Mem Cogn (2015) 43:193–205
DOI 10.3758/s13421-014-0462-6

Multiple-choice tests stabilize access to marginal knowledge

Allison D. Cantor · Andrea N. Eslick · Elizabeth J. Marsh · Robert A. Bjork · Elizabeth Ligon Bjork

Published online: 9 September 2014
© Psychonomic Society, Inc. 2014

Abstract Marginal knowledge refers to knowledge that is stored in memory, but is not accessible at a given moment. For example, one might struggle to remember who wrote *The Call of the Wild*, even if that knowledge is stored in memory. Knowing how best to stabilize access to marginal knowledge is important, given that new learning often requires accessing and building on prior knowledge. While even a single opportunity to restudy marginal knowledge boosts its later accessibility (Berger, Hall, & Bahrick, 1999), in many situations explicit relearning opportunities are not available. Our question is whether multiple-choice tests (which by definition expose the learner to the correct answers) can also serve this function and, if so, how testing compares to restudying given that tests can be particularly powerful learning devices (Roediger & Karpicke, 2006). In four experiments, we found that multiple-choice testing had the power to stabilize access to marginal knowledge, and to do so for at least up to a week. Importantly, such tests did not need to be paired with feedback, although testing was no more powerful than studying. Overall, the results support the idea that one's knowledge base is unstable, with individual pieces of information coming in and out of reach. The present findings have implications for a key educational challenge: ensuring that students have continuing access to information they have learned.

A. D. Cantor (✉) · E. J. Marsh
Department of Psychology and Neuroscience, Duke University, Box 90086, Durham, NC 27708-0086, USA
e-mail: allison.cantor@duke.edu

A. N. Eslick
Department of Social Sciences, Wartburg College, 100 Wartburg Blvd., Waverly, IA 50677-0903, USA

R. A. Bjork · E. L. Bjork
Department of Psychology, University of California, Los Angeles, CA, USA

Keywords Memory · Knowledge · Testing effect

To say that we have an impressive capacity to store information in memory is a gross understatement. Various ways of estimating the storage capacity of human memory (e.g., Landauer, 1986) suggest memory is virtually unlimited from a storage standpoint. Conversely, to say that we can always retrieve our memories is a gross overstatement. Using Tulving and Pearlstone's (1966) classic distinction, the point is that information may be *available* in that it is stored in memory but not *accessible* at a given time. In other words, the amount recalled is often an underestimate of what is stored in memory; estimates of remembering will change depending on the particular retrieval cues available.

Although these ideas were originally developed to describe episodic memory, this distinction between availability and accessibility can be made for facts, vocabulary words, concepts, and other information not tied to a particular past experience. A prime example involves tip-of-the-tongue (TOT) states (e.g., Brown, 1991; Brown & McNeil, 1966), whereby people report having knowledge that is very close to being retrievable but is not currently accessible. Supporting their claims of knowing, people in TOT states can often report the first letter and number of syllables of the correct response (e.g., Brown & McNeil, 1966; Yarmey, 1973). Additionally, with time, people often regain access to this knowledge (e.g., Cohen & Faulkner, 1986; Read & Bruce, 1982).

More generally, people often have a sense of whether knowledge is stored. Hart (1965) had participants answer general-knowledge questions, making "feeling-of-knowing" judgments (FOK) when they were unable to answer. That is, participants reported whether they felt they knew each correct answer well enough to recognize it from a list of possible answers. Participants were much more likely to select the correct answer following a positive than a negative FOK

judgment. Many studies have confirmed that FOK judgments are quite accurate indicators of stored knowledge (e.g., Bahrick & Phelps, 1988; Gruneberg, Smith, & Winfrow, 1973; Nelson, Gerler, & Narens, 1984).

Regardless of their metacognitive state, people have information stored that may be inaccessible at a given moment; Berger, Hall, and Bahrick (1999) used the term *marginal knowledge* to describe this phenomenon and devised a clever methodology to demonstrate its existence. They developed two parallel sets of general-knowledge questions, one real and the other fictitious (i.e., the authors made up the latter questions, which had no factual basis). The real and fictitious questions were matched on topic, sentence length, and sentence structure; see Table 1 for examples. The logic was that people could have marginal knowledge for real but not fictitious questions. To the extent that manipulations stabilize access to marginal knowledge, they should only influence performance on real questions, with any improvements for fictitious questions reflecting new learning.

To test this idea, Berger and colleagues examined the benefit of providing feedback following retrieval failures, to see if it would be more effective at promoting later retention of real than fictitious facts (given feedback could only activate marginal knowledge for real facts). After answering each question, participants saw the answer for five seconds. Of interest was performance for initially failed items on a re-test occurring up to nine days later. Feedback was more helpful for real questions, as many of those answers likely existed in memory. It was less effective for fictitious questions, where there was no marginal knowledge to activate; this new learning was more quickly forgotten over time, consistent with Jost's first law (see Bjork & Bjork, 1992, and Wixted, 2004, for discussions of Jost's laws).

Table 1 Examples of Real (R) and Yoked Fictitious (F) Questions with Targets and Lures

Question	Target	Lures
(R) What is the long process by which a dead organism turns to stone?	Petrification	Decomposition Ossification Rigor Mortis Torgisation
(F) What is the long process by which a live organism's hair turns white?	Pigmosis	Albination Decromatization Ostoresis Transcoloration
(R) What open-air public theater was home to William Shakespeare's theatrical company?	Globe	Aavion Avon Haven World
(F) What open-air public theater was home to Baruch's theatrical company?	Roche	Confair Dutch Renaissance Strobe

Providing feedback is the only documented method for stabilizing access to marginal knowledge, short of preventing information from being lost in the first place (Bahrick & Hall, 1991). In the present work, we explored a possible new way to recover marginal knowledge: answering multiple-choice questions. Multiple-choice tests are likely to help for two reasons: (1) they expose learners to correct answers, even if not explicitly labeled as such, and (2) they require information to be retrieved from memory, a process known to boost long-term retention (e.g., Roediger & Karpicke, 2006). In experiments conducted for other purposes, multiple-choice tests have generally provided enough retrieval practice to boost learning, even without feedback (e.g., Little, Bjork, Bjork, & Angello, 2012; Roediger & Marsh, 2005; see Marsh & Cantor, 2014, for a review). Therefore, multiple-choice testing might be as good as, or even better than, studying in reactivating marginal knowledge. The possibility of multiple-choice tests serving this purpose is appealing, as they are commonly used in education and beyond (e.g., by the DMV, in job training).

However, will multiple-choice testing also have a negative side effect? Specifically, multiple-choice testing might yield a negative testing effect, whereby learners reproduce multiple-choice lures on later tests of memory (Marsh, Roediger, Bjork, & Bjork, 2007; Roediger & Marsh, 2005). To be clear, correct responding on the multiple-choice test should stabilize access to marginal knowledge but endorsement of lures on that same test could yield a negative testing effect—the question is whether the benefits will outweigh the costs.

To summarize, we examined whether multiple-choice testing boosted accessibility of marginal knowledge and how it compared to studying. We utilized Bahrick and colleagues' logic that marginal knowledge would only be possible for real but not fictitious questions. Therefore, if multiple-choice testing is re-establishing marginal knowledge rather than teaching new knowledge, it should benefit real more than fictitious questions. We used an initial general-knowledge test to establish what knowledge was accessible and then examined the impact of multiple-choice testing on later ability to retrieve knowledge that was initially inaccessible.

We also manipulated delay to the final test, both because it is educationally relevant and because new learning should be forgotten more quickly than stabilized knowledge. The longest delay used was 10 minutes in Experiment 1, 48 hours in Experiments 2 and 3a, and one week in Experiment 3b. To determine whether marginal knowledge could be recovered without explicitly stating the answer, no feedback was provided in Experiment 1. After successfully demonstrating that multiple-choice tests stabilized access to knowledge, in Experiment 2 we asked whether receiving answer feedback boosted these benefits. In Experiment 3, we directly compared

the benefits of multiple-choice testing (without feedback) and studying.

Experiment 1

Method

Participants Thirty-six Duke University undergraduates participated for course credit.

Design A 2 (stabilizing activity: MC test, none) × 2 (item-type: real, fictitious) × 3 (delay: 0, 5, 10 min) design was employed. All variables were manipulated within-subjects.

Materials and counterbalancing Materials were taken from Berger et al. (1999): 150 real general-knowledge questions and 150 parallel fictitious questions. Each question paired the target with four plausible lures, which Berger and colleagues selected to be from the same category as the target and approximately the same level of familiarity (see Table 1). Because the fictitious questions had no correct answer, the target was an arbitrarily chosen lure.

To identify marginal knowledge in Duke students, 37 pilot participants answered the 150 real questions in short-answer and then multiple-choice format. For each question, we determined the percentage of participants who demonstrated marginal knowledge, failing to correctly answer the short-answer question but selecting the target on the multiple-choice test. We chose the 84 questions most likely to elicit marginal knowledge (range: 22 % to 68 %; M = .39, SD = .11). For each question, the yoked fictitious question was also taken from Berger et al. (1999). Thus, there were 84 question-pairs, each containing a real and corresponding fictitious question.

For each participant, 42 question-pairs were assigned to the real condition and 42 pairs to the fictitious condition. An item's assignment was the same across tests: if an item appeared in fictitious format on the initial test, it also appeared that way on the multiple-choice and final test. A given participant never saw both the real and fictitious versions of the same item.

The initial short-answer test consisted of the 84 critical questions and ten easy filler questions (five fillers always appeared first and the computer randomly ordered all subsequent questions). Half of the critical items appeared on the multiple-choice test, which paired each question with the target and four lures. One-third of the critical questions appeared on each of the three final short-answer tests, meaning that each test contained 14 real and 14 fictitious items. Twelve conditions were required to fully counterbalance item-type, stabilizing activity, and delay across participants.

Procedure After giving informed consent, participants took the initial knowledge measure. Participants were instructed to try their best to answer questions but not to be discouraged if they couldn't answer them and to type "don't know." Feedback was not given during this or any other phase.

In the stabilization phase, participants answered 42 randomly ordered multiple-choice questions.

In the final retention phase, participants completed three 28-item short-answer tests, with items randomly ordered on each test and with the same instructions given as on the initial test. The first test occurred immediately after the multiple-choice test; participants then solved Sudoku puzzles for five minutes before completing the second test. Five more minutes of puzzles separated the second and third tests.

Results

Two coders scored all short-answer responses as correct or incorrect; no partial credit was given. Inter-rater reliability was very high on the initial (κ = .98) and final tests (κ = .96); a third coder resolved all discrepancies.

All analyses were restricted to items that were *not* retrieved on the initial test (at the individual participant level), as knowledge that was retrieved was already stable. This constraint restricted the analyses to 64% (SD = .18) of real and 100% (SD = .00) of fictitious items. Significance was determined at the $p < .05$ level for all experiments unless otherwise noted.

Benefits of multiple-choice testing Did multiple-choice testing help participants retrieve information that they initially failed to produce? A 2 (stabilizing activity: MC Test, none) × 2 (item-type: real, fictitious) × 3 (delay: 0, 5, 10 min) ANOVA was computed on proportion of final-test questions answered with targets. The data appear in Table 2; the analyses were conducted on the top two rows of data (to aid the reader, the bottom row depicts benefits of testing over no activity). We observed a testing effect: participants answered more final-test questions with targets if those items had appeared on the multiple-choice test, $F(1, 35) = 269.01$, MSE = .03, η^2 = .27. More importantly, this benefit of testing was much larger for real than fictitious questions, $F(1, 35) = 171.64$, MSE = .03, η^2 = .17, and this pattern held for all three delays, $F < 1$.

Performance on the multiple-choice test was as expected, with participants demonstrating underlying knowledge for real questions. After a retrieval failure, participants were more than twice as likely to select the target for real (M = .58) than for fictitious questions (M = .24), $t(35) = 9.42$, SEM = .04.

To pinpoint the benefits of multiple-choice testing on later knowledge, we conducted an analysis on targets that were

Table 2 Proportion of Final Test Questions Answered with Targets (given Initial Test Failure) in Experiment 1

	Real Questions			Fictitious Questions		
	No Delay	5 Min	10 Min	No Delay	5 Min	10 Min
MC Test	.52 (.05)	.49 (.05)	.46 (.03)	.07 (.02)	.03 (.02)	.06 (.01)
No Activity	.03 (.01)	.03 (.02)	.02 (.01)	.00 (.00)	.01 (.01)	.00 (.00)
MC Test – No Activity	.49 (.05)	.46 (.05)	.44 (.05)	.07 (.02)	.02 (.01)	.06 (.01)

Note. Standard errors are presented in parentheses. Third row depicts benefit of multiple-choice testing over no activity

selected on the multiple-choice test. We restricted the analysis to target selections, as participants almost never produced targets on the final test if they were not selected on the multiple-choice test ($M = .01$, $SD = .02$). Instead, the benefits were due to the reactivation of marginal knowledge. A 2 (item-type: real, fictitious) × 3 (delay: 0, 5, 10 min) ANOVA was computed on proportion of final-test questions answered with targets (given target multiple-choice selection). Only 22 participants were included in the analysis, as the others did not have observations in all cells (because of the low rate of selecting the target for fictitious questions). As illustrated in Fig. 1 (panel A), the benefits of selecting targets were much greater when there was underlying marginal knowledge to be activated. That is, participants retained many more targets selected on the multiple-choice test for real than fictitious items, $F(1, 21) = 122.15$, $MSE = .09$, $\eta^2 = .51$. Furthermore, there was a marginal item-type × delay interaction in that retention of real targets remained relatively stable while retention of fictitious targets diminished at a delay, $F(2, 42) = 2.77$, $MSE = .08$, $p = .07$, $\eta^2 = .02$.

Costs of multiple-choice testing Did multiple-choice testing increase the likelihood that participants answered final-test questions with (incorrect) multiple-choice lures? We computed a 2 (stabilizing activity: MC Test, none) × 2 (item-type: real, fictitious) × 3 (delay: 0, 5, 10 min) ANOVA on proportion of final-test questions answered with lures. Table 3 shows a negative testing effect: participants answered more final-test questions with lures if the items had appeared on the multiple-choice test, $F(1, 35) = 93.07$, $MSE = .03$, $\eta^2 = .24$. This negative testing effect was similar for real and fictitious items, $F < 1$. Overall, the negative testing effect diminished as time passed, $F(2, 70) = 6.40$, $MSE = .02$, $\eta^2 = .02$, and it did so similarly for real and fictitious questions, $F < 1$.

Discussion

Multiple-choice testing did indeed stabilize access to marginal knowledge. Given an initial retrieval failure, multiple-choice testing was considerably more beneficial for questions targeting marginal knowledge (real questions) than new learning (fictitious questions). Participants demonstrated marginal knowledge for many of the initially failed real questions, selecting the target alternative much more frequently for real than fictitious questions. Crucially, selecting the target was powerful enough to recover marginal knowledge even though participants were never explicitly told the correct answer. While multiple-choice testing did increase reproduction of multiple-choice lures on the final test, this negative testing effect was much smaller than the benefits of multiple-choice testing. Critically, the negative testing effect was similar for real and fictitious questions and decreased with time—suggesting that the effect represented new learning.

Experiment 1 used very short delays but most real-world situations require retrieval of information months or years later (e.g., Bahrick, 1984; Bahrick, Bahrick, & Wittlinger, 1975). Benefits of testing are known to last days (Kang, McDermott, & Roediger, 2007) or even months (Roediger,

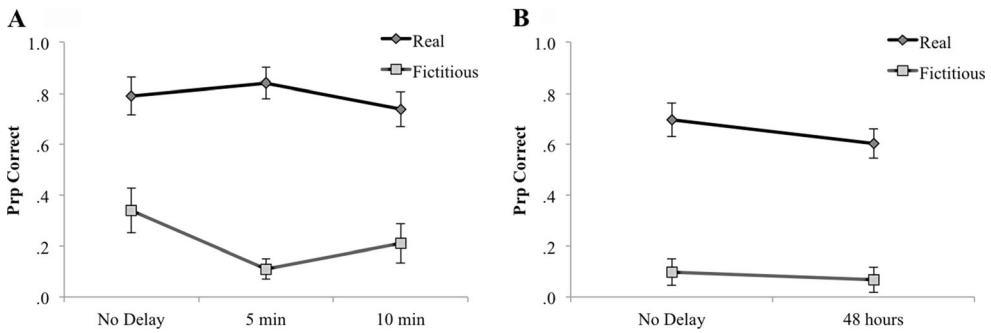

Fig. 1 Production of previously marginal knowledge: proportions of final-test questions answered with targets given initial test failure and target multiple-choice selection. Data are from Experiment 1 (**A**) and Experiment 2 (**B**). Error bars represent standard error of the mean

Table 3 Proportion of Final-Test Questions Answered with Lures (given Initial Test Failure) in Experiment 1

	Real Questions			Fictitious Questions		
	No Delay	5 Min	10 Min	No Delay	5 Min	10 Min
MC Test	.22 (.04)	.23 (.04)	.17 (.03)	.25 (.03)	.15 (.03)	.14 (.02)
No Activity	.01 (.01)	.08 (.03)	.06 (.02)	.02 (.01)	.02 (.01)	.01 (.01)
MC Test – No Activity	.21 (.04)	.15 (.05)	.11 (.03)	.23 (.03)	.13 (.03)	.13 (.03)

Note. Standard errors are presented in parentheses. Third row depicts difference between multiple-choice testing and no-activity conditions

Agarwal, McDaniel, & McDermott, 2011; see Agarwal, Bain, & Chamberlain, 2012, for a review), suggesting the possibility of long-lasting benefits of multiple-choice testing on access to marginal knowledge. Thus, in Experiment 2, we extended the delay to two days.

Experiment 2 also investigated whether receiving feedback would boost the benefits of multiple-choice testing for stabilizing knowledge as well as reduce the negative testing effect (Butler & Roediger, 2008). Experiment 2 was very similar to Experiment 1, except that half of the multiple-choice questions were paired with feedback, and delay to the final test was immediate or 48 hours.

Experiment 2

Method

Participants Forty-eight Duke University undergraduates participated in exchange for monetary compensation.

Design Experiment 2 had a 3 (stabilizing activity: MC test with feedback, MC test without feedback, none) × 2 (item-type: real, fictitious) × 2 (delay: none, 48 h) design. All variables were manipulated within-subjects.

Materials and counterbalancing Materials were the same as in Experiment 1. Assignment of items to stabilizing activity condition was counterbalanced across participants so a total of 56 questions appeared on the multiple-choice test and participants received feedback for 28 of them. Each final short-answer test consisted of 42 questions (21 real, 21 fictitious), and assignment of items to delay was counterbalanced across subjects. Experiment 2 required 12 conditions to fully counterbalance item-type, stabilizing activity, and delay across participants.

Procedure The procedure was the same as Experiment 1 with the following exceptions: after making their selections, participants received answer feedback for half of the multiple-choice questions, with the target appearing for five seconds. For the remaining questions, "No Feedback," appeared for five seconds. The final tests occurred immediately after the multiple-choice test or 48 hours later.

Results

Two coders scored all responses independently. Inter-rater reliability was very high (initial test: κ = .99; final tests: κ = .97). A third coder resolved all discrepancies.

Again, the analyses only included items that participants failed to retrieve initially (64 % [SD = .18] of real items and 100% [SD = .00] of fictitious items across participants).

Effects of delay on benefits of multiple-choice testing To determine whether the testing effect observed in Experiment 1 generalized to a delay of 48 hours, we restricted analyses to the same stabilizing activity conditions as used in Experiment 1. We computed a 2 (stabilizing activity: MC test without feedback, none) × 2 (item-type: real, fictitious) × 2 (delay: none, 48 h) ANOVA on proportion of final-test questions answered with targets. The data appear in Table 4 (second and third rows). Multiple-choice testing led to higher scores on the final test, $F(1, 47) = 124.47$, $MSE = .03$, $\eta^2 = .21$. This testing effect was much larger for real than fictitious items, $F(1, 47) = 135.41$, $MSE = .02$, $\eta^2 = .16$. Whereas the size of the testing effect was consistent across the shorter delays in Experiment 1, it decreased slightly after two days, $F(1,47) = 6.48$, $MSE = .02$, $\eta^2 = .01$.

Table 4 Proportion of Final-Test Questions answered with Targets (given Initial Test Failure) in Experiment 2

	Real Questions		Fictitious Questions	
	No Delay	48 Hours	No Delay	48 Hours
Feedback	.81 (.03)	.51 (.04)	.40 (.03)	.10 (.02)
MC Test	.49 (.04)	.39 (.03)	.03 (.01)	.02 (.02)
No Activity	.04 (.01)	.06 (.02)	.00 (.00)	.00 (.00)
MC Test – No Activity	.45 (.05)	.33 (.04)	.03 (.01)	.02 (.01)
Feedback – MC Test	.32 (.04)	.12 (.05)	.37 (.03)	.08 (.02)

Note. Standard error is noted in parentheses. Fourth row depicts benefit of multiple-choice testing over no activity. Fifth row depicts benefit of feedback over multiple-choice testing alone

Multiple-choice test performance was consistent with the hypothesis that participants had marginal knowledge for real but not fictitious items. Following an initial recall failure, participants chose the target more than twice as often for real ($M = .63$) than for fictitious questions ($M = .22$), $t(47) = 13.50$, $SEM = .03$, $d = 2.66$. To examine the retention of these target selections, we computed a 2 (item-type: real, fictitious) × 2 (delay: none, 48 h) ANOVA on proportion of final-test questions answered with targets (given target multiple-choice selection). This analysis included the 26 participants who had observations in all cells; incorrect selections were not analyzed as participants rarely produced targets on the final test if they were not selected on the prior multiple-choice test ($M = .01$, $SD = .02$).

As shown in Fig. 1 (panel B), participants were more likely to retain target selections and produce them on the final test for real than fictitious items, $F(1, 25) = 122.85$, $MSE = .07$, $\eta^2 = .70$. While Fig. 1 depicts a small decline in retention of targets over time, this effect of delay did not reach significance, $F(1,25) = 1.95$, $MSE = .04$, $p = .17$, $\eta^2 = .01$.

Effects of feedback on benefits of multiple-choice testing We also examined whether the provision of feedback boosted the benefits of multiple-choice testing. Participants were equally likely to select targets on the multiple-choice test regardless of whether or not they received feedback ($t < 1$), allowing us to isolate the effect of receiving feedback. We computed a 2 (stabilizing activity: MC test with feedback, MC test without feedback) × 2 (item-type: real, fictitious) × 2 (delay: none, 48 h) ANOVA on proportion of final-test questions answered with targets. These data appear in Table 4 (first and second rows). Performance was much higher on final-test questions that were previously tested with feedback, $F(1,47) = 120.00$, $MSE = .04$, $\eta^2 = .14$. However, this boost was similar for real and fictitious items, $F < 1$, and dropped dramatically as time passed, $F(1,47) = 64.86$, $MSE = .02$, $\eta^2 = .04$. Therefore, this benefit of feedback likely represented new learning.

A similar conclusion was reached when we examined the retention of selected targets. We computed a 2 (stabilizing activity: MC test with feedback, MC test without feedback) × 2 (item-type: real, fictitious) × 2 (delay: none, 48 h) ANOVA on the proportion of final-test questions answered with targets (given target multiple-choice selection). This analysis included 22 participants; the data appear in Fig. 2. Although feedback helped participants retain target selections, $F(1, 21) = 8.39$, $MSE = .140$ $\eta^2 = .03$, this benefit was similar for real and fictitious questions, $F < 1$. Overall, retention of targets diminished slightly over time, $F(1, 21) = 5.37$, MSE = .03, $\eta^2 = .01$.

Negative testing effects To examine whether the longer delay influenced the negative testing effect, we computed a 2 (stabilizing activity: MC test without feedback, none) × 2 (item-type: real, fictitious) × 2 (delay: none, 48 h) ANOVA on proportion of final-test questions answered with lures. These data appear in Table 5 (second and third rows). Participants were more likely to answer final-test questions with lures if the item had appeared on the multiple-choice test, $F(1,47) = 44.64$, $MSE = .02$, $\eta^2 = .12$. This negative testing effect was similar for real and fictitious items ($F < 1$) and decreased with time, $F(1,47) = 4.38$, $MSE = .02$, $\eta^2 = .01$ for both real and fictitious items, $F < 1$.

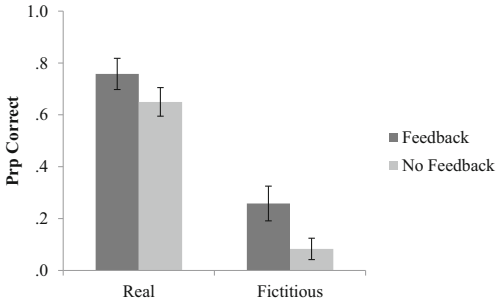

Fig. 2 Production of previously marginal knowledge: proportion of final-test questions answered with targets given initial test failure and target multiple-choice selection in Experiment 2 (MC test with feedback and MC test without feedback conditions). Error bars represent standard error of the mean

Consistent with prior research (Butler & Roediger, 2008; Marsh, Fazio, & Goswick, 2012), feedback reduced the negative testing effect. We computed a 2 (stabilizing activity: MC test with feedback, MC test without feedback) × 2 (item-type: real, fictitious) × 2 (delay: none, 48 h) ANOVA on proportion of lures produced on the final test. The data appear in Table 5 (first and second rows). Participants were less likely to produce lures on the final test if they had received feedback, $F(1,47) = 54.79$, $MSE = .01$, $\eta^2 = .12$. This benefit of feedback was similar for real and fictitious items, $F(1,47) = 1.47$, $MSE = .01$, $p = .23$, $\eta^2 = .01$, and diminished with time, $F(1,47) = 6.04$, MSE = .02, $\eta^2 = .02$.

Table 5 Proportion of Final-Test Questions Answered with Lures (given Initial Test Failure) in Experiment 2

	Real Questions		Fictitious Questions	
	No Delay	48 Hours	No Delay	48 Hours
Feedback	.01 (.01)	.07 (.02)	.05 (.01)	.04 (.01)
MC Test	.14 (.02)	.13 (.02)	.14 (.02)	.07 (.01)
No Activity	.04 (.01)	.06 (.02)	.02 (.01)	.01 (.01)
MC Test – No Activity	.10 (.03)	.07 (.03)	.12 (.02)	.06 (.02)
Feedback – MC Test	−.13 (.02)	−.06 (.03)	−.09 (.02)	−.03 (.02)

Note. Standard error is noted in parentheses. Fourth row depicts difference between multiple-choice testing and no-activity conditions. Fifth row depicts benefit of feedback over multiple-choice testing

Discussion

In two experiments, multiple-choice testing stabilized access to marginal knowledge, and the benefits were robust over 48 hours. Although access dropped slightly over this delay, 60 % of the correct multiple-choice selections were maintained over time. The careful reader may have noticed that delay had statistically different effects on the data appearing in Table 4 (positive testing effect) and Fig. 1 (panel B) (retention of correct multiple-choice selections), with the effect of delay only reaching significance for the former. The results in the figure, however, are based on a subset of the data, and thus that analysis had less power. We simply note that the pattern is the same across the two analyses: there is a tendency for slight forgetting over time, which is expected. It was not expected that reactivated knowledge would stay accessible forever. This reactivated knowledge should stay accessible longer than newly learned information (Jost's first law), but both will become inaccessible over time without further activation.

While feedback boosted the positive testing effect, this boost was similar for real and fictitious items and dropped dramatically over 48 hours. Therefore, the benefits of feedback were likely due to new learning. The only added benefit of feedback was that it reduced the negative testing effect, but we remind the reader that the negative testing effect was relatively small and decreased naturally with time. We see no problem recommending the use of multiple-choice tests alone, even without feedback, to reactivate access to marginal knowledge.

However, it is unclear whether multiple-choice tests are better, equivalent to, or worse at stabilizing access to marginal knowledge than simply reviewing the correct answers. Berger and colleagues (1999) showed that five seconds of exposure to targets helped recover marginal knowledge, but this method has never been directly compared to the benefits of recognizing marginal knowledge. In Experiment 3, we directly compared the two methods of reactivating marginal knowledge, to evaluate whether the retrieval practice involved in answering a multiple-choice question makes that method more powerful than seeing the fact. Because the advantage of retrieval practice over studying is typically observed after a delay (e.g., Roediger & Karpicke, 2006), we compared the benefits of these two methods on immediate and delayed final tests.

The inclusion of a study condition raises a tricky methodological issue; namely, that the comparison will only be fair if one can limit analyses to marginal knowledge. To the extent that participants do not know the critical facts, the study condition (which exposes participants to all facts) will have an advantage over the multiple-choice testing condition. Participants will have the opportunity to learn 100 % of unknown facts in the study condition, but only a 25 % chance of selecting (and learning) unknown facts when faced with multiple-choice questions.

The ideal solution would be to use only marginal items, so that the study condition did not mix new learning with activation of marginal knowledge. However, the marginal knowledge rates observed in our experiments (M = .38, SD = .11) were not nearly high enough. We tried to increase the rate of marginal knowledge by recruiting subjects from Amazon Mechanical Turk (a website that allows researchers to recruit a diverse population to participate in online experiments; the data collected are consistent with laboratory data; e.g., Buhrmester, Kwang, & Gosling, 2011). However, a pilot study (N = 103) did not find enough items that were consistently marginal across participants to create a study list containing only marginal knowledge or even 75 % marginal items (to be clear, people demonstrated considerable marginal knowledge, just not on the same items).

We reasoned that the best way to determine marginal knowledge was to ask participants to predict their own marginal knowledge. As described in the introduction, people are very good at predicting whether they will be able to recognize answers to general-knowledge questions (e.g., Bahrick & Phelps, 1988; Gruneberg et al., 1973; Nelson et al., 1984). To ensure this method would work, we conducted another pilot (N = 100) where participants predicted their own marginal knowledge. After answering each real short-answer question, participants made a binary decision as to whether they would be able to select the correct answer out of four choices. To test the accuracy of these predictions, the same participants also answered each question in multiple-choice format. Participants were quite accurate in their predictions, selecting the correct multiple-choice alternative more often after positive (M = .84) than negative FOK judgments (M = .58), $t(85)$ = 7.70, SEM = .03, $p < .001$, d = .99. Although not perfect, FOK judgments provided a means of estimating marginal knowledge that could be used for both the study and multiple-choice testing conditions.

Experiment 3 was similar to the earlier experiments: we identified participants' marginal knowledge, assigned them to stabilizing activities, and measured knowledge after varying delays. The main difference was how marginal knowledge was identified in Experiment 3, with participants making FOK judgments to estimate their own marginal knowledge. To ensure we had enough items in each cell, we manipulated stabilizing activity between-subjects, with some participants doing a filler task (playing Tetris), others studying critical facts, and the third group answering multiple-choice questions. Participants took two final tests: one immediate and one delayed (48 hours in Experiment 3a; 1 week in Experiment 3b).

Experiment 3a

Method

Participants 174 participants completed this two-session experiment via MTurk and were compensated $3 (an additional 37 completed the first session but not the second). Six were excluded due to technical difficulties (e.g., computer froze), and 24 for indicating they looked up or wrote down answers during the experiment. Thus, a total of 144 participants were included in the analyses.

Design Experiment 3a had a 3 (stabilizing activity: MC test, study, none) × 2 (item-type: real, fictitious) × 2 (delay: none, 48 h) mixed-factorial design. Stabilizing activity was manipulated between subjects and the others were manipulated within-subjects.

Materials and counterbalancing We selected the 84 items from the MTurk pilot described earlier that yielded the highest rates of marginal knowledge. Marginal knowledge was defined as the inability to answer the short-answer question correctly followed by a positive FOK judgment (i.e. not contingent upon correct multiple-choice performance). The selected questions were rated as marginal for 28 % to 67 % (M = .39, SD = .09) of our pilot sample. For the 84 selected general-knowledge questions, we used the yoked fictitious question from Berger et al. (1999), yielding a total of 84 question pairs.

For each participant, 42 question-pairs were assigned to the real condition and the other 42 to the fictitious condition. An item's assignment was the same across all phases. The computer randomly ordered the questions within each test.

The initial short-answer test contained 104 questions: 84 critical questions and 20 easy filler questions. The stabilizing activity varied depending upon experimental condition (MC test, study, none). The multiple-choice test contained the 84 critical questions, each of which paired the target with three plausible lures. The study list contained the critical questions paired with targets (no lures). Half of the 84 critical questions appeared on each of the final short-answer tests.

Twelve conditions were required to fully counterbalance item-type, stabilizing strategy, and delay across participants.

Procedure All participants began with the initial test, with the same instructions as earlier experiments. Immediately after answering each short-answer question, participants (regardless of condition) responded "yes" or "no" to the question, "Do you think you would be able to select the correct answer out of 4 choices?"

Next, participants took a multiple-choice test, studied, or played Tetris, depending on their randomly assigned condition. Participants in the multiple-choice testing condition answered the 84 critical questions without feedback. To equate time on task and spacing, participants in the study condition viewed each critical short-answer question and the target for six seconds (time was determined through the pilot).

Session 1 ended with the immediate final test. Approximately 48 hours later, participants were emailed a link to complete Session 2, which began with the delayed final test. Participants were then asked whether they had looked up or written down any answers, with explicit reassurance that their response would have no impact on their compensation.

Results

Two coders scored all short-answer responses and inter-rater reliability was very high (initial test: κ = .98, final tests: κ = .98). Discrepancies were resolved via discussion.

Determining Marginal Knowledge We used participants' FOK judgments to determine their marginal knowledge for our key analysis comparing multiple-choice testing and studying in their ability to stabilize access to marginal knowledge. These estimates were quite accurate in the multiple-choice testing condition (it was impossible to assess accuracy for the study and no-activity conditions), with participants correctly selecting the target (for real questions) more frequently after a positive (M = .86) than after a negative FOK judgment (M = .54), $t(43) = 7.81$, SEM = .04, d = 1.54. Four participants did not make any negative FOK judgments for real questions and were not included in this analysis.

Again, our focus is on items that subjects failed to retrieve initially, and all analyses that follow are restricted to these items (60 % [SD = .20] of real and 100 % [SD = .004] of fictitious items).

Comparing the benefits of studying and multiple-choice testing If one cannot produce information but believes one would recognize it, what is the best way to reactivate that knowledge and ensure later access? Because of the methodological problem outlined earlier, we do not report the analysis comparing all studied to all tested items (since the study condition allowed more new learning), although we note that analysis yielded the same pattern of results. Instead, we compared the benefits of multiple-choice testing and studying on later ability to produce estimated marginal knowledge (real items given a positive FOK) and newly learned information (fictitious items). We computed a 2 (stabilizing activity: study, MC test) × 2 (item-type: estimated marginal, newly learned) × 2 (delay: none, 48 h) ANOVA on proportion of final-test questions answered with targets. Table 6 (first and second rows) depicts the data; to preview, it did not matter whether the student reactivated her knowledge through study or multiple-choice testing – both were effective methods for stabilizing access to marginal knowledge.

Table 6 Proportion of Final-Test Questions Answered with Targets (given Initial Test Failure). Data are from Experiment 3a. Estimated Marginal Knowledge Comprises Real Items given a Positive FOK and New Learning Comprises all Fictitious Items

	Estimated Marginal Knowledge		New Learning	
	No Delay	48 Hours	No Delay	48 Hours
Study	.69 (.04)	.60 (.04)	.34 (.03)	10 (.01)
MC Test	.63 (.04)	.57 (.04)	.04 (.03)	.01 (.01)
No Activity	.07 (.03)	.13 (.03)	.01 (.02)	.01 (.01)
Study – MC Test	.06	.03	.30	.09

Note. Standard error is noted in parentheses. Fourth row depicts benefit of studying over multiple-choice testing

There was a benefit of studying over multiple-choice testing that existed largely for newly learned fictitious information, especially when it was assessed immediately (3-way interaction: $F(1,94) = 12.77$, $MSE = .02$, $\eta^2 = .01$). Studying was better than multiple-choice testing when new learning was assessed on the immediate final test, $t(50.76) = 7.68$, $SED = .04$, $d = 1.57$.[1] While this benefit also held for the delayed final test, $t(52.96) = 5.39$, $SED = .02$, $d = 1.10$, the effect size dropped across the delay, indicating that much of this new learning was quickly forgotten.[1] However, multiple-choice testing was equally beneficial when it came to stabilizing access to (estimated) marginal knowledge. Crucially, there were no differences between multiple-choice testing and studying in production of marginal knowledge on the immediate or delayed final test, ts < 1.

Because FOK judgments were not perfect estimates of marginal knowledge, we also conducted an analysis on only participants in the multiple-choice testing condition so that we could define marginal knowledge in the traditional sense: failure to produce the target initially followed by successful recognition. The patterns were almost identical to those obtained with FOK judgments in the current experiment and those from Experiments 1 and 2. As shown in Fig. 3 (panel A), participants were more likely to retain target selections for real than fictitious items, $F(1,47) = 792.86$, $MSE = .02$, $\eta^2 = .83$, and this pattern was consistent across the 48-h delay, $F < 1$.

Negative testing effects As expected, participants who answered multiple-choice questions ($M = .07$) were more likely than participants in the no-activity condition ($M = .02$) to produce lures on the final test, $F(1, 94) = 67.71$, $MSE = .01$, $\eta^2 = .42$. However, this effect was similar for real and fictitious items, $F < 1$, and decreased after a delay, $F(1,94) = 22.52$, $MSE = .01$, $\eta^2 = .06$. One advantage of studying over multiple-choice testing was that studiers reproduced fewer lures on the final test ($M = .01$), $F(1, 94) = 139.15$, MSE = .01, $\eta^2 = .60$. Importantly, this effect was similar for real and fictitious items, $F(1, 94) = 3.13$, $MSE = .01$, $p = .08$, $\eta^2 = .01$ and decreased with time, $F(1,94) = 29.46$, $MSE = .01$, $\eta^2 = .08$.

[1] Levene's Test for Equality of Variances was violated so we report corrected values.

Discussion

Experiment 3a found that taking a multiple-choice test was just as effective in stabilizing access to marginal knowledge as studying the answers. This result is particularly impressive considering that it was not possible to restrict analyses to 100 % marginal items. Feeling-of-knowing judgments were good but not perfect, causing some unknown items to be included. For unknown items, study-participants should have a large advantage, especially on the immediate test. As noted earlier, we obtained the same pattern of results even when the analyses included all items (and were not restricted to positive FOK judgments). This comparison certainly included many unknown items, giving study-participants an even larger advantage, yet multiple-choice testing still measured up. Thus, multiple-choice testing might prove even more powerful than studying if it were possible to consider only pure marginal knowledge.

While including unknown items in the analyses gave study-participants an advantage, this advantage should diminish with time. Indeed, participants' new learning of fictitious items decreased dramatically over 48 hours. However, after 48 hours, participants still retained 10 % of fictitious targets. At longer delays, the retention of new learning should drop toward zero and the comparison between multiple-choice testing and studying should become fairer. Thus, Experiment 3b was an exact replication of Experiment 3a except that we extended the delay to one week.

Experiment 3b

Method

Participants 173 participants completed this two-session experiment via MTurk and were compensated $4.00 total (an additional 30 subjects started the experiment but did not complete it). Twelve participants were excluded due to technical difficulties. Seventeen participants indicated that they looked up or wrote down answers and were excluded, meaning that 144 participants were included in the analyses.

Design, Materials, and Procedure The materials, design, and procedure were identical to those used in Experiment 3a, except that delay was operationalized as one week instead of two days.

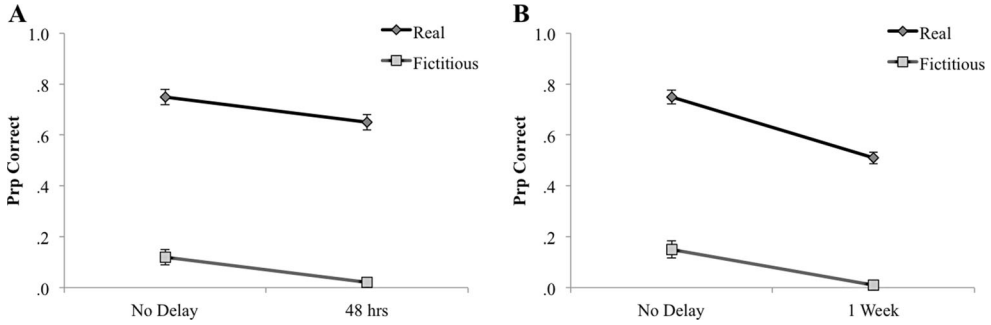

Fig. 3 Production of previously marginal knowledge: proportions of final-test questions answered with the target given initial test failure and target multiple-choice selection in Experiment 3a (**A**) and Experiment 3b (**B**). Error bars represent standard error of the mean

Results

Two coders independently scored all short-answer responses; inter-rater reliability was very high (initial test: $\kappa = .98$; final test: $\kappa = .97$). All discrepancies were resolved through discussion.

Determining marginal knowledge Participants were quite accurate at predicting their own marginal knowledge. Participants in the multiple-choice testing condition were much more likely to select the correct multiple-choice answer (for real questions) following a positive ($M = .86$) than a negative FOK judgment ($M = .61$), $t(47) = 6.73$, $SEM = .04$, $d = 1.17$. Thus, for our analysis comparing multiple-choice testing and studying, we again estimated which real items were marginal using participants' FOK judgments.

All analyses that follow are restricted to items which participants failed to retrieve the target on the initial test (60 % [SD = .21] of real and 100 % [SD = .00] of fictitious items).

Effects of delay on benefits of multiple-choice testing Before turning to our key comparison of multiple-choice testing with studying, we examined whether the benefits of multiple-choice testing persisted over a longer delay. We computed a 2 (stabilizing activity: MC test, none) × 2 (item-type: real, fictitious) × 2 (delay: none, 1 week) ANOVA on the proportion of final-test questions answered with targets. The data appear in Table 7. Consistent with our earlier results, we found a positive testing effect that was much larger for real than fictitious items $F(1,94) = 118.89$, $MSE = .02$, $\eta^2 = .18$. It did decrease across the delay $F(1,94) = 58.45$, $MSE = .01$, $\eta^2 = .04$, however, even after one week, the advantage of multiple-choice testing was quite large, $t(85.36) = 6.04$, $SED = .04$, $d = 1.23$.[1]

Table 7 Proportion of Final-Test Questions Answered with Targets (given Initial Test Failure). Data are from Experiment 3b (from MC Test and No Activity Conditions Only)

	Real Questions		Fictitious Questions	
	No Delay	1 Week	No Delay	1 Week
MC Test	.55 (.03)	.40 (.03)	.06 (.02)	.01 (.01)
No Activity	.07 (.03)	.16 (.03)	.00 (.01)	.00 (.01)
MC Test – No Activity	.48	.24	.06	.01

Note. Standard error is noted in parentheses. Third row depicts benefit of multiple-choice testing over no activity

Comparing multiple-choice testing and studying at a longer delay As in Experiment 3a, we restricted real items in this analysis to those rated as marginal by participants (a positive FOK judgment). We computed a 2 (stabilizing activity: study, MC test) × 2 (item-type: estimated marginal, newly learned) × 2 (delay: none, 1 week) ANOVA on the proportion of final-test questions answered with targets. The data appear in Table 8 (first and second rows).

Replicating Experiment 3a, the benefits of studying over multiple-choice testing existed mainly for new learning that was assessed immediately; (significant stabilizing activity × item-type × delay interaction, $F(1,94) = 15.21$, $MSE = .02$, $\eta^2 = .01$). Studying had the advantage over multiple-choice testing for new learning on the immediate final test, $t(60.96) = 9.81$, $SED = .27$, $d = 2.00$.[1] While this advantage persisted across the week delay, $t(48.90) = 4.55$, $SED = .05$, $d = .93$, the effect size decreased dramatically (from a Cohen's d of 2.00 to .93) suggesting that most new learning was forgotten.[1] Most

Table 8 Proportion of Final-Test Questions Answered with Targets (given Initial Test Failure). Data are from Experiment 3b. Estimated Marginal Knowledge comprises real items given a positive FOK and New Learning comprises all fictitious items

	Estimated Marginal Knowledge		New Learning	
	No Delay	1 Week	No Delay	1 Week
Study	.70 (.03)	.52 (.04)	.33 (.02)	.05 (.01)
MC Test	.61 (.03)	.44 (.04)	.06 (.02)	.01 (.01)
No Activity	.07 (.03)	.18 (.03)	.01 (.02)	.01 (.01)
Study – MC Test	.09	.08	.27	.04

Note. Standard error is noted in parentheses. Fourth row depicts benefit of studying over multiple-choice testing

importantly, multiple-choice testing and studying reactivated similar levels of estimated marginal knowledge on both the immediate $t(94) = 1.83$, $SED = .04$, $p = .07$, and delayed final tests, $t(94) = 1.53$, $SED = .05$, $p = .13$.

While FOK judgments were good predictors of marginal knowledge, they were not perfect. Thus, we re-did the analysis in the multiple-choice testing condition, using the traditional definition of marginal knowledge (initial retrieval failure followed by recognition). The pattern of results mirrored that of the FOK data from the current experiment and the results from the earlier experiments. As illustrated in Fig. 3 (panel B), participants retained many more real than fictitious target selections, $F(1,47) = 588.70$, $MSE = .02$, $\eta^2 = .74$. Production of targets dropped over one week, with a floor effect for fictitious items leading to a significant interaction between item-type and delay, $F(1,47) = 6.90$, $MSE = .02$, $\eta^2 = .01$.

Negative testing effects Again, participants who answered multiple-choice questions ($M = .08$) were more likely to reproduce lures on the final test than those in the no-activity condition ($M = .02$), $F(1, 94) = 139.15$, $MSE = .01$, $\eta^2 = .30$. This negative testing effect was similar for real and fictitious items, $F < 1$, and diminished with time, $F(1,94) = 26.13$, $MSE = .01$, $\eta^2 = .09$. In fact, after a week, the negative testing effect disappeared entirely for real items, $t(94) = 1.25$, $SED = .01$, $p = .22$. As in Experiment 3a, participants who studied the targets produced fewer lures on the final test ($M = .01$), $F(1,94) = 4.95$, $MSE = .01$, $\eta^2 = .46$. However, this effect operated similarly for real and fictitious items, $F(1,94) = 3.30$, $MSE = .01$, $p = .07$, $\eta^2 = .01$, and decreased with time, $F(1,94) = 36.06$, $MSE = .01$, $\eta^2 = .13$.

Discussion

After one week, studiers forgot most (but not all) new learning, permitting an (almost) unbiased comparison of studying and multiple-choice testing. Under these fairer conditions, Experiment 3b confirmed the findings from Experiment 3a. The main advantages of studying over multiple-choice testing were that studiers learned more new information (albeit false in this case) and were protected from a negative testing effect. After a week, however, the majority of the new learning was forgotten and there was no longer a negative testing effect to be protected from. Crucially, in terms of stabilizing access to (estimated) marginal knowledge, studying and multiple-choice testing were equally beneficial.

Given the vast literature on testing effects, it might come as a surprise that multiple-choice testers did not outperform studiers via a retrieval practice effect. One possibility is that when it comes to marginal knowledge, all one needs to do in order to ensure future access is to be exposed to the target. Whether that exposure consists of being told the target or recognizing it out of a list may not make a difference. A second possibility is that our multiple-choice tests did not offer very much retrieval practice. While the literature is mixed, there are several studies that find smaller testing effects with multiple-choice tests than short-answer tests and attribute this to less retrieval effort (e.g., Butler & Roediger, 2007; Kang et al., 2007; McDaniel, Anderson, Derbish, & Morrisette, 2007). Indeed, if knowledge is marginal, it ought to be recognized immediately and perhaps little retrieval effort is exerted. A final possibility is that multiple-choice testing might win out at longer delays. This explanation does not seem likely considering we saw no differences between two days and one week, but it is certainly an open question for future research.

General Discussion

Building on the small literature about marginal knowledge (mostly from Bahrick's laboratory), all of our experiments provided strong support for the concept of marginal knowledge: following a retrieval failure, participants were much more likely to recognize the answer for questions targeting marginal knowledge than new learning. All four experiments also showed that marginal knowledge can easily be reactivated, through multiple-choice testing or (in Experiments 3a and 3b) re-exposure. Overall, these findings fit well with the idea that most learning happens across many trials rather than one (e.g., Rawson, Dunlosky, & Sciartelli, 2013).

Providing feedback after multiple-choice testing did not further increase access to marginal knowledge, as compared to multiple-choice testing alone. Feedback did minimize the negative testing effect; however, we do not feel the need to recommend feedback just for this purpose, since the negative testing effect tends to be small and naturally diminishes with time. Of course, if the goal is to teach new knowledge as well as stabilize access to old knowledge, feedback is recommended.

In two experiments, multiple-choice testing and re-exposure were equally effective strategies for stabilizing access to marginal knowledge. While we were initially surprised by these results (given the power of retrieval practice in other contexts), the results highlight how relatively simple it is to reactivate knowledge. It is noteworthy that stabilization persisted over a week, given that multiple-choice testing is a pervasive tool used not only in the classroom, but also for other high-stake purposes, such as gaining admission to a desired training program or becoming certified to drive or practice law. In most cases, multiple-choice tests are used for assessment purposes, but more recently cognitive psychologists are encouraging their use as a learning tool (e.g., Little et al., 2012; Marsh & Cantor, 2014). Our experiments document a new benefit of multiple-choice testing: such tests can help students maintain access to information that might otherwise be lost. We believe multiple-choice testing could be

particularly helpful at the beginning of a new course or topic where students have to draw on knowledge that has not been retrieved recently. Educators may feel the need to review and re-teach this information and we encourage them to utilize multiple-choice testing as one tool for this purpose.

Theoretically, these ideas and data can be interpreted within Bjork and Bjork's (1992) New Theory of Disuse. That is, the problem with marginal knowledge is not with storage strength, but with retrieval strength that is low at a given moment. Marginal knowledge likely has low retrieval strength because it is unlikely to have been retrieved recently. In addition, other related items (with stronger retrieval strength) may compete with it, and/or the necessary cues are not available to increase retrieval strength. Multiple-choice testing works to stabilize access to marginal knowledge because it increases the retrieval strength of the target information, consistent with the Theory's assumptions that successfully retrieving an item from memory should increase the retrieval strength of that item and that this increase in retrieval strength is an increasing function of existing storage strength. In addition, increases in storage strength are assumed in the Theory to be greater the *lower* the current retrieval strength, so the increments in storage strength should be large for items in marginal knowledge, which will then retard the loss of retrieval strength over the subsequent delay.

More broadly, the present research highlights the instability of our knowledge base. While an impressive amount of information is stored in memory, individual items can fluctuate in accessibility: a word or name that is unavailable at one time is not necessarily gone forever – it can be reactivated via re-exposure or answering a multiple-choice question. People have a good sense of what they know and what can be reactivated, even if they cannot currently access it. Our results highlight how important it is to consider how the knowledge that individuals can report at any moment is likely a vast underestimate of their stored knowledge.

Author Note This research was supported by a Collaborative Activity Award from the James S. McDonnell Foundation's 21st Century Science Initiative in Bridging Brain, Mind and Behavior (EJM). We would like to thank Sarah Cox for her help with coding. We also thank the members of the Marsh Lab for their helpful comments and suggestions on earlier drafts of the manuscript.

References

Agarwal, P. K., Bain, P. M., & Chamberlain, R. W. (2012). The value of applied research: Retrieval practice improves learning and recommendations from a teacher, a principal, and a scientist. *Educational Psychology Review, 24,* 437–448.

Bahrick, H. P. (1984). Semantic memory content in permastore: Fifty years of memory for Spanish learning in school. *Journal of Experimental Psychology: General, 113,* 1–29.

Bahrick, H. P., & Hall, L. K. (1991). Preventative and corrective maintenance of access to knowledge. *Applied Cognitive Psychology, 5,* 1–18.

Bahrick, H. P., & Phelps, E. (1988). The maintenance of marginal knowledge. In U. Neisser & E. Winograd (Eds.), *Remembering Reconsidered: Ecological and traditional approaches to the study of memory* (pp. 178–192). Cambridge: Cambridge University Press.

Bahrick, H. P., Bahrick, P. O., & Wittlinger, R. P. (1975). Fifty years of memory for names and faces: A cross-sectional approach. *Journal of Experimental Psychology: General, 104,* 54–75.

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

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes* (Vol. 2, pp. 35–67). Hillsdale: Erlbaum.

Brown, A. S. (1991). A review of the tip-of-the-tongue experience. *Psychological Bulletin, 109,* 204–223.

Brown, R., & McNeil, D. (1966). The "tip of the tongue" phenomenon. *Journal of Verbal Learning and Verbal Behavior, 5,* 325–337.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6*, 3–5.

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

Butler, A. C., & Roediger, H. L., III. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition, 36,* 604–616.

Cohen, G., & Faulkner, D. (1986). Memory for proper names: Age differences in retrieval. *British Journal of Developmental Psychology, 4,* 187–197.

Gruneberg, M. M., Smith, R. L., & Winfrow, P. (1973). An investigation into response blockaging. *Acta Psychological, 37,* 187–196.

Hart, J. T. (1965). Memory and the feeling-of-knowing experience. *Journal of Educational Psychology, 56,* 208–215.

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

Landauer, T. K. (1986). How much do people remember? *Cognitive Science, 10,* 477–493.

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

Marsh, E. J., & Cantor, A. D. (2014). Learning from the Test: Do's and Don'ts for Using Multiple-Choice Tests. Chapter to appear. In M.A. McDaniel & R.F. Frey, S.M. Fitzpatrick, & H.L. Roediger (Eds.), *Integrating Cognitive Science with Innovative Teaching in STEM Disciplines. (in press)*

Marsh, E. J., Roediger, H. L., Bjork, R. A., & Bjork, E. L. (2007). The memorial consequences of multiple-choice testing. *Psychonomic Bulletin & Review, 6,* 194–199.

Marsh, E. J., Fazio, L. K., & Goswick, A. E. (2012). Memorial consequences of testing school-aged children. *Memory, 20,* 899–906.

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

Nelson, T. O., Gerler, D., & Narens, L. (1984). Accuracy of feeling-of-knowing judgments for predicting perceptual identification and relearning. *Journal of Experimental Psychology: General, 113,* 282–300.

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

Read, J. D., & Bruce, D. (1982). Longitudinal tracking of difficult memory retrievals. *Cognitive Psychology, 14,* 280–300.

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

Roediger, H. L., III, & Marsh, E. J. (2005). The positive and negative consequences of multiple choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 1155–1159.

Roediger, H. L., III, Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: Long-term improvements from quizzing. *Journal of Experimental Psychology: Applied, 17,* 382–395.

Tulving, E., & Pearlstone, Z. (1966). Availability versus accessibility of information in memory for words. *Journal of Verbal Learning and Verbal Behavior, 5,* 381–391.

Wixted, J. T. (2004). On Common Ground: Jost's (1897) Law of Forgetting and Ribot's (1881) Law of Retrograde Amnesia. *Psychological Review, 11,* 864–879.

Yarmey, A. D. (1973). I recognize your face but I can't remember your name: Further evidence on the tip-of-the-tongue phenomenon. *Memory & Cognition, 1,* 287–290.