Journal of Experimental Psychology:
Learning, Memory, and Cognition
2018, Vol. 44, No. 4, 540–552

© 2017 American Psychological Association
0278-7393/18/$12.00 http://dx.doi.org/10.1037/xlm0000473

The Efficacy of Learners' Testing Choices

Jonathan G. Tullis
University of Arizona

Joshua L. Fiechter and Aaron S. Benjamin
University of Illinois at Urbana-Champaign

Practice tests provide large mnemonic benefits over restudying, but learners judge practice tests as less effective than restudying. Consequently, learners infrequently utilize testing when controlling their study and often choose to be tested only on well-learned items. In 5 experiments, we examined whether learners' choices about testing and restudying are effective for improving subsequent memory performance. Learners studied a list of word pairs and chose which items to restudy and which to test. Some of learners' choices were honored (by assigning those items to the chosen activity) and some of learners' choices were dishonored (by assigning those items to the opposite study activity). Surprisingly, and in contrast with all work to date on the metacognitive monitoring of testing effects, honoring learners' testing choices consistently resulted in better memory performance than dishonoring choices. This effect occurred principally because learners often chose to restudy difficult items, and those items did not benefit from testing. The effectiveness of learners' choices about testing casts the metacognition of testing in a new light: learners may not appreciate the benefits of testing, but they do have an understanding of circumstances in which the benefits of testing are minimal.

Keywords: testing effect, retrieval practice, metamemory, metacognitive control

Retrieval benefits memory performance more than additional study of the same information; this phenomenon is known as the testing effect. The testing effect is reliably found across different types of material (Carpenter & Delosh, 2005; Roediger & Karpicke, 2006; Tullis, Finley, & Benjamin, 2013), across different memory tests (Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008; Butler & Roediger, 2007; Glover, 1989; Putnam & Roediger, 2012), and in classroom settings (Benjamin & Pashler, 2015; Leeming, 2002; McDaniel, Agarwal, Huelser, McDermott, & Roediger, 2011; Rowland, 2014). Research suggests, however, that learners underutilize testing when choosing study activities (Karpicke, 2009) because they fail to recognize the mnemonic benefits that testing provides (Karpicke & Roediger, 2008; Tullis et al., 2013). However, little prior research has explicitly tested whether and to what degree learners' testing choices impede learning in circumstances in which they choose their own study (and test) activities. In the current experiments, we examined whether learners' choices ultimately help or hinder their mnemonic performance compared to a variety of control conditions. In other words, while almost all prior research has focused on how accurately learners *monitor* the mnemonic impact of restudying and testing, here we analyze how effectively learners *control* their allocation of restudying and testing. In contrast with prior work, which has demonstrated an apparent reluctance to employ testing as a means of enhancing memory, learners in the current experiments benefited from the opportunity to control which items they would be tested on.

Learners' sophisticated use of metacognitive control, which is often based upon private access to personal knowledge (e.g., an individual's fluency of processing words; Benjamin, Bjork, & Schwartz, 1998), often boosts mnemonic performance. Learners effectively allocate study time across items (Tullis & Benjamin, 2011; Tullis, Benjamin, & Liu, 2014), effectively choose which items will benefit from restudy (Atkinson, 1972; Kornell & Metcalfe, 2006; cf. Tullis & Benjamin, 2012), distribute repetitions in time in a productive way (Benjamin & Bird, 2006; Son, 2004; Toppino, Cohen, Davis, & Moors, 2009), and tailor their encoding to fit the expected demands of an upcoming test (Finley & Benjamin, 2012). However, research suggests that learners inefficiently utilize retrieval practice (i.e., self-tests) to enhance their memory (Karpicke, 2009; Tullis et al., 2013). When asked whether they employ testing during self-controlled study, a minority of students responded affirmatively (Karpicke, Butler, & Roediger, 2009). Only 42% of students reported that they use self-testing while studying for exams, even when there is a chance for restudy after the self-test. In circumstances when restudy after a test is not allowed, only 18% of learners reported utilizing self-testing.

Researchers have argued that learners underutilize self-testing because of two failures of their metacognitive monitoring. First, learners fail to grasp the mnemonic benefits that testing provides relative to restudying, especially when they judge how well items are learned on an item-by-item basis. Learners almost always rate tested items as less memorable than restudied items (Agarwal et al., 2008; Tullis et al., 2013), even when provided corrective feedback after each practice test (Karpicke, 2009). In fact, in order to prompt learners to correctly assess tested material as better learned than restudied material, learners need comprehensive sum-

This article was published Online First November 2, 2017.

Jonathan G. Tullis, Department of Educational Psychology, University of Arizona; Joshua L. Fiechter and Aaron S. Benjamin, Department of Psychology, University of Illinois at Urbana-Champaign.

Correspondence concerning this article should be addressed to Jonathan G. Tullis, Department of Educational Psychology, 1430 E. Second Street, Tucson, Arizona 85721. E-mail: jonathantullis@gmail.com

mary feedback about their mnemonic performance that explicitly compares memory for previously restudied and previously tested material (Tullis et al., 2013). All of this evidence suggests that the mnemonic benefit provided by testing relative to restudy is not readily understood by the average learner.

Second, learners fail to appreciate the benefits of testing as a general strategy for learning. That is, not only do learners fail to monitor the benefits of retrieval over restudy following an experience with those procedures, most hold a theory of memory in which testing does not affect memory at all. Learners explain that they use testing only as a means of assessing their memory rather than improving it. When asked why they test themselves, two thirds of college students report that they test themselves only to figure out what they do and do not know (Kornell & Bjork, 2007; Kornell & Son, 2009). Only 20% of students report that they use testing because it causes them to learn more than restudying. Clearly, even when learners self-test, they do not recognize the benefits of testing as a strategy for remembering information. Further, when learners choose how many times they should test themselves, learners prematurely choose to drop items after a single successful practice test (rather than continue to practice their retrieval), and this dropping strategy impairs their ultimate recall (Kornell & Bjork, 2008). This evidence suggests that learners largely use testing as a means of assessing—rather than enhancing—learning, and will therefore stop self-testing once information is successfully retrieved.

Taken together, learners' misguided perceptions of self-testing, and their failure to appreciate the benefits of retrieval relative to restudy, suggest that they will use testing inefficiently during learning. That is, allowing learners to control the allotment of items to be tested (vs. restudied) may not benefit mnemonic performance at all, and may even negatively affect memory. Prior work has evaluated learners' choices about how many times to self-test when learners are already engaged in retrieval practice (Kornell & Bjork, 2008). However, little prior research has examined whether learners effectively choose when to apply self-testing, whether learners underutilize testing during self-regulated study, and whether giving learners control over testing upon initial encoding impairs their mnemonic performance. We, therefore, pose two major questions: (a) Even in the face of a clear lack of appreciation for the value of testing, can learners effectively choose which items to self-test? and (b) Can learners' memory benefit from utilizing testing more often than they spontaneously choose to?

When learners do utilize testing, they selectively test well-learned items. Learners choose to eliminate the best-learned items from their study list, restudy the worst-learned items, and test the items in between (Karpicke, 2009). Son (2005) showed that even first graders selectively choose to restudy the difficult items and be tested on the easiest items. Testing the easiest items is unlikely to enhance memory much for those items, but testing more difficult items, which have the most be to be gained by testing, may result in the best overall performance. Therefore, applying testing to well-learned items may reduce the overall mnemonic improvement that testing can provide.

Alternatively, when no feedback is provided during retrieval practice, testing only the well-learned items may be the most effective strategy because testing only benefits memory when learners successfully retrieve the target (Izawa, 1967; Kornell, Bjork, & Garcia, 2011; Landauer & Bjork, 1978; Storm, Friedman, Murayama, & Bjork, 2014). According to the bifurcation model of testing, memories for tested items that are successfully retrieved become stronger, but nonretrieved items receive no mnemonic benefits (Kornell et al., 2011). Restudying strengthens all items, although to a much lesser degree than does successful retrieval. Therefore, if a learner knows she will not successfully recall an item on a practice test, she should choose to restudy the item in order to receive the modest benefit that restudying provides. In other words, when initial learning is weak, it may be smarter to restudy than to test an item when no feedback is provided. Incorrectly answering a test query does not improve memory performance; in fact, in some cases, practice tests could lead to worse later performance. Errors of commission—in which learners retrieve incorrect material—may promote sustained deleterious effects (Henkel, 2007; Mathias, Marsh, & Dougherty, 2002; McDermott, 1996, 2006). A decision to be selective about what items to test may reduce errors of commission. When only testing the easiest or most well-learned items, learners guarantee they can correctly retrieve the targets and therefore benefit from the test (though that benefit may be minimal). In circumstances without feedback, by choosing to restudy the more difficult items, learners avoid the possibility of incorrectly answering an initial test query, guarantee another study opportunity for the item, and ensure they receive a modest mnemonic benefit resulting from an added exposure. Even though learners are not sensitive to the quite broad benefits of testing, they may well be sensitive to the cost of testing memory, relative to restudying, when an answer is not known. If so, selective testing choices might ensure mnemonic benefits for all items and result in better memory performance than alternative strategies, particularly when memory for items is weak following an initial presentation and the probability of recalling these items is low.

Throughout the experiments presented here, we investigated the effectiveness of learners' restudy and test choices for improving final mnemonic performance. We questioned whether learners' use of testing on only the easier items can be an effective strategy and whether learners can benefit from more broadly utilizing testing. Learners studied a list of heterogeneously difficult word pairs, chose whether to restudy or be tested on each item, restudied or were tested on the word pairs, and took a cued-recall test two days later on all of the word pairs. To assess the effectiveness of a learner's choices, we compared subsequent memory performance between honored and dishonored choices. Learners' choices were honored by giving them the study activity (restudy or test) that they chose; learners' choices were dishonored by presenting them with the activity opposite from the one they chose. The honor/dishonor paradigm has been used to test the effectiveness of other metacognitive choices and has revealed that learners are effective about choosing which items to restudy (Kimball, Smith, & Muntean, 2012; Kornell & Metcalfe, 2006; cf. Tullis & Benjamin, 2012) and when to space repetitions in time (Son, 2010; cf. Mulligan & Peterson, 2014). However, the mnemonic consequences of honoring or dishonoring learners' choices about which items to test have not been thoroughly explored.

Experiment 1

In the first experiment, learners studied a list of word pairs and were required to choose half of the items to restudy and half to test. The choices from half of each learner's categories were honored, and half were dishonored. Cued recall was compared between honored and dishonored items after a 2-day delay to determine whether learners effectively selected which activity to allocate to which items.

Method

Participants. Twenty-three introductory-level psychology students at the University of Illinois, Urbana-Champaign participated in exchange for partial course credit.

Materials. Forty word pairs were collected from the University of South Florida Word Association, Rhyme, and Word Fragment Norms (Nelson, McEvoy, & Schreiber, 1998). The word pairs were weakly associated, such that the cue had a small forward association to the target ($M = 0.02$, $SD = 0.006$). Variability in difficulty of word pairs was introduced for each subject by keeping half of the weakly associated pairs intact and randomly pairing the cues and the targets of the other half. This manipulation produced 20 unassociated word pairs and 20 associated word pairs for each subject. Unassociated word pairs are more difficult to remember than weakly associated word pairs (Arbuckle & Cuddy, 1969), and subjects are metacognitively aware of this difference in memorability (Koriat, 1997).

Design. The experimental design here, and in following experiments, was approved by the University of Illinois, Urbana-Champaign IRB (Protocol Number: 05653). The experiment utilized a 2 (study choice: restudy or test) × 2 (actual study activity) × 2 (item difficulty: associated or unassociated) within-subjects design. Subjects were required to select half of the items to restudy and half of these choices were honored; subjects were required to select the other half of the items to test and half of these choices were honored.

Procedure. Subjects completed the experiment, which was programmed with MATLAB using the Psychophysics Toolbox extensions (Brainard, 1997), on personal computers in individual rooms. Subjects first read detailed instructions about the procedure, including clear warnings about the nature of the practice tests (detailed below), the lack of feedback after tests, and the 2-day delay before the final memory test. Subjects were told they would study 40 total word pairs; they would choose to restudy exactly half (20 pairs) and to be tested on the other half of these pairs. Subjects then began the study and choice phase, in which each word pair was shown in the center of the screen for 1 s in 32-point, black Arial font. After each pair was shown for 1 s, it was removed from the screen and subjects chose whether to restudy or be tested on the specific word pair later. A tally of how many items had been chosen for each study activity and how many items remained was displayed during every choice screen so that subjects did not have to mentally keep track of what choices they had already made (cf. Tullis et al., 2012). Once a study choice had been used 20 times, subjects had to select the other choice for the remaining items.

After completing the study and choice phase, subjects restudied and were tested on the items. Half of each subject's choices were honored and half were dishonored. Subjects restudied 10 items they had chosen to restudy and 10 items they had chosen to be tested on; subjects were tested on 10 items they had chosen to be tested on and on 10 items they had chosen to restudy. The order of the items was randomized under the constraint that the same study activity was not used for more than three consecutive items. Before each item, subjects saw a 1-s warning about the type of study activity that was going to be next so that they could prepare for the activity. Subjects then either restudied the pair for 6 s or were given the cue and asked to type in the target item. Tests were self-paced and included no feedback about the correct answers. Subjects were dismissed when they finished the restudy and test trials, and returned 2 days later to take the final test. At test, cues were presented one at a time in a new random order and subjects were asked to type in the appropriate target before moving on to the next item.

Results

In this and all subsequent experiments, subjects' responses were considered correct only if they matched the target exactly. Subjects chose to restudy 76% ($SD = 13\%$) of the unassociated word pairs and to be tested on 76% ($SD = 13\%$) of the associated word pairs, indicating a preference, consistent with prior research, to reserve testing for easier materials. Only one subject chose to be tested on more unassociated than associated pairs. During the initial test on the first day, learners recalled more targets from the pairs that they chose to be tested on than from the pairs they chose to restudy ($M_{\text{tested}} = 0.64$, $SD = 0.19$; $M_{\text{restudy}} = 0.29$, $SD = 0.15$; $t(22) = 8.76$, $p < .001$, Cohen's $d = 1.87$).

Cued recall performance is displayed in Figure 1, and the comparison between honor and dishonor conditions is shown in Table 1. Cued recall performance for honored items was higher than for dishonored items, $t(22) = 2.51$, $p = .02$; 14 participants had higher performance on honored items and only five had higher performance on dishonored items. Further, a 2 (study activity choice) × 2 (actual study activity) repeated measures ANOVA on cued recall performance revealed a significant interaction between choice and actual study activity, $F(1, 22) = 6.31$, $p = .02$, $\eta^2_{\text{partial}} = 0.23$, a main effect of learners' choices, $F(1, 22) = 43.96$, $p <$

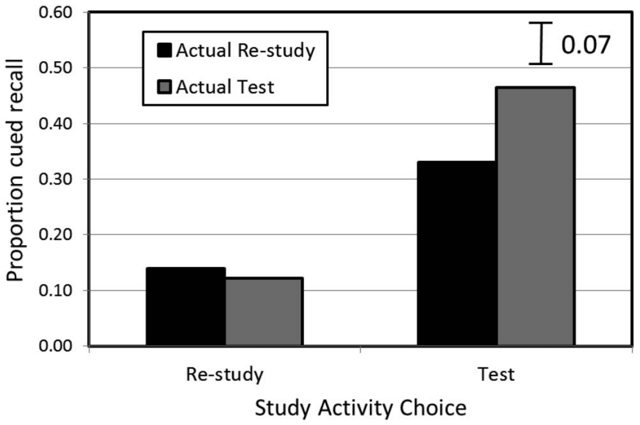

Figure 1. Cued recall performance in Experiment 1 conditionalized upon study choice and actual study activity. The width of the error bars here and on all subsequent graphs indicate the within-subjects 95% confidence interval across conditions (Loftus & Masson, 1994).

Table 1
Means, Standard Deviations, Cohen's d Effect Sizes, and the Number of Subjects Showing Better Performance When Choices Were Honored or Dishonored for the Final Memory Performance Across all Honored and Dishonored Items in Experiments 1–5

Experiment	Honor	Dishonor	Cohen's *d*	Subjects w/hon > dis	Subjects w/dis > hon
Experiment 1	.30 (.12)	.23 (.14)	.54	14	5
Experiment 2	.25 (.17)	.22 (.17)	.29	26	14
Experiment 3	.35 (.18)	.23 (.15)	1.09	57	13
Experiment 4	.25 (.18)	.18 (.18)	.53	38	10
Experiment 5	.35 (.22)	.26 (.23)	.63	45	16

.001, $\eta^2_{\text{partial}} = 0.67$, and a main effect for actual study activity, $F(1, 22) = 5.00$, $p = .04$, $\eta^2_{\text{partial}} = 0.19$. Follow-up *t* tests indicated that tested items were recalled at a higher level than restudied items for items chosen to be tested, $t(22) = 2.73$, $p = .01$, $d = 0.58$, but no difference in recall was found for items chosen to be restudied, $t(22) = 0.63$, $p = .63$, $d = 0.13$. Further, we examined whether practice recall performance was related to final recall performance, with the understanding that easier items are more likely to be recalled on both the practice and final tests. For items correctly recalled on the practice test, 62% were recalled on the final test; for items not correctly recalled during the practice test, only 1% were recalled on the final test. Finally, we examined how long subjects spent on practice tests that they chose compared with those that they did not choose (all restudy opportunities were 6 s long). Subjects spent less time on average for practice tests that they chose ($M = 6.44$ s, $SD = 2.57$) than practice tests that they did not choose ($M = 9.81$ s, $SD = 3.00$; $t(22) = 6.28$, $p < .001$, $d = 1.34$).

Discussion

Learners chose to be tested on the easier items and to restudy the more difficult items. Consequently, during the practice test, learners recalled significantly more items that they had chosen to be tested on than on those chosen to be restudied. Learners' choices about study activities reflect how well each item is learned, as shown in prior studies (Karpicke, 2009; Son, 2005).

Honoring learners' choices significantly improved final memory performance. More honored items were recalled than dishonored items, which suggests that learners employed effective strategies when allocating restudy and testing among the items. More specifically, honoring or dishonoring restudy choices did not affect ultimate performance, but honoring or dishonoring testing choices did. The items chosen to be tested showed a testing effect: they were recalled better than restudied items of roughly the same difficulty (though the lack of experimental control over individual allocation policies makes an exact comparison impossible). The lack of an effect of honoring or dishonoring restudy choices may indicate that restudy trials did not provide substantial, long-lasting mnemonic benefits. Restudying a poorly learned item should provide mnemonic benefits that failed retrieval attempts do not; however, the mnemonic benefits of an additional study may be so small in our study that restudy does not boost memory any more than a failed retrieval attempt with no feedback.

When learners wanted to restudy an item, testing did not improve cued recall for that item. Analogously, the mnemonic benefits of spaced presentations over massed presentations are only apparent on items that learners choose to space (Son, 2010). However, further research revealed that spacing benefits learning regardless of learners' metacognitive choices and the lack of a spacing effect in Son's study was due to item selection effects (Mulligan & Peterson, 2014). Here, we argue, the lack of a testing effect for items chosen to be restudied similarly arose because those items were more difficult and the practice test queries for those items resulted in failed retrievals, which did not improve later cued recall (as only 1% of items failed on the practice test were retrieved on the final test). Testing, like spacing, may only be advantageous when learners have sufficiently encoded the stimuli during initial study. Through their conservative choices, learners appeared to grasp the concept that testing without feedback is of little value when memory for items is weak, and allocated their study strategies accordingly. Learners appeared to recognize the limitations of testing by only choosing to test items that are likely to be recalled. Allowing them to control their study strategies was therefore beneficial to mnemonic performance.

Honoring learners' testing choices led to significant improvements in recall, whereas honoring learners' restudy choices did not ultimately affect recall. Successful retrievals on practice tests engender large and long-lasting mnemonic benefits compared with restudying; consequently, allowing learners to engage in retrieval on those items that can be retrieved is essential.

Experiment 2

Experiment 1 suggests that learners' metacognition about when to selectively apply retrieval practice is somewhat effective. However, constraining learners' choices by requiring half of the items to be restudied and half to be tested may have altered learners' preferred study strategies. Those artificial constraints ensured a greater degree of compatibility between the conditions for analysis, but may have forced subjects to make choices they did not want to make and to use strategies they would not have freely chosen. For some items, learners may have chosen testing only because that choice was required. In the subsequent experiments, learners were allowed to choose to restudy and test as many word pairs as they desired. Further, in Experiment 2, learners could choose to be done with a word pair instead of restudying or testing it. If learners selected the *done* option, the word pair was neither restudied nor tested. Prior literature suggests that learners utilize the done option primarily for items that are easy or judged to be well-learned (Son, 2004; Son, 2010; Toppino & Cohen, 2010; Toppino, Cohen, Davis, & Moors, 2009); in addition, learners are overconfident in their mnemonic abilities and choose to be done studying even though additional study would improve their performance (Kornell, 2009; Kornell & Bjork, 2008). Granting learners this extra freedom helps emulate the large amount of control learners exercise when guiding their own study.

Method

Participants. Forty-seven subjects participated in this experiment. Twenty-nine introductory-level psychology students at Indiana University and 18 introductory-level psychology students from the University of Illinois, Urbana-Champaign participated in exchange for partial course credit.

Design. Subjects chose between three study options for each word pair: done, restudy, and test. Half of subjects' restudy choices were honored (subjects restudied the items) and half of the restudy choices were dishonored (subjects were tested on the items). Similarly, half of subjects' test choices were honored (subjects were tested on the items) and half were dishonored (subjects restudied the items). All *done* choices were honored and subjects did not see those items again until the final test. The word pairs were the same as in Experiment 1.

Procedure. In addition to the previous approval from the University of Illinois, we obtained approval of the experimental procedures from the Indiana University IRB (Protocol #: 0801000097: 05–9950). The experiment followed almost exactly the same procedure as in Experiment 1. Two differences were that subjects could choose to be "done" with a word-pair and did not have to select exactly half of the pairs to restudy and half to test. On individual computers, subjects read detailed instructions about how they could choose to be done, restudy, or be tested on each of the 40 word pairs. With regards to the "done" option, subjects were instructed: "DONE items will not be shown again at all" and "For DONE items, you will not get any further exposure to the items until they are tested during the final memory test." No instructions were given about when or why subjects should use this option. No restrictions were placed on the amount of each option subjects could choose. If learners chose an odd number of items to be tested or restudied, the program was biased to honor one more choice than to dishonor.

Results

Subjects' study choices are displayed in Figure 2. Subjects chose more unassociated pairs to restudy (M = 0.57, SD = 0.33) than associated pairs (M = 0.31, SD = 0.31; $t(46)$ = 6.22, $p <$ .001, d = 0.92). Conversely, subjects chose to test themselves on more associated pairs (M = 0.47, SD = 0.34) than unassociated pairs (M = 0.22, SD = 0.26; $t(46)$ = 4.93, $p <$ .001, d = 0.76). Overall, subjects chose to be done with similar proportions of associated (M = 0.22, SD = 0.30) and unassociated (M = 0.21, SD = 0.28) word pairs, $t(46)$ = 0.26, p = .80, d = 0.01. During the initial test during the first day, subjects recalled more of the items they chose to be tested on (M = 0.60, SD = 0.31) than those they chose to restudy (M = 0.35, SD = 0.27; $t(37)$ = 6.11, $p <$ .001, d = 1.00). As before, subjects chose to be tested on the easier items and restudy the more difficult items. Additionally, subjects spent more time per test engaged in practice tests that they did not choose (M = 8.69 s, SD = 3.32) than on practice tests that they chose (M = 6.80 s, SD = 2.80; $t(37)$ = 3.73, $p <$ .001, d = 0.61); only eight subjects spent more time on practice tests that they chose than those they did not choose.

Twenty-seven subjects utilized the *done* option; we honored all of these choices, and performance on these items was very low (M = 0.07, SD = 0.10). As shown in Table 1, for the items chosen to be restudied or tested, cued recall performance on honored items was higher than on dishonored items, $t(45)$ = 1.98, p = .05. Final cued recall performance conditionalized upon choice and actual study activity is displayed in Figure 3. A 2 (actual study activity: restudy vs. test) × 2 (study choice: restudy vs. test) repeated measures ANOVA on final cued recall performance revealed a significant interaction, $F(1, 35)$ = 8.66, p = .006, η^2_{partial} = 0.20, a main effect of learners' choices, $F(1, 35)$ = 28.80, $p <$ .001, η^2_{partial} = 0.45, and a main effect of actual study activity, $F(1, 35)$ = 5.63, p = .02, η^2_{partial} = 0.14. Follow-up t tests showed that testing led to better cued recall performance than restudying for items selected to be tested, $t(38)$ = 3.98, $p <$ .001, d = 0.65, but not for items selected to be restudied, $t(42)$ = 0.27, p = .79, d = 0.04. The interaction between study choice and actual study activity replicated Experiment 1 and indicated that honoring a learner's choices improves performance. Testing largely benefited learners when they chose testing, but did not benefit learners when they chose restudying. Twenty-six subjects performed better when their choices were honored than dishonored; 14 subjects performed better when their choices were dishonored. Similar to Experiment 1, 59% of items recalled on the practice test were recalled on the final test, but only 2% of items not recalled on the practice test were ultimately recalled.

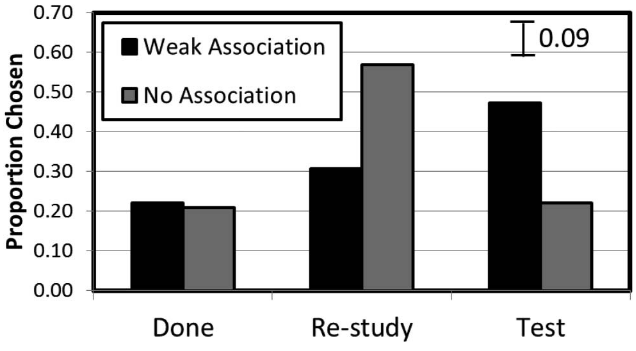

Figure 2. Proportion of word pairs selected for each study activity in Experiment 2 conditionalized upon the strength of association.

Discussion

As in the prior study, learners were conservative about their testing choices, choosing to restudy the difficult items and be tested on the easy items. Even when given full control over their study activity allocation, learners chose to be tested on more of the associated items than the unassociated items. This decision enabled learners to recall more of the items chosen to be tested on the initial test. Testing on the easier items was an effective strategy: honoring learners' choices resulted in greater recall than dishonoring those choices. Specifically, honoring learners' testing choices boosted recall performance on those items, but honoring restudy choices did not greatly change learners' ultimate mnemonic performance. This suggests that learners selectively and effectively utilized testing to benefit performance. A lack of testing

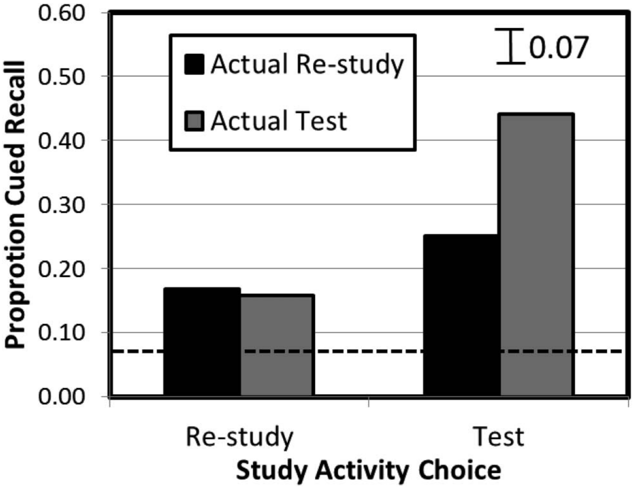

Figure 3. Cued recall performance in Experiment 2 conditionalized upon study choice and actual study activity. The dashed line indicates performance on items chosen to be "done."

effect was found for items chosen to be restudied. As before, this lack of an effect may reflect the fact that more difficult items were chosen for testing, that learners had a low rate of retrieval for these items on the practice test, that no feedback was provided, and finally that items not retrieved during the practice test were not subsequently recalled on the final test.

Learners performed very poorly on items chosen for the done option, as in prior research (e.g., Son, 2004). However, unlike prior research, learners did not selectively utilize the done option for the easier, associated items. Several plausible explanations exist, though they cannot be evaluated with the data at hand. Perhaps the difference in difficulty between the classes of items was not as dramatic as in other work. There may be idiosyncratic effects in our sample that led subjects to judge the items they selected to be done with as easy, regardless of their putative difficulty class. Alternatively, learners may have judged some items as too difficult to learn within the context of the long study list and abandoned them to reduce wasted time and effort (see Metcalfe & Kornell, 2005). Research that collects learners' judgments of learning could help distinguish between these explanations. And, of course, the lack of a difference could simply be a false negative error.

Although learners used testing choices effectively, learners may nonetheless underutilize testing. That is, even though learners appreciate the limitations of testing, they may not optimally discriminate between items that are encoded strongly enough for testing and items that would benefit from restudy. In the third experiment, we investigated whether learners underutilize testing during their learning. Some argue that testing needs to be applied more broadly than learners choose in order to maximize memory (Agarwal et al., 2008). We explored this idea in Experiment 3 by testing a subset of the word pairs during retrieval practice regardless of subjects' study choices. Performance on this subset of items relative to the subset of items on which subjects' choices were honored should reveal whether learners underutilize testing. Additionally, we included a condition in which study activities were randomly assigned regardless of learners' study choices for those items (similar to Kornell & Metcalfe, 2006). Performance in the random condition relative to the honor condition indicates whether learners' allocation of restudy and test choices is better than a purely random allocation of the same quantity of restudy and test choices, and is a more conservative baseline to use as a control group than the dishonor group.

Experiment 3

Method

Participants. Seventy subjects participated in Experiment 3. Twenty-six introductory-level psychology students at the University of Illinois, Urbana-Champaign and 44 introductory-level psychology students at Indiana University participated in exchange for partial course credit.

Materials. Forty new word pairs from the University of South Florida Word Association, Rhyme, and Word Fragment Norms were added to the 40 used in the prior two experiments. Variability in the difficulty of word pairs was introduced for each subject by randomly pairing half of the cues with other, unassociated target items, just as in the prior two experiments. Each subject studied 40 unassociated word pairs and 40 associated word pairs.

Design. Twenty word pairs were assigned to each of four different within-subjects conditions. The first two conditions were the same as the prior two studies: Learners' choices were either honored or dishonored. The third condition was a random condition. For the random condition, once all of a learner's choices were made, the program tallied the number of test choices and restudy choices across the subset of items in the random condition. The program then randomly assigned the test and restudy choices across the subset of pairs in that condition. Therefore, learners restudied the same number of items that they chose to restudy, but the specific restudied items were randomly selected. Similarly, learners were tested on the same number of items that they chose to be tested on, but which exact items they were tested on were assigned randomly. Finally, in the fourth subset of word pairs, learners were tested on all items, regardless of their study choices. These four within-subject conditions were assigned to individual word pairs before learners made their choices, such that a random assortment of 10 high association and 10 low association word pairs were assigned to each condition. The items from the four conditions were randomly ordered throughout the study and test lists.

Procedure. Subjects completed the experiment just as in Experiment 1, but their choices were not restricted to half restudy and half test. In order to ensure there were 20 word pairs in each condition, subjects did not have a "done" option in this experiment.

Results

Subjects chose to test more of the associated word pairs than unassociated word pairs, as in the prior two experiments. Subjects chose to test themselves on 86% (SD = 16%) of the associated pairs and 46% (SD = 31%) of the unassociated word pairs. Three subjects chose to be tested on all of the items, and no subjects chose to restudy all of the items. During the initial test on the first day, subjects recalled more targets from the pairs that they chose to be tested on than to restudy (M_{tested} = 0.51, SD = 0.21,

$M_{\text{restudy}} = 0.20$, $SD = 0.14$, $t(66) = 20.22$, $p < .001$, $d = 2.45$). As in the prior two experiments, subjects chose to self-test on the easier items. Further, subjects spent more time per test engaged in practice tests that they did not choose ($M = 8.79$ s, $SD = 3.77$) than practice tests that they did choose ($M = 6.48$ s, $SD = 1.82$; $t(66) = 5.36$, $p < .001$, $d = 0.66$); only 10 (out of 70) subjects spent more time on the practice tests that they chose than the ones they did not choose.

Cued recall performance across the four different conditions is displayed in Figure 4. A repeated measures ANOVA revealed a significant effect of condition, $F(3, 207) = 5.13$, $p = .002$, $\eta^2_{\text{partial}} = 0.07$. Across conditions, learners showed greatest cued recall in the honor condition and worst cued recall in the dishonor condition. Tukey post hoc tests showed that both the honor and test-all conditions significantly outperformed the dishonor condition (at $p < .05$), but no other pairwise comparisons reached significance.[1]

Across all four conditions, we analyzed performance when learners' choices were honored and dishonored and the results are shown in Table 1. Learners recalled more when their choices were honored than when their choices were dishonored, $t(69) = 9.07$, $p < .001$. Cued recall performance across the four conditions and conditionalized upon actual study activity is displayed in Figure 5. Five subjects performed best in the dishonor condition, six performed best in the random condition, 15 performed best in the test-all condition, and 23 performed best in the honor condition.

As in the prior two experiments, a 2 (study activity choice) × 2 (actual study activity) repeated-measures ANOVA on cued recall performance revealed a significant interaction between choice and actual study activity, $F(1, 59) = 18.81$, $p < .001$, $\eta^2_{\text{partial}} = 0.24$. Further, cued recall performance was higher for items chosen to be tested than restudied, $F(1, 59) = 269.30$, $p < .001$, $\eta^2_{\text{partial}} = 0.82$, and higher for items actually tested than restudied, $F(1, 59) = 3.90$, $p = .05$, $\eta^2_{\text{partial}} = 0.06$. Follow-up t tests showed that, as in prior studies, items chosen to be tested benefited from testing ($M_{\text{test}} = 0.44$, $SD = 0.25$) compared with restudying ($M_{\text{restudy}} = 0.36$; $t(69) = 2.84$, $p = .006$, $d = 0.34$), whereas restudying ($M = 0.14$, $SD = 0.16$) and testing ($M = 0.14$, $SD = 0.19$) were equally effective for items chosen to be restudied, $t(69) = 0.61$, $p = .55$, $d = 0.07$. Items correctly recalled during the practice test showed high levels of recall on the final test ($M = 78\%$), but items not recalled during the practice test showed very low levels of recall on the final test ($M = 2\%$).

Combined Analysis

Finally, we examined the combined data from all subjects across the first three experiments. Subjects recalled more when their choices were honored ($M = 0.31$, $SD = 0.17$) than dishonored ($M = 0.22$, $SD = 0.16$; $t(138) = 7.30$, $p < .001$, $d = 0.62$). A 2 (study activity choice) × 2 (actual study activity) repeated-measures ANOVA on final cued recall performance showed a significant interaction, $F(1, 118) = 29.89$, $p < .001$, $\eta^2_{\text{partial}} = 0.20$, an advantage for items chosen to be tested, $F(1, 118) = 238.94$, $p < .001$, $\eta^2_{\text{partial}} = 0.67$, and an advantage for testing over restudying, $F(1, 118) = 13.76$, $p < .001$, $\eta^2_{\text{partial}} = 0.10$. Follow-up paired t tests revealed that testing improved memory for items chosen to be tested, $t(131) = 5.97$, $p < .001$, $d = 0.51$, but not for items chosen to be restudied, $t(125) = 1.11$, $p = .27$, $d = 0.07$.

Discussion

As in the previous experiments, subjects chose to test themselves on the easier items and to restudy the difficult items. This choice enabled learners to recall more of their chosen items during the retrieval practice phase. Testing on the easier items proved once again to be an effective strategy: Honoring learners' choices resulted in greater recall than dishonoring those choices. Even testing all of the word pairs did not improve final mnemonic performance more than just testing the items learners selected. These results suggest that learners recognize the limitations of testing in situations with no feedback and control their study such that only sufficiently well mastered items (in this case, the easier items) are selected for testing, thereby ensuring that items that will be successfully recalled during testing achieve the most gain.

An important question remains: Do learners choose to self-test effectively when they will receive feedback about the correct answers after each retrieval practice trial? In the three prior experiments, if learners did not correctly retrieve the target, they did not get any reexposure to the correct answer. When learners choose to self-test during study in the real world (e.g., flashcards or practice tests), feedback about the correct answers is often present. Therefore, in the last two experiments, we included feedback after practice tests. In addition to being an additional study opportunity, feedback after practice tests may change the effectiveness of learners' choices because an attempted but failed retrieval can potentiate memory of the target item more than restudy (Arnold & McDermott, 2013; Kornell, Hays, & Bjork, 2009). Learners may benefit more by self-testing on the more difficult items than the easier items because the more difficult items could receive the greater potentiation by failed retrieval attempts followed by feedback than successful retrieval attempts of easier items.

Experiment 4

The three previous experiments showed that learners selectively and effectively choose self-testing to improve final mnemonic performance. Experiment 4 explores this same question in conditions where feedback is provided during practice tests.

Method

Participants. Fifty-six introductory-level psychology students at Indiana University participated in exchange for partial course credit. Using the average effect size for comparing honoring choices to dishonoring choices from Experiments 1–3, we conducted a power analysis to determine the needed sample size with alpha at 0.05 and high power (power = 0.95).

Materials. Sixty-four weakly associated word pairs were collected from the University of South Florida Word Association,

[1] It is interesting to note that the honor condition did not significantly out-perform the random condition. The performance in the random condition must lie between the dishonor and the honor conditions, because it is a mix of those two conditions. Within the random condition of Experiment 3, 65% of subjects' choices were honored (due to subjects' unequal selections of restudy and practice test trials). When subjects' choices were honored within the random condition, their recall performance ($M = .32$, $SD = .20$) was greater than when their choices were dishonored ($M = .23$, $SD = .20$; $t(61) = 3.57$, $p < .001$, $d = 0.47$).

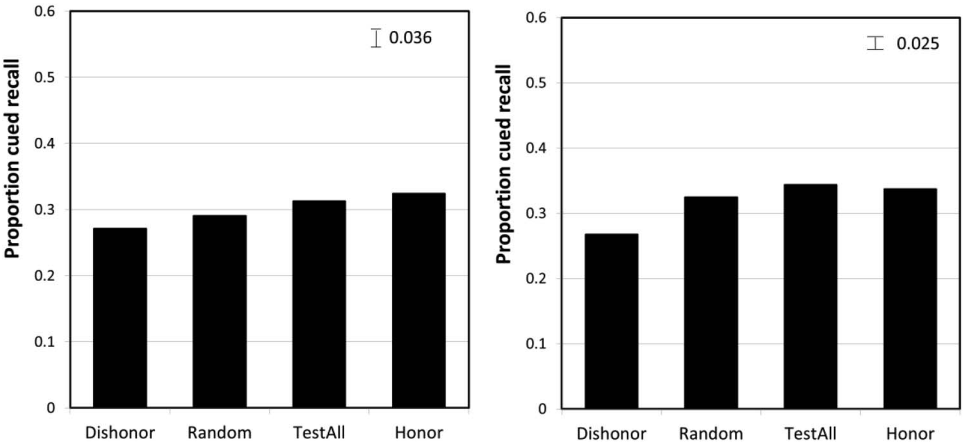

Figure 4. Proportion of word pairs recalled for each of the four conditions in Experiment 3 (left panel, no feedback) and Experiment 5 (right panel, with feedback).

Rhyme, and Word Fragment Norms (Nelson, McEvoy, & Schreiber, 1998), and, as in the previous experiments, half were studied intact and half were randomly paired.

Design. The experiment utilized a 2 (study choice: restudy or test) × 2 (actual study activity) × 2 (item difficulty: associated or unassociated) × 2 (half-required or unlimited choices) mixed design. Like Experiment 1, subjects in the *half-required* condition were required to select half of the items to restudy; subjects were required to select the other half of the items to test. Like Experiments 2 and 3, subjects in the *unlimited-choices* condition could choose to self-test as many word pairs as they wanted. Half of subjects' choices were honored in each condition. The half-required condition was included to ensure that subjects chose adequate numbers of items to restudy and to test.

Procedure. The procedure for this experiment was similar to that of Experiment 1. Subjects studied 64 word pairs presented in the middle of the computer screen and chose to restudy or test each pair. Half of subjects' choices were honored and half were dishonored. One group of subjects (the half-required group) had to choose exactly half of the items to self-test and half to restudy. The other group of subjects (the unlimited-choices group) could choose as many self-test and restudy attempts as they wanted. The only significant change from prior experiments was the addition of feedback after each practice test. When subjects entered their responses for each practice test, they were shown the correct word pair for as long as they wanted to view it. The presence of feedback after each practice test was emphasized multiple times during the initial instructions.

Results

Subjects in the unlimited group chose to be tested on 53% (SD = 26%) of the word pairs, which did not reliably differ from the mandatory 50%-selection policy in the half-required group, $t(29) = 0.56$, $p = .58$. Across every analysis, consistent patterns were found between the half-required and unlimited choices groups and there were no significant main effects or interactions with condition. Therefore, to simplify reporting, we present analyses combined across these conditions. Subjects chose to self-test more frequently on associated word pairs ($M = 0.67$, $SD = 0.25$) than unassociated word pairs ($M = 0.36$, $SD = 0.24$; $t(55) = 7.22$, $p < .01$, $d = 0.97$). Only two subjects (out of 30) in the unlimited choices group chose not to self-test on any items, and no subjects chose to self-test all items. During the initial practice test, subjects recalled more items that they selected to self-test ($M = 0.49$, $SD = 0.31$) than those they selected to restudy ($M = 0.21$, $SD = 0.17$;

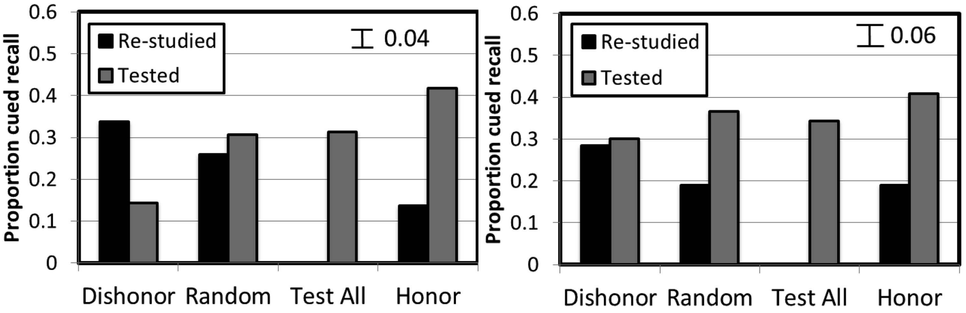

Figure 5. Proportion of word pairs recalled for each of the four conditions in Experiment 3 (left panel) and Experiment 5 (right panel), as a function of actual study activity. Experiment 3 did not have corrective feedback, while Experiment 5 did.

$t(52) = 7.81$, $p < .01$, $d = 1.08$). Sixty-nine percent of the items correctly recalled during the practice test were recalled during the final test; even with corrective feedback, only 10% of items not correctly recalled during the practice test were subsequently recalled.

Final cued recall performance is displayed in Figure 6, and the comparison across honor and dishonor conditions is shown in Table 1. Honoring subjects' study choices led to greater final cued recall than dishonoring their choices, $t(55) = 3.91$, $p < .01$. In fact, 38 subjects performed better when their choices were honored; only 10 subjects performed better when their choices were dishonored. Further, a 2 (study activity choice) × 2 (actual study activity) repeated-measures ANOVA on final cued recall performance revealed a significant interaction between choice and actual study activity, $F(1, 51) = 21.08$, $p < .001$, $\eta^2_{\text{partial}} = 0.29$, a main effect of learners' choices, $F(1, 51) = 46.92$, $p < .001$, $\eta^2_{\text{partial}} = 0.48$, and a main effect for actual study activity, $F(1, 51) = 67.11$, $p < .001$, $\eta^2_{\text{partial}} = 0.57$. Follow-up t tests showed that testing led to better cued recall performance than restudying both for items selected to be restudied, $t(54) = 4.84$, $p < .001$, $d = 0.66$ and for items selected to be tested, $t(52) = 7.18$, $p < .001$, $d = 1.00$. Finally, we examined the time spent during practice tests and restudying across the two honor conditions. Subjects spent similar amounts of total practice time (including practice tests, feedback, and restudy time) in the dishonor condition ($M = 182$ s, $SD = 133$) as in the honor condition ($M = 164$ s, $SD = 101$; $t(55) = 1.33$, $p = .19$, $d = 0.18$).

Discussion

Some of the central results from prior experiments, which did not include feedback during retrieval practice, replicated in Experiment 4, which did include feedback during retrieval practice. First, subjects chose to self-test more frequently on easy word pairs than on difficult word pairs. Second, honoring subjects' choices led to greater final cued recall performance than dishonoring those choices. Practice tests on the easier items led to greater final memory performance than practice tests on the more difficult items, even when feedback was provided after every test. The benefits of honoring subjects' choices were driven by a larger testing effect for the pairs chosen to be tested than for the pairs chosen to be restudied.

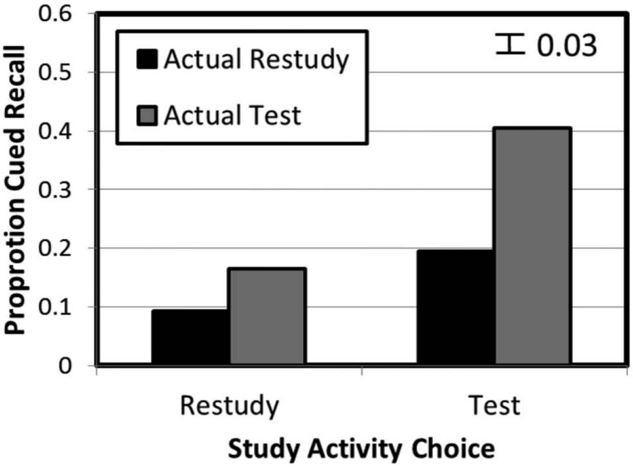

Figure 6. Cued recall performance in Experiment 4 conditionalized upon study choice and actual study activity.

A major difference between the results from Experiment 4 and the prior experiments is that testing led to greater performance than restudying even for pairs that were chosen to be restudied. When subjects have feedback, a failed practice test provides an additional study opportunity, may potentiate learning from feedback (Arnold & McDermott, 2013), and ultimately provides larger mnemonic benefits than restudying. In the final experiment presented here, we test whether testing subjects on all items benefits memory more than honoring testing selections in situations where feedback is provided. If testing with feedback is more beneficial than restudying in all cases, more testing should always be advantageous, as has been argued previously (Agarwal et al., 2008).

Experiment 5

Method

Participants. Participants for Experiment 5 were recruited online through Amazon Mechanical Turk. They were paid $1.50 for completing the first day and an additional $1.50 for completing the second day. Eighty subjects started participation, but only 63 subjects completed both the first and second days. We analyzed the data only from subjects who completed both days.

Design. The design, materials, and procedure replicated those of Experiment 3, but participants were given feedback after every practice test. Procedure of Experiment 5 was approved by the University of Arizona IRB (15–007-EDP). As in Experiment 3, participants studied 80 word pairs, with 20 word pairs assigned to each of four different within-subjects conditions (honor, dishonor, random, and test-all). Participants were explicitly told that they would get corrective feedback about the correct word pair after each practice test several times throughout the instructions. After participants typed their answer to each practice test, they were shown the correct word pair for as much time as they wanted to study.

Results

As in all prior experiments, subjects chose to test themselves on more of the associated pairs ($M = 74\%$, $SD = 33\%$) than unassociated pairs ($M = 57\%$, $SD = 42\%$; $t(62) = 4.51$, $p < .001$, $d = 0.57$). Eighteen subjects chose to be tested on all of the items, and four subjects chose to restudy all of the items. During the initial test on the first day, subjects recalled more targets from the pairs that they chose to be tested on than the pairs they chose to restudy ($M_{\text{tested}} = 0.48$, $SD = 0.24$; $M_{\text{restudy}} = 0.30$, $SD = 0.25$; $t(37) = 4.09$, $p < .001$, $d = 0.67$). Even though corrective feedback was provided, subjects chose to self-test more frequently on the easier items, as in all prior experiments. Further, of those items that were correctly recalled on the practice test, 65% were recalled on the subsequent final test; of those items not correctly recalled on the practice test, only 17% were recalled on the final test.

Cued recall performance across the four different conditions is displayed in right panel of Figure 4. A repeated-measures ANOVA revealed a significant effect of condition, $F(3, 186) = 7.74$, $p < .001$, $\eta^2_{\text{partial}} = 0.11$. Subjects showed similar levels of recall across the honor, test-all, and random conditions, but worse recall in the dishonor condition. Tukey post hoc tests showed that the dishonor condition performed significantly worse than any other condition (at $p < .05$), but no other pairwise comparisons reached significance.[2] Fifteen subjects showed best performance in the honor condition, 17 in the test all condition, six in the dishonor condition, and eight in the random proportion condition.

We continued to compare the effectiveness of the four conditions by analyzing the average amount of time taken by subjects to complete the practice tests (and study the test feedback) and restudy attempts; these data are shown in Figure 7. A repeated-measures ANOVA on the total time taken on practice tests, feedback, and restudy attempts revealed a significant effect of condition, $F(3, 186) = 21.35$, $p < .001$, $\eta^2_{\text{partial}} = 0.26$. Tukey post hoc tests showed that the test-all condition required significantly more time than all other conditions (at $p < .05$) and that the dishonor condition required significantly less time than all other conditions, but honor and random-proportion conditions did not differ. These data show that testing all of the items took considerably more time during practice but did not lead to higher performance than allowing subjects to choose their own subset of items on which to be tested.

We analyzed performance when subjects' choices were honored and dishonored, collapsing across all four conditions. Results are displayed in Table 1: Subjects recalled more when their choices were honored than when their choices were dishonored, $t(62) = 4.75$, $p < .001$. Cued recall performance across the four conditions and split by actual study activity is displayed in the right panel of Figure 5.

A 2 (study activity choice) × 2 (actual study activity) repeated-measures ANOVA on cued recall performance revealed a significant interaction between choice and actual study activity, $F(1, 32) = 11.03$, $p = .002$, $\eta^2_{\text{partial}} = 0.26$. Cued recall was higher for items chosen to be tested than restudied, $F(1, 32) = 11.78$, $p = .002$, $\eta^2_{\text{partial}} = 0.27$, and higher for items actually tested than restudied, $F(1, 32) = 37.49$, $p < .001$, $\eta^2_{\text{partial}} = 0.54$. t tests on all items across conditions showed that items chosen to be tested benefited from testing ($M = 0.41$, $SD = 0.22$) compared with restudying ($M = 0.26$, $SD = 0.21$; $t(57) = 7.58$, $p < .001$, $d = 1.00$), and, like Experiment 4, items chosen to be restudied also benefited from testing ($M = 0.24$, $SD = 0.27$) compared to restudying ($M = 0.18$, $SD = 0.22$; $t(37) = 3.28$, $p = .002$, $d = 0.55$). Feedback ensured that testing was beneficial for all items, regardless of learners' chosen study activity, which adds qualifications to claims about when self-chosen testing choices result in better learning than testing all items.

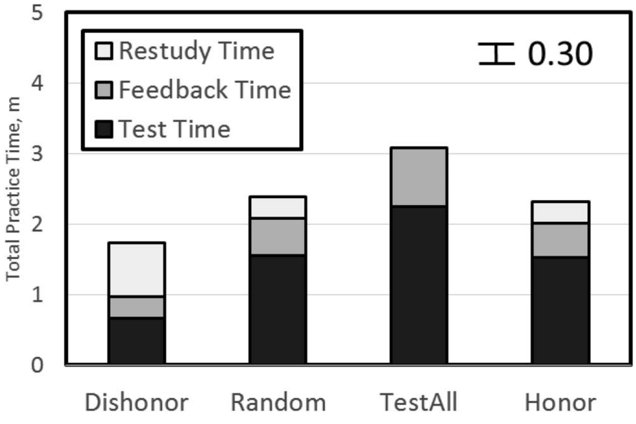

Figure 7. Average amount of time spent during practice in Experiment 5 in minutes. Feedback time indicates how much time learners spent studying the correct answer after being tested on that item.

Combined Analysis

Finally, we examined the combined data from all subjects across experiments with feedback after practice tests (Experiments 4 and 5). Subjects recalled more when their choices were honored ($M = 0.30$, $SD = 0.21$) than dishonored ($M = 0.22$, $SD = 0.21$; $t(118) = 6.02$, $p < .001$, $d = 0.57$). A 2 (study activity choice) × 2 (actual study activity) repeated measures ANOVA on final cued recall performance showed a significant interaction, $F(1, 84) = 32.38$, $p < .001$, $\eta^2_{\text{partial}} = 0.28$, an advantage for items chosen to be tested, $F(1, 84) = 53.87$, $p < .001$, $\eta^2_{\text{partial}} = 0.39$, and an advantage for testing over restudying, $F(1, 84) = 105.62$, $p < .001$, $\eta^2_{\text{partial}} = 0.56$. The interaction between study choice and actual study activity arose because there was a larger testing effect for items chosen to be tested than for items chosen to be restudied; however, we caution against strong interpretations of nondisordinal interactions, especially given the differences in the difficulty of items in each condition. Follow-up paired t tests revealed that testing improved memory both for items chosen to be tested, $t(110) = 10.39$, $p < .001$, $d = 0.99$ and for items chosen to be restudied, $t(93) = 5.81$, $p < .001$, $d = 0.60$.

Discussion

As with Experiment 4, subjects elected to test themselves on more easy pairs than difficult pairs, even when corrective feedback was provided. Reducing the costs of failed practice tests by providing corrective feedback had little influence on learners' conservative study choices. Their conservative approach did not appear to negatively impact their learning, as testing on all pairs did not provide an advantage over testing only the learners' choices. This result obtained in both Experiments 3 and 5. However, practice time in the test-all condition was much longer than in the honor condition, suggesting that testing all of the items comes with significant costs to efficiency. Unlike Experiment 3, honoring subjects' study choices did not improve performance relative to a random allocation of practice activities, nor did it result in less study time than a random allocation of practice activities. This finding shows that learners' global tendency to test more often on easy pairs—reflected in both the honor and random conditions—

[2] In Experiment 5, 80% of choices in the random proportion group turned out to be honored, due to large imbalances in restudy and test choices within subjects. Twenty-nine subjects had all of their choices for the random condition items honored; only seven subjects had less than half of their choices honored. Given that the random condition is highly similar to the honor condition, it is not surprising that we find no big differences between these two conditions. However, from the 34 subjects who had some choices honored and dishonored within the random condition, honoring choices led to higher final recall ($M = 0.35$, $SD = 0.26$) than dishonoring choices ($M = 0.19$, $SD = 0.22$; $t(33) = 2.19$, $p = .04$, $d = 0.39$).

advantageously resulted in more efficient practice time without hurting performance relative to testing on all items.

General Discussion

Across five experiments, learners selectively utilized testing for easier items and reserved restudying for the more difficult items. Selectively utilizing testing for easier items was effective and efficient for improving memory, as honoring learners' choices consistently led to better memory performance than dishonoring them. Memory performance on items chosen to be restudied did not benefit when they were tested rather than restudied in situations where no feedback was provided. By selectively using testing, learners boosted overall levels of memory performance and reduced the overall amount of time spent practicing. Forcing learners to be tested on all of the word pairs (regardless of the presence of corrective feedback) did not enhance performance relative to only utilizing testing for items chosen to be tested, and required more total time than selectively testing only the items subjects chose.

In the current experiments, learners demonstrated effective and efficient control of restudy and test decisions, contrary to what might be expected given students' reluctance to endorse testing as an effective study strategy. When given unlimited options to choose between restudying, testing, and being done, learners chose testing more frequently than restudying (across Experiments 2 and 3, $M_{\text{restudy}} = 0.38$, $M_{\text{test}} = 0.54$). Learners consistently rate restudying as more mnemonically beneficial than testing (Agarwal et al., 2008) and consider it one of their top study activities (Karpicke, Butler, & Roediger, 2009). However, as seen in this and prior research (Kornell & Son, 2009), learners freely choose to be tested on a large proportion of items, which indicates they believe that testing can effectively enhance their memory performance more than restudying in some circumstances, or can provide some other benefit. In fact, across the 117 subjects from Experiments 2 and 3 (where no feedback about practice test results were provided), only six subjects never chose testing.

Learners profess that restudying is more effective than testing, but still choose testing over restudying on a considerable proportion of word pairs. This finding replicates and extends claims made by Kornell and Son (2009), which showed that learners chose to self-test 50% or more of the time, despite thinking that restudying was more effective for memory. Those subjects reported that they engaged in testing to determine what they had learned and what they had not learned. Learners may choose self-testing not because they believe it will enhance their memories, but because it provides other information that is valuable to them. Additionally, the highly regimented ways in which the field has measured metacognitive monitoring may not capture the complexity of learners' knowledge about the benefits and limitations of testing for improving mnemonic performance. Learners may understand some of the effects that testing can have on memory but may focus on the limitations; specifically, learners may recognize that tests will not help memory for items than are not successfully recalled on those tests.

The experiments reported here remind us of the limitations of retrieval practice and reveal a surprising metacognitive awareness of those limitations. When no feedback is given, the benefits of testing are limited to items correctly recalled during practice (Izawa, 1967; Kornell et al., 2011). Testing, therefore, should be utilized selectively on items that will be recalled during practice. Broadly applying testing across all items in situations without feedback may not be beneficial for memory when compared with self-controlled testing. In fact, testing can promote sustained errors of commission (Henkel, 2007; Mathias, Marsh, & Dougherty, 2002; McDermott, 1996, 2006), a circumstance that we observed in our own data. Across the first three experiments, on trials where learners made an error on the initial test, learners produced the same error on the final test on about one third of all trials ($M = 0.35$, $SD = 0.21$). Even in situations with feedback, learners repeated one fifth of their errors from the practice test on the final test (Experiment 4: $M = 0.18$, $SD = 0.12$; Experiment 5: $M = 0.23$, $SD = 0.14$). Learners may underestimate the value of testing, as shown throughout the prior literature, but may still appreciate the limitations of testing—that is, learners know that testing is only beneficial when practice test retrieval is successful and understand that testing poorly learned items is not only an inefficient use of their time, but also a catalyst for sustained retrieval errors.

In circumstances where feedback is withheld, restudying the more difficult items appears to be more helpful than testing. In contrast, the presence of feedback during retrieval practice ensures that all items benefit from testing, at least as much as they do from restudy. It is therefore somewhat surprising that learners' choices did not seem to be much affected by the expectation of feedback. Comparing learners' choices across Experiments 3, 4, and 5, the proportion of practice tests was not substantially higher when feedback was given. Perhaps learners believe that incorrect tests with feedback do not bolster memory any more than a simpler restudy attempt; if tests require more time or effort than restudy, learners may choose to gain the benefits from the easier restudy trial. In other words, learners show no evidence of beliefs about possible test-potentiated learning (Arnold & McDermott, 2013).

The results presented here add perspective to the real-life survey data about self-testing during self-regulated learning. College students may not utilize self-testing often because they need to devote their effort to learning the material well enough before it can benefit from testing or because they view testing everything as an inefficient use of their time. Learners who do not first learn material well enough to ensure successful practice retrieval will probably receive no benefit from testing. Failed retrievals, with no feedback, would be time and effort that is largely wasted, and may affect motivation adversely. Devoting time and effort to restudying poorly learned material may actually be the best use of learners' time, which may be one reason why the majority of college students rate reading the textbook and their notes as their primary method of studying (Karpicke et al., 2009). Alternatively, college students may not accurately judge their learning in a complex college course or feel confident in their estimates of what they know and what they do not know. Without accurate estimates of their own knowledge, selectively applying retrieval practice to boost learning would be difficult and prone to costly errors. Students, then, may fall back on restudying, which is guaranteed to provide mnemonic benefits, albeit small ones.

A significant limitation of these experiments is that overall performance during the practice tests is rather low. In order to maximize the differences between honoring and dishonoring testing choices, we selected experimental conditions that constrained cued recall on the practice tests to around 50%. With better-learned stimuli, the metacognitive control exercised across these three experiments might be ineffective at supporting memory. For example, if memory before the first practice test were robust, long-term retention would likely benefit the most by testing the difficult items or by maximizing the amount of testing.

These experiments add to the growing literature suggesting that trusting learners to control their own learning results in better mnemonic performance than restricting that control (Finley, Tullis, & Benjamin, 2009). The prior literature shows that learners effectively choose items for restudy (Kornell & Metcalfe, 2006; Tullis & Benjamin, 2012) and allocate study time across items (Tullis & Benjamin, 2011; Tullis, Benjamin, & Liu, 2014) to improve mnemonic performance. Learners can choose items for restudy and allocate study time across items effectively because the mnemonic consequences of additional study time are intuitive and well understood by learners (Koriat, 1997). Here we investigated learners' implementation of testing because every existing study has shown that learners' monitoring about the relative benefits of testing and restudy are inaccurate (Agarwal et al., 2008; Karpicke, 2009; Roediger & Karpicke, 2006; Tullis et al., 2013). Despite these misapprehensions, some research has shown that under some circumstances learners will choose testing more frequently than restudying (Kornell & Son, 2009). In the current series of experiments, allowing learners to control their self-testing—a study activity for which they have demonstrated metacognitive illusions—still benefitted mnemonic performance. This finding provides an interesting dissociation between metacognitive monitoring and control; learners value restudying more than testing for improving their memories, but will still effectively utilize testing, particularly under conditions that allow the limitations of retrieval practice to be revealed (e.g., when no feedback is provided). While authors have reasonably suggested that the metacognitive illusions associated with testing and restudy would result in "dire consequences" for learners who monitor and control their own learning (Agarwal et al., 2008), the current results show that learners can actually be quite effective at allocating testing and restudy. Even for a study activity with misunderstood benefits, such as testing, learners deploy a selective and sophisticated use of metacognitive control that ultimately benefits their performance.

References

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger, H. L., III, & McDermott, K. B. (2008). Examining the tested effect with open- and closed-book tests. *Applied Cognitive Psychology, 22,* 861–876. http://dx.doi.org/10.1002/acp.1391

Arbuckle, T. Y., & Cuddy, L. L. (1969). Discrimination of item strength at time of presentation. *Journal of Experimental Psychology, 81,* 126–131. http://dx.doi.org/10.1037/h0027455

Arnold, K. M., & McDermott, K. B. (2013). Test-potentiated learning: Distinguishing between direct and indirect effects of tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39,* 940–945. http://dx.doi.org/10.1037/a0029199

Atkinson, R. C. (1972). Optimizing the learning of a second-language vocabulary. *Journal of Experimental Psychology, 96,* 124–129. http://dx.doi.org/10.1037/h0033475

Benjamin, A. S., & Bird, R. D. (2006). Metacognitive control of the spacing of study repetitions. *Journal of Memory and Language, 55,* 126–137. http://dx.doi.org/10.1016/j.jml.2006.02.003

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General, 127,* 55–68. http://dx.doi.org/10.1037/0096-3445.127.1.55

Benjamin, A. S., & Pashler, H. (2015). The value of standardized testing: A perspective from cognitive psychology. *Policy Insights from the Behavioral and Brain Sciences, 2,* 13–23. http://dx.doi.org/10.1177/2372732215601116

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10,* 433–436. http://dx.doi.org/10.1163/156856897X00357

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. http://dx.doi.org/10.1080/09541440701326097

Carpenter, S. K., & DeLosh, E. L. (2005). Application of the testing and spacing effects to name learning. *Applied Cognitive Psychology, 19,* 619–636. http://dx.doi.org/10.1002/acp.1101

Finley, J. R., & Benjamin, A. S. (2012). Adaptive and qualitative changes in encoding strategy with experience: Evidence from the test-expectancy paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38,* 632–652. http://dx.doi.org/10.1037/a0026215

Finley, J. R., Tullis, J. G., & Benjamin, A. S. (2009). Metacognitive control of learning and remembering. In M. S. Khine & I. M. Saleh (Eds.), *New science of learning: Cognition, computers and collaboration in education*. New York, NY: Springer Science & Business Media.

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

Henkel, L. A. (2007). The benefits and costs of repeated memory tests for young and older adults. *Psychology and Aging, 22,* 580–595. http://dx.doi.org/10.1037/0882-7974.22.3.580

Izawa, C. (1967). Function of test trials in paired-associate learning. *Journal of Experimental Psychology, 75,* 194–209. http://dx.doi.org/10.1037/h0024971

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

Karpicke, J. D., Butler, A. C., & Roediger, H. L., III. (2009). Metacognitive strategies in student learning: Do students practise retrieval when they study on their own? *Memory, 17,* 471–479. http://dx.doi.org/10.1080/09658210802647009

Karpicke, J. D., & Roediger, H. L., III. (2008). The critical importance of retrieval for learning. *Science, 319,* 966–968. http://dx.doi.org/10.1126/science.1152408

Kimball, D. R., Smith, T. A., & Muntean, W. J. (2012). Does delaying judgments of learning really improve the efficacy of study decisions? Not so much. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38,* 923–954. http://dx.doi.org/10.1037/a0026936

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126,* 549–570. http://dx.doi.org/10.1037/0096-3445.126.4.349

Kornell, N. (2009). Optimising learning using flashcards: Spacing is more effective than cramming. *Applied Cognitive Psychology, 23,* 1297–1317. http://dx.doi.org/10.1002/acp.1537

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

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

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language, 65,* 85–97. http://dx.doi.org/10.1016/j.jml.2011.04.002

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 989–998. http://dx.doi.org/10.1037/a0015729

Kornell, N., & Metcalfe, J. (2006). Study efficacy and the region of proximal learning framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32,* 609–622. http://dx.doi.org/10.1037/0278-7393.32.3.609

Kornell, N., & Son, L. K. (2009). Learners' choices and beliefs about self-testing. *Memory, 17,* 493–501. http://dx.doi.org/10.1080/09658210902832915

Landauer, T. K., & Bjork, R. A. (1978). Optimum rehearsal patterns and name learning. In M. M. Gruneberg, P. E. Morris, & R. N. Skykes (Eds.), *Practical aspects of memory* (pp. 625–632). London, UK: Academic Press.

Leeming, F. C. (2002). The exam-a-day procedure improves performance in psychology classes. *Teaching of Psychology, 29,* 210–212. http://dx.doi.org/10.1207/S15328023TOP2903_06

Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review, 1,* 476–490.

Mathias, C. W., Marsh, D. M., & Dougherty, D. M. (2002). Reliability estimates for the immediate and delayed memory tasks. *Perceptual and Motor Skills, 95,* 559–569. http://dx.doi.org/10.2466/pms.2002.95.2.559

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L. (2011). Test-enhanced learning in a middle school science classroom: The 4 effects of quiz frequency and placement. *Journal of Educational Psychology, 103,* 399–414. http://dx.doi.org/10.1037/a0021782

McDermott, K. B. (1996). The persistence of false memories in list recall. *Journal of Memory and Language, 35,* 212–230. http://dx.doi.org/10.1006/jmla.1996.0012

McDermott, K. B. (2006). Paradoxical effects of testing: Repeated retrieval attempts enhance the likelihood of later accurate and false recall. *Memory & Cognition, 34,* 261–267. http://dx.doi.org/10.3758/BF03193404

Metcalfe, J., & Kornell, N. (2005). A region of proximal learning model of study time allocation. *Journal of Memory and Language, 52,* 463–477. http://dx.doi.org/10.1016/j.jml.2004.12.001

Mulligan, N. W., & Peterson, D. J. (2014). The spacing effect and metacognitive control. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40,* 306–311. http://dx.doi.org/10.1037/a0033866

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). *The University of South Florida word association, rhyme, and word fragment norms*. Retrieved from http://www.usf.edu/FreeAssociation/

Putnam, A. L., & Roediger, H. L. (2012). Does response mode affect amount recalled or the magnitude of the testing effect? *Memory & Cognition*. Advance online publication.

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

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140,* 1432–1463. http://dx.doi.org/10.1037/a0037559

Son, L. K. (2004). Spacing one's study: Evidence for a metacognitive control strategy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 601–604. http://dx.doi.org/10.1037/0278-7393.30.3.601

Son, L. K. (2005). Metacognitive control: Children's short-term versus long-term study strategies. *The Journal of General Psychology, 132,* 347–363. http://dx.doi.org/10.3200/GENP.132.4.347-364

Son, L. K. (2010). Metacognitive control and the spacing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 255–262. http://dx.doi.org/10.1037/a0017892

Storm, B. C., Friedman, M. C., Murayama, K., & Bjork, R. A. (2014). On the transfer of prior tests or study events to subsequent study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40,* 115–124. http://dx.doi.org/10.1037/a0034252

Toppino, T. C., & Cohen, M. S. (2010). Metacognitive control and spaced practice: Clarifying what people do and why. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 1480–1491. http://dx.doi.org/10.1037/a0020949

Toppino, T. C., Cohen, M. S., Davis, M. L., & Moors, A. C. (2009). Metacognitive control over the distribution of practice: When is spacing preferred? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 1352–1358. http://dx.doi.org/10.1037/a0016371

Tullis, J. G., & Benjamin, A. S. (2011). On the effectiveness of self-paced learning. *Journal of Memory and Language, 64,* 109–118. http://dx.doi.org/10.1016/j.jml.2010.11.002

Tullis, J. G., & Benjamin, A. S. (2012). Consequences of restudy choices in younger and older learners. *Psychonomic Bulletin & Review, 19,* 743–749. http://dx.doi.org/10.3758/s13423-012-0266-2

Tullis, J. G., Benjamin, A. S., & Liu, X. (2014). Self-pacing study of faces of different races: Metacognitive control over study does not eliminate the cross-race recognition effect. *Memory & Cognition, 42,* 863–875. http://dx.doi.org/10.3758/s13421-014-0409-y

Tullis, J. G., Finley, J. R., & Benjamin, A. S. (2013). Metacognition of the testing effect: Guiding learners to predict the benefits of retrieval. *Memory & Cognition, 41,* 429–442. http://dx.doi.org/10.3758/s13421-012-0274-5

Received December 14, 2016
Revision received May 25, 2017
Accepted July 16, 2017 ■