Understanding How Knowledge Fluctuates in Accessibility

by

Allison Cantor Black-Maier

Department of Psychology & Neuroscience
Duke University

Date:_______________________

Approved:

___________________________

Elizabeth Marsh, Supervisor

___________________________

David Rubin

___________________________

Makeba Wilbourn

___________________________

Felipe De Brigard

Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Psychology & Neuroscience in the Graduate School of Duke University

2017
ABSTRACT

Understanding How Knowledge Fluctuates in Accessibility

by

Allison Cantor Black-Maier

Department of Psychology & Neuroscience
Duke University

Date:_______________________

Approved:

___________________________
Elizabeth Marsh, Supervisor

___________________________
David Rubin

___________________________
Makeba Wilbourn

___________________________
Felipe De Brigard

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Psychology & Neuroscience in the Graduate School of Duke University

2017
Abstract

While an impressive amount of knowledge is stored in memory, individual items can fluctuate in accessibility: Two attempts to retrieve the same knowledge (e.g., the US States) often yield somewhat inconsistent results. In this way, knowledge is unstable. The first goal of my dissertation was to document these fluctuations and to examine whether delay and testing impact this instability. Second, since knowledge is often described as being more stable relative to event memory, I directly compared fluctuations in access for the two memory types. The substantial fluctuations in knowledge documented in my first two experiments should have key implications for education, where students are regularly expected to draw upon prior knowledge. Accordingly, my third goal was to examine the power of multiple-choice testing as a tool for reactivating knowledge that has become inaccessible. In the laboratory, I compared multiple-choice testing to studying the target knowledge. My final goal was to investigate these issues in a real classroom; I evaluated the effectiveness of multiple-choice testing intervention in reactivating background course knowledge and promoting the acquisition of new material. Overall, my results highlight the instability of the knowledge base, with individual pieces of information coming in and out of reach.
Contents

Abstract ........................................................................................................................................ iv

List of Tables .................................................................................................................................. x

List of Figures ................................................................................................................................. xi

Acknowledgements ......................................................................................................................... xiii

1. Introduction ............................................................................................................................... 1
   1.1 Defining Stability vs. Instability ............................................................................................. 3
   1.2 Demonstrations of Inaccessible Knowledge ........................................................................... 5
   1.3 Factors Influencing Knowledge Accessibility ......................................................................... 8
       1.3.1 The Role of Memory Strength ......................................................................................... 10
       1.3.2 The Role of Recent Experience ....................................................................................... 11
       1.3.3 The Role of Retrieval Cues ............................................................................................ 12
       1.3.4 The Role of Interference and Inhibition ......................................................................... 14
       1.3.5 Summary ........................................................................................................................ 17
   1.4 Knowledge Instability in Education ....................................................................................... 18
       1.4.1 Reactivating Marginal Knowledge .................................................................................. 19
   1.5 Overview of Current Experiments ......................................................................................... 21

2. Experiments 1A & 1B ............................................................................................................... 24
   2.1 Method ................................................................................................................................... 25
       2.1.1 Participants ....................................................................................................................... 25
       2.1.2 Design ........................................................................................................................... 26
2.1.3 Materials and Counterbalancing ................................................................. 27
2.1.4 Procedure ................................................................................................... 28

2.2 Results ........................................................................................................... 30
2.2.1 Coding ........................................................................................................ 30
2.2.2 Stability in Total Exemplars Retrieved ....................................................... 31
2.2.3 Instability in Specific Exemplars Retrieved ................................................ 32
2.2.4 Effects of Delay .......................................................................................... 33
2.2.5 Effects of Testing ....................................................................................... 34

3. Experiment 2 .................................................................................................... 36
3.1 Method ............................................................................................................ 36
3.1.1 Participants ................................................................................................. 36
3.1.2 Design ......................................................................................................... 37
3.1.3 Materials and Counterbalancing ................................................................. 38
3.1.4 Procedure .................................................................................................. 39

3.2 Results ............................................................................................................ 41
3.2.1 Replication of Experiment 1 ........................................................................ 41
3.2.2 Continuously Cumulating Meta-Analyses ................................................... 41
3.2.3 Knowledge vs. Event Memory ..................................................................... 44
3.2.3.1 Coding ................................................................................................... 44
3.2.3.2 Stability in Total Exemplars Retrieved .................................................. 44
3.2.3.3 Instability in Specific Exemplars Retrieved ............................................ 45
3.2.3.4 Effects of Delay ...................................................................................... 46
3.2.3.5 Effects of Testing........................................................................................................ 47
3.2.3.6 Item Analyses ............................................................................................................. 48
3.3 Discussion of Experiments 1A, 1B, and 2....................................................................... 51
4. Experiments 3A & 3B ......................................................................................................... 56
4.1 Experiment 3A .................................................................................................................. 58
4.1.1 Method .......................................................................................................................... 58
4.1.1.1 Participants .............................................................................................................. 58
4.1.1.2 Design ...................................................................................................................... 58
4.1.1.3 Materials and Counterbalancing ............................................................................ 58
4.1.1.4 Procedure ............................................................................................................... 60
4.1.2 Results ........................................................................................................................ 61
4.1.2.1 Coding .................................................................................................................... 61
4.1.2.2 Determining Marginal Knowledge ....................................................................... 61
4.1.2.3 Benefits of Multiple-Choice Testing ..................................................................... 62
4.1.2.4 Comparing the Benefits of Studying and Multiple-Choice Testing ................. 64
4.2 Experiment 3B ................................................................................................................ 66
4.2.1 Method ........................................................................................................................ 67
4.2.1.1 Participants ............................................................................................................ 67
4.2.1.2 Design .................................................................................................................... 67
4.2.1.3 Materials and Counterbalancing .......................................................................... 67
4.2.1.4 Procedure ............................................................................................................... 67
4.2.2 Results ........................................................................................................................ 68
Biography
List of Tables

Table 1: Continuously Cumulating Meta-Analyses (CCMAs) for the effects of delay and testing on losses and gains in Experiments 1A, 1B, and 2 (knowledge condition only). . 43

Table 2: Examples of real (R) and fictitious (F) questions with targets and lures used in Experiments 3A and 3B............................................................. 59

Table 3: Proportion of final-test questions answered with targets (given initial test failure) in Experiments 3A and 3B. ................................................................. 63

Table 4: 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. ................. 65

Table 5: 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. ............................ 70

Table 6: Examples of prior course (P) and new course (N) material questions with answers for the six topics used in Experiment 4. .................................................. 78

Table 7: Proportion of prior course and new course material questions answered correctly on the mid-course and end-of-course assessments (given failure on the baseline assessment). Data are from Experiment 4. ................................................................. 84

Table 8: Correlations between performance on prior and new course material questions by topic for the baseline, mid-course, and end-of-course assessments. Data are from Experiment 4. ............................................................................. 86

Table 9: Average number of exemplars lost and gained from the initial test to the immediate retest, delayed retest, and delayed retest with an intervening retest. Data are from Experiments 1A, 1B, and 2 (knowledge condition only). .................................................. 96

Table 10: Average number of exemplars lost and gained from the initial test to the immediate retest, delayed retest without an intervening retest, and delayed retest with an intervening retest. Data are from Experiment 2. ............................................................................. 96
List of Figures

Figure 1: Schematic of Experiments 1A and 1B. To examine the effects of delay, compare changes in recall from A to B (immediate) with changes in recall from A to E (delayed). To examine the effects of testing, compare changes in recall from A to E (not tested) with changes in recall from A to D (tested). ................................................................. 27

Figure 2: A schematic of scoring in Experiments 1A and 1B .................................................. 30

Figure 3: Average number of unique responses retrieved on the initial test, immediate retest, and delayed retests (with and without an intervening retest) for a single category in Experiments 1A, 1B, and 2 (knowledge condition only). .............................................. 31

Figure 4: Proportion of items lost (A) and gained (B) from the initial test to immediate and delayed retest (without an intervening retest). Data are from Experiments 1A, 1B, and 2 (knowledge condition only). Error bars represent standard error of the mean...... 33

Figure 5: Proportion of items lost (A) and gained (B) from the initial test to week-delayed retests with and without an intervening immediate retest. Data are from Experiments 1A, 1B, and 2 (knowledge condition only). Error bars represent standard error of the mean................................................................. 34

Figure 6: Schematic of Experiment 2. To examine the effects of delay, compare changes in recall from A to B (immediate) with changes in recall from A to E (delayed). To examine the effects of testing, compare changes in recall from A to E (not tested) with changes in recall from A to D (tested). .............................................................................. 38

Figure 7: Forest plots for the effect of delay on losses (A), the effect of delay on gains (B), the effect of testing on losses (C) and the effect of testing on gains (D). Each line is an experiment for which the box size represents study weighting and the line width represents 95% confidence interval. The diamond represents the overall effect, with the edges marking the 95% confidence interval. ................................................................. 42

Figure 8: Average number of responses (out of a possible 40) retrieved for the initial test, immediate retest, and delayed retests (with and without an intervening retest) for a single category in Experiment 2 ...................................................................................... 44

Figure 9: Proportion of items lost (A) and gained (B) from the initial test to immediate and delayed retest (without an intervening retest) for event memory and knowledge
conditions. Data are from Experiment 2. Error bars represent standard error of the mean.

Figure 10: Proportion of items lost (A) and gained (B) from the initial test to delayed retests with and without an intervening immediate retest for event memory and knowledge conditions. Data are from Experiment 2. Error bars represent standard error of the mean.

Figure 11: Proportion of items recalled on the initial test for the event memory (A) and knowledge (B) conditions as a function of the item’s rank (1 = least common, 40 = most common). Data are from Experiment 2.

Figure 12: Proportion of immediate event losses (A), delayed event losses (B), immediate knowledge losses (C), and delayed knowledge losses (D) from the initial test to the retests as a function of the item’s rank (1 = least common, 40 = most common). Data are from Experiment 2.

Figure 13: Proportion of immediate event gains (A), delayed event gains (B), immediate knowledge gains (C), and delayed knowledge gains (D) from the initial test to the retests as a function of the item’s rank (1 = least common, 40 = most common). Data are from Experiment 2.

Figure 14: 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.

Figure 15: Schematic of Experiment 4.
Acknowledgements

First and foremost, I would like to thank my advisor, Elizabeth Marsh, for being such a dedicated mentor. You have been incredibly supportive, always helping me to develop my ideas, suggesting different perspectives, and offering careful feedback on countless paper drafts. As I’ve progressed through graduate school, you have challenged me to grow as an independent thinker, writer, and researcher. I appreciate all the time and effort you invested in me. I’d also like to thank my committee members, David Rubin, Felipe De Brigard, and Makeba Wilbourn for your thoughtful feedback and enthusiastic encouragement. To the Marsh lab members, past and present, thank you for being great friends in addition to being collaborators. You have made Duke a happy place for me. To my family, particularly my parents, I would not be here without your unconditional love and support. And finally, Eric, thank you for my biggest supporter, my closest confidant, and my best friend.
1. Introduction

Human knowledge (e.g., vocabulary, facts, schemas) and event memories (e.g., episodes from one’s personal past) are often treated as different kinds of memory: They differ in phenomenology, show different time-courses over development, and are often measured and assessed via unique methods. These differences were captured in Tulving’s (1972) distinction between semantic memory (knowledge) and episodic memory (event memory). Recent work suggests that a sharp division between the two memories is no longer appropriate. While such a distinction may be “heuristically useful” (Baddeley, 1984, p. 238), many researchers have rejected the idea of two separate brain-based systems (e.g., Cabeza & Moscovitch, 2013; Ryan, Hoscheidt, & Nadel, 2008). Instead, event memories and knowledge are constructed from common content and are drawn from a single system (Rubin & Umanath, 2015), a position supported by both behavioral and neuroimaging data highlighting the interactive and complementary nature of retrieving episodes and knowledge (Nelson & Shiffrin, 2013). Behaviorally, knowledge supports episodic retrieval; for example, it is easier to learn and remember new information about familiar versus unfamiliar people (e.g., Kole & Healy, 2007). Conversely, one can draw on one’s personal memories when retrieving knowledge, such as when one imagines oneself in one’s kitchen when generating exemplars of “kitchen utensils” (Vallee-Tourangeau, Anthony, & Austin, 1998; see also Westmacott &

---

1 Parts of this chapter are under review for publication in a larger manuscript at Psychological Bulletin.
Moscovitch, 2003). Furthermore, neuroimaging evidence finds that traditionally episodic and semantic tasks (e.g., category cued recall versus category exemplar production) yield overlapping patterns of activation, even in the hippocampus, which was previously believed to be involved only in event memory retrieval (e.g., Ryan, Cox, Hayes, & Nadel, 2008; see also Burianova & Grady, 2007; Rajah and McIntosh, 2005).

Even as the field emphasizes the overlap and interactions between knowledge and event memories, vestiges remain from Tulving’s (1972) arguments. The distinction was originally based on differences in function and phenomenology (e.g., with events “remembered” but knowledge “known”). Among other differences, Tulving also argued that “forgetting appears to be more readily produced in the episodic than in the semantic system” (Tulving, 1972, p. 391). Variations on this claim are still prevalent today, such as assertions in textbooks and papers that “episodic memories are more fragile than semantic memories” (Kalat, 2014, p. 227; see also Coon & Mitterer, 2013; Eysenck, 2009; etc.). It is often assumed that an individual’s knowledge “would not change over a short delay” (Sitzman, Rhodes, & Tauber, 2014, p. 86). Similarly, recent debates about retrieval and reconsolidation have focused on event memories, given that “once that information is well-learned, the semantic traces remain stable and can be accessed without affecting the representation” (Aue, Criss, & Prince, 2015, p. 1803). This purported difference between knowledge and event memories was the focus of my dissertation.
The purpose of my dissertation was to evaluate this claim that human knowledge is relatively stable. In four experiments, I examined variables that influence the degree to which knowledge is stable versus unstable, directly compared the stability of knowledge with that of event memory, and investigated methods for reestablishing “forgotten” knowledge in both the laboratory and the classroom. In the sections that follow, I begin by defining exactly what I mean by “stable” and review examples of how knowledge – like event memory – can be considered quite unstable under this definition. Next, I identify factors that determine whether specific knowledge is currently accessible, which will have key implications for the conditions under which knowledge is stable vs. unstable. Finally, I discuss the issue of knowledge instability with respect to education, focusing on how students can regain access to knowledge that has become inaccessible.

1.1 Defining Stability vs. Instability

What does it mean for memory to be “stable”? One might focus on the stability of the total amount of information stored in memory, for example, highlighting that an individual’s IQ score remains relatively constant over time (e.g., Craik, 2000). Instead, my interest is the stability of the individual pieces of information stored in memory. Critically, demonstrations of the stability of overall test scores do not require the stability of the individual items contributing to those scores – a score could stay the same even if some information was forgotten so long as new information was recovered.
For example, consider the student who is asked to name the fifty United States on two separate tests. On both tests, her overall score could be exactly 42 states, but the individual states recalled might be inconsistent. On the first test, she might recall Wyoming and fail to recall Delaware but on the second test she recalls Delaware but not Wyoming. Brown (1923) was likely the first to demonstrate the inconsistencies between two successive attempts to retrieve information from knowledge. In this simple study, students were asked to name the United States twice, 30 minutes apart. Brown found that several states that were inaccessible on the first test became accessible on the second, which is termed reminiscence (Ballard, 1913). Additionally, a few states that were recalled on the first test were “forgotten” and not recalled on the second. Similar effects have been found many times, including with the production of other category exemplars (e.g., sports, birds, US Presidents, etc.; Roediger, Payne, Gillespie, & Lean, 1982) and with retrieval of answers to general knowledge questions (e.g., What is the name of a young sheep?; Umanath, 2016).

How inconsistent are two retrievals from knowledge? In an attempt to quantify the reliability of retrieval from stored knowledge, Bellezza (1984a; 1984b; 1984c; 1987; 1988) asked participants to repeatedly retrieve category exemplars, noun meanings, information about friends and family, scripts, and self-information. Participants completed an initial generation test and then returned one week later to take the exact same test. Across these various types of information, Bellezza calculated modest, but not
very high, within-individual reliability between two retrievals from stored knowledge. For example, participants were more consistent over time when naming musical instruments \((r = .80\) between the two generation attempts\) than when defining abstract nouns \((r = .43\) (Bellezza, 1984a; 1984c). This variability is not limited to basic facts, but also extends to higher level processes, such as the organization of knowledge. For example, categories are often “fuzzy” in nature rather than well-defined. The result is that participants shift their classifications over time, categorizing a tomato as a fruit initially but as a vegetable one month later (McCloskey & Glucksberg, 1978; see Barsalou, 1987 for a review).

### 1.2 Demonstrations of Inaccessible Knowledge

Given the inconsistency between two retrieval attempts, there must be knowledge that is stored but is not retrieved at a particular moment (marginal knowledge; Bahrick & Phelps, 1988). Returning to the United States retrieval example, Delaware was clearly stored at the time of the first test (since it was later recalled on the second test without any intervening study session) but was not successfully retrieved. Capturing this idea, Tulving and Pearlstone (1966) proposed a distinction between availability and accessibility – items that are stored in memory are available but only items that can be retrieved are currently accessible. To demonstrate this important distinction, the researchers had participants study categorized word lists (e.g., professions – engineer, lawyer, doctor) and then take a free recall test either with the category name
(professions) as a cue or without a cue. Performance on this recall test was much higher when participants were cued with the category name, indicating that additional items were available in memory, but not accessible without further cues.

While Tulving and Pearlstone (1966) conceptualized the availability versus accessibility distinction in an event memory paradigm, this distinction also holds for knowledge. For example, people typically demonstrate more knowledge on recognition (multiple-choice) tests than they can produce on recall (short answer or essay) tests. This pattern holds broadly: It occurs in education when examining students’ long-term retention of course content learned in school (see Semb & Ellis, 1994 for a review), in everyday life when people remember names and faces of former classmates (Bahrick, Bahrick, & Wittlinger, 1975), in ethnography research where interviewers often use “free listing” as a tool to better understand cultural knowledge (e.g., Brewer, 2002; Hutchinson 1983), as well as in recent work on collective memory where people recognize far more of the US Presidents than they can produce from memory (Roediger & DeSoto, 2014; 2016).

Metacognitive measures indicate that people are typically aware that they have marginal knowledge. For example, when participants failed to successfully recall the target answer to a general knowledge question, Hart (1965) asked them to make a “feeling-of-knowing” (FOK) judgment: To predict whether or not they would be able to successfully recognize the target answer out of a list of choices. He demonstrated that
FOK judgments were quite accurate predictors of storage (see also Gruneberg & Monks, 1974; Nelson, Gerler, & Narens, 1984). Furthermore, sometimes people are highly confident that knowledge is stored and they feel painfully close to being able to retrieve it – such instances are special cases of marginal knowledge and are referred to as “tip-of-the-tongue” (TOT) states (e.g., Brown & McNeil, 1966; see Brown, 2012 for a review). Psychologists at least as far back as William James (1893) have acknowledged such retrieval “gaps” that are “intensely active… making us at moments tingle with the sense of our closeness” (James, 1893, p. 251). The TOT experience is nearly universal, documented in everyday settings (e.g., Burke, MacKay, Worthley, & Wade, 1991) and reliably induced in the laboratory (e.g., Brown & McNeil, 1966).

But what is the evidence that knowledge was actually stored at the time of the retrieval failure, as opposed to being learned later? The relative accuracy of FOK judgments suggests that the answer was in fact stored in memory, but it could also be the case that related accessible information helped one to determine the likelihood of winnowing down the choices to the correct one. Such explanations are much less likely when examining fluctuations in retrieval of over-learned concepts like US states – although possible, it is highly unlikely that people learned additional states in the interval between tests. Additional evidence comes from TOT studies, which show that people sometimes have partial access to the inaccessible word, correctly naming its first letter and number of syllables (e.g., Brown & McNeil, 1966).
The most definitive demonstration of inaccessible knowledge compared hypothesized marginal knowledge to fictional items, for which it was not possible to have marginal (or any) knowledge (Berger, Hall, & Bahrick, 1999). Parallel sets of questions were created – one real, the other fictitious. The fictitious questions (e.g., *Who was Captain Child’s greatest enemy?*) shared the same structure as real questions (e.g., *Who was Robin Hood’s greatest enemy?*) but were entirely made up by the researchers. The logic was that manipulations aimed at reestablishing access to marginal knowledge should be much more beneficial for real than fictitious items. Any improvement in performance on fictitious items would have to reflect new learning as there would be no prior knowledge to reactivate. The actual experiment consisted of an initial short-answer test, to ascertain what participant could not retrieve. The question was whether a brief exposure to the answer would be equally beneficial for real and fictitious questions answered incorrectly on the initial test; it was not. Participants benefitted much more from exposure to answers of real items, indicating that they had marginal knowledge stored in memory for at least some of the real items.

### 1.3 Factors Influencing Knowledge Accessibility

The first major goal of my dissertation was to understand the factors that determine whether knowledge is stable versus unstable. In other words, what influences the consistency of underlying items across repeated attempts to retrieve one’s knowledge? What conditions promote the “forgetting” of Wyoming and the “recovery”
of Delaware? In order to address these questions, data with two time points is needed. That is, participants’ knowledge must be tested twice so that one can measure the items lost and the items gained from the first test to the retest. Using this type of paradigm, one can manipulate various factors (e.g., the delay between the two tests) to identify factors that influence the degree of stability. However, such a database does not currently exist.

Instead, most experiments involving knowledge retrieval contain only a single time point. For example, most TOT studies consist of one experimental session where participants are presented with definitions of words and are tasked with retrieving the target word. Experimenters have manipulated various factors (e.g., recent exposure to the target word, whether the target words are high vs. low frequency, etc.) in order to identify when TOT states are more or less likely. While such data do not directly address questions about knowledge stability per se, they do point to factors that determine whether knowledge is accessible at a single moment in time. Understanding the factors influencing momentary knowledge accessibility will be critical in generating predictions about fluctuations in accessibility. Specifically, to the extent that such factors change across retrievals, knowledge should be more likely to fluctuate in accessibility. For example, since interference from competing knowledge impacts momentary knowledge accessibility (see below), access to target knowledge should be more likely to fluctuate to the extent that the amount of interference changes from one retrieval to the next. Thus,
as a starting point, I first identified key factors that determine whether specific
knowledge is accessible at a given moment. In the following sections, I briefly outline
four factors that affect the current accessibility of knowledge while drawing parallels to
the event memory literature.

1.3.1 The Role of Memory Strength

Some memories are more likely to stand the test of time. For example, hundreds
of studies have shown that events are more likely to be retrieved following meaning-
based encoding (deep processing) than encoding focused on perceptual features
(shallow encoding; i.e., encoding a word by counting the number of syllables it contains)
(levels of processing; Craik & Lockhart, 1972). Pictures are more likely to be remembered
than words (e.g., Paivio & Csapo, 1973) and emotional events are more likely to be
remembered than neutral ones (e.g., Buchanan, 2007).

Asking the parallel question about knowledge, however, it is trickier –
oftentimes there is no measure how well knowledge was originally learned. Initial
evidence comes from classroom data which shows that student grades significantly
predict long-term retention of course material, even after controlling for other relevant
factors such as taking additional higher-level courses (e.g., Bacon & Stewart, 2006;
Bahrick, 1984). Further evidence for the role of memory strength comes from TOT states,
which are more common for low-frequency words than high-frequency words (e.g.,
Burke et al., 1991; Harley & Bown, 1998). Of course, in addition to greater storage
strength, high frequency words are also likely to have been encountered more recently – which also increases accessibility, as will be discussed next. One way around this problem is to examine words that were acquired early in life (Age of Acquisition; see Juhasz, 2005 for a review), which are argued to have greater storage strength than words acquired later in life because they are encountered more times over the lifespan (Cumulative Frequency Hypothesis; Carroll & White, 1973). Importantly, independent of word frequency in the language (thereby avoiding the confound of recent access), words acquired earlier in life are less likely to elicit TOT states than are words acquired later in life (Navarrete, Pastore, Valentini, & Peressotti, 2015).

1.3.2 The Role of Recent Experience

All else being equal, the more recently information has been used (i.e. retrieved, studied, reactivated), the more accessible it will be. Decades of event memory research have confirmed Ebbinghaus’ (1885/1964) classic experiment demonstrating that retention decreases as a function of time (e.g., Rubin, Hinton, & Wenzel, 1999). Similar findings occur with knowledge. My first example comes from the Berger et al. (1999) study on marginal knowledge described earlier, where the benefits of seeing a target word (reactivation of marginal knowledge) decreased over time: Immediately after knowledge was reactivated, performance was almost perfect but the likelihood of retrieval dropped continuously over time to 49% nine days later. A second example comes from the TOT literature, where recent exposure to a word lessens the chance of a
TOT state (e.g., Cleary, 2006) and increases the likelihood that an answer will be retrieved on a general knowledge test (e.g., Kelley & Lindsay, 1993). A third example involves semantics and interpretations of ambiguous words; recent encounters with a particular meaning of an ambiguous word (e.g., bank) promote accessibility of that specific meaning (e.g., Rodd et al., 2016). Of course, the mere passage of time is not the true cause of changes in accessibility. Instead, time is likely correlated with changes in retrieval cues (Estes, 1955) and exposure to other information that may interfere with retrieval of the target (McGeoch, 1932), both of which will be discussed in subsequent sections.

1.3.3 The Role of Retrieval Cues

Whether or not event memories are retrieved depends crucially on the nature of the retrieval cues available at test. The most effective cues reinstate the original event to a greater degree. For example, viewing pictures taken automatically during one’s day (via a SenseCam) dramatically boosts memory for that day (Sellen et al., 2007). The daily photos provide a wealth of cues: People, objects, events, etc. from the original viewpoint. Among other cues, music (e.g., Belfi, Karlan, & Tranel, 2016), smells (Willander & Larsson, 2006), spatial contexts (Robin, Wynn, & Moscovitch, 2016), and objects (Uzer, Lee, & Brown, 2012) can all support the retrieval of event memories.

Not surprisingly, knowledge is also more likely to be retrieved with more cues, which is one reason why students often prefer multiple-choice exams over essay tests.
People are more likely to generate target words (e.g., horse) given multiple converging cues: The first letter of the target (h), a word that rhymed with the target (coarse), or a word semantically associated with the target (chariot). A single cue was not much better than no cues at all, but the combination of two or three cues greatly promoted target generation (Solso & Biersdorff, 1975). Similarly, providing the first letter of a forgotten but desired word often resolves tip-of-the-tongue states (e.g., Brown & Knight, 1990).

Finally, the desired knowledge can serve as a retrieval cue, albeit an exact one – we are grateful when our conversational partner provides a forgotten word. For instance, when knowledge is marginal, simply seeing it is enough to reactivate it (e.g., Berger et al., 1999).

While semantic and phonological cues clearly help people retrieve knowledge, the relevance of other cues is less clear. In particular, contextual cues (physical or imagined locations) are known to promote event memory retrieval (e.g., Godden & Baddeley, 1975; see Smith & Vela, 2001 for a meta-analysis) but may not be important or even relevant for retrieving knowledge. Knowledge is often extracted across multiple episodes, potentially reducing memory for source (Watkins & Kerkar, 1975) and leading to a feeling of “knowing” rather than “remembering” a specific experience (e.g., Barber, Rajaram, & Marsh, 2008; Conway, Gardiner, Perfect, Anderson, & Cohen, 1997). To the extent that knowledge is decontextualized and sourceless, the learning context should not be a relevant cue for retrieval. Empirically, the data are mixed. One way to approach
this question is with classroom experiments where an exam’s location is manipulated to be either one’s original classroom or an alternative classroom. While some studies found that students perform better when tested in their original classroom (e.g., Metzger, Boschee, Haugen, & Schnobrich, 1979; Van Der Wedge & Barry, 2008), others found no differences (e.g., Corballis, 1988; Saufley, Otaka, & Bavaresco, 1985). Such classroom studies are not the strongest test of context-dependent memory, however, as most classrooms are fairly similar and are unlikely to provide distinctive cues. Furthermore, classroom studies do not control the qualities of to-be-remembered information, likely mixing information encountered in multiple contexts (lecture, notes, readings) with classroom demonstrations or other information uniquely associated with the classroom (for which a context-dependency effect would be more likely). One final concern is whether course learning really qualifies as “knowledge” given that it is new learning, and may still be associated with memories of the original learning episode; to that end, classroom studies are likely not the best test of the role of contextual cues in retrieving knowledge. At this point in time, it is clear that physical context matters less when retrieving knowledge than when remembering events; less clear is whether it plays a small role or none at all.

1.3.4 The Role of Interference and Inhibition

As noted earlier, time is correlated with exposure to competing information (McGeoch, 1932), which can impair access to one’s memories. In the event memory
literature, for example, researchers have utilized an A-B, A-D paradigm where participants first learn the A-B associations until they can recall them perfectly. Compared to control participants who next study the unrelated C-D pairs, participants who study the A-D pairs are later impaired at retrieving the original A-B pairs (see Roediger, Weinstein, & Agarwal, 2010 for a review).

Knowledge retrieval also suffers when competing information is learned. For example, consider the ability to name streets from one’s childhood neighborhood; naming decreases as a function of the number of times one moved in childhood, even after controlling for factors such as the number of years spent in that neighborhood (i.e. memory strength) and the number of years since the individual moved away (i.e. recency of access) (Schmidt, Peeck, Paas, & van Breukelen, 2000). With more moves, more competitors (new addresses and streets) were stored and strengthened in memory, blocking access to earlier addresses. A similar phenomenon is observed in TOT studies, where many people essentially report blocking by competitors (i.e. being “stuck” on related terms). For instance, as one tries to retrieve “the navigational instrument used in measuring angular distances, especially the altitude of sun, moon, and stars at sea”, more accessible related words like, “compass”, “protractor”, and “secant”, can come to mind and block access to the target “sextant” (Brown, 2012). This blocking process can be simulated in the laboratory by exposing participants to competitors immediately before retrieval. For example, priming orthographically similar words reduces access to
other similar words. That is, reading ANALOGY makes it harder to later complete the fragment A_L_ _ GY, whereas reading UNICORN does not (Logan & Balota, 2003; Smith & Tindell, 1997).

I also note a possible role for inhibition in knowledge fluctuations. Sometimes retrieval of a memory appears to require suppression of competitors – rendering them temporarily inaccessible (Levy & Anderson, 2002). It can be difficult to separate interference and inhibition, unless a memory remains inaccessible when a strong cue is provided. Consider a study by Johnson and Anderson (2004), who had participants generate category exemplars in response to a category cue plus fragment (e.g., seasoning – n). Participants practiced a subset of the exemplars, from some but not all of the categories (e.g., practice retrieving nutmeg to the cue seasoning - n). Retrieval-induced forgetting occurred, with poorer subsequent memory for non-practiced competitors (e.g., salt) from practiced categories relative to items from non-practiced categories (e.g., lamp). Critically, Johnson and Anderson (2004) used independent cues that were unrelated to the tested category or the other exemplars, but would normally cue the competitors (e.g., popcorn – s). Such cues should free participants from interference; only if “salt” were suppressed would one expect reduced generation of it to popcorn-s (nutmeg should not block salt since nutmeg is not related to popcorn). The results were consistent with the inhibition account, with participants less likely to recall salt than they were to generate exemplars of another non-practiced category. To be clear, this study
does not rule out interference as a cause of knowledge fluctuations, but it suggests an additional role for inhibition.

**1.3.5 Summary**

Like event memory, knowledge is more likely to be accessible to the extent that it was well-learned and experienced recently, given the presence of relevant retrieval cues, and if interference and inhibition are minimal. One important difference between event memory and knowledge seems to be the nature of the cues that contribute to fluctuations. Contextual cues appear relatively less important for retrieving knowledge, in contrast to the support they provide when attempting to remember memories of specific events.

As previewed earlier, the four factors reviewed in this section are critical in generating predictions about when knowledge is stable versus unstable. That is, knowledge should be more likely to fluctuate in accessibility to the extent that these factors change across retrievals. For example, knowledge should fluctuate more when the available retrieval cues change drastically between the two attempts. A change in retrieval cues should promote both loss of items recalled initially as well as recovery of items that were initially inaccessible. As will be explained more later, these factors provided the basis for my investigating how knowledge fluctuations might be influenced by delay and testing in Experiments 1 and 2.
1.4 Knowledge Instability in Education

While one could imagine many real-world situations where knowledge instability could have an impact, my particular interest lies in education. Students often forget previously learned material, much to the frustration of teachers (e.g., Willingham, 2009). For example, students forget roughly one month of instruction over summer break, as estimated by performance on achievement tests (e.g., Cooper 2003). All students suffer these losses, regardless of grade level, subject matter, or family income (Cooper, Nye, Charlton, Lindsay, & Greathouse, 1996). Learning loss, however, is not limited to the summer: Most students quickly forget a portion of course content once the course ends (e.g., Bahrick, Hall, & Baker, 2013; Custers, 2010). The problem is larger than the mere loss of information – losing background knowledge should hinder the acquisition of new information, because topics and courses build upon one another (e.g., Bloom, 1976; Dochy, Segers, & Buehl, 1999). For instance, AP Biology assumes knowledge of concepts from Introductory Biology, just like an upper-level Spanish course assumes a certain level of fluency. Thus, there is great value in identifying simple interventions that might help students regain access to forgotten knowledge.

Educators often treat the problem as if knowledge has been completely lost from memory and, as a result, spend three to four weeks re-teaching previous material (National Summer Learning Association, 2013). However, the research on knowledge instability reviewed earlier suggests that some “forgotten” knowledge may still be
stored in memory – it has not been lost from the system. Instead, the knowledge may have become inaccessible (i.e. marginal) with disuse. This distinction between knowledge that is lost versus simply inaccessible matters – namely, it has implications for how the information can be reinstated. To the extent that knowledge is still stored, it only needs to be made accessible again, which should be faster and easier than learning new material (e.g., Ebbinghaus, 1964).

1.4.1 Reactivating Marginal Knowledge

How can the learner regain access to knowledge that has become marginal? Berger and colleagues (1999) have shown the power of a brief exposure to the target knowledge in promoting access to previously marginal knowledge. That is, for participants who could not answer, “What is the capital of New York?”, studying “Albany” for five seconds drastically improved their ability to answer correctly up to nine days later. As explained earlier, the researchers developed a creative method to ensure that the benefit of their intervention reflected reactivated knowledge, rather than new learning. That is, they utilized both real and fictitious general knowledge questions to show that the benefit of studying inaccessible targets (i.e. ones that were not recalled on an initial test) was much larger for real than fictitious questions, indicating that the participants had marginal knowledge for many of the inaccessible targets.

Recently, my colleagues and I investigated whether multiple-choice testing could also serve the purpose of reactivating marginal knowledge (Cantor, Eslick, Marsh, Bjork,
& Bjork, 2015). There were two motivations for this particular intervention. First, multiple-choice tests are ubiquitous in the classroom and can also serve as formative assessments, which would be an added benefit to educators. Second, the rapidly growing testing effect literature finds many sizable benefits of practicing retrieving information from memory over studying, both in laboratory and classroom learning (see Roediger & Karpicke, 2006a for a review). 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.

We adopted Berger and colleagues’ clever methodology of using both real and parallel fictitious questions. For questions that participants could not answer on an initial test, we found that taking a multiple-choice test without feedback (Experiment 1) greatly promoted performance on the final test. Critically, this testing effect was much larger for real than fictitious questions, supporting the idea that multiple-choice testing served to reactivate marginal knowledge, rather than teach new information. In Experiment 2, we asked whether receiving correct answer feedback after the multiple-choice test would boost this effect. Interestingly, we found that receiving feedback boosted final test performance, but it did so equally for real and fictitious questions, indicating that the main function of feedback was to teach new information. Overall, our
research suggested that multiple-choice testing could be used as an effective tool for reestablishing access to marginal knowledge in the classroom.

1.5 Overview of Current Experiments

The first goal of my dissertation was to understand the conditions that influence the stability of knowledge across retrieval attempts. I decided to investigate the potential impacts of delay and testing on knowledge instability in Experiments 1A and 1B for two reasons. First, delay and testing have well-documented effects on event memory, so investigating them here would allow for a comparison across memory types. Second, the previously outlined factors influencing the current accessibility of knowledge suggested hypotheses for both delay and testing. As explained more in the next section, I predicted that time should make knowledge less stable and that testing might act to stabilize knowledge. To test these predictions, I had participants repeatedly generate well-learned category exemplars (e.g., fruits) and measured the within-participant consistency between the exemplars generated. To measure the effect of delay, generation attempts were separated by either 15 minutes or one week. I also manipulated whether or not participants took an immediate retest after the initial generation attempt to look for a testing effect.

The results of Experiments 1A and 1B demonstrated clear effects of delay and testing on knowledge instability. Since the results also hinted at similarities and differences between knowledge and event memory, Experiment 2 served to directly
compare fluctuations of the two memory types. In an effort to make the knowledge and event memory conditions as similar as possible, participants either repeatedly generated category exemplars or first studied a list of exemplars and then repeatedly retrieved the studied items. As in Experiment 1, I manipulated both delay and testing and examined the within-participant consistency between the exemplars retrieved at each test.

My first two experiments highlighted that even over-learned knowledge like the US states can fluctuate in accessibility. As explained earlier, this instability has important educational implications where students are regularly expected to draw upon prior knowledge. The third goal of my dissertation was to apply the issue of knowledge instability to education and answer the question: How can we reactivate knowledge that has become inaccessible? My prior work found that multiple-choice tests could serve this role, but it was unclear how this method measured up to just telling students the correct answer. Experiments 3A and 3B directly compared multiple-choice testing and studying in their ability to reestablish access to marginal knowledge.

Transitioning from the laboratory to the classroom, Experiment 4 asked whether this multiple-choice testing intervention would successfully reactivate inaccessible background knowledge in a real college course. First, I needed to establish that for at least some of the prior course material that students “forgot”, the information had actually just become inaccessible. Second, Experiment 4 examined whether a multiple-choice testing intervention would reestablish access to this background knowledge.
Finally, I investigated whether reactivation of the prior course knowledge via the multiple-choice intervention promoted acquisition of the new course material.
2. Experiments 1A & 1B

The purpose of Experiment 1 was to investigate two conditions that might influence knowledge instability: Delay and testing. As mentioned earlier, I selected these two variables in part because they are well-known to affect the stability of event memories. Specifically, increasing the delay between two event memory retrievals decreases stability by promoting forgetting (e.g., Rubin et al., 1999). Furthermore, as discussed earlier, retrieval practice is one powerful way to curb event memory forgetting (see Roediger & Karpicke, 2006a for a review). For example, Roediger and Karpicke (2006b, Experiment 2) had participants read prose passages and then either study them three more times (SSSS), study them twice more and then take a practice recall test (SSST), or take three practice recall tests (STTT); participants took a final recall test that was either immediate or delayed by one week. While performance for all conditions was lower when the final test was delayed versus immediate, participants in the STTT group showed the least forgetting after one week. Since the literature clearly documents the effects of delay and testing on event memory, investigating them here would allow for a comparison between fluctuations in event memory and knowledge.

The previously outlined factors affecting momentary knowledge accessibility suggest hypotheses for both variables. There is particularly strong reason to predict that delay should make knowledge less stable. With increasing time, retrieval cues are more likely to change and interference is more likely to have occurred. While the prediction is
less clear for testing, the literature on retrieval-induced forgetting suggests that testing may reduce competition from related information, which should promote stability. To test these predictions, I had participants repeatedly generate category exemplars; generation attempts were separated by 15 minutes or one week. I also manipulated whether participants took an immediate intervening retest after the initial generation attempt to examine the possibility that doing so stabilized knowledge (by preventing forgetting) on the week-delayed retest.

2.1 Method

2.1.1 Participants

Participants were workers from Amazon Mechanical Turk (MTurk, 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). A total of 115 MTurk workers completed the first session of Experiment 1A. Of the original 115, I excluded 24 participants for failing to complete the second session, 11 participants for failing to follow instructions (e.g., looking up answers as indicated on the final survey), five participants whose data did not record completely, and one participant who restarted the experiment mid-way through. The remaining 74 participants were not evenly split between the two counterbalances (35 vs. 39) so I excluded four participants (randomly, without looking at the data) to achieve even
counterbalancing, yielding a total of 70 participants (35 females, 35 males; $M$ age = 35, range = 20-68).

A total of 99 MTurk workers completed the first session of Experiment 1B. Of the original 99, I excluded 17 participants for failing to complete the second session, seven participants for failing to follow instructions, six participants whose data did not record completely, and two participants who restarted the experiment mid-way through. The remaining 67 participants were not evenly split between the two counterbalances (34 vs. 33) so I excluded one participant to achieve even counterbalancing, yielding a total of 66 participants (40 females, 26 males; $M$ age = 38, range = 20-64). Participants in both Experiments 1A and 1B were compensated a total of $4 for completing the study.

2.1.2 Design

Both Experiments 1A and 1B used a 2 (delay: immediate, delayed) X 2 (intervening retest: tested, not tested) within-participants designs. See Figure 1 for a schematic.
Figure 1: Schematic of Experiments 1A and 1B. To examine the effects of delay, compare changes in recall from A to B (immediate) with changes in recall from A to E (delayed). To examine the effects of testing, compare changes in recall from A to E (not tested) with changes in recall from A to D (tested).

2.1.3 Materials and Counterbalancing

The materials for Experiment 1A consisted of four categories taken from the Van Overschelde, Rawson, and Dunlosky (2004) norms: Four-footed animals, US States, fruits, and musical instruments. In Experiment 1B, I attempted to equate the chosen categories on category size (defined as the number of exemplars reported in the norms) and so I selected the following categories: Vegetables, insects, trees, and fish. The chosen categories have 25, 23, 25, and 27 exemplars, respectively, listed in the norms. In both
experiments, the materials were counterbalanced such that across participants, categories were evenly assigned to each level of the delay and intervening retest conditions.

2.1.4 Procedure

After giving informed consent, Experiment 1A participants were instructed that they would be taking knowledge tests in which they were to name as many instances of a specified category as they could in five minutes. Participants were asked to spend the entire five minutes trying to generate new instances. I also specifically asked participants to only consult their own memories and not to utilize any outside sources. Each participant took a total of ten tests across the two sessions. In the first session, they were tested on all four categories initially, and then two of those categories were tested a second time. In the second session, participants began by taking a second test for the two categories that were only tested once during session one, and then took a third test for the two categories that were tested twice during session one.

The first test began by asking participants to name as many four-footed animals (fruits, in the other counterbalance) as they could in five minutes. After typing an exemplar, participants pressed “Enter” and the response appeared in a list that remained on the side of the screen. Once the five minutes had elapsed, participants were no longer able to view the responses they had typed, and repeated the same procedure for US States (musical instruments), then fruits (four-footed animals), and then musical
instruments (US States). After taking one test on each of the four categories, participants were told they would take a second test on four-footed animals (fruits) in which they were to name as many instances as they could in five minutes again. They then took a second test for US States (musical instruments) to complete the first session of the experiment.

One week later, participants began with the instructions that they would complete a second test on fruits (four-footed animals) and then musical instruments (US States) in which they, again, were to recall as many category instances as they could in five minutes. Finally, they completed a third test on four-footed animals (fruits) and then US States (musical instruments) with the same instructions as prior tests. Before finishing the experiment, I asked participants a few questions to determine whether they had followed instructions; participants were explicitly told that their honest answers would not influence their compensation. Such questions included whether participants had looked up any answers at any point during the experiment and whether they attempted to recall as many category instances as they could (as opposed to only new instances, or only ones recalled previously). Answers to these questions were used to exclude participants on the basis of cheating or otherwise not following instructions.

The procedure for Experiment 1B was the same as Experiment 1A except for the different categories.
2.2 Results

2.2.1 Coding

<table>
<thead>
<tr>
<th>Initial Test</th>
<th>Retest</th>
<th>Losses &amp; Gains</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asparagus</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Broccoli</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Bok Choy</td>
<td>✓</td>
<td>Gain</td>
</tr>
<tr>
<td>Carrot</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Celery</td>
<td>✓</td>
<td>Gain</td>
</tr>
<tr>
<td>Corn</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Eggplant</td>
<td>✓</td>
<td>Loss</td>
</tr>
<tr>
<td>Kale</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Leek</td>
<td>✓</td>
<td>Gain</td>
</tr>
<tr>
<td>Mushrooms</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Okra</td>
<td>✓</td>
<td>Loss</td>
</tr>
<tr>
<td>Onion</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Pea</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Pepper</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Potato</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Asparagus</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Mushroom</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Squash</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Pepper</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Spinach</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Kale</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Leek</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Bok Choy</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Totals</td>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 losses</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 gains</td>
</tr>
</tbody>
</table>

Figure 2: A schematic of scoring in Experiments 1A and 1B.

For each participant on each test, every unique response was given credit.

Responses repeated by a participant within a single test were only counted once. For
each participant on each test, I scored every possible response (i.e. the total set of unique responses for that category) as present (1) or absent (0). See Figure 2 for a depiction. This method of scoring allowed me to examine both changes from the initial test to the retests in total number of responses, but also any changes (i.e. losses and gains) in the individual exemplars retrieved.

### 2.2.2 Stability in Total Exemplars Retrieved

First I examined the total amount of knowledge generated across the repeated retrieval attempts. As depicted in Figure 3, there was considerable stability in the total number of exemplars retrieved across the repeated tests in both Experiments 1A and 1B. That is, people tended to generate the same total number of vegetables for each test, for example.

![Graph showing stability in total exemplars retrieved](image)

**Figure 3:** Average number of unique responses retrieved on the initial test, immediate retest, and delayed retests (with and without an intervening retest) for a single category in Experiments 1A, 1B, and 2 (knowledge condition only).
2.2.3 Instability in Specific Exemplars Retrieved

However, relative stability of overall scores does not necessitate that the individual items contributing to those scores were consistent. To examine fluctuations in the specific knowledge accessible at each test, I calculated the number of exemplars lost and gained from the initial test to the retests. Note that the term “loss” is not meant to imply that the item was lost from memory, but that it lost accessibility from the initial test to the retest. Conversely, when a participant gained an item, this does not reflect new learning (since there was no intervening study), but rather a gain in accessibility from the initial test to the retest.

Of course, the number of exemplars someone lost or gained is fairly meaningless without knowing how many items they recalled initially – failing to recall seven fruits on a retest that were named on an initial test is more dramatic when the person only named nine fruits initially versus 30 fruits. Thus, I report the proportion of items lost and gained from the initial test to the retests, with the total number of responses from the initial test as the denominator. Collapsing across delay and testing conditions, there was substantial instability in the specific items retrieved across tests, both in terms in losses (1A $m = .22$; 1B $m = .23$), and gains (1A $m = .27$; 1B $m = .31$) from the initial test to the retests.

---

1 For completeness, the data are also reported as the total number of items gained and lost in Appendix A. Analyzing these data yields similar patterns of results.
2.2.4 Effects of Delay

Did knowledge become less stable with a greater delay between the two generation attempts? I compared the proportion of items lost and gained from the initial test to the immediate retest with the proportion of items lost and gained from the initial test to the week-delayed retest (without an intervening retest). Figure 4 depicts the data. Indeed, participants in both Experiments 1A and 1B showed greater losses in accessibility with a longer delay, $t(69) = 3.20, SEM = .02, p = .002, d = .38$ and $t(65) = 4.79, SEM = .02, p < .001, d = .59$, respectively. The gains mirrored this pattern of losses: Participants were more likely to generate new exemplars after a one-week delay, $t(69) = 2.15, SEM = .03, p = .035, d = .26$, and $t(65) = 5.38, SEM = .04, p < .001, d = .66$.

![Figure 4: Proportion of items lost (A) and gained (B) from the initial test to immediate and delayed retest (without an intervening retest). Data are from Experiments 1A, 1B, and 2 (knowledge condition only). Error bars represent standard error of the mean.](image)
2.2.5 Effects of Testing

![Proportion of items lost (A) and gained (B) from the initial test to week-delayed retests with and without an intervening immediate retest. Data are from Experiments 1A, 1B, and 2 (knowledge condition only). Error bars represent standard error of the mean.](image)

Figure 5: Proportion of items lost (A) and gained (B) from the initial test to week-delayed retests with and without an intervening immediate retest. Data are from Experiments 1A, 1B, and 2 (knowledge condition only). Error bars represent standard error of the mean.

Did taking an immediate retest influence the likelihood that a particular exemplar would be repeated or that a new item would be generated? I compared the proportion of items lost and gained from the initial test to the week-delayed retest without an intervening retest with the proportion of items lost and gained from the initial test to the week-delayed retest with an immediate intervening retest. The data are shown in Figure 5. In contrast to the well-documented benefits of retrieval practice for event memories, testing did not appear to influence shifts in knowledge accessibility. In both Experiments 1A and 1B, participants lost a similar proportion of items with and without an intervening retest, $t < 1$ and $t(65) = 1.74$, $SEM = .02$, $p = .086$, $d = .21$, respectively, and gained a similar proportion of items with and without an intervening
retest, $t(69) = 1.95, SEM = .03, p = .055, d = .23$ and $t(65) = 1.23, SEM = .03, p = .224, d = .15$, respectively.
3. Experiment 2

The prior experiments suggested that access to knowledge fluctuates similarly to event memory in that it becomes less stable with time. On the other hand, knowledge seems to behave differently from event memory in that it does not stabilize with retrieval practice. To better understand these similarities and differences, Experiment 2 directly compared the fluctuations of knowledge and event memory. I selected a design that kept the procedure and materials as similar as possible so that any differences between the event memory and knowledge conditions could be attributed to differences in the memory types. To do so, I manipulated whether participants studied category exemplars and took repeated tests (event memories) or repeatedly generated category exemplars (knowledge). To control the content across conditions, I utilized the data from Experiment 1B to select the most commonly recalled category exemplars for the event memory group to study. To examine the impact of delay and testing, all participants were retested immediately on half of the categories and were also retested on all categories one week later.

3.1 Method

3.1.1 Participants

One-hundred and eighty MTurk workers completed the first session. I excluded 31 participants for failing to complete the second session and 35 participants for failing to follow instructions (e.g., indicating on the final survey that they looked up answers).
The remaining 114 participants were not evenly split between the four counterbalances (24, 36, 27, and 27) so I excluded 18 participants to achieve even counterbalancing, yielding a total of 96 participants (57 females, 39 males; $M_{\text{age}} = 37$, age range = 20-74).

Participants were compensated a total of $5 for completing the experiment.

### 3.1.2 Design

I employed a 2 (memory type: event, knowledge) X 2 (delay: immediate, delayed) X 2 (intervening retest: tested, not tested) mixed design. The first variable was manipulated between-participants and the other two variables were manipulated within-participants. See Figure 6 for a schematic.
3.1.3 Materials and Counterbalancing

The categories from Experiment 1B were used: Vegetables, insects, trees, and fish. To allow fair comparisons across the two memory types, I decided that participants from the two conditions would need to be judged on the same set of responses. That is,
while participants in the knowledge condition might generate exemplars not studied by
the event memory condition, including them in the analyses might suggest differences
between the two memory types that are would be driven by the different content. For
example, it is possible that fluctuations are more likely for more obscure exemplars (e.g.,
jackfruit), and if the knowledge group generated such items, the results would indicate
greater fluctuations for knowledge than event memory. Therefore, I selected items that,
based on data from Experiment 1B, participants in the knowledge condition would be
most likely to name. Furthermore, I reasoned that the comparison would be fair to the
extent that participants recall approximately the same number of responses on the initial
test. Therefore, I calculated the average number of responses generated by participants
\( m = 18 \) per category in Experiment 1B and conducted a pilot to determine that the event
memory group needed to study 40 exemplars per category to match the performance of
Experiment 1B participants on the initial test. The materials were counterbalanced the
same way as in Experiment 1B.

**3.1.4 Procedure**

The procedure for the knowledge condition was the same as Experiment 1B,
except that at the very beginning of the experiment, participants solved puzzles for 10
minutes (time determined by pilot testing) as a filler task to equate overall time (since
the event memory group needed to study exemplars).
The procedure for the event memory condition was designed to mirror the knowledge condition, except that participants first studied exemplars and were instructed to recall only studied exemplars on the tests. They began with the study phase: Participants were told that they would be studying four lists of words, with the words within a given list belonging to the same category. When beginning a list, they were told the category name (e.g., Vegetables). As they studied each word, they were asked to judge how typical or representative that item is for that category on a five-point scale of highly atypical to highly typical. The study phase was self-paced. The reason for a self-paced orienting task was the online nature of the experiment – I wanted to require that participants actively respond to each item to ensure that they actually encoded the study list. Once participants had studied all words on all four lists, they solved puzzles for three minutes to ensure that no responses were held in short-term memory. They then completed the same tests in the same order as the knowledge condition, except with different instructions. Event memory participants were instructed to name only items that were studied earlier in the experiment, even though they may know many more items belonging to that category. Finally, for the event memory condition, I modified the questions at the end of the experiment that assess whether participants followed instructions. Specifically, participants were asked whether they attempted to recall only previously studied items or whether they also tried to generate additional non-studied items.
3.2 Results

3.2.1 Replication of Experiment 1

Before comparing the knowledge and event memory conditions, I first analyzed the knowledge condition separately to confirm that the results replicated the findings from Experiments 1A and 1B. To do so, I coded the data from the knowledge condition in the same manner as Experiments 1A and 1B. The key patterns of results were replicated: The total number of exemplars generated was relatively stable over generation attempts (see Figure 3) but there was considerable underlying instability between retrievals in terms of both losses ($m = .28$) and gains ($m = .38$). Additionally, time increased both losses (see Figure 4), $t(47) = 3.27$, $SEM = .02$, $p = .002$, $d = .47$, and gains, $t(47) = 2.18$, $SEM = .04$, $p = .035$, $d = .31$, while testing had no impact on either losses or gains (see Figure 5), both $ts < 1$.

3.2.2 Continuously Cumulating Meta-Analyses

Given that I had three experiments examining knowledge fluctuations, I conducted a continuously cumulating meta-analysis (CCMA; Braver, Thoemmes, & Rosenthal, 2014) to provide a more precise estimate of the effect sizes. Specifically, I conducted four separate CCMAs using Comprehensive Meta-Analysis software (Borenstein, Hedges, Higgins, & Rothstein, 2005) to measure the effects of delay and testing on both losses and gains. Table 1 reports summary statistics and Figure 7 presents forest plots for each effect. The effect of delay on instability was robust,
increasing both losses \((p < .001, d = .48, 95\% \text{ CI } [.32, .63])\) and gains \((p < .001, d = .40, 95\% \text{ CI } [.25, .56])\). While individual experiments hinted at effects of testing, the CCMA indicated that testing did not significantly impact losses \((p = .25, d = -.09, 95\% \text{ CI } [-.23, .06])\), or gains \((p = .40, d = .06, 95\% \text{ CI } [-.08, .21])\).

Figure 7: Forest plots for the effect of delay on losses (A), the effect of delay on gains (B), the effect of testing on losses (C) and the effect of testing on gains (D). Each line is an experiment for which the box size represents study weighting and the line width represents 95\% confidence interval. The diamond represents the overall effect, with the edges marking the 95\% confidence interval.
Table 1: Continuously Cumulating Meta-Analyses (CCMAs) for the effects of delay and testing on losses and gains in Experiments 1A, 1B, and 2 (knowledge condition only).

<table>
<thead>
<tr>
<th></th>
<th>Mean diff</th>
<th>Std dev</th>
<th>t</th>
<th>p</th>
<th>Std diff in means (d)</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Effect of Delay on Losses</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experiment 1a</td>
<td>.07</td>
<td>.18</td>
<td>3.20</td>
<td>.002</td>
<td>.38</td>
<td>3.09</td>
</tr>
<tr>
<td>Experiment 1b</td>
<td>.09</td>
<td>.16</td>
<td>4.79</td>
<td>&lt;.001</td>
<td>.59</td>
<td>4.42</td>
</tr>
<tr>
<td>Experiment 2</td>
<td>.07</td>
<td>.14</td>
<td>3.27</td>
<td>.002</td>
<td>.47</td>
<td>3.11</td>
</tr>
<tr>
<td>CCMA results</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Effect of Delay on Gains</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experiment 1a</td>
<td>.06</td>
<td>.25</td>
<td>2.15</td>
<td>.035</td>
<td>.26</td>
<td>2.11</td>
</tr>
<tr>
<td>Experiment 1b</td>
<td>.20</td>
<td>.30</td>
<td>5.38</td>
<td>&lt;.001</td>
<td>.66</td>
<td>4.87</td>
</tr>
<tr>
<td>Experiment 2</td>
<td>.09</td>
<td>.27</td>
<td>2.18</td>
<td>.035</td>
<td>.31</td>
<td>2.12</td>
</tr>
<tr>
<td>CCMA results</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Effect of Testing on Losses</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experiment 1a</td>
<td>-.01</td>
<td>.16</td>
<td>-.71</td>
<td>.48</td>
<td>-.08</td>
<td>-.71</td>
</tr>
<tr>
<td>Experiment 1b</td>
<td>-.03</td>
<td>.13</td>
<td>-1.74</td>
<td>.09</td>
<td>-.21</td>
<td>-1.72</td>
</tr>
<tr>
<td>Experiment 2</td>
<td>.01</td>
<td>.11</td>
<td>.59</td>
<td>.56</td>
<td>.09</td>
<td>.59</td>
</tr>
<tr>
<td>CCMA results</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Effect of Testing on Gains</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experiment 1a</td>
<td>.06</td>
<td>.26</td>
<td>1.95</td>
<td>.06</td>
<td>.23</td>
<td>1.93</td>
</tr>
<tr>
<td>Experiment 1b</td>
<td>-.04</td>
<td>.25</td>
<td>-1.23</td>
<td>.22</td>
<td>-.15</td>
<td>-1.22</td>
</tr>
<tr>
<td>Experiment 2</td>
<td>.03</td>
<td>.25</td>
<td>.76</td>
<td>.45</td>
<td>.11</td>
<td>.75</td>
</tr>
<tr>
<td>CCMA results</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
3.2.3 Knowledge vs. Event Memory

3.2.3.1 Coding

To allow a fair comparison between the two memory conditions, I only scored the 40 target responses for each category (i.e. items studied in the event memory condition). For each participant on each test, each target item was scored as present (1) or absent (0). Responses other than the target items were not scored and not included in any of the following analyses.

3.2.3.2 Stability in Total Exemplars Retrieved

Figure 8 depicts the average number of items participants retrieved for a single category across the multiple tests. The knowledge condition replicated the pattern of overall stability observed in Experiments 1A and 1B, while the event memory condition exhibited an overall decline with the week delay.

![Figure 8: Average number of responses (out of a possible 40) retrieved for the initial test, immediate retest, and delayed retests (with and without an intervening retest) for a single category in Experiment 2.](image)
As a reminder, my goal was for event memory participants to retrieve approximately 18 studied items per category on the initial test since that was the average number of exemplars generated by Experiment 1B participants. While participants in the event memory condition did recall approximately the goal number ($M = 19.63$), because participants in the knowledge condition were being scored on only the 40 target responses per category (whereas in Experiment 1B, all responses were scored), they generated significantly fewer items ($M = 15.17$), $t(78.53) = 4.46$, $SED = 1.05$, $p < .001$, $d = .87^1$. Therefore, when comparing across the two conditions in terms of item fluctuations, it will be particularly crucial to use proportions of (rather than the total number of) items lost and gained from one test to the next$^2$.

### 3.2.3.3 Instability in Specific Exemplars Retrieved

To better understand any fluctuations underlying the total scores, I calculated the number of items lost and gained from the initial test to the retests. As in Experiments 1A and 1B, I report the proportion of items lost or gained from the initial test to the immediate and delayed retests, with the total number of responses from the initial test serving as the denominator. Before determining whether the effects of delay and testing on fluctuations differed for knowledge and event memory, I first explored differences between the two memory types in their overall amount of item fluctuations. There was

---

$^1$ Levene's test for equality of variances was violated so I report corrected values.

$^2$ For completeness, the data are also reported as the total number of items gained and lost in Appendix A. Analyzing these data yields similar patterns of results.
substantial instability for both event memory and knowledge, but the overall levels of losses and gains clearly differed by memory type. Whereas the event memory group showed greater overall losses than the knowledge group (.31 vs. .23), $t(94) = 3.49$, $SED = .02$, $p = .001$, $d = .71$, the knowledge group exhibited much larger gains overall (.30 vs. .14), $t(71.19) = 6.58$, $SED = .02$, $p < .001$, $d = 1.34$.

3.2.3.4 Effects of Delay

![Graph](image)

Figure 9: Proportion of items lost (A) and gained (B) from the initial test to immediate and delayed retest (without an intervening retest) for event memory and knowledge conditions. Data are from Experiment 2. Error bars represent standard error of the mean.

Did delay differentially effect fluctuations in event memory and knowledge?

Figure 9 depicts the data. Clearly the largest impact of delay was on losses from event memory: Delay drastically increased losses from event memory, $t(47) = 12.41$, $SEM = .02$, $p < .001$, $d = 1.79$, while only moderately increasing losses from knowledge, $t(47) = 2.27$, $p = .03$, $d = .56$. 
SEM = .02, p = .028, d = .33. Additionally, whereas delay decreased event memory gains, 
$t(47) = 3.58, SEM = .02, p = .001, d = .52$, participants showed greater gains in knowledge 
after a delay, $t(47) = 2.32, SEM = .04, p = .025, d = .33$.

### 3.2.3.5 Effects of Testing

Next, I examined the impact of testing on the instability of event memories and knowledge. Figure 10 depicts the data. Testing had opposing effects on losses and gains from event memory but no effect on knowledge whatsoever. That is, testing prevented event memory losses, $t(47) = 3.31, SEM = .02, p = .002, d = .48$, while promoting event memory gains, $t(47) = 2.73, SEM = .02, p = .009, d = .39$. On the other hand, testing did not impact losses or gains for knowledge, both $ts < 1$.

![Figure 10](image_url)

**Figure 10:** Proportion of items lost (A) and gained (B) from the initial test to delayed retests with and without an intervening immediate retest for event memory and knowledge conditions. Data are from Experiment 2. Error bars represent standard error of the mean.
3.2.3.6 Item Analyses

I conducted a series of item analyses to answer the question of whether highly typical items were less likely to fluctuate. That is, are we less likely to lose access to “carrot” than “okra”? Items that are more typical should have a higher storage strength and are more likely to have recent exposure, so they should be less prone to losses in accessibility. To operationalize item typicality, I utilized Experiment 1B data (where the same categories were used) and ranked exemplars in order of how frequently they were recalled. Each exemplar within each category received a unique score between one and 40 where the most frequently recalled exemplar received a score of 40 and the least frequently recalled exemplar received a score of one.

Figure 11: Proportion of items recalled on the initial test for the event memory (A) and knowledge (B) conditions as a function of the item’s rank (1 = least common, 40 = most common). Data are from Experiment 2.
First, I examined how item typicality affected whether items were recalled or not on the initial test. As depicted in Figure 11, more common items were more likely to recalled on the initial test for both the event memory (panel A; $r = .75, p < .001$) and knowledge (panel B; $r = .95, p < .001$) conditions, although the relationship was stronger for knowledge.

![Figure 12](image)

**Figure 12:** Proportion of immediate event losses (A), delayed event losses (B), immediate knowledge losses (C), and delayed knowledge losses (D) from the initial test to the retests as a function of the item’s rank (1 = least common, 40 = most common). Data are from Experiment 2.
Next I asked whether more common items were less likely to lose accessibility from the initial test to the retests. The data are depicted in Figure 12. For event memory, item typicality did not matter when the delay was immediate (panel A; \( r = -.01, p = .946 \)), but over one week, less typical items were more likely to be forgotten (panel B; \( r = -.59, p < .001 \)). For knowledge, common items were less likely to be lost regardless of the delay, although the pattern was stronger when the delay was one week (panel D; \( r = -.84, p < .001 \)) compared to when it was immediate (panel C; \( r = -.34, p = .034 \)).

As mentioned above, more common items were more likely to be recalled on the initial test than less common items. However, given that an item was not recalled initially, does item typicality influence whether or not that item is newly recalled on a retest? As illustrated in Figure 13, items across conditions were more likely to be recovered if they were more common. This pattern held for immediate (panel A; \( r = .65, p < .001 \)) and delayed event memory gains (panel B; \( r = .77, p < .001 \)), as well as immediate (panel C; \( r = .71, p < .001 \)) and delayed knowledge gains (panel D; \( r = .90, p < .001 \)).
Figure 13: Proportion of immediate event gains (A), delayed event gains (B), immediate knowledge gains (C), and delayed knowledge gains (D) from the initial test to the retests as a function of the item’s rank (1 = least common, 40 = most common). Data are from Experiment 2.

3.3 Discussion of Experiments 1A, 1B, and 2

A focus on the total amount of knowledge and event memories retrieved suggests a characterization of memory that is consistent with prevailing assumptions – That knowledge is relatively stable while event memory declines over time. However, a closer look at the specific items accessible at each time point paints a very different
picture. Much like event memory, specific pieces of knowledge fluctuate in accessibility. These experiments highlight the importance of analyzing the data at the level of individual responses. This is not to say that the total amount of information retrieved is unimportant. Rather, we should consider both perspectives when characterizing the stability of memory.

For both event memory and knowledge, access to specific knowledge fluctuated more with increasing delay, both in terms of certain items losing and others gaining accessibility. The nature of this instability, however, was quite different for the two memory types. Specifically, knowledge saw modest increases in both losses and gains whereas event memory saw dramatic increases in losses and modest decreases in gains. One key difference between event memory and knowledge that might account for these different fluctuation patterns is the role of context. While event memories are typically associated with a particular context (e.g., remembering when and where the event occurred), knowledge is typically decontextualized, presumably due to encounters in multiple different contexts (e.g., Watkins & Kerkar, 1985). Indeed, the instructions utilized in the experiment highlighted this difference: Event memory participants were explicitly instructed to think back to the initial study episode while knowledge participants were told to just generate as many exemplars as they could. Critically, since context is more likely to change with increasing time, it makes sense that event memories should be more affected by delay.
Additionally, I replicated prior research documenting retrieval practice effects in event memory (e.g., Roediger & Karpicke, 2006a), but found no effect of testing on knowledge fluctuations. Again, we can draw upon the differential role of context in retrieving knowledge and event memories to explain these results. According to the Episodic Context Account of the testing effect (Karpicke, Lehman, & Aue, 2014), when learners retrieve event memories, they attempt to reinstate the temporal context of the initial learning episode (context A). When information is retrieved in the present (context B), an updated composite context combining features of contexts A and B is created. This new composite context is unique and thus promotes future recall by allowing learners to restrict their search set. Thus, an event memory may benefit from testing because it is tied to a particular context whereas knowledge is relatively decontextualized.

Experiment 2 also suggested that not all knowledge fluctuates equally – the more common items were more likely to be consistently recalled. Carrot was more likely to be generated than okra in the first place, and given that it was recalled, it was less likely to be forgotten on a later test. Furthermore, when a more common item was initially inaccessible, it was more likely to become accessible on a later retrieval attempt. These typicality effects are likely due to the fact that more common items are both better learned and are more likely to have been encountered recently; these data do not allow me to separate the influence of memory strength and recency of access. Interestingly,
these patterns held for event memory as well (except for immediate losses) but the effects were consistently less strong. As mentioned earlier, event memory retrieval is not divorced from knowledge – it is easier to learn new information that is supported by prior knowledge (e.g., Kole & Healy, 2007). However, event memory retrieval is influenced by other factors as well, namely the ability to think back to the original learning episode. Thus, it makes sense that event memory was influenced by item typicality, but to a lesser extent than the knowledge condition.

Finally, I acknowledge limitations in Experiment 2’s comparison of event memories and knowledge. I opted for a design that kept the procedure and materials as similar as possible: All participants were tasked with repeatedly naming the same category exemplars, with instructions to either generate from knowledge or to think back to a prior study phase. Certainly this comparison was not perfect – the knowledge and event memory conditions likely did not measure “pure” knowledge and event memories, respectively. For example, generating an exemplar is an event itself; on the repeated tests, participants in the knowledge condition might have thought back to that event (even though instructed otherwise). Similarly, when participants in the event memory condition struggled to retrieve additional studied exemplars, they may have relied on their prior knowledge to infer what was likely studied (again, despite explicit instruction). Acknowledging these potential sources of contamination, the design of
Experiment 2 allowed for as direct a comparison of knowledge and event memory fluctuations as possible.
4. Experiments 3A & 3B¹

My first two experiments documented the considerable instability in access to specific knowledge. My next two experiments applied this phenomenon to education and asked: How can we best reestablish access to inaccessible knowledge? Prior research has identified two effective techniques: Studying the answer (Berger et al., 1999) and taking a multiple-choice test (Cantor et al., 2015). The purpose of Experiment 3 was to compare these two methods and determine whether one was more effective than the other. Based on the vast testing effect literature that finds many benefits of testing over studying (Roediger & Karpicke, 2006a), I predicted that multiple-choice testing might prove even more effective than studying as a method for reactivating marginal knowledge.

In Experiment 3A, I directly compared multiple-choice testing and studying as interventions for regaining access to marginal knowledge. Such a comparison raised a critical methodological issue – to the extent that the participants’ inability to answer general knowledge questions reflected no knowledge rather than marginal knowledge, the studying condition would have a large advantage over the multiple-choice testing condition (since studiers are told the correct answers while multiple-choice testers have

a 25% chance of guessing correctly for unknown items). Thus, I needed a method of restricting my analyses to strictly marginal knowledge.

Since my prior work showed that there was substantial between-participant variability in which knowledge was marginal, I recognized it would not be possible to select a set of questions that would be marginal for all or even the majority of participants. Instead, I drew on the finding that people are quite accurate at estimating their own marginal knowledge. As reviewed earlier, when people cannot retrieve the answer to a short-answer general knowledge question, they are very good at predicting whether or not they will be able to recognize the correct answer when given several alternatives (feeling-of-knowing (FOK); Hart, 1965). Accordingly, my idea was to estimate which knowledge was marginal for all participants by asking them to judge whether or not they would be able to recognize the correct answer for each short-answer question.

To confirm the validity of this method, I ran a pilot (N = 86) where, after answering each (real) short-answer question (taken from Berger et al., 1999), MTurk participants made a binary decision about whether they would be able to choose the correct answer out of four alternatives. Next, the same participants answered each question in multiple-choice format, allowing me to assess the accuracy of the FOK judgments. Participants were quite accurate, 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 = .83.
While certainly not perfect, FOK judgments allowed me to estimate marginal knowledge equally for the study and multiple-choice testing conditions.

4.1 Experiment 3A

4.1.1 Method

4.1.1.1 Participants

One hundred seventy-four MTurk participants completed the experiment (an additional 37 completed the first session but not the second). Six participants were excluded due to technical difficulties (e.g., computer froze), and 24 for indicating they looked up or wrote down answers during the experiment, yielding a total of 144 participants (88 females, 56 males; M age = 34 years, range = 19-68 years). Participants were compensated a total of $4 for completing the experiment.

4.1.1.2 Design

I employed a 3 (activity: MC test, study, none) x 2 (item-type: real, fictitious) x 2 (delay: none, two days) mixed-factorial design. Activity was manipulated between participants and the others were manipulated within-participants.

4.1.1.3 Materials and Counterbalancing

The materials consisted of 84 pairs of real and corresponding fictitious questions taken from Berger et al. (1999). See Table 2 for sample materials. Based on the pilot described earlier, I selected the 84 real questions that were most likely to be marginal. I defined marginal knowledge as the failure to correctly answer the short-answer question
followed by a positive FOK judgment (i.e. not contingent upon correct multiple-choice performance). The 84 selected items were judged to be marginal for 28% to 67% ($M = .39$, $SD = .09$) of pilot participants. For each participant, half of the question pairs were assigned to the real condition and the other half to the fictitious condition; an item’s assignment remained consistent throughout the entire experiment.

Table 2: Examples of real (R) and fictitious (F) questions with targets and lures used in Experiments 3A and 3B.

<table>
<thead>
<tr>
<th>Question</th>
<th>Target</th>
<th>Lures</th>
</tr>
</thead>
<tbody>
<tr>
<td>(R) What is the long process by which a dead organism turns to stone?</td>
<td>Petrification</td>
<td>Decomposition, Rigor Mortis, Torgisation</td>
</tr>
<tr>
<td>(F) What is the long process by which a live organism's hair turns white?</td>
<td>Pigmosis</td>
<td>Decromatization, Ostoresis, Transcoloration</td>
</tr>
<tr>
<td>(R) What open-air public theater was home to William Shakespeare's theatrical company?</td>
<td>Globe</td>
<td>Aavion, Haven, World</td>
</tr>
<tr>
<td>(F) What open-air public theater was home to Baruch's theatrical company?</td>
<td>Roche</td>
<td>Confair, Dutch, Strobe</td>
</tr>
</tbody>
</table>

The initial short-answer test consisted of 84 critical questions and 20 easy filler questions. The 84 critical questions also comprised the multiple-choice test, which paired the target with three plausible lures. The study list paired each critical question with its target (no lures). Half of the 84 critical questions appeared on the immediate final short-
answer test and the other half on the final short-answer test. I used a total of 12 conditions to counterbalance item-type, activity, and delay across participants.

4.1.1.4 Procedure

All participants began with the initial short-answer general knowledge test; question order was determined at random by the computer. They were instructed to try their best but not be discouraged if they couldn’t answer them and to type “don’t know.” Immediately after submitting an answer to a short-answer question, participants made binary FOK judgments, answering “yes” or “no” to the question, “Do you think you would be able to select the correct answer out of 4 choices?”

Next, depending on their activity condition, participants took the multiple-choice test, studied the study list, or played Tetris. Participants in the multiple-choice testing condition answered each of the real and fictitious questions at their own pace. There was no feedback given in the multiple-choice condition since my prior work found that feedback only promoted new learning and was not necessary for the reactivation of marginal knowledge (Cantor et al., 2015 Experiment 2). Participants in the study condition viewed each question paired with the target answer (for both real and fictitious items) for six seconds, which was the average time participants spent answering each multiple-choice test question in the pilot.

To complete Session 1, all participants took the immediate final short-answer test. Unlike the initial test, participants just answered questions on the final tests and did
not make FOK judgments. Approximately 48 hours after completing Session 1, participants were emailed a link to complete Session 2 in which they completed the delayed final short-answer test. Finally, given that the experiments were conducted online, I asked participants whether they had looked up or written down any answers; I explicitly told them that their honest responding was valued and thus would not impact their compensation.

4.1.2 Results

4.1.2.1 Coding

All short-answer responses were scored by two independent coders. Inter-rater reliability was very high (initial test: \( \kappa = .98 \), final tests: \( \kappa = .98 \)) and any discrepancies were resolved via discussion.

4.1.2.2 Determining Marginal Knowledge

To determine which knowledge was marginal for individual participants, I used participants’ FOK judgments. Replicating the pilot, I found that participants in the multiple-choice testing condition were quite accurate. Participants were more likely to select the correct alternative for real questions after a positive (\( M = .86 \)) than after a negative FOK judgment (\( M = .54 \)), \( t(43) = 7.81, SEM = .04, p < .001, d = 1.17^2 \). While it was not possible to assess the accuracy of FOK judgments for the study and control

---

\(^2\) Four participants were not included in this analysis as they did not make any negative FOK judgments for real questions.
conditions, I verified that the proportion of positive FOK judgments was similar for all three conditions, \( F(2, 141) = 2.44, MSE = .09, p = .091. \)

I restricted all subsequent analyses to items that were not retrieved on the initial test (at the individual participant level), as questions answered correctly indicate that knowledge was already accessible. This constraint restricted the analyses to 60% (\( SD = .20 \)) of real and 100% (\( SD = .004 \)) of fictitious questions.

### 4.1.2.3 Benefits of Multiple-Choice Testing

Before comparing studying and multiple-choice testing, I first confirmed that my prior research replicated – that multiple-choice testing was a successful intervention. A 2 (activity: MC Test, none) x 2 (item-type: real, fictitious) x 2 (delay: none, two days) ANOVA was computed on the proportion of final-test questions answered with targets. The data appear in Table 3. Consistent my prior research, I observed a testing effect: Participants were more likely to answer final test questions with targets when that question appeared on the multiple-choice test, \( F(1, 94) = 189.32, MSE = .03, p < .001, \eta^2 = .67 \). Critically, this testing effect was substantially larger for real than fictitious questions, \( F(1, 94) = 169.92, MSE = .02, p < .001, \eta^2 = .25 \), indicating that it reflected reactivated knowledge rather than new learning. While the size of the testing effect did diminish slightly over the two day delay, \( F(1, 94) = 14.53, MSE = .01, p < .001, \eta^2 = .01 \), the
The advantage of multiple-choice testing for real items was still quite large two days later, $t(63.60) = 10.64, SED = .04, p < .001, d = 2.17$.

Table 3: Proportion of final-test questions answered with targets (given initial test failure) in Experiments 3A and 3B.

<table>
<thead>
<tr>
<th></th>
<th>Real Questions</th>
<th>Fictitious Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No Delay</td>
<td>2 Days</td>
</tr>
<tr>
<td>MC Test 3A</td>
<td>.56 (.03)</td>
<td>.49 (.03)</td>
</tr>
<tr>
<td>Control 3A</td>
<td>.07 (.03)</td>
<td>.10 (.01)</td>
</tr>
<tr>
<td>MC Test 3B</td>
<td>.55 (.03)</td>
<td>--</td>
</tr>
<tr>
<td>Control 3B</td>
<td>.07 (.03)</td>
<td>--</td>
</tr>
</tbody>
</table>

Note. Standard error is noted in parentheses.

To pinpoint the benefits of multiple-choice testing, I conducted an analysis on targets that were selected on the multiple-choice test. I restricted this analysis to target selections, as participants almost never produced targets on the final test if they were not selected on the multiple-choice test ($M = .02, SD = .02$). Instead, the benefits were due to the reactivation of marginal knowledge. I computed a 2 (item-type: real, fictitious) x 2 (delay: none, two days) ANOVA on the proportion of final-test questions answered with targets (given target multiple-choice selection). As illustrated in Figure 14 (panel A), the

---

3 Levene's test for equality of variances was violated so I report corrected values.
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, $F(1,47) = 776.09$, $MSE = .02$, $p < .001$, $\eta^2 = .84$, and this pattern was consistent across the two-day delay, $F < 1$.

**Figure 14:** 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.

### 4.1.2.4 Comparing the Benefits of Studying and Multiple-Choice Testing

As explained earlier, I do not report the analysis comparing all studied items to all tested items, since the study condition allowed more new learning. However, I note that this analysis yielded the same pattern of results as presented below. Instead, my key analysis compared the power of studying and multiple-choice testing in promoting future recall of estimated marginal knowledge (real items with a positive FOK) and newly learned information (fictitious items). To do so, I computed a 2 (activity: study, MC test) X 2 (item-type: estimated marginal, newly learned) X 2 (delay: none, two days)
ANOVA on the proportion of final-test questions answered with targets. The data are presented in Table 4.

Table 4: 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.

<table>
<thead>
<tr>
<th></th>
<th>Estimated Marginal Knowledge</th>
<th>New Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No Delay</td>
<td>48 Hours</td>
</tr>
<tr>
<td>Study</td>
<td>.69 (.04)</td>
<td>.60 (.04)</td>
</tr>
<tr>
<td>MC Test</td>
<td>.63 (.04)</td>
<td>.57 (.04)</td>
</tr>
<tr>
<td>Control</td>
<td>.07 (.03)</td>
<td>.13 (.03)</td>
</tr>
</tbody>
</table>

Note. Standard error is noted in parentheses.

A significant three-way interaction indicated that there was a benefit of studying over multiple-choice testing mainly for the newly-learned fictitious information, particularly when the final test was immediate, $F(1,94) = 12.77, MSE = .02, p = .001, \eta^2 = .01$. Studying promoted higher final test performance than multiple-choice testing for fictitious items on both the immediate final test, $t(50.76) = 7.68, SED = .04, p < .001, d = 1.57$, and delayed final test, $t(52.96) = 5.39, SED = .02, p < .001, d = 1.10^3$. Note that the decrease in effect size across the delay highlights that much of this new learning was quickly forgotten. Most importantly, multiple-choice testing and studying promoted equal production of targets for (estimated) marginal knowledge on both immediate and delayed final tests, both $ts < 1$. Thus, the two strategies were equally powerful in
reestablishing access to marginal knowledge; the advantage of studying appeared to be restricted entirely to new learning.

4.2 Experiment 3B

Experiment 3A demonstrated that multiple-choice testing was just as powerful a tool as studying the answers for regaining access to marginal knowledge. Note that multiple-choice testing measured up even though it was impossible to restrict analyses to strictly marginal knowledge. That is, participants’ FOK judgments were fairly accurate, but not perfect, meaning that the analyses inevitably included a few items which participants did not know. For such unknown items, the studying condition had a large advantage; as shown with the fictitious items, studying was more effective when it came to new learning. However, any bias toward the studying condition should decrease over time since, as also demonstrated with the fictitious items, new learning dropped dramatically with a delay. As shown in Table 4, studiers retained 34% of fictitious targets on an immediate test, but that performance dropped to 10% two days later. I reasoned that with an even longer delay, new learning should drop even closer to zero, leveling the playing field between studying and multiple-choice testing. Thus, Experiment 3B was an exact replication of Experiment 3A, except that the delayed final test occurred one week after Session 1.
4.2.1 Method

4.2.1.1 Participants

One hundred seventy-three MTurk participants completed Experiment 3B (an additional 30 participants started the experiment but did not complete it). Twelve participants were excluded due to technical difficulties and 17 were excluded for indicating that they looked up or wrote down answers, yielding a total of 144 participants (78 females, 65 males, 1 did not report gender; $M$ age = 34 years, range = 19-72 years). Participants were compensated a total of $4 for completing the experiment.

4.2.1.2 Design

Experiment 3B used the same design as Experiment 3A except that delay was operationalized as one week instead of two days.

4.2.1.3 Materials and Counterbalancing

The same materials and counterbalancing methods as Experiment 3A were used.

4.2.1.4 Procedure

The same procedure as Experiment 3A was used except that participants were emailed the link to complete Session 2 approximately one week after Session 1.
4.2.2 Results

4.2.2.1 Coding

All short-answer test responses were scored by two independent coders. Inter-reliability was quite high for both the initial (κ = .98) and final test (κ = .97). Any discrepancies were resolved through discussion.

4.2.2.2 Determining Marginal Knowledge

Replicating Experiment 3A, participants were quite good at estimating whether knowledge was marginal. Participants in the multiple-choice testing condition were more likely to successfully select the target for real questions if they made a positive FOK judgment ($M = .86$) than if they made a negative FOK judgment, ($M = .61$), $t(47) = 6.73$, $SEM = .04$, $p < .001$, $d = .97$. Thus, I again utilized FOK judgments to estimate which items were marginal for my analysis comparing multiple-choice testing with studying.

Again, because I was not interested in knowledge that was already accessible, all analyses that follow were 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).

4.2.2.3 Benefits of Multiple-Choice Testing

Experiment 3B allowed me to confirm the benefits of multiple-choice testing observed in Experiment 3A and examine whether those benefits persisted over a longer delay of one week. To do so, I computed a 2 (activity: MC test, none) X 2 (item-type: real, fictitious) X 2 (delay: none, 1 week) ANOVA on the proportion of final-test questions
answered with targets. The data appear in Table 3. Replicating Experiment 3A, there
was positive testing effect, $F(1, 94) = 127.18$, $MSE = .03$, $p < .001$, $\eta^2 = .58$. Critically, this
benefit of multiple-choice testing was much larger for real than fictitious items $F(1, 94) = 
118.89$, $MSE = .02$, $p < .001$, $\eta^2 = .18$. While the size of the testing effect did decrease across
the delay, $F(1, 94) = 58.45$, $MSE = .01$, $p < .001$, $\eta^2 = .04$, even after one week, the advantage
of multiple-choice testing was quite large, $t(85.36) = 6.04$, $SED = .04$, $p < .001$, $d = 1.23^3$.

Next, I again conducted an analysis restricted to correct selections on the
multiple-choice test. I computed a 2 (item-type: real, fictitious) X 2 (delay: none, two
days) ANOVA on the proportion of final test questions answered with targets given
accurate responding on the multiple-choice test. As illustrated in Figure 14 (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, $F(1, 47) =
6.90$, $MSE = .02$, $\eta^2 = .01$.

4.2.2.4 Comparing the Benefits of Studying and Multiple-Choice Testing

To reduce the bias toward the studying condition, as in Experiment 3A, the
analysis comparing studying and multiple-choice testing was restricted to items that
were estimated as marginal by participants (a positive FOK for real items; all fictitious
items were used). I computed a 2 (activity: study, MC test) X 2 (item-type: estimated
marginal, newly learned) X 2 (delay: none, one week) ANOVA on the proportion of final-test questions answered with targets. Table 5 contains the relevant data.

Table 5: 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.

<table>
<thead>
<tr>
<th></th>
<th>Estimated Marginal Knowledge</th>
<th>New Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No Delay</td>
<td>1 Week</td>
</tr>
<tr>
<td>Study</td>
<td>.70 (.03)</td>
<td>.52 (.04)</td>
</tr>
<tr>
<td>MC Test</td>
<td>.61 (.03)</td>
<td>.44 (.04)</td>
</tr>
<tr>
<td>Control</td>
<td>.07 (.03)</td>
<td>.18 (.03)</td>
</tr>
</tbody>
</table>

*Note. Standard error is noted in parentheses.*

Replicating Experiment 3A, there was a significant three-way interaction, indicating that the advantage of studying over multiple-choice testing existed mainly for new learning that was assessed immediately, $F(1,94) = 15.21, MSE = .02, p < .001, \eta^2 = .01$. The benefit of studying for new learning was quite large on the immediate final test, $t(60.96) = 9.81, SED = .27, p < .001, d = 2.00$, and was still present, but dropped dramatically in size on the delayed final test, $t(48.90) = 4.55, SED = .05, p < .001, d = .93^3$. Indeed, studiers retained only 5% of newly learned fictitious items a week later, suggesting that most new learning was forgotten. Thus, at a week delay, the comparison between studying and multiple-choice testing was almost (but still not perfectly) fair. Critically, multiple-choice testing was just as powerful as studying in reactivating
estimated marginal knowledge on both the immediate $t(94) = 1.83$, $SED = .04$, $p = .07$, $d = .37$, and week-delayed final tests, $t(94) = 1.53$, $SED = .05$, $p = .13$ $d = .31$.

4.2.3 Discussion of Experiments 3A and 3B

Building on the small literature about marginal knowledge (mostly from Bahrick’s laboratory), Experiments 3A and 3B 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. Both experiments also confirmed that marginal knowledge can easily be reactivated, through multiple-choice testing or re-exposure. These interventions had a lasting impact – although the benefit was largest on an immediate test, most of the reactivated knowledge remained accessible one week later.

Critically, multiple-choice testing was just as effective as studying the correct answers in reestablishing access to marginal knowledge. This finding was impressive given that the studying condition had a clear advantage: For information that was unknown rather than marginal, studying the answers afforded more new learning. Experiment 3B helped level the playing field by extending the delay to one week, where studiers forgot most (but not all) new learning. Under these fairer circumstances, I replicated Experiment 3A’s conclusion that multiple-choice testing and studying reactivated similar levels of marginal knowledge.
Given that the vast testing effect literature finds large advantages of testing over studying, it might come as a surprise that multiple-choice testing was merely equivalent to studying in these experiments. Why did the retrieval practice offered on the multiple-choice test not promote even greater long-term retention? One possibility is that the multiple-choice tests did not offer enough retrieval practice to see such an advantage. Although the literature is mixed, several studies find that multiple-choice tests yield smaller testing effects than short-answer tests, presumably because they require less retrieval effort (e.g., Butler & Roediger, 2007; but see Little et al., 2012; see Marsh & Cantor, 2014 for a review). Alternatively, it could be the case that reactivating marginal knowledge simply requires re-exposure to the target, and that it does not matter whether that re-exposure takes the form of being told the answer or recognizing it on a list. 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. Regardless, the finding that multiple-choice testing was just as powerful as studying means that an educator looking to reactive key background knowledge can do so while simultaneously measuring her students’ knowledge.
5. Experiment 4

While the availability-accessibility distinction is commonly drawn upon in the laboratory, this idea is not typically applied to education. Learning loss is typically characterized as a loss from the memory system (e.g., “It is clear that students do forget mathematics material over the summer” Cooper, 2003, p. 3) and educators typically treat the problem as such by spending weeks re-teaching background material (National Summer Learning Association, 2013). Accordingly, it was an open question whether students who “forget” course material still have at least some of that information stored in memory. Supporting this possibility, students taking Spanish exams years after course completion perform better on questions requiring recognition (multiple-choice) than those requiring recall (short-answer; Bahrick, 1984). Thus, the first goal of Experiment 4 was to confirm that inaccessibility is part of the learning loss problem. If so, Experiment 3 suggests that a simple intervention consisting of multiple-choice questions should help students regain access to forgotten classroom knowledge. In Experiment 4, I implemented this intervention in a classroom with educational materials to determine whether these findings generalized.

While recovering forgotten background knowledge is valuable in its own right, it is also important to consider how reestablishing this knowledge influences the acquisition of upper-level material. For educators who typically reteach the foundational concepts, a multiple-choice testing intervention should free up time (since questions can
be administered quickly and outside of class), allowing them to dedicate more time to the new material. Furthermore, compared to situations where review is not standard, recovering background knowledge should promote learning of the new material (e.g., Bloom, 1976; Dochy et al., 1999). Accordingly, Experiment 4 also examined whether reestablishing access to forgotten background knowledge via the multiple-choice testing intervention promoted learning of new material that built upon this knowledge.

To test these ideas, I collaborated with a Pharmacy professor at University of North Carolina Chapel Hill to implement a multiple-choice testing intervention in his classroom. The goal was to take the topics that the instructor was going to be teaching in the course and identify critical foundational knowledge. This foundational knowledge was intended to be information that 1) students should have learned in prior courses but 2) may have become inaccessible and 3) would be necessary to know in order to acquire the new course information. To preview, the results suggested that the selected knowledge was not actually critical to learning the new course material, so I will refer to the information as “prior course material” rather than “foundational.”

At the beginning of the semester, students took baseline short-answer tests for all topics to determine what information was already available and accessible. Immediately afterwards, all students completed the multiple-choice testing intervention for half of the topics. The instructor proceeded to teach the new course material without explicitly re-teaching any of the prior course material. We measured the reactivation of the prior
course material twice: Once during the middle of the semester before the corresponding new course material was taught and once at the very end of the semester.

5.1 Method

5.1.1 Participants

Participants were first-year students enrolled in the four-week Pharmacy Bridging Course before the Fall 2016 semester at the University of North Carolina Eshelman School of Pharmacy. This bridging course was designed to help students transition into the graduate pharmacy program. All enrolled students completed the experiment as a part of the course, but only students who consented to participate in the study had their de-identified data released for analysis. Out of the 141 students enrolled in the course, five students did not consent to have their data released for analysis and six students did not complete all of the critical assessments. The remaining 130 participants were not evenly split between the two counterbalances (68 vs. 62) so I excluded six participants (randomly, without looking at the data) to achieve even counterbalancing, yielding a total of 124 participants (83 females, 39 males, 2 did not report gender; M age = 23 years, range = 19-52 years). Participants were not compensated.
5.1.2 Design

I used a within-participants design where all students received the multiple-choice testing intervention for half (three) of the target prior course material topics. See Figure 15 for a schematic.

![Figure 15: Schematic of Experiment 4.](image)

5.1.3 Materials, Counterbalancing, and Procedure

Seven topics were taught in the course: Cardiology, renal, hematology, central nervous system, gastrointestinal, respiratory, and endocrine. See Table 6 for sample materials. For each of these topics, the instructor created short-answer questions about the material that would be explicitly taught in the course (new course material). There were between six and 10 questions per topic for a total of 61 new course material questions. As previewed earlier, the intention was to also identify critical background information for each topic that students were expected to have learned about in prerequisite courses. However, since the results suggest that this knowledge was not truly foundational, I refer to this information as prior course material. For each of the
seven topics, the instructor created five short-answer questions that measured the prior course material. For each prior course material question, he also created four response options (the correct answer and three plausible lures) for when the question appeared on the multiple-choice testing intervention. The new course material and prior course material questions were used during the baseline, mid-course, and end-of-course assessments; prior course material questions were also used in the multiple-choice testing intervention.

An even number of topics was required to evenly counterbalance the topics across conditions, so the cardiology questions were not included in the experiment. This decision was made based on the instructor’s prior data showing that student performance was highest for cardiology questions. Thus, a total of six topics were included in the experiment. Materials were counterbalanced so that across students, each topic was assigned equally to the intervention and control conditions.
Table 6: Examples of prior course (P) and new course (N) material questions with answers for the six topics used in Experiment 4.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNS</td>
<td>(P) The ___ is where the axon is joined to the cell. It is from here that the electrical firing known as the action potential occurs.</td>
<td>Axon Hillock</td>
</tr>
<tr>
<td>CNS</td>
<td>(N) When potassium channels are open, the neuron cannot fire again. This is called the ___.</td>
<td>Absolute Refractory Period</td>
</tr>
<tr>
<td>Endo</td>
<td>(P) This class of hormones are synthesized on demand and are bound to carrier proteins in the blood (e.g., albumin). ___</td>
<td>Steroids</td>
</tr>
<tr>
<td>Endo</td>
<td>(N) An increase of circulation thyroid hormones will most likely cause the overall metabolic rate to ___.</td>
<td>Increase</td>
</tr>
<tr>
<td>GI</td>
<td>(P) ___ cells are specialized macrophages located in the liver.</td>
<td>Kupffer Cells</td>
</tr>
<tr>
<td>GI</td>
<td>(N) What blood vessel carries blood from the intestinal tract into the liver? ___</td>
<td>Portal Vein</td>
</tr>
<tr>
<td>Heme</td>
<td>(P) In blood, this is the component that does not contain blood cells or clotting factors. ___</td>
<td>Serum</td>
</tr>
<tr>
<td>Heme</td>
<td>(N) This white blood cell type is ultimately responsible for secreting antibodies. ___</td>
<td>B-Cell</td>
</tr>
<tr>
<td>Renal</td>
<td>(P) This cell is within the Bowman's capsule that wrap around capillaries of the glomerulus and the slits between the foot-like projects allow blood to be filtered. ___</td>
<td>Podocyte</td>
</tr>
<tr>
<td>Renal</td>
<td>(N) Passive reabsorption of solutes by the kidney is depends on urine pH/drug pKA, lipophilicity and ___.</td>
<td>Urine Flow</td>
</tr>
<tr>
<td>Resp</td>
<td>(P) The ___ nerve branches to the bronchial nerves which innervate the lungs to control breathing</td>
<td>Vagus</td>
</tr>
<tr>
<td>Resp</td>
<td>(N) Oxygen and carbon dioxide exchange in the lungs occurs by ___.</td>
<td>Diffusion</td>
</tr>
</tbody>
</table>
5.1.3.1 First Day of Course

All students began the course by taking two baseline assessments: The new course material and prior course material assessments. Students completed both assessments on individual computers. The baseline new course material assessment was comprised of the 61 new course material questions and the baseline prior course material assessment consisted of the 35 prior course material questions. The purpose of these assessments was to establish which information was accessible versus inaccessible (or not known) at the beginning of the course. The two assessments were administered in immediate succession and the students were not informed about any differences between the two types of questions. For all questions, students were instructed to work independently and to answer all questions to the best of their ability. No feedback was given at any point.

The students began with the new course material assessment. Each new course material question was presented individually in short-answer format and students were instructed to type their answer. Students were required to submit an answer before moving on to the next question; backward navigation was not permitted.

After answering all the new course material questions, students completed the prior course material assessment and multiple-choice testing intervention. The intervention was implemented during the assessment: For the 15 questions pertaining to the three topics that were assigned to the multiple-choice testing condition, the multiple-
choice version of the question immediately followed the short-answer version. For these questions, the short-answer version appeared first and students typed their response. Once the response was submitted, the multiple-choice version was presented and students were instructed to select the best answer out of the four response options. Again, backward navigation was not permitted so students were not able to change their short-answer response upon viewing the multiple-choice options. For the remaining 20 questions not assigned to the multiple-choice testing condition (15 control, 5 Cardiology questions not included in analyses), students only answered them in short-answer format.

5.1.3.2 During the Course

The instructor taught the new course material pertaining to the seven topics named earlier. One topic was taught each class meeting. Students were instructed to prepare for each class meeting by studying the topic that would be taught. The instructor began each class meeting by having the students complete the mid-course new course material and prior course material assessments. These assessments consisted of the questions corresponding to the topic that would be taught in that class. For example, when the instructor taught his lesson on hematology, the mid-course new course material and prior course material assessments contained only the hematology questions. These questions were the same ones that appeared on the baseline assessments and were presented only in short-answer format. Students were again
instructed to answer each question to the best of their ability on their own without consulting outside sources. No feedback was given. The purpose of the prior course material assessment was to measure the persistence of any reactivation of the prior course knowledge due to the multiple-choice testing intervention. The new course material assessment was included to evaluate whether any reactivation of the prior course material translated into improved acquisition of the new material. Although the instructor had not given his lecture at this point, students were expected to study the new topic prior to the lecture.

The instructor then proceeded to teach the new course material via lecture. Importantly, the lectures focused on the new course material and did not explicitly reteach the prior course material.

5.1.3.3 End of Course

Finally, the end-of-course new course material and prior course material assessments for all seven topics were administered at the end of the course. These assessments included the same questions from the baseline and mid-course assessments. Like the mid-course assessments, questions were only presented in short-answer format. The instructions were the same as the prior assessments. The prior course material assessment was included to evaluate whether any benefits of the intervention on reactivating prior course knowledge persisted until the end of the course. The new
course material assessment measured whether the intervention promoted learning of new course material.

5.2 Results

5.2.1 Coding

The instructor scored all short-answer responses as correct or incorrect. No partial credit was given.

5.2.2 Determining Marginal Knowledge

First I determined whether students had prior course material stored in memory that was not accessible at the beginning of the course. To do so, I examined questions that students did not answer correctly on the baseline short-answer prior course material assessment. The key question was whether students were able to select the correct answer on the multiple-choice test at a rate higher than chance (i.e. more than 25% of the time), indicating that at least some of the knowledge was actually stored. Note that this analysis could be conducted for all participants but only for the three only topics assigned to the multiple-choice testing intervention condition. Indeed, when students failed to answer the short-answer question correctly (which was most of the time; \( M = .90, \ SD = .11 \)), they selected the correct multiple-choice alternative 46% of the time, a rate substantially greater that 25%, \( t(123) = 13.56, p < .001, d = 1.24 \). Thus, for almost half of the knowledge students could not retrieve, they were able to recognize it – some of this knowledge was stored in memory but it was not accessible.
5.2.3 Reactivating Marginal Knowledge

Clearly some of students’ prior course material was marginal – stored but inaccessible. The laboratory data from Experiment 3 suggests that the multiple-choice testing intervention should reestablish access to this marginal knowledge. First I examined whether any reactivation of marginal prior course knowledge persisted until the mid-course assessment. Because the timing of the assessment varied depending on topic, these assessments occurred between six and 19 days after the intervention. Table 7 contains the relevant data. For questions that participants could not answer correctly on the baseline short-answer assessment, taking the multiple-choice test improved performance on the mid-course assessment relative to control questions not on the multiple-choice test, $t(123) = 5.24$, $SEM = .02$, $p < .001$, $d = .47$. This benefit of the multiple-choice testing intervention appeared to be driven largely by participants’ success on that test: Students performed much better on the mid-course assessment after answering the multiple-choice test question correctly ($m = .41$) than incorrectly ($m = .14$), $t(119) = 12.25$, $SEM = .02$, $p < .001$, $d = 1.12$.

The multiple-choice testing intervention did successfully reactivate marginal knowledge as demonstrated on the mid-course assessment. Next I examined whether this reactivation persisted until the end-of-course assessment that occurred approximately three weeks after the intervention. Indeed, the results from the mid-

---

1 Four participants were not included in this analysis as they did not answer any multiple-choice questions correctly that were incorrectly answered in short-answer format.
course assessment were replicated on the post-course assessment. First, the benefit of multiple-choice testing extended until the end of the course. That is, for questions not answered correctly on the baseline short-answer assessment, students performed better on the end-of-course assessment when the questions appeared on the multiple-choice test, \( t(123) = 5.19, SEM = .02, p < .001, d = .47 \). Again, students performed far better on the end-of-course assessment after selecting the correct multiple-choice answer (\( m = .40 \)) relative to selecting an incorrect alternative (\( m = .12 \)), \( t(119) = 12.72, SEM = .02, p < .001, d = 1.16 \).

Table 7: Proportion of prior course and new course material questions answered correctly on the mid-course and end-of-course assessments (given failure on the baseline assessment). Data are from Experiment 4.

<table>
<thead>
<tr>
<th></th>
<th>Prior Course Material</th>
<th>New Course Material</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mid-Course</td>
<td>End-Of-Course</td>
</tr>
<tr>
<td>Intervention</td>
<td>.27 (.02)</td>
<td>.26 (.02)</td>
</tr>
<tr>
<td>Control</td>
<td>.17 (.01)</td>
<td>.18 (.01)</td>
</tr>
</tbody>
</table>

*Note. Standard error is noted in parentheses.*

### 5.2.4 Acquiring New Course Material

My original intention was to investigate the power of reactivating necessary background knowledge on the acquisition of new material. As I’ll explain below, the data suggest that the prior course material was not truly foundational. Nevertheless, I examined whether reactivation of prior course knowledge via the multiple-choice
testing intervention promoted acquisition of the corresponding new course material. To answer this question, I looked to performance on the mid-course and end-of-course new course material assessments for questions that participants could not answer correctly on the baseline new material assessment. The data appear in Table 7. The multiple-choice testing intervention did not promote performance relative to the control condition on the mid-course assessment, $t(123) = 1.06, SEM = .01, p = .290, d = .10$, or the end-of-course assessment, $t < 1.$

Why did reactivation of prior course knowledge not translate to improved acquisition of the new course material? I consider a few possible explanations. First, there appeared to be a ceiling effect on the end-of-course assessment, in that performance was at 93% for both intervention and control topics. Given that students performed so well on this assessment, any potential differences between the intervention and control topics could have been masked. Of course, this explanation does not apply to the mid-course assessment. Second, the overall rates of marginal knowledge were relatively modest. While clearly some of the knowledge was marginal, at least half of it was not: For short-answer questions students could not answer, they selected the incorrect multiple-choice alternative approximately half of the time, suggesting they did not know this material.

The most likely possibility is that the prior course material identified by the instructor was not truly “foundational” – it was not actually necessary to know in order
to learn the new material. A few data points support this possibility. Namely, as highlighted in Table 8, most of the correlations between performance on the prior course material and new course material questions for a given topic were quite low, particularly on the mid-course and end-of-course assessments. Providing further support for this idea, Table 7 indicates that students reached almost perfect performance on the new material by the end of the course, but failed to master even a third of the prior course material. Given that the prior course knowledge in this study did not seem to be truly critical for the acquisition of the new material, this data should not be interpreted as evidence that reactivating marginal knowledge does not promote new learning.

Table 8: Correlations between performance on prior and new course material questions by topic for the baseline, mid-course, and end-of-course assessments. Data are from Experiment 4.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Baseline</th>
<th>Mid-Course</th>
<th>End-of-Course</th>
</tr>
</thead>
<tbody>
<tr>
<td>Central Nervous System</td>
<td>.54**</td>
<td>.15</td>
<td>.14</td>
</tr>
<tr>
<td>Hematology</td>
<td>.49**</td>
<td>.18**</td>
<td>.09</td>
</tr>
<tr>
<td>Renal</td>
<td>.40**</td>
<td>.20*</td>
<td>.17</td>
</tr>
<tr>
<td>Endocrine</td>
<td>.25**</td>
<td>.19**</td>
<td>.12</td>
</tr>
<tr>
<td>Gastrointestinal</td>
<td>.20**</td>
<td>.15</td>
<td>.17</td>
</tr>
<tr>
<td>Respiratory</td>
<td>.39**</td>
<td>.30**</td>
<td>.25**</td>
</tr>
</tbody>
</table>

*p < .05. **p < .01.
5.3 Discussion

Applying the availability-accessibility distinction to the classroom, Experiment 4 first confirmed that students do have prior course material stored in memory that is momentarily inaccessible. This finding suggests a different way for educators to conceptualize student “forgetting.” Rather than assuming that students who cannot produce the prerequisite information have forgotten it completely, educators should consider that the information may have become inaccessible. As such, it may not be necessary to use class time to reteach this background information. To the extent that background material has become inaccessible, it only needs to be reactivated rather than relearned from scratch.

Indeed, I found that multiple-choice testing successfully reactivated such marginal knowledge. This simple intervention appeared to have long-lasting effects: Reactivated knowledge remained accessible at least three weeks later. Anecdotally, the instructor administered a surprise follow-up test for the prior course material at the beginning of the Spring semester. Five months after the intervention, students still were more likely to recover initially inaccessible knowledge if it was tested on the multiple-choice test.

Finally, to the extent that marginal knowledge is truly foundational, reactivation may promote the acquisition of new material. I did not find support for this conclusion here, most likely because the marginal knowledge was not in fact critical to the learning
of the new course material. Future research should utilize the appropriate materials to fully address this important question.
6. General Discussion

My experiments do not speak to the impressiveness of human knowledge – we know that chess experts have upwards of 50,000 game board patterns stored in memory (Gobet & Charness, 2006) and that adult English speakers are estimated to know tens of thousands of words (Brysbaert, Stevens, Mandera, & Keuleers, 2016). Classic work by Bahrick and colleagues (see Bahrick et al., 2013 for a review) shows that the total amount of information (e.g., Spanish language, city maps, high school classmates’ names, etc.) often remains relatively stable over time, following a period of initial forgetting. Furthermore, such knowledge appears to be quite robust in the face of amnesia (Parkin, 1997) and old age (see Umanath & Marsh, 2014 for a review). My point is about fluctuations in access to specific knowledge; one may know a lot and yet be momentarily unable to retrieve a desired name, citation, or fact.

While it is not hard to find examples of specific knowledge fluctuating in accessibility, there is no larger database that allows us to make predictions about when knowledge will be stable versus unstable. Thus, the first goal of my dissertation was to begin to build such a database where participants repeatedly generate knowledge. Experiments 1A and 1B found that despite relative consistency in the total amount of knowledge accessible, individual pieces of knowledge fluctuated in accessibility. This instability was quite substantial: After only 15 minutes, about one fifth of initially accessible knowledge became inaccessible and a roughly equivalent amount of
previously inaccessible information was recovered. These considerable fluctuations are even more impressive when considering that the knowledge being generated (e.g., US States) was very well learned. Knowledge became even less stable over time, with about one quarter of knowledge becoming inaccessible. Interestingly, taking an intervening retest between generation attempts did not help stabilize knowledge. Of course, these experiments are only a starting point to understanding the factors that influence knowledge instability. Future research should continue to probe additional factors, for example, examining whether certain types of knowledge are more or less stable.

Given that knowledge is often described as stable in relation to event memory, the second goal of my dissertation was to directly compare the two types of memories. Experiment 2 highlighted a few key ways that event memory fluctuations differed from those in knowledge. Time had a greater impact on event memory, causing participants to lose access to about 40% of previously accessible information. The most striking difference between event memory and knowledge pertained to the effect of testing. Whereas testing did not affect knowledge, testing prevented participants from losing access to event memories and promoted the recovery of previously inaccessible ones.

These differences between event memory and knowledge are likely due (at least in part) to the differential role that context plays in memory retrieval. As discussed earlier, event memories are tied to their encoding contexts whereas knowledge retrieval appears to depend less upon an encoding-retrieval context match. Since context is more
likely to change with increasing time, it makes sense that event memories should be affected more by delay. Furthermore, to the extent that the testing effect relies upon a reinstatement of the initial learning context (Karpicke et al., 2014), testing should stabilize event memory but not knowledge. Such explanations fit well with theories that identify context as a key aspect of event memory. For example, Rubin and Umanath (2015) argued that the defining property of event memory – the one that contrasts it with knowledge – is that event memories are retrieved with a spatial context. More broadly, given that the differences between event memory and knowledge were observed when content was held constant across conditions, these data fit better with theories positing that the manner with which content is retrieved (rather than the content itself) distinguishes the two memory types (e.g., Klein 2013; Markowitsch & Staniloiu 2011). The data are less consistent with content-based and system-based distinctions between event memory and knowledge (e.g., Dickerson & Eichenbaum, 2010; Hutchinson & Turk-Browne, 2012).

Given how important context seems to be in explaining differences between event memory and knowledge, future research should further elucidate the process by which knowledge becomes sourceless. Thus far, there is evidence that over time, information changes from being “remembered” to known” (e.g., Conway et al., 1997) and that testing attenuates this shift (Barber et al., 2008). Furthermore, there is some evidence that when an item is presented multiple times via multiple sources, memory
for the source of any individual presentation is impaired (Watkins & Kerkar, 1975).

Therefore, encoding variability may play a key role in knowledge becoming sourceless.

While many theories posit that encoding variability promotes long-term retention (e.g., Bower, 1972; Glenberg, 1979; Martin, 1968), the empirical findings are mixed (see Huff & Bodner, 2014). Future research should clarify how variable encoding influences the creation of knowledge.

The substantial knowledge fluctuations I documented have clear implications for education, as stored knowledge provides the foundation for new learning (Bransford, Brown, & Cocking, 2000). While the focus of educational research is often on new learning, that new learning depends upon retention, retrieval, and transfer of prior learning, making understanding how to reestablish access to that foundation a laudable goal. Accordingly, the third goal of my dissertation was to examine methods for regaining access to marginal knowledge. I found that when knowledge became inaccessible, multiple-choice testing and studying were equally effective in reinstating access. The fourth and final goal of my dissertation was to apply these ideas in a real classroom. Experiment 4 demonstrated that some of students’ prior course knowledge had become marginal and that multiple-choice testing could be used to reactivate this information. This simple intervention had a lasting impact – previously inaccessible educational content remained accessible three weeks later.
Experiments 3 and 4 highlighted the prevalence of marginal knowledge, both in the laboratory and in the classroom. These findings indicate that knowledge accessibility is at least part of the learning loss problem, suggesting a different way for educators to conceptualize and remediate student forgetting. As I have demonstrated, reactivating knowledge that has become inaccessible can be as simple as administering a multiple-choice test. Multiple-choice tests can be administered and graded easily and can be completed in or outside the classroom. Considering their simplicity, the interventions were quite powerful and long-lasting. It is noteworthy that reactivation persisted weeks later, 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.

These findings also contribute to a broader literature examining the effects of retrieval practice on memory. 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., Marsh & Cantor, 2014). Multiple-choice tests have been shown to promote the retention of tested material (e.g., Roediger & Marsh, 2005) and the ability to transfer that information to new contexts (McDaniel, Thomas, Agarwal, McDermott, & Roediger, 2013). My experiments contribute to this literature by highlighting a new benefit of multiple-choice testing: Such tests can help students maintain access to information that might otherwise be lost. Multiple-choice testing
should 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. These findings also fit well with the idea that most learning happens across many trials rather than one (e.g., Rawson, Dunlosky, & Sciartelli, 2013).

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 reestablish 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.

Appreciating the fluctuations of individual pieces of knowledge allows for a more nuanced and complete characterization of knowledge as simultaneously stable on
a macro level but relatively unstable on a micro level. 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 re-activated, even if they cannot currently access it. My 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.
### Appendix A

Table 9: Average number of exemplars lost and gained from the initial test to the immediate retest, delayed retest, and delayed retest with an intervening retest. Data are from Experiments 1A, 1B, and 2 (knowledge condition only).

<table>
<thead>
<tr>
<th></th>
<th>Losses</th>
<th>Gains</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Immediate</td>
<td>Delayed (Tested)</td>
</tr>
<tr>
<td>Exp 1A</td>
<td>4.89 (.56)</td>
<td>6.16 (.43)</td>
</tr>
<tr>
<td>Exp 1B</td>
<td>3.33 (.32)</td>
<td>4.83 (.33)</td>
</tr>
<tr>
<td>Exp 2</td>
<td>4.32 (.41)</td>
<td>5.42 (.42)</td>
</tr>
</tbody>
</table>

*Note.* Standard error is noted in parentheses.

Table 10: Average number of exemplars lost and gained from the initial test to the immediate retest, delayed retest without an intervening retest, and delayed retest with an intervening retest. Data are from Experiment 2.

<table>
<thead>
<tr>
<th></th>
<th>Losses</th>
<th>Gains</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Immediate</td>
<td>Delayed (Tested)</td>
</tr>
<tr>
<td>Event</td>
<td>2.29 (.20)</td>
<td>8.08 (.55)</td>
</tr>
<tr>
<td>Knowledge</td>
<td>2.93 (.31)</td>
<td>3.56 (.25)</td>
</tr>
</tbody>
</table>

*Note.* Standard error is noted in parentheses.
References


National Summer Learning Association (2013). Teachers confirm time wasted due to summer learning loss. Baltimore, MD.


Biography

Allison Cantor Black-Maier was born April 8, 1990 in Evanston, Illinois. She later moved to Rye, New York and graduated from Rye High School in 2008.

Allison attended Washington University in St. Louis from 2008-2012 and completed a bachelor’s degree in the interdisciplinary Philosophy-Neuroscience-Psychology program. She worked as a research assistant in Dr. Kathleen McDermott’s Memory Laboratory and conducted a thesis on test-potentiated learning. Allison graduated Summa Cum Laude and Phi Beta Kappa.

Allison began her graduate studies at Duke University in the Psychology and Neuroscience Department in 2012. She conducted research on human learning and memory with Dr. Elizabeth Marsh while completing her PhD in Cognitive Psychology. Allison earned her Master’s in Psychology in 2015 and will graduate with her PhD in May of 2017.