in the SRL and criterion groups. A total of 104 undergraduates participated in exchange for partial course credit (67% female; 70% white, 6% black, 3% Hispanic or Latino, 3% First Nations, 1% Asian; 44% were in their first year of college [$M = 1.82$, $SD = 1.10$]; age [$M = 19.09$, $SD = 1.30$]; 20% were psychology majors).

Materials Materials included 60 Lithuanian-English word pair translations. Half of the translations were normatively easy and half were normatively difficult, based on the first retrieval success norms reported by Grimaldi et al. (2010). These norms represent the percentage of participants who successfully retrieved the English item on their first cued recall test trial after one initial study trial. For the current experiment, we selected the 30 items with the highest first retrieval success rates as the

easy items ($M = .40$, $SD = .10$, range .29–.76) and the 30 items with the lowest first retrieval success rates as the difficult items ($M = .09$, $SD = .03$, range .03–.13). These word pairs were split into two blocks of 30 items, each containing 15 easy and 15 difficult items. In the criterion group, each block contained five easy items assigned to each of the three criterion levels and five difficult items assigned to each of the three criterion levels.

Memory for the word pairs was assessed using a cued recall test and a multiple-choice associative recognition test. On each trial of the cued recall test, participants were presented with a Lithuanian word and were asked to type the correct English translation. On each trial of the recognition test, participants were presented with a Lithuanian word and five English words as the answer choices (the correct answer plus four lures). Lures were English words from other items of the same difficulty level.

Procedure All instructions and tasks were administered via computer at an in-person lab setting. Participants were instructed that they would be learning 60 items, in two blocks of 30, and their goal was to learn all 60 items for a memory test 2 days later in which they would see the Lithuanian word and be asked to type the corresponding English translation. Participants started with their predetermined first block of 30 items (counterbalanced across participants) and completed all learning for that block before starting the second block of items.

In the SRL group, participants were instructed that they would start with an initial study trial for each item and that after the initial trial, they would choose what they wanted to do next with that item—study, test, or drop. Participants were told that choosing “study it again later” meant the item would be placed at the end of the item list for them to study again after the remaining items had been presented. Participants were told that choosing “take a practice test” meant the item would be placed at the end of the list, and after the remaining items had been presented, they would see the Lithuanian word and be asked to recall the English translation. They were told that after attempting to recall the answer, they would be shown the correct English word before moving on to the next item. Participants were told that choosing “drop it from the list” meant the item would be removed from the list and not presented again for any additional study or practice tests. Finally, participants were told they could choose to study or test items as many times as they wanted before dropping them and could make a new choice for items each time they were presented.

Participants then started the learning task, and items were presented one at a time in a random order. The first time an item was presented, participants started by making an ease-of-learning (EOL) judgment. Both the Lithuanian and English words were visible simultaneously and 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 Lithuanian word?” 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, participants moved to a self-paced initial study trial for that item. After participants clicked a button to indicate they were done with the study

trial, they saw a screen with buttons pertaining to their three choices for the item and descriptions of the options (see Figure 1). When participants chose to study, the item was placed at the end of the item list for self-paced restudy 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 cued recall practice test the next time it was presented. On these test trials, the Lithuanian word was presented as a cue along with a prompt to type in the English translation. After participants clicked a button to indicate they were done trying to recall the answer, the correct English translation was displayed for self-paced restudy. When participants chose to drop, 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. 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 Lithuanian word?” 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 continued engaging in study or test trials and continued making new choices each time an item was presented until they chose to drop all items in the first block. The number of study 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 30 items.

In the criterion group, participants were told that they would start with an initial study trial for each item and then would complete practice tests for the items. They were told that each item would continue to be presented for practice tests until it was correctly recalled a predetermined number of times (1, 3, or 5), at which point it would be dropped from the learning list and not presented again.

Participants in the criterion group started with an EOL and initial study trial for each item, involving the same procedure as in the SRL group. After EOLs and initial study trials were complete for all 30 items in the first block, participants started cued recall practice test trials for the items. These practice test trials were identical to those in the SRL group, in which participants saw the Lithuanian word, were asked to type the English translation, and were then shown the correct answer for self-paced restudy. If an item was answered incorrectly, it was placed at the end of the

Fig. 1 Screenshot from the self-regulated learning group, showing participants' three choices for what they could do next with an item

item list and presented again after the remaining items had been tested. If an item was answered correctly but had not yet reached its predetermined 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. When an item was ready to be removed, participants made a JOL using the same prompt and slider scale as the SRL group. Participants continued cycling through the items until all 30 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 30 items.

In both groups, Session 1 ended after participants finished learning the second block of items or after 55 min had elapsed. To foreshadow, this 55-min time limit was too short for some participants to fully complete the learning task and was removed in Experiment 2.

Participants returned 2 days later for the final memory tests. All participants started with a cued recall test for the 60 items. The items were presented one at a time, in a random order, and all test trials were self-paced. Participants then completed a multiple-choice recognition test for the same 60 items, with items presented one at a time, in a random order, and trials were self-paced. Participants then completed a brief post-experimental questionnaire for exploratory purposes that we do not discuss further here.

Results

Analyses excluded data for seven participants who did not return to complete Session 2, for one participant who reported they knew the Lithuanian words used in the experiment, and for one participant who completed less than 25% of Session 1. The 55-min time limit for Session 1 was too short for some participants; 14 participants completed less than 50% of second block of items before 55 min elapsed. For these 14 participants, analyses exclude data for items from the second block. The final sample included 95 participants (SRL group $n = 44$, criterion group $n = 51$). In the SRL group, some of the participants' Session 1 learning choices were not normally distributed, and we will report the medians for those variables when relevant. All Cohen's d values were computed using pooled standard deviations (Cortina & Nouri, 2000).

The design and methods of Experiment 1 yielded a rich data set with many outcome variables. Below, we focus on those most relevant to our main questions of interest. For archival purposes, other outcomes that may be of interest to readers are reported in Appendix 1.

Primary Outcomes To revisit, the primary research questions were: Will students differentially regulate their use of self-testing for easy and difficult items? And how effective is students' use of self-testing for achieving long-term retention of both easy and difficult items?

Fig. 2 Mean number of test trials per item and mean number of correct recalls (criterion) per item during practice in the self-regulated learning group in Experiment 1. Error bars report standard error of the mean. Dashed lines represent the medians

Fig. 3 Mean final cued recall performance for the self-regulated learning (SRL) group and for each of the three criterion levels in the criterion group in Experiment 1. Easy and difficult refer to normative item difficulty. Error bars report standard error of the mean

Concerning SRL participants' use of self-testing, Figure 2 reports the average number of test trials and average criterion reached for easy and difficult items. Participants engaged in more test trials for difficult items than easy items, $t(43) = 6.01$, $p < .001$, $d = .90$. However, they reached a similar criterion level for both easy and difficult items, $t(43) = .62$, $p = .54$, $d = .09$. Thus, participants differentially regulated in terms of choosing more test trials for difficult items, but they did not continue testing difficult items until they achieved multiple correct recalls per item. Instead, participants dropped items after about one correct recall regardless of item difficulty. Concerning the effectiveness of students' use of self-testing, these learning choices led to lower performance on the delayed cued recall test for difficult items compared to easy items (27% versus 52%, as shown in Figure 3), $t(43) = 9.62$, $p < .001$, $d = 1.53$.

A second way to assess effectiveness was comparing delayed cued recall performance from participants in the SRL group to that of participants in the criterion group. As shown in Figure 3, participants in the SRL group could have performed better on the difficult items if they had continued practicing difficult items to a higher criterion during learning. For example, performance on the difficult items was considerably worse for participants in the SRL group compared to performance on the Criterion 5 items in the criterion group (27% vs. 67%), $t(93) = 9.29$, $p < .001$, $d = 1.91$. (A full ANOVA and descriptive statistics for the criterion group are reported in Appendix 1 for interested readers.)

Of secondary interest, retention on the delayed multiple-choice recognition test followed a similar pattern (see Table 1). In the SRL group, delayed multiple-choice performance was lower for difficult items than for easy items, $t(43) = 9.85$, $p < .001$, $d = 1.52$. Additionally, participants in the SRL group could have performed better if they had continued practicing to a higher criterion. For example, multiple-choice performance for the difficult items was lower for participants in the SRL group compared to performance on the Criterion 5 items in the criterion group, $t(93) = 3.02$, $p = .003$, $d = 1.52$.

Exploratory Outcomes The analyses above show that participants in the SRL group did not adjust the criterion they reached for easy and difficult items, but why? One possibility is that participants did not identify which items were normatively easy and which were normatively difficult. We explored this possibility by assessing differences in participants' EOLs and use of restudying for easy versus difficult items (see Table 2). Participants' EOLs for difficult items were significantly lower than for easy items [$t(43) = 8.92$, $p < .001$, $d = 1.34$], indicating that participants judged the normatively difficult items would be more challenging to learn. Additionally, participants chose more restudy trials for difficult items than for easy items [$t(43) = 4.30$, $p < .001$, $d = .65$], providing further evidence that they were aware of differences in item difficulty.

EOLs in the analysis above were on an overall group level, averaged across all easy items and all difficult items. However, EOLs did not always match normative item difficulty. Accordingly, we conducted a parallel set of analyses based on idiosyncratic item difficulty. For each participant, we identified the subset of 20 items that received the participant's highest EOLs and the 20 items that received the

Table 1 Percent correct on final multiple-choice test in Experiment 1

	Easy items		Difficult items	
	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>
SRL group	82	3	63	3
Criterion group				
Criterion 1 items	82	3	56	4
Criterion 3 items	89	3	68	4
Criterion 5 items	91	2	78	3

M mean, *SE* standard error of the mean

lowest EOLs. Mean EOLs for the resulting subsets of idiosyncratically difficult and easy items were 20 and 54, respectively. For analyses based on these items, the main findings did not change; participants in the SRL group still engaged in more test trials for difficult items ($p = .02$, $d = .38$), reached a similar criterion for easy and difficult items ($p = .76$, $d = .04$), and had lower delayed cued recall performance for difficult items ($p < .001$, $d = .64$). Thus, a lack of awareness of which items are easier or more difficult is likely not the reason why participants failed to adjust the criterion reached during learning.

We are primarily interested in students' testing choices for easy and difficult items, but another aspect of students' self-regulated learning that can be examined is the time spent practicing items. Practice time includes time spent on initial study trials, restudy trials, and test trials. It does not include time spent on judgment and decision tasks (EOLs, JOLs, or decisions in the SRL group). Table 2 reports the mean practice time for easy and difficult items in the SRL and criterion groups (reported as seconds per item). Consistent with the broader literature on learning choices for easy and difficult items (e.g., Son & Metcalfe, 2000), participants in the SRL group spent more time practicing the difficult items compared to the easy items ($p < .001$, $d = 1.14$). Additionally, participants in the SRL group spent less time per item than participants in the criterion group spent on criterion 3 and 5 items ($ps < .001$, $ds > .76$). Consider these outcomes in relation to the corresponding boost in performance associated with reaching a higher criterion for difficult items. For example, the additional test trials associated with reaching criterion 5 for difficult items took about 28 s more per item and were associated with a 40 percentage point increase in delayed cued recall performance, compared to participants in the SRL group. If participants in the SRL group had reached the higher criterion of 5 correct

Table 2 Descriptive statistics for additional Session 1 outcomes in the self-regulated learning group and criterion group in Experiment 1

	Easy items		Difficult items	
	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>
SRL group				
Ease of learning judgment magnitude	42	2	31	2
Number of restudy trials per item before first test choice for that item	.5	.1	.8	.1
Number of restudy trials per item	.8	.1	1.2	.2
Practice time per item (seconds)	17	1	24	2
Criterion group				
Practice time per item (seconds)				
Criterion 1 items	17	2	29	2
Criterion 3 items	26	2	45	3
Criterion 5 items	35	2	52	3

Ease of learning judgment magnitude could range from 0 (very difficult) to 100 (very easy)

M mean, *SE* standard error of the mean

recalls for all difficult items, it would add about 14 min to their total practice time but would also result in large boost in performance for those difficult items.

Experiment 2

In Experiment 1, participants in the SRL group did not adjust the criterion they reached for easy versus difficult items. The agenda-based regulation framework described earlier motivated four explanations for why students may not effectively regulate their learning. Experiment 2 aimed to test each of these hypotheses with the addition of three new self-regulated groups.

The first new group builds upon the SRL group from Experiment 1 by adding item difficulty labels to each item. This *labels* group evaluates the hypothesis that participants did not know which items were normatively easy and which items were difficult. Exploratory outcomes from Experiment 1 weigh against this possibility, but the labels group afforded more direct evaluation of the hypothesis. According to this hypothesis, participants in the labels group will achieve a higher criterion for difficult items than for easy items. The second new group builds upon the labels group by also providing participants with a specific goal to learn items well enough to achieve at least 70% correct for the easy items and 70% correct for the difficult items on the delayed memory test. This *goal* group evaluates the hypothesis that participants did not have the same learning goal for easy and difficult items and were fine with lower performance on the difficult items. This hypothesis predicts that the goal instructions combined with the item difficulty labels will help participants in the goal group achieve a higher criterion for difficult items than for easy items. The third new group adds to the goal group by also asking participants questions regarding how they will achieve this goal for easy and difficult items, including how many times they think they should correctly recall the items before removing them from practice. This *plan* group evaluates the hypothesis that participants think the difficult items do not need to be practiced beyond one correct recall during learning. The prediction here is that students will plan to correctly recall the difficult items about one time. If we disconfirm this hypothesis and find that participants report that difficult items should be practiced beyond one correct recall, then the plan group will allow us to evaluate whether participants implement their plan during the learning task. Thus, this plan group can also evaluate the hypothesis that participants do not fully implement their plan during learning. In particular, a prediction here is that students will achieve a lower criterion than they planned to reach for the difficult items. The hypotheses, procedures, and data analysis plan for this experiment were preregistered on the Open Science Framework (<https://osf.io/b3j2n/>).

Methods

Design and Participants The experiment used a 4 (SRL group: basic, labels, goal, plan) $\times$ 2 (item difficulty: easy vs. difficult) mixed design. Participants were

randomly assigned to one of the four SRL groups, and item difficulty was manipulated within participant.

The target sample size was 136 participants, based on an a priori power analysis conducted using G*Power 3.1.9.6 (Faul et al., 2009) to detect an interaction between SRL group and item difficulty in a repeated measures ANOVA with power set at .90, $\alpha = .05$, and $f = .20$. A total of 180 undergraduates participated in exchange for partial course credit (78% female; 73% white, 8% black, 7% Asian, 6% Hispanic or Latino, 3% First Nations, 1% Native Hawaiian or Pacific Islander; 39% were in their first year of college [$M = 2.12$, $SD = 1.14$]; age [$M = 20.1$, $SD = 4.65$]; 33% were psychology majors). Forty-five participants completed the experiment at an in-person lab setting, and 135 completed it online remotely (this shift in context was due to the COVID-19 pandemic).

Materials Materials included 40 items from the list of 60 Lithuanian-English word pairs used in Experiment 1. Based on the norms established by Grimaldi et al. (2010), the 20 items with the highest first retrieval success rate were utilized as the easy items ($M = .44$, $SD = .10$, range .33–.76) and the 20 items with the lowest first retrieval success rate were utilized as the difficult items ($M = .07$, $SD = .02$, range .03–.10). Items were no longer split into two blocks; instead, participants learned all 40 items in one block. Participants learned the items in a randomized order.

Procedure All four groups engaged in self-regulated learning. The labels group, goal group, and plan group each included all features of their preceding group, plus a new defining feature represented by the new group's name. The procedure for the basic group was identical to the SRL group in Experiment 1. In the labels group, after receiving the initial instructions explaining the learning task, participants were instructed that 20 of the items were easy and 20 items were difficult. Participants were also instructed that they would see a label at the top of the screen telling them whether the current item was easy or difficult. As they completed the learning task, this item difficulty label (i.e., “easy item” or “difficult item”) was displayed at the top of the screen during all study trials, practice test trials, choice screens, and JOL screens; labels were not displayed on EOL judgment screens. In the goal group, participants received these same instructions and item difficulty labels, plus additional goal instructions before they began the learning task. Participants were told, “On the final memory test two days from now, your goal is to get at least 70% correct for the easy items and at least 70% correct for the difficult items.” This message was also displayed during choice screens to remind participants of their goal while they were completing the learning task.

Participants in the plan group received the same item difficulty instructions, item labels, and learning goal. In addition, before beginning the learning task, participants were asked a series of questions about what they believe an *optimal* student should do to achieve the learning goal. Participants were asked whether they believe a student should restudy items, take practice tests for items, the criterion they should reach for items, and how they should decide when to drop items. For each question, participants indicated their answer separately for easy items and difficult items. The order

of question prompts for easy and difficult items was counterbalanced across participants, such that half of the participants indicated their answer for easy items first and half indicated their answer for difficult items first. After completing these questions in terms of what an optimal student *should* do to achieve the learning goal, participants were asked the same series of questions about what they *plan to do* as they complete the learning task (plan group instructions and questions are in Appendix 2).

In all groups, participants proceeded through the learning task in the same manner as the SRL group from Experiment 1. Participants started with EOLs, initial study trials, and choices (study, test, drop) for each item and continued cycling through the items as many times as they wanted until they chose to drop all items. Unlike Experiment 1, there was no time limit, although all participants finished Session 1 in less than 1 h. Two days later, participants completed a final cued recall test for all 40 items followed by a multiple-choice test for all items.

Results

Analyses excluded data for 19 participants who did not return to complete Session 2, for 15 participants who completed less than 50% of the learning task in Session 1, for 3 participants who did not make any learning choices during Session 1, and for 3 participants who indicated they knew more than 50% of the Lithuanian words used in the experiment. These higher rates of exclusion can be attributed to the necessary shift to online data collection. For example, of the 19 participants who did not return to complete Session 2, 17 were online participants. Additionally, all 15 of the participants who completed less than 50% of the learning task in Session 1 were online participants who stopped after a few trials (total number of learning trials completed $M = 11.1$, including initial study trials). The final sample included 139 participants (basic group $n = 37$, labels group $n = 35$, goal group $n = 32$, and plan group $n = 35$). All Cohen's d values were computed using pooled standard deviations (Cortina & Nouri, 2000).

The interactions between group and item difficulty were of primary interest for all comparisons. To check whether outcomes could be attributed to our shift to online data collection, AVOVAs were conducted with location (in-person vs. online) as an additional independent variable. The results indicated that our conclusions did not differ across location (all three-way interaction F s < 1.21). Thus, we have collapsed across in-person and online participants for the main analyses presented below.

To revisit, the two main goals of Experiment 2 were to replicate the novel outcomes from the SRL group in Experiment 1 and to evaluate four hypotheses for why participants did not adjust the criterion they reached for easy and difficult items. To evaluate these explanations, we focus the discussion below on the criterion that participants reached for easy and difficult items and on final cued recall performance. Descriptive statistics for other variables are in Table 3.

Replication Analyses Participants' learning choices and memory performance in the basic group replicated the findings from Experiment 1. Participants engaged in more

Table 3 Descriptive statistics for additional Session 1 outcomes in each of the four groups in Experiment 2

	Easy items			Difficult items		
	<i>M</i>	<i>SE</i>	Med.	<i>M</i>	<i>SE</i>	Med.
Basic group						
Ease of learning judgment magnitude	50	2		37	2	
Number of test trials per item	1.9	.3	1.7	2.7	.4	2.0
Number of restudy trials per item before first test choice for that item	.7	.1	.5	1.1	.3	.9
Number of restudy trials per item	1.0	.2	.9	1.5	.3	1.1
Practice time per item (seconds)	17	2		24	3	
Final multiple-choice test performance (%)	84	4		68	5	
Labels group						
Ease of learning judgment magnitude	48	3		36	2	
Number of test trials per item	1.5	.2	1.5	2.2	.3	1.5
Number of restudy trials per item before first test choice for that item	.4	.1	.2	.7	.2	.3
Number of restudy trials per item	.6	.1	.2	1.4	.4	.6
Practice time per item (seconds)	13	2		24	5	
Final multiple-choice test performance (%)	82	4		62	4	
Goal group						
Ease of learning judgment magnitude	49	2		37	2	
Number of test trials per item	1.4	.2	1.4	2.6	.6	1.8
Number of restudy trials per item before first test choice for that item	.3	.1	.1	.5	.1	.1
Number of restudy trials per item	.5	.1	.2	.9	.2	.5
Practice time per item (seconds)	15	2		24	4	
Final multiple-choice test performance (%)	86	4		72	5	
Plan group						
Ease of learning judgment magnitude	54	2		36	2	
Number of test trials per item	2.0	.3	2.0	3.0	.4	2.9
Number of restudy trials per item before first test choice for that item	.8	.1	.6	1.4	.2	1.0
Number of restudy trials per item	1.2	.2	1.0	2.1	.3	2.0
Practice time per item (seconds)	18	2		30	4	
Final multiple-choice test performance (%)	84	4		64	5	

Ease of learning judgment magnitude could range from 0 (very difficult) to 100 (very easy). *M* (mean), *SE* (standard error of the mean), and *Med.* (median), which are reported only for the Session 1 learning choices that were not normally distributed

test trials for difficult versus easy items [Table 3; $t(36) = 4.21$, $p < .001$, $d = .69$] but reached a similar criterion level for difficult and easy items [Figure 4; $t(36) = .68$, $p = .50$, $d = .11$]. As a result, participants' final cued recall performance was lower for the difficult items compared to easy items (Figure 5; $p < .001$, $d = 1.09$), and their final multiple-choice performance was also lower for difficult versus easy items (Table 3; $p < .001$, $d = .84$).

Fig. 4 Left panel displays the mean number of correct recalls (criterion) per item during practice in the first three groups from Experiment 2. The right panel displays outcomes from the plan group, including participants' mean response to what an optimal student should do (left bars), what they planned to do in the current learning task (middle bars), and the mean criterion per item that plan group participants actually reached in the learning task (right bars). Error bars represent standard errors. Dashed lines represent the medians

Extension Analyses: Why Did Participants Not Reach a Higher Criterion for Difficult Items? We turn now to the three extension groups, which were designed to test four hypotheses regarding why participants did not differentially regulate the criterion reached for easy and difficult items. First, did participants reach a similar criterion because they could not discriminate between easy and difficult items? Outcomes from the labels group indicate the answer is no. Despite explicitly labeling each item as easy or difficult, participants in the labels group still reached a similar criterion level for easy and difficult items (left panel of Figure 4) and still had lower final cued recall performance for the difficult items compared to easy items (Figure 5). Compared to the basic group, these item difficulty labels did not lead to changes in the criterion reached or in final cued recall performance. (The full ANOVAs for these comparisons and other analyses below are presented in Table 4.) Because these two

Fig. 5 Mean performance on the final cued recall test in Experiment 2. Error bars represent standard errors

groups did not differ on either of the primary variables of interest, they are functionally equivalent and will be collapsed for the following analyses.

Second, did participants reach a similar criterion because they did not start with the same learning goal for easy and difficult items? Outcomes from the goal group indicate the answer is no. Participants in the goal group were given the same item difficulty labels as the labels group; additionally, they were given a specific learning goal to achieve the same level of retention (70%) on the final test for both easy and difficult items. However, they still reached a similar criterion for easy and difficult items (left panel of Figure 4). This criterion level was also similar to the criterion achieved by participants in the combined basic/labels group (Table 4). Because participants in the goal group did not adjust criterion based on item difficulty, their final cued recall performance was lower for difficult versus easy items (Figure 5) and was similar to performance in the combined basic/labels group (Table 4). Given that

Table 4 ANOVAs for criterion reached in Session 1 and final cued recall performance in Session 2

	<i>df</i>	<i>F</i>	MSE	<i>p</i>	η_p^2
Basic group vs. labels group					
Criterion reached					
Main effect of group	1, 70	1.45	.14	.23	.02
Main effect of item difficulty	1, 70	5.24	.09	.03	.07
Group * item difficulty	1, 70	1.95	.09	.17	.03
Final cued recall performance					
Main effect of group	1, 70	2.99	1365.17	.09	.04
Main effect of item difficulty	1, 70	84.97	203.75	< .001	.55
Group * item difficulty	1, 70	2.59	203.75	.11	.04
Basic/labels vs. goal group					
Criterion reached					
Main effect of group	1, 102	< 1			
Main effect of item difficulty	1, 102	< 1			
Group * item difficulty	1, 102	1.85	.42	.18	.02
Final cued recall performance					
Main effect of group	1, 102	3.83	1487.03	.05	.04
Main effect of item difficulty	1, 102	107.54	189.79	< .001	.51
Group * item difficulty	1, 102	< 1			
Basic/labels/goal vs. plan group					
Criterion reached					
Main effect of group	1, 137	3.68	3.69	.05	.03
Main effect of item difficulty	1, 137	< 1			
Group * item difficulty	1, 137	< 1			
Final cued recall performance					
Main effect of group	1, 137	< 1			
Main effect of item difficulty	1, 137	150.38	178.82	< .001	.52
Group * item difficulty	1, 137	< 1			

these groups did not differ on either of the two variables of primary interest, we collapsed them for the final ANOVAs reported below.

Next, did participants reach a similar criterion for easy and difficult items because they started the task with an ineffective learning plan? Outcomes from the plan group suggest the answer is no. Participants in the plan group received item difficulty labels and the same learning goal as the goals group; additionally, they answered questions about what an *optimal* student should do to reach the learning goal and what they *plan* to do in the current learning task. In the right panel of Figure 4, the first set of bars shows that participants had a good idea of what would be an optimal criterion for easy and difficult items. Participants indicated that an optimal student should reach a higher criterion for difficult items than for easy items, $t(34) = 2.57, p = .02, d = .43$. Additionally, participants thought an optimal student should reach a criterion of about 5 for the difficult items. Results from the criterion group in Experiment 1 suggest this is what participants should do to correctly answer 70% of the difficult items on the final cued recall test. Thus, participants started out with a good idea of what would be an optimal plan.

Participants may have understood what an optimal student should do without ever planning to execute those choices themselves, which is why we also asked participants what they *planned* to do in the current learning task. In the right panel of Figure 4, the second set of bars shows a numerical trend toward planning to reach a higher criterion for difficult items, $t(34) = 1.68, p = .10, d = .28$. Although there was not a significant difference, participants were starting out with a relatively effective plan of reaching a high criterion for all items.

Given that participants planned to achieve multiple correct recalls during practice, did they fail to properly implement their plan during the actual task? Outcomes from the plan group indicate the answer is yes. Participants planned to reach a high criterion and believed that a higher criterion for difficult versus easy items would be optimal. In contrast, the final set of bars in the right panel of Figure 4 shows the actual criterion that participants in the goal group reached during the learning task. Participants reached a lower criterion than they planned for both easy items ($p < .001, d = 1.04$) and difficult items ($p < .001, d = 1.20$). Thus, participants did not execute their plan effectively. Additionally, participants in the plan group reached a similar criterion for easy and difficult items ($p = .89, d = .02$). Given that the actual criterion reached was similar for easy and difficult items, the finding that final cued recall performance was lower for difficult versus easy items is not surprising (Figure 5).

Exploratory Outcomes Participants in all groups reached a similar criterion for easy and difficult items, but it can also be helpful to compare the time spent practicing items to investigate whether this time differed across groups. Table 3 reports the mean practice time for easy and difficult items in each of the four groups (reported as seconds per item). Consistent with Experiment 1, and the broader literature on learning choices for easy and difficult items, participants in all groups spent more time practicing the difficult items compared to the easy items ($ps < .01, ds > .40$). Additionally, participants in all four groups spent a similar amount of time practicing difficult items ($ps > .25$). In other words, the different combinations of

item difficulty labels, a specific learning goal, and creating a learning plan did not increase the time that participants spent practicing the difficult items.

General Discussion

Retrieval practice is one of the most effective learning strategies for improving long-term memory, but research on how students use testing during self-regulated learning is scant. The current project is the first to examine students' self-testing choices for normatively easy-to-learn and difficult-to-learn items when students could test or reread items as many times as they would like. Across two experiments, we investigated two main research questions: Will students differentially regulate their use of self-testing for easy and difficult items? And how effective is students' use of self-testing for achieving long-term retention of both easy and difficult items? Each of these questions will be discussed in the following sections, starting with a summary of the relevant outcomes.

Do Students Differentially Regulate Their Use of Self-Testing for Easy and Difficult Items?

The answer to this question depends on how students' use of self-testing was measured. In particular, students did not differentially regulate their use of self-testing for easy versus difficult items in terms of *criterion reached*, but they did differentially regulate in terms of *number of test trials*. We will discuss each of these in turn.

Criterion Reached In Experiment 1, participants in the SRL group obtained a similar (and relatively low) criterion for easy and difficult items, and in Experiment 2, we investigated why they did not adjust criterion with three new SRL groups. These extension groups included different scaffolds that were aimed at minimizing potential barriers to effective regulation. For instance, if students had problems judging item difficulty or did not develop an explicit goal to perform well on the difficult items, then providing item difficulty labels or giving students an appropriate learning goal was expected to increase the criterion achieved for difficult items. In contrast to expectations, these scaffolds were not successful in encouraging students to reach a higher criterion for the difficult items. However, results from the plan group indicated that when participants were given item difficulty labels, a specific learning goal, and asked to create a plan for achieving this goal, participants started with an effective plan to practice items until they were correctly recalled multiple times, but they did not execute this plan during the learning task.

This finding may be surprising, particularly because Ariel and Karpicke (2018) found that informing participants of the benefits of testing items until they are correctly recalled three times was effective at getting participants to reach this suggested criterion. Why might these participants have effectively implemented the plan that was suggested to them, whereas participants in the current plan group did not effectively implement a plan they created on their own? Methodological differences may have contributed to this difference in findings. First, in the current experiments,

participants made their next learning choice immediately after completing a trial for an item. Imagine a participant who just retrieved a correct answer on a practice test trial—when participants are making their next learning choice immediately after this correct retrieval, they may be overconfident in their current level of knowledge and decide to drop the item after this first correct recall. In contrast, Ariel and Karpicke (2018) had participants make learning choices in choice blocks in which participants saw all the remaining items and selected which items to test again, restudy, or drop. If a participant tested an item and recalled the correct answer, there would be a delay before making another learning choice for that item, caused by cycling through the remaining items before reaching the next choice block. After this delay, participants may be less confident in their memory of an item and realize they could benefit from more test trials. Although this is speculative, it could be evaluated in future research. This difference in the timing of choices also represents another difference in how students made their learning choices. In the current experiments, participants made their learning choices for each item individually in a sequential format, but Ariel and Karpicke (2018) had participants make their learning choices for all items at the same time in a simultaneous format. Executing a study plan can be more difficult using a sequential format compared to simultaneous format (Ariel et al., 2009; Dunlosky & Thiede, 2004; Middlebrooks & Castel, 2018), which could be another contributing factor to why participants in the current experiment's plan group failed to properly execute the plan they created. Additionally, in the current Experiment 2, participants were learning all 40 items in one block. Learning 40 items at one time could make it difficult to track the criterion that has been reached for each item. Thus, even though participants planned to continue testing items until they reached three correct recalls, it may have been difficult for them to track their progress. In contrast, Ariel and Karpicke (2018) used blocks of 20 items. With these smaller blocks of items, it may be easier to mentally track the number of correct recalls, which likely contributed to their participants' success in executing the suggested plan.

Why might participants in the current research have planned to reach a higher criterion but did not execute their plan effectively? Participants may have been aware that testing would have a direct effect on their learning, but after starting the learning task, they may have used testing for a different purpose. In particular, students report using tests to monitor whether they have learned items and not because testing itself has a direct effect on learning; that is, they use testing as a monitoring tool instead of a learning tool (for a review of the accumulating evidence consistent with this hypothesis, see Rivers, 2020). If students' prepotent bias is to use testing as a monitoring tool, they presumably would use a practice test to evaluate whether they knew an item well enough to pass the final test. If the student answers the practice test trial correctly, it would be evidence that they had learned the item well enough and hence they would then remove that item from further practice. In the present case for participants in the plan group, they may have planned to reach a high criterion, but when they were engaged with the learning task, their prepotent bias to use testing as a monitoring tool may have overridden their initial plan (for a demonstration of prepotent biases overriding effective learning plans, see Dunlosky & Thiede, 2004).

Number of Test Trials Across all SRL groups, students chose more test trials for difficult items than for easy items. This finding may appear inconsistent with results from Toppino et al. (2018) and Tullis et al. (2018), who found that students chose self-testing more often for easy items than for difficult items. Tullis et al. (2018) stated, “Across five experiments, learners selectively utilized testing for easier items and reserved restudying for the more difficult items” (p. 550). The different conclusions may arise from a methodological difference between the current project and the prior studies. In the current project, participants in the SRL groups could choose as many additional learning trials as they wanted, whereas students in both prior studies could choose only one additional learning trial per item. On the initial trial, participants may choose to restudy difficult items, but when given the opportunity to continue engaging with an item across multiple trials, they may eventually test themselves more often on the difficult items. However, if participants are only allowed to engage with an item one time, this shift to testing difficult items would not be captured.

This possibility leads to a hypothesis that can be empirically evaluated by a post hoc analysis of the present data. On the initial trial for each item of the learning task, participants in the SRL groups should choose to restudy difficult items more often than easy items, which would differ from their choices to test difficult items more than easy items (reported in Figure 2 and Table 3).

To evaluate this prediction, we analyzed participants’ initial learning choices from both experiments and calculated the proportion of difficult items in which participants’ first choice was to restudy (or to test) the item. These proportions are presented in Table 5 separately for difficult and easy items. Participants’ restudy choices for difficult versus easy items are of primary interest, but test choices are also included for completeness. When analyzing participants’ initial choice for items, they restudied a higher proportion of difficult items than easy items (Experiment 1: $p < .001$, $d = .71$; Experiment 2: $p < .001$, $d = .48$). This leads to the largely redundant finding that participants also tested a lower proportion of difficult items than easy items (Experiment 1: $p < .001$, $d = .62$; Experiment 2: $p < .001$, $d = .31$).

The full results of the current experiments have shown that this initial preference for restudying difficult items is not representative of participants’ overall learning choices. Thus, the apparent differences in conclusions across studies can be attributed to a critical moderator: Students tend to restudy difficult items when they can either study or test themselves (but not both, as in the prior research), whereas when given the opportunity to engage with difficult items across multiple trials (in the current research), they begin by restudying and then eventually shift to using practice tests.

Why might this factor moderate students’ self-regulated use of testing for difficult items? One explanation is again offered by the hypothesis that students primarily use testing as a monitoring tool, which was discussed in the previous section. If students use testing to monitor their learning, they would wait to test themselves until they think they could answer the question correctly—that is, why check your knowledge if you know you will fail? When students are making their initial learning choice for a difficult item, they may judge that they do not know the item well enough after the initial presentation and hence choose to restudy that item. When students are

Table 5 Mean proportion of initial learning choices (restudy or test) for easy items and difficult items

	Easy items		Difficult items		
	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>	<i>p</i>
Experiment 1					
Restudy choice	.32	.04	.43	.04	< .001
Test choice	.58	.04	.48	.04	< .001
Experiment 2					
Overall					
Restudy choice	.38	.04	.48	.04	< .001
Test choice	.42	.03	.36	.03	< .001
Basic group					
Restudy choice	.46	.07	.54	.07	.01
Test choice	.37	.06	.30	.06	.03
Labels group					
Restudy choice	.26	.05	.38	.07	.005
Test choice	.52	.06	.42	.07	.05
Goal group					
Restudy choice	.27	.07	.34	.07	.02
Test choice	.45	.07	.43	.07	.74
Plan group					
Restudy choice	.51	.07	.66	.07	< .001
Test choice	.35	.07	.28	.06	.01

Mean proportions do not add up to 1 because some items were dropped on the initial trial

SE standard error of the mean

making their initial learning choice for an easy item, they may judge that they have already learned it well enough to pass the upcoming test and hence decide to evaluate whether they will succeed.

Do Students Effectively Regulate Their Use of Self-Testing for Easy and Difficult Items?

The first way to answer this question was comparing SRL participants' final memory performance on easy versus difficult items. Across all SRL groups, final memory performance was significantly lower for difficult items than for easy items, indicating that students do not effectively regulate their learning for difficult items. The second way to answer this question was comparing final memory performance in the SRL group to the criterion group in Experiment 1. The criterion group illustrated that participants in the SRL group could have performed better on the difficult items if they had practiced to a higher criterion during learning, which supports the same conclusion that students' use of testing was not effective for difficult items.

Participants' choices were ineffective for *single session* learning, which was used in the current set of experiments. However, if students made similar choices across multiple learning sessions, they would reach a higher overall criterion, and these choices could be effective for long-term retention (for an example of students' self-testing choices across multiple learning sessions, see Janes et al., 2018). Investigating how choices for easy and difficult items may change across learning sessions could be a promising direction for future research.

Closing Remarks

Students are responsible for regulating much of their learning, so revealing how they use various strategies—and whether they do so effectively—will be important for understanding how to potentially improve their achievement. In the present research, our focus was on how students use self-testing for normatively difficult-to-learn versus easy-to-learn items. Although difficult items require a higher criterion than easy items to achieve the same level of retention (Figure 3), participants on average did not regulate effectively and obtained a criterion of about one for both easy and difficult items—even when they thought that reaching a higher criterion would be beneficial. These outcomes are consistent with a small but growing body of evidence that indicates students do not fully capitalize on the benefits of practice testing and tend to use it mainly to monitor their learning. Important avenues for research involve discovering techniques that will encourage students to use testing more fully and to explore how students use testing in classroom contexts and with more complex materials.

Appendix 1. Descriptive statistics and ANOVA for criterion group from Experiment 1

Table 6, Table 7

Table 6 Descriptive statistics from criterion group in Experiment 1

	Easy items		Difficult items	
	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>
Test trials to reach criterion (per item)				
Criterion 1 items	2.8	.2	5.0	.4
Criterion 3 items	4.9	.3	7.8	.6
Criterion 5 items	7.2	.3	9.7	.6
JOL				
Criterion 1 items	51	3	44	3
Criterion 3 items	60	3	50	3
Criterion 5 items	81	3	54	3

Judgment-of-learning (JOL) magnitude could range from 0 (0% likely to correctly recall on the memory test 2 days from now) to 100 (100% likely to correctly recall)

SE standard error of the mean

Table 7 ANOVAs for final cued recall performance and final multiple-choice performance for the criterion group in Experiment 1

	<i>df</i>	<i>F</i>	MSE	<i>p</i>	η_p^2
Final cued recall performance					
Main effect of criterion ^a	1, 74, 86.94	118.34	497.11	< .001	.70
Main effect of item difficulty	1, 50	75.61	294.43	< .001	.60
Group * item difficulty	2, 100	28.66	184.12	< .001	.36
Final multiple-choice performance					
Main effect of criterion ^a	1, 80, 90.22	24.73	266.98	< .001	.33
Main effect of item difficulty	1, 50	92.48	320.17	< .001	.65
Group * item difficulty	2, 100	4.47	214.17	.01	.08

^aThe assumption of sphericity was violated and the Huyn-Feldt corrected values are reported

Appendix 2. Instructions and questions from plan group in Experiment 2. The information below was presented on separate screens for participants but is presented contiguously here for ease of reading

For the items you are going to study, 20 items will be **easy** and 20 will be **difficult**. On the final memory test **2 days** from now, your goal is to get **at least 70% correct** for the easy items and **at least 70% correct** for the difficult items.

Imagine a group of students who have been given this same goal, to learn the items well enough that they will be able to get at least 70% correct for the easy items and at least 70% correct for the difficult items on a final test two days from now. Answer the following questions in terms of what **you believe these students should do** in order to make sure the learning goal is achieved.

Do you believe these students should complete practice tests for the **easy** items? Yes/No

Do you believe these students should complete practice tests for the **difficult** items? Yes/No

Imagine that students decided to test themselves by seeing the Lithuanian word and trying to recall the English translation. How many times should they get the answer correct?

For **easy** items, they should do practice tests until they get the answer correct _____ times

Explain why:

For **difficult** items, they should do practice tests until they get the answer correct _____ times

Explain why:

Do you believe that students should restudy the **easy** items? Yes/No

Do you believe that students should restudy the **difficult** items? Yes/No

How many times should students restudy an **easy** item **before** they start taking practice tests for that item? ____

How many times should students restudy a **difficult** item **before** they start taking practice tests for that item? ____

How many times should students restudy an **easy** item **after** they start taking practice tests for that item? ____

How many times should students restudy a **difficult** item **after** they start taking practice tests for that item? ____

How should students decide when they are done learning an item and are ready to drop an **easy** item?

How should students decide when they are done learning an item and are ready to drop a **difficult** item?

Now, answer the following questions in terms of what **you plan to do** with the Lithuanian-English word pairs you will be learning today. Remember that your goal is to learn the items well enough that you will be able to get at least 70% correct for the easy items and at least 70% correct for the difficult items on a final test two days from now.

One option is to take a practice test for an item. When taking a practice test, you are shown the Lithuanian word and asked to practice recalling the English translation. After trying to recall the English word, you would be shown the right answer.

Do you plan to complete practice tests for the **easy** items? Yes/No

Do you plan to complete practice tests for the **difficult** items? Yes/No

Assuming you decide to test yourself by seeing the Lithuanian word and trying to recall the English translation, how many times do you want to get the answer correct?

For **easy** items, I plan to do practice tests until I get the answer correct ____ times

Explain why:

For **difficult** items, I plan to do practice tests until I get the answer correct ____ times

Explain why:

Do you plan to restudy the **easy** items? Yes/No

Do you plan to restudy the **difficult** items? Yes/No

How many times will you restudy an **easy** item **before** you start taking practice tests for that item? ____

How many times will you restudy a **difficult** item **before** you start taking practice tests for that item? ____

How many times will you restudy an **easy** item **after** you start taking practice tests for that item? ____

How many times will you restudy a **difficult** item **after** you start taking practice tests for that item? ____

How will you decide when you are done learning an item and are ready to drop an **easy** item?

How will you decide when you are done learning an item and are ready to drop a **difficult** item?

References

- Ariel, R., Dunlosky, J., & Bailey, H. (2009). Agenda-based regulation of study-time allocation: When agendas override item-based monitoring. *Journal of Experimental Psychology: General*, *138*, 432–447.
- Ariel, R., & Karpicke, J. D. (2018). Improving self-regulated learning with a retrieval practice intervention. *Journal of Experimental Psychology: Applied*, *24*, 43–56.
- Blasiman, R. N., Dunlosky, J., & Rawson, K. A. (2017). The what, how much, and when of study strategies: Comparing intended versus actual study behaviour. *Memory*, *25*, 784–792.
- Cortina, J. M., & Nouri, H. (2000). *Effect size for ANOVA designs*. CA: Sage.
- de Jonge, M., & Tabbers, H. K. (2013). Repeated testing, item selection, and relearning: The benefits of testing outweigh the costs. *Experimental Psychology*, *60*, 206–212.
- Dunlosky, J., & Ariel, R. (2011). Self-regulated learning and the allocation of study time. *Psychology of Learning and Motivation*, *54*, 103–140.
- 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., & Thiede, K. W. (2004). Causes and constraints of the shift-to-easier-materials effect in the control of study. *Memory & Cognition*, *32*, 779–788.
- 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.
- Grimaldi, P. J., Pyc, M. A., & Rawson, K. A. (2010). Normative multitrial recall performance, meta-cognitive judgments, and retrieval latencies for Lithuanian–English paired associates. *Behavior Research Methods*, *42*, 634–642.
- Hartwig, M. K., & Dunlosky, J. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement? *Psychonomic Bulletin & Review*, *19*, 126–134.
- 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.
- Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General*, *138*, 469–486.
- Kornell, N., & Bjork, R. A. (2008). Optimizing self-regulated study: The benefits and costs of dropping flashcards. *Memory*, *16*, 125–136.
- Middlebrooks, C. D., & Castel, A. D. (2018). Self-regulated learning of important information under sequential and simultaneous encoding conditions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*, 779–792.
- 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*(4), 437–447.
- Rivers, M. L. (2020). Metacognition about practice testing: A review of learners' beliefs, monitoring, and control of test-enhanced learning. *Educational Psychology Review*, 1–40.
- Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 204–221.
- Toppino, T. C., LaVan, M. H., & Iaconelli, R. T. (2018). Metacognitive control in self-regulated learning: Conditions affecting the choice of restudying versus retrieval practice. *Memory & Cognition*, *46*, 1164–1177.
- Tullis, J. G., Fiechter, J. L., & Benjamin, A. S. (2018). The efficacy of learners' testing choices. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*, 540–552.
- 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.
- Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 277–304). LEA.
- Wissman, K. T., Rawson, K. A., & Pyc, M. A. (2012). How and when do students use flashcards? *Memory*, *20*, 568–579.