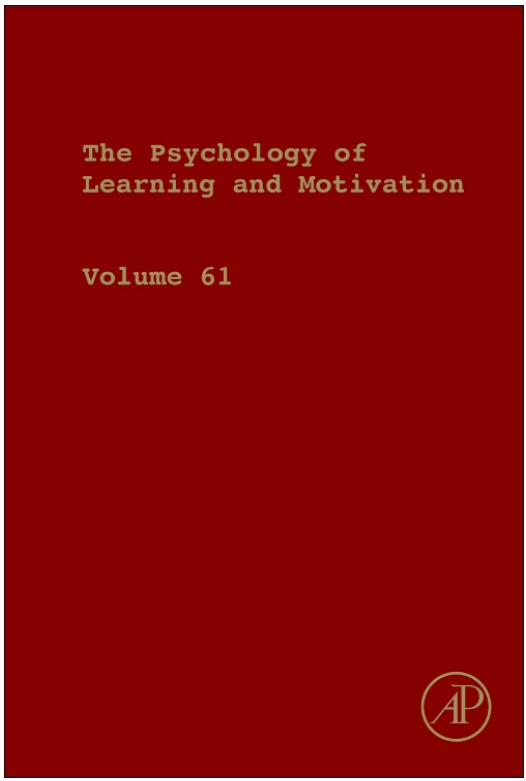

From: Kenneth J. Malmberg, Melissa Lehman, Jeffrey Annis,
Amy H. Criss and Richard M. Shiffrin. Consequences of Testing Memory.
In Brian H. Ross editor: *The Psychology of Learning and Motivation,* Vol. 61,
Burlington: Academic Press, 2014, pp. 285-313.
ISBN: 978-0-12-800283-4

Academic Press

CHAPTER EIGHT

Consequences of Testing Memory

Kenneth J. Malmberg[*,1], **Melissa Lehman**[†], **Jeffrey Annis**[*], **Amy H. Criss**[‡], **Richard M. Shiffrin**[§]

[*]Department of Psychology, University of South Florida, Tampa, Florida, USA
[†]Department of Psychological Sciences, Purdue University, West Lafayette, Indiana, USA
[‡]Department of Psychology, Syracuse University, Syracuse, New York, USA
[§]Department of Brain and Psychological Sciences, Indiana University, Bloomington, Indiana, USA
[1]Corresponding author: e-mail address: malmberg@usf.edu

Contents

1. Introduction 285
2. Benefits of Memory Testing 287
   2.1 Generalization of the Testing Effect 288
   2.2 Why Does Testing Improve or Harm Memory? 290
3. Costs of Testing Memory: Output Interference 298
4. The Influence of One Test on the Next: Sequential Dependencies 302
5. Past Decisions Influence Future Decisions: Shifts in Bias 306
6. Conclusions 308
References 308

Abstract

Studies using a wide variety of conditions and a diverse set of procedures show that testing memory affects future behavior. The studies have used differing terminology and have been ascribed to differing specialty areas of the literature. Partly, for this reason, the various phenomena have been described in ways, suggesting they differ in substance. In this chapter, we relate many of these phenomena and show that they might be due to a set of common memory processes, processes that can act through conscious, strategic or unconscious, implicit means. The critical strand that links the phenomena is that memory is a continuous process that constantly stores and retrieves information.

1. INTRODUCTION

In typical laboratory investigations of human memory, learning occurs when one is asked to study a set of to-be-remembered items. During study, participants are either left to their own devices or assigned a variety of tasks that guide the learning in a particular direction. In these tasks, it is typical for

Psychology of Learning and Motivation, Volume 61
ISSN 0079-7421
http://dx.doi.org/10.1016/B978-0-12-800283-4.00008-3

memory to be tested in one or more ways after some delay. The three phases are referred to as study, retention, and test. Ebbinghaus (1885) established that the amount of information forgotten increases with increases in the length of retention interval, and the literature developed since then has focused primarily on how the conditions during study are related to performance when memory is tested. In much of the early literature, it was assumed that only study, not testing, impacts future memory. Evidence to the contrary was often obscured by designs, randomizations, and analyses that made it hard or impossible to see the effects of testing. However, recent years have seen an upsurge of investigations examining and demonstrating remarkably strong effects of testing. This should not be a surprise given that we can obviously remember what was tested, observe what transpires during testing, and learn from the results of testing. The general rule is that the act of remembering affects subsequent learning and retrieval. Many of the effects are beneficial, but some are harmful. In this chapter, we review both and discuss how memory models can explain them.

There is actually a fairly long history of studies of the effects of testing, but most of the older studies confounded testing and studying. A set of older studies and models by Izawa (1970, 1971) examined the "potentiating" effects of testing, but did not at the time engender further research. Roediger and Karpicke (2006) showed that testing a memory can produce better learning than studying a second time, sparking a renewed interest in the consequences of testing memory. The studies have inspired both systematic investigations into applications and a better understanding of basic processes involved in learning, remembering, and forgetting. Because the various findings are not well organized by a common theoretical framework and are found in different areas of psychological science, we aim in this chapter to relate the various findings with the help of memory theory.

As just mentioned, a driving force behind the recent investigations is the finding that eventual retrieval of some initially studied information can be increased more by an intermediate act of retrieval than by an intermediate act of study. This phenomenon is often referred to as *the testing effect* or *the retrieval practice effect*. Such benefits of retrieving information from memory are robust, occurring for a range of materials and test types (see Roediger & Karpicke, 2006 for a review). In fact, the layperson may have benefited from similar procedures, such as flashcards, when studying for exams (albeit typically such use involves both testing and study).

However, there are actually quite a large number of "testing effects." Some are positive as in the many studies of learning, and the testing effects

just mentioned. Others are negative. For instance, retrieval from memory has negative consequences for the future retrieval of other information, generally termed *interference* and studied extensively in list learning experiments in the 1950s and 1960s (see Crowder, 1976 for a review). Another set of retrieval tasks producing negative memory effects are referred to as *retrieval-induced forgetting* (Anderson, Bjork, & Bjork, 1994, 2000; Jakab & Raaijmakers, 2009; Raaijmakers & Jakab, 2013a, 2013b; there is an ongoing debate concerning the degree to which such effects are due to *active suppression* or *competition*). Yet other paradigms demonstrate *output interference* during the course of successive testing (Criss, Malmberg, & Shiffrin, 2011; Raaijmakers & Shiffrin, 1980; Wickens, 1970).

Other test effects can be either positive or negative. For example, decisions made on the basis of what is retrieved from memory often affect in the future what we believe we have or have not experienced (Malmberg & Annis, 2012). In addition, *sequential dependencies* occur for both events that were and were not experienced; these consequences of testing memory are robust and tend to produce systematic mnemonic bias rather than changes in overall levels of accuracy.

These brief citations are enough to suggest that a number of processes are at work during testing. It is abundantly clear that testing and retrieving has significant implications for what we will remember in the future. We shall see that major factors are the storage of new information during testing (producing both positive and negative effects), and strategic changes induced by learning from the results of testing. Specifically, we consider three classes of processes by which memory testing can alter performance on subsequent tests: (a) storing new memory traces during testing; (b) enhancing, modifying, or updating existing memory traces; and (c) altering learning and retrieval strategies. Examining how these factors work is an aim of this chapter. We first review some of what is known about the consequences of testing memory.

2. BENEFITS OF MEMORY TESTING

Starting in the late 1800s (e.g., Ebbinghaus, 1885), there has been a long history of studies of learning and forgetting. Such learning produces memory for recent events, such as lists of words or words pairs, and also produces knowledge. Nelson and Shiffrin (2013) describe how a common set of memory processes can produce both types of learning. Many articles and books have dealt with learning and memory, typically dealing with tasks in which testing involves both testing and studying. This chapter instead

focuses on more recent tasks in which the effects of testing are separated (at least partially) from the effects of studying.

2.1. Generalization of the Testing Effect

Studies examining the "testing effect" on retention often utilize free recall tests, whereby a list of to-be remembered items is studied and after a retention interval, one is asked to recall as many items from the study list in any order. In the control condition, one is given a second chance to study the items. The key question concerns whether testing memory provides any substantial improvement in memory beyond the benefits of additional study, as measured during a subsequent round of free recall. This final measure of memory retention is referred to as a *criterion test*. When the retention interval between the first and second round of testing is lengthened, forgetting of the original material increases in both conditions, but the rate of forgetting is greater for items given additional study compared to items that were recalled in the first round of testing (Roediger & Karpicke, 2006).

Of course, we often want to remember specific events, in contrast to remembering an entire class of events as in free recall. For instance, in cued recall participants study pairs of items and are tested with one member of the pair provided as a cue (a testing procedure somewhat like a short answer or fill-in-the-blank test found in educational settings). In recognition tests, participants must decide whether a tested item had (a "target") or had not (a "foil") been studied on a prior list (a situation analogous to a true–false test or a multiple-choice test found in educational and legal settings, such as identifying a suspect from a lineup).

Researchers have asked whether the benefits found for testing in free recall extend to cued recall and recognition. They report that initial free recall testing reduces the rate of forgetting more than initial cued recall or yes/no recognition testing (Carpenter & DeLosh, 2006; Glover, 1989). In fact, it appears that initial yes/no recognition has little effect compared to restudying regardless of the task used on the final test of memory (Carpenter & DeLosh, 2006). Likewise, evidence for an effect of initial free recall on final recognition testing is mixed (Carpenter & DeLosh, 2006; Darley & Murdock, 1971; Jones & Roediger, 1995; Roediger & McDermott, 1995). Hence, recognition testing appears to be much less effective than free recall testing in reducing the subsequent rate of forgetting, and recognition testing appears to benefit much less strongly than free recall testing when used as the criterion task.

Some of the variability reported in studies of test effects may be due to differing processes used in different tasks, differences likely caused by differing tasks and materials (see Gillund & Shiffrin, 1984; Malmberg, 2008 for a review of the way theory can take such differences into account). For instance, Chan and McDermott (2007) speculated that retrieval practice increases performance on recognition tests that rely on "recollective" processes (Mandler, 1980). Recollective processes are often associated with or defined by memory for details of an encoding event and may be due to the use of processes during recognition that are also required for cued or free recall. Chan and McDermott found that retrieval practice increased performance on a list discrimination test and increased "remember" responses, although there was little effect of retrieval practice on final yes–no recognition hit rates. They concluded that retrieval practice increased the tendency to make a recognition decision based on the retrieval of episodic details of the events and decreased the tendency to make a recognition decision based on the familiarity of the test item. Whether recollective processes are used to a significant degree during recognition testing is a debatable issue (Dunn, 2004; Malmberg, 2008; Malmberg, Zeelenberg, & Shiffrin, 2004), and likely depends on procedural details including timing, incentives, and instructions, because the use of recollective processes to make recognition judgments surely requires more time and effort than judgments made on the basis of familiarity.

It is also likely that the benefits of testing do not apply equally to all aspects of to-be-remembered events. Brewer, Marsh, Meeks, Clark-Foos, and Hicks (2010) presented participants with two lists consisting of words spoken by both male and female speakers. Half of the participants completed a free recall test after each list, and the others completed an arithmetic task. A final source memory task followed in both conditions, which required participants to identify either the original study list from which the word was read or the gender of the speaker that read the item. Initial free recall testing increased final list discrimination performance, suggesting that the free recall practice improves the encoding and/or access to the temporal aspects of the events, but it did not increase gender discrimination performance, indicating that other aspects of the event, such as the representation of the source, are not enhanced by free recall testing. Taking this result together with the finding that recognition and cued recall do not benefit from testing to the extent that free recall does, it appears likely that more of the benefit imparted by retrieval is the result of the storage of additional features representing the context in which the testing occurred than the storage of additional perceptual features about the items to be remembered.

2.2. Why Does Testing Improve or Harm Memory?

Theories are constrained by the findings that testing provides more benefits for eventual free recall of a given item than benefits for eventual cued recall and recognition, especially when the initial testing also uses free recall. One hypothesis holds that free recall testing makes the memory in question more accessible by altering or adding to its contextual representation (Karpicke, Lehman, & Aue, 2014). Indeed, free recall is heavily dependent on the use of context information (Lehman & Malmberg, 2013; Malmberg & Shiffrin, 2005), and changes in context have been implicated in forgetting going back decades (Lehman & Malmberg, 2009; McGeoch, 1942; Mensink & Raaijmakers, 1989). Another hypothesis holds that testing causes an increase in item information and/or inter-item information than does study alone. For instance, there is a large amount of evidence that people are usually overconfident in their ability to recall specific items in the future, and they underestimate the amount of time it takes to learn pairs of words in anticipation of a cued recall test of memory, but when given the opportunity to test their learning prior to the criterion test, subjects are able to increase the amount of time allocated to studying the most difficult to remember pairs, at least when there is ample time to do so (Koriat & Bjork, 2005, 2006; Nelson & Leonesio, 1988; Son & Metcalfe, 2000). And of course, one should consider encoding models that assume free recall testing results in the storage of various combinations of these forms of information.

Another hypothesis holds that free recall testing may lead the subject to improve their retrieval strategy. Retrieval strategies are especially important in performing free recall, and they may be less important in item recognition or cued recall. However, it is quite difficult to think about changes in retrieval strategies without considering related changes in encoding. For instance, a more (or less) effective retrieval strategy may be implemented during or after the course of a single series of recall or recognition trials. Presumably, if it was adopted at some point during a series of tests, the new retrieval strategy would have some effect during the current round of testing and an effect during a subsequent round of testing, take for instance, associative recognition, which requires subjects to discriminate between pairs of items that were studied together (intact pairs) from pairs of items that were studied but not studied as part of the same pair (rearranged pairs). It is possible that the subject begins an initial round of testing with a strategy that utilizes information representing the familiarity of the test pair, but after a small number of associative recognition trials that a subject realizes that a

recollective strategy, such as recall to reject, may improve their accuracy on subsequent trials (cf. Malmberg, Holden, & Shiffrin, 2004; Malmberg & Xu, 2007; Xu & Malmberg, 2007). If this strategy were also used during criterion testing and if similar encoding did not take place in the control condition, then the benefits of testing would be apparent. However, the switch in retrieval strategy may co-occur with a new encoding strategy to support it. Consider again the associative recognition task. Experience with a few trials of the associative recognition task may inspire the subject to focus on encoding of associative information during the course of initial testing, strengthening the inter-item associations originally acquired during study (cf. Palmeri & Flanery, 2002, e.g., from categorization literature). If so, should one attribute the benefits of testing to the change in retrieval or the change in encoding? This question is difficult to answer without careful experimentation.

It seems reasonable to assume that both the retrieval strategy and the information retrieved will affect what is encoded during the test trial and the subsequent effects of testing. For instance, tasks like cued recall and associative recognition may encourage the encoding of inter-item associative information, but tasks or strategies that require retrieval of temporal context may result in more extensive encoding of temporal context during the course of testing. If the criterion task also requires access to temporal context, then test effects should be observed, in a manner akin to transfer-appropriate processing (Morris, Bransford, & Franks, 1977). However, if the criterion task requires access to different information, say associative information for a recall task, then encoding of temporal context features would be less beneficial. Information stored that is particular to a given memory task is what distinguishes in memory the performance of different tasks, and we will return to the effects of switches in task switching when we discuss output interference in a later section of this chapter. For now, we note that the benefits of testing are sometimes diminished when there is a conflict in the memory task performed during initial testing (e.g., free recall) and criterion testing (e.g., recognition).

However, it is also noted that matches in task context (e.g., recognition) do not necessarily predict benefits of memory testing (Carpenter & DeLosh, 2006). To receive benefits from memory testing, it is necessary to encode or store information that is not easily stored during the course of studying. In free recall, the items "tested" are only those generated by subject. According to some models, the cue set used to probe memory is determined in large part by the inter-item associations stored during study (cf. Lehman &

Malmberg, 2013; Raaijmakers & Shiffrin, 1980). It then follows that the order in which items are retrieved during free recall is decidedly nonrandom during free recall, and this provides a potentially rich reflection of the organization of memory traces stored during study and prior testing, but free recall does not provide a perfect image of the list of studied items, and in these cases, there is an opportunity to improve memory. Especially, for longer lists of items, there is likely to be mismatch between the study list—the order in which items were studied and/or the contents of the list itself—and what is retrieved during memory testing. This points to a possible component of the benefits to free recall of testing; use of inter-item associations during memory testing may increase the strength of those associations formed during initial learning. In addition, if the traces are also updated with temporal context features, then the benefits may persist over a period of time. If the same associations are strengthened during study in the control condition, similar encoding benefits could be available. However, the study conditions need to be just right; otherwise, there is a risk of adding new weak traces, especially when the order in which items are initially studied is different from the order in which items are restudied, because having many weak associative cues may be less effective than having single strong cues when performing free recall.

The fact that recognition testing and cued recall testing often involve all of the material on the study list distinguishes these tasks from free recall. Similarly, in the control condition, all items are studied. For paired associate cued recall, one might expect memory testing would provide the subject with knowledge about which pairs have been learned and which have not been learned because subjects are quite poor at predicting how likely they are to remember a cue–target combination in the future, unless some retention interval intervenes (Nelson & Dunlosky, 1991). Testing should therefore impart some benefit through the storage of item and/or associative information over and above study, especially if feedback in the form of the correct answer is provided when mistakes are made. In the Carpenter and DeLosh (2006) experiment, however, paired associate cued recall was not utilized in order to create a common study condition, in which single items were studied. Rather, subjects were cued with the first letter of each target word, and this version of cued recall is more similar to item recognition insofar as associations between items are not important in order to complete the task, and since feedback was not provided, it is unclear what advantage testing memory would have over additional study.

Assessing the benefits of recognition testing is complicated by the fact that unstudied items are tested in addition to the studied items. Although there may be some benefits of encoding additional item or contextual features during testing that distinguish targets from foils, the storage of traces representing the foil test trials will cause some additional interference if these traces are accessed during the criterion test. We will have much more to say concerning the consequences of recognition testing in subsequent sections.

Improvements in learning attributable to testing that are due to contextual storage and to cognitive control may both occur. Two studies by Lehman and Malmberg were aimed to study and perhaps provide separate evidence for these factors. In the published study (Lehman & Malmberg, 2013), participants completed several study–test cycles for free recall, with different words in each cycle. There were improvements in free recall over the course of eight study–test cycles. Thus, whatever harm might have been caused by interference from the storage of words studied and tested on earlier lists was overcome by factors that improved memory as cycles continued, such as improved storage and retrieval processes and strategies. Interestingly, when individual differences were analyzed, the improved performance was due almost entirely to the subjects who had the highest overall rate of free recall. The advantage of the high performers over the low performers extended throughout the free-recall serial position curve, a not surprising result showing that overall encoding and retrieval were better for high performers, but also showing that the advantage was not limited to short-term memory (recency portions of the serial position curve) or long-term memory (the rest of the serial positions). A more detailed analysis showed that the high performers were increasingly likely to begin retrieval with the final item on the list over cycles, whereas the low performers were more likely to begin retrieval with the item in the first serial position and less likely to switch this strategy. This finding indicates that the retrieval cues used to probe memory were different in the two groups, and that the high performers were increasingly using a short-term-memory-first retrieval strategy as cycles continued.

The unpublished results from a similar recognition memory study (Lehman, in press) were that the gains found over cycles of free recall did not appear when recognition was used (there was a slight decrease over cycles). This could suggest that the gain found in free recall was not due to better storage of items, but there are several other possibilities. For example, participants could store co-rehearsed items increasingly well over cycles of free recall, but such storage might not much improve recognition when recognition

judgments are based on item familiarity. In this case, interference due to storage and test of prior lists could dominate performance.

2.2.1 On the Efficiency of Test Taking Strategies

Test takers can be quite adaptable, even in the absence of explicit feedback, altering their retrieval strategies to improve overall performance. K. J. Malmberg and J. Annis (unpublished) were interested in the extent to which default test-taking strategies were efficient (see Malmberg, 2008 for a discussion of the efficiency of recognition memory). Efficiency is a critical issue for many testing situations such as standardized one-chance testing with a time deadline.

Figure 8.1 shows the results of an experiment that utilized an associative recognition procedure to test memory. Subjects studied five lists of pairs of words, each list with different words. On a given list, eight word pairs were studied one, two, or six times with random spacing. The associative recognition test involved discriminating pairs of words that were studied together (intact pairs) from pairs of words that were studied but not studied together (rearranged pairs). Subjects were asked to respond "old" to intact pairs and "new" to rearranged pairs. The task is difficult because items comprising intact and rearranged pairs were in fact studied, and therefore "familiar," but only the intact pairs should be endorsed.

We were interested in the efficiency of task performance over all test trials for a given list. Efficiency is a joint function of the speed and accuracy with which the test trials are completed. The efficient test taker maximizes accuracy while minimizing the amount of time allotted to the task (see Malmberg, 2008 for a discussion of the interaction of subjective goals and efficiency). In one condition, subjects performed the testing at their own pace, as in most laboratory experiments. In another condition, subjects were given a 36-s deadline to complete the same 24 test trials. Testing with a time limit is typical of the way that most exams are given in educational or standardized testing settings. The 36-s deadline was deemed to be sufficiently challenging based on the reaction time observations of dozens of subjects in prior experiments (e.g., Malmberg & Xu, 2007; Xu & Malmberg, 2007).

Figure 8.1 shows the results, panels (A), (B), (C), and (D) showing data averaged across the five lists. Panel (A) plots hit rates and false alarm rates, neither of which were significantly different in self-paced versus deadline conditions. Panel (B) shows that self-paced subjects took longer to complete the testing. This suggests that under conditions commonly found in the laboratory subjects tend to perform this task relatively inefficiently. Panel (C)

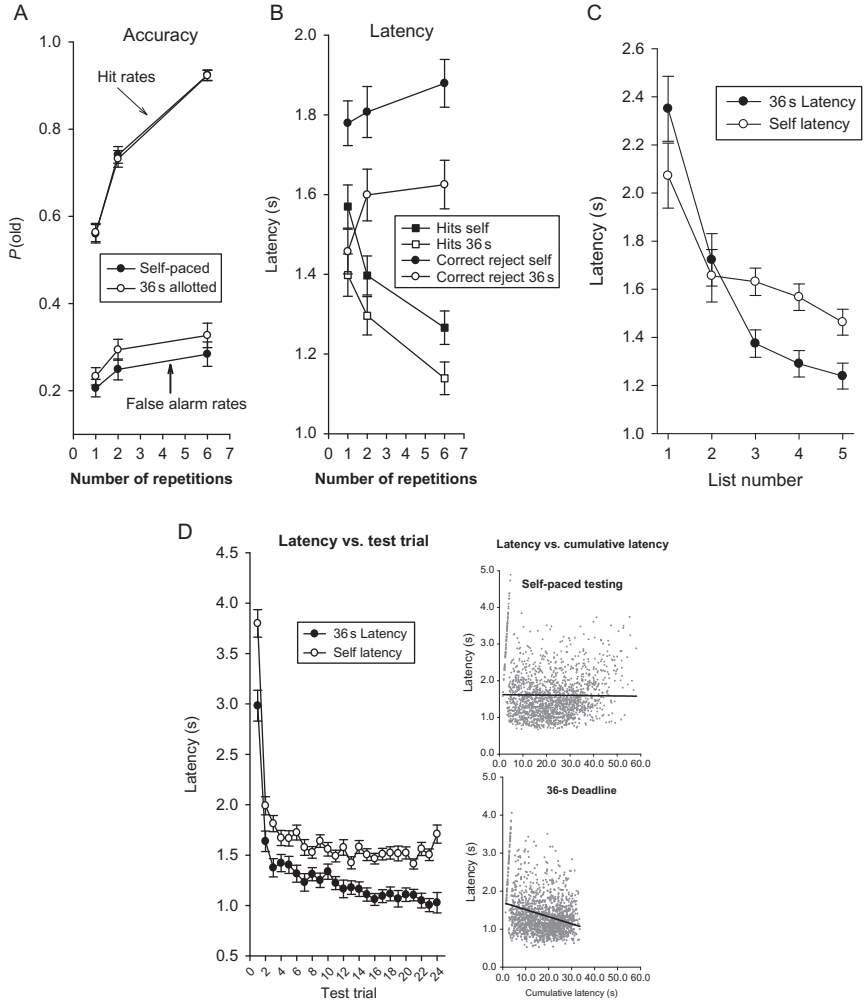

Figure 8.1 The effects of a deadline on associative recognition performance.

shows the average response latency for each list. After two study–test cycles, subjects who were given a deadline began to perform the task more quickly. Panel (D) shows that the gain in efficiency was achieved over all 24 test trials in a given cycle; subjects did not simply begin to respond more quickly as the deadline neared. These results indicate the typical subject with a deadline learns to perform the memory tests more efficiently than the typical subject without a deadline, but practice performing the memory tasks is required in order to do so.

2.2.2 Memory Testing Affects Metacognitive Judgments

One factor by which experience leads participants to change their learning and retrieval strategies involves metacognitive judgments. In one experiment, K. J. Malmberg and T. O. Nelson (unpublished) asked subjects to classify words as being relatively "easy" to remember or relatively difficult to remember (an ease-of-learning judgment or EOL). The words varied in the frequency with which they are used in natural language. On average, common words (high frequency or HF) were judged to be easier to remember than rare words (low frequency or LF) about 64% of the time [$t(57) = 4.24$]. Table 8.1 shows that when given the opportunity to study these words, along with words that were not given EOL judgments, subjects tended to study the words judged as relatively difficult to remember longer, [$F(1,57) = 4.31$, MSE = 1.34], but there was little difference in the amount of time subjects allocated to studying high common and rare words [$F < 1$].

Figure 8.2 shows that when memory was subsequently tested using yes–no recognition procedure, hit rates were greater for rare words than for common words, but the patterns of false alarm rates were dependent on the EOL judgment given. The outcome of EOL judgments (i.e., easy vs. difficult) reliably affected both hit rates and false alarm rates [$F(1,57) = 42.30$, MSE = 0.26]. The interaction of word frequency and EOL judgment was not significant [$F(1,49) = 0.22$]. By contrast, the simple effect of normative word frequency on the false alarm rate was significant only for words judged difficult to learn [$t(57) = 3.31$] and not for words judged easy to learn [$t(57) = 0.29$], and the interaction between word frequency and EOL judgment was significant [$F(1,57) = 4.95$, MSE = 0.13]. Thus, although rare words were better recognized than common words, only those words that were judged to be relatively difficult to learn produced the mirror-patterned word-frequency effect, and amount of time spent studying a word was predicted by the metacognitive judgment assessing it as easy to be learned.

Table 8.1 Amount of Self-Paced Study Time (s) Allocated to Words

	Normative Word Frequency		
EOL	**LF**	**HF**	$\overline{X}$
"Easy"	2.80 (0.20)	2.90 (0.23)	2.85 (0.22)
"Difficult"	3.05 (0.20)	2.95 (0.22)	3.00 (0.21)
None	2.93 (0.21)	2.97 (0.20)	2.94 (0.21)
$\overline{X}$	2.91 (0.20)	2.94 (0.22)	

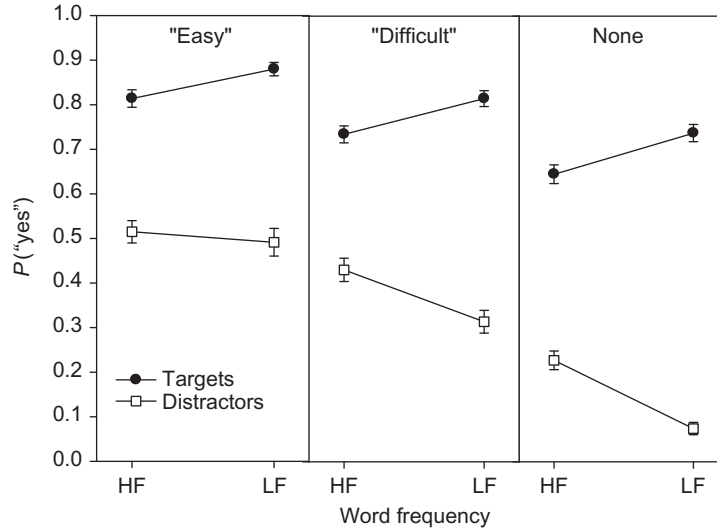

Figure 8.2 The relationship between ease-of-learning judgments given and normative word frequency.

Importantly, Benjamin (2003) reported that LF words were judged easier to learn subsequent to recognition testing, suggesting the possibility that recognition testing could be used to improve the allocation of study time.

In summary, there appear to be three classes of processes by which memory testing can alter performance on subsequent tests: (a) storing new memory traces during testing; (b) enhancing, modifying, or updating existing memory traces; and (c) altering learning and retrieval strategies.

2.2.3 What Is Stored During Memory Testing?

When an item is tested, it is important to know whether a new trace will be stored, or whether a prior trace for the tested item will be updated or modified. This issue has been raised to explain "differentiation" in the articles by Ratcliff, Clark, and Shiffrin (1990) and Shiffrin, Ratcliff, and Clark (1990) and modeled thereafter (e.g., Shiffrin & Steyvers, 1997, 1998). Those articles dealt with repeated study events rather than testing and argued that extra item storage in an existing trace decreased similarity to other items, thereby improving performance for other items, whereas storage of a new trace added noise and decreased performance for other items. Their list-strength results suggested that repeated study events often resulted in adding information to existing traces. The same issues arise when one considers storage of test events, though this has not yet been the subject of empirical research.

Considering the cases when testing an item does add information to its existing episodic trace, the kinds of information added may be information about the item itself (e.g., its spelling or meaning), information linking the item to others, or information linking the item to context. The ways that this added information determines later performance will depend on the relative proportions of storage of each of these types, the ways the information is used in a given task, and obviously on whether the subsequent testing is of the item itself (in which case storage during testing is generally beneficial) or of other items (in which case storage during testing will be generally harmful).

It is also possible that storage during testing will not just add information but modify existing information, especially if the information stored initially was stored inaccurately (e.g., initial storage of a word may have involved a misspelling, but upon testing, the spelling in the existing trace is corrected). Related to modification of existing traces is a recent literature on "reconsolidation," suggesting that reactivation of an existing recent trace places it in a malleable state that then allows it to be stored again in a different form (Lattal & Wood, 2013; Schiller & Phelps, 2011; see Rodriguez-Ortiz & Bermúdez-Rattoni, 2007 for a review): Injection of neuroinhibitors (in nonhumans) after reminders call a trace to mind can cause the initially stored memory to be lost. Here, testing has negative consequences, but due to the chemical treatment following retrieval rather than the retrieval itself. The mechanisms by which the memories are harmed are not entirely clear. It is possible, for example, that the treatments add a great deal of context noise to the existing memory. It is quite reasonable and consistent with prior research that the retrieval of an item's memory trace may add test context to existing context and existing item information. This should have especially significant benefits when the criterion test involves free recall and/or source memory since these are tasks that depend heavily on the use of context as retrieval cues (Lehman & Malmberg, 2013; Malmberg & Shiffrin, 2005). Tasks like recognition and cued recall may be less affected because item information is a more important component of the retrieval probe.

3. COSTS OF TESTING MEMORY: OUTPUT INTERFERENCE

Testing typically provides benefits to the items that are tested, but typically has costs to memory for items that are not tested, and sometimes costs to items tested but not retrieved. Such costs are mostly due to storage of

information associated with the test and thus are similar for study events, test events, or the situations in which both occur. The costs of testing memory were very important to researchers during the verbal-learning heyday (Crowder, 1976). To take just one example of many findings and studies, Brown (1958) and Peterson and Peterson (1959) showed that forgetting of a small list of items increases with the duration of a retention interval filled with another task or activity (such as arithmetic). This intervening activity has two important effects: it suppresses rehearsal and adds to memory the intervening material. Keppel and Underwood (1962) proposed that forgetting is directly related to the amount of testing (and studying) that occurs between study and test of a given item. In these paradigms, the deleterious effects of testing are evidenced by loss of what is usually termed short-term memory. The nature, time course, and cause of short-term memory loss are issues being studied by scientists today, but cannot be reviewed in this chapter.

Of course, interference effects are legion in long-term memory, as seen in over a century's worth of list learning studies. These costs associated with memory testing and studying of some items upon memory for other items are typically attributed to the storage of new traces and/or to the strengthening of existing traces and most often explained by competition. The idea is that a probe of memory tends to produce activation of similar traces: New material and/or new traces are stored during testing. At least some of the resultant traces tend to have increased similarity to the item that is the target of a later test. Increased activation of those more similar traces decreases activation of the target trace because retrieval from memory is a competitive process.

The forgetting caused by memory storage and retrieval has been traditionally termed *interference*; when the focus is on testing, the term *output interference* has been used. Researchers in recent years have also termed such interference *retrieval-induced forgetting*. This term is neutral concerning the causes of forgetting, but researchers using this term have tended to favor an explanation involving suppression or inhibition of traces that would otherwise interfere. One form of evidence comes from studies using cued recall (e.g., Anderson, Bjork, & Bjork, 1994, 2000): Lists of pairs of words are studied, each with a common cue, such as a category label (FURNITURE), and the targets of memory testing are exemplars of that category (CHAIR, TABLE, BED, etc.). Following the study list, practice is given for the retrieval of some of the targets by presenting the cue that specifies the target of retrieval (FURNITURE-B_). Such practice is

intended to produce inhibition of activation of the traces of the category members whose first letter is not B. When memory for all category members is then tested with the category cue, the practiced items are better recalled and the unpracticed items are worse recalled. A variety of conditions and controls are used in an attempt to demonstrate that such findings are due to active suppression of unwanted memories rather than competition from the practiced items. However, this issue is highly contentious because essentially all the results can be explained by some form of competition (e.g., Raaijmakers & Jakab, 2013a, 2013b). At the time of this writing, it appears that the greatest portion of the costs associated with retrieving other information is due to competition but the degree to which active suppression occurs as well is not yet known. Whatever the causes, it is accepted that the costs of testing are due to storage of the tested traces, whether it acts directly by increasing strength of the tested items or indirectly by causing suppression of the traces that would otherwise interfere.

It is of course the case that costs and benefits associated with storage and retrieval are found in many paradigms (Lehman & Malmberg, 2009). An interesting phenomenon in this regard is that known as "part-list cuing" (e.g., Slamecka, 1968, 1969). After study of a list of items, some of the list items are provided at test, supposedly as cues that might help free recall of the remaining items, but in fact, free recall is harmed rather than helped. Because it is well known that associations between items occur during list study, intuition suggests that the part-list cues will help free recall by fostering use of associative links. Raaijmakers and Shiffrin (1980) showed this intuition to be incorrect: Their SAM model did store associations and did make use of these during free recall, but nonetheless reliably predicted the observed results regardless of the choices of parameters for the model. The reasons are complex, but a major factor is the fact that associated items tend to fall into different groups. If it could be arranged that one item were provided from each group, then recall would indeed be helped, as often observed in categorized recall (e.g., see Raaijmakers & Shiffrin, 1981). However, the random choice of part-list cues too often results in cues that fall within a single group, thereby harming free recall.

Although it is generally the case that storage and retrieval of items other than the target of a later test harm performance, there is at least one important exception, known as the list-strength effect (Ratcliff et al., 1990; Shiffrin et al., 1990). Strengthening some list items during study helps rather than harms recognition of other list items. The theoretical account is based on *differentiation*: Making a trace stronger and more accurate decreases its

similarity to a later test probe of a different item. This account requires that additional study causes a strengthened trace (most of the time) rather than storage of an additional trace. It is interesting to note that testing of some list items harms rather than helps subsequent recognition testing of other items (described in more detail below), suggesting that testing after list study causes storage of a separate test trace rather than a combined single trace, though it is perfectly possible and perhaps likely that the separate test trace will contain information about both study and test for the tested item.

The accounts above of differentiation and storage of additional information in the same trace or different traces leave out a critical factor that has been important in all modeling of the empirical phenomena: The difference between context cues and content cues. In short, context cues are common to all list items, so additional storage of context makes all traces more similar, in contrast to storage of content which makes different traces less similar. Thus, the observed effects are a balance of these two opposing factors. A critical link in understanding how this works was found by Malmberg and Shiffrin (2005) who obtained evidence that repeated massed study produces just "one shot of context," but increasing storage of content. This idea has not been explored in testing.

As mentioned briefly above, output interference is observed not only in free and cued recall but also in recognition. In fact, output interference in recognition memory testing is particularly important as a constraint on theory. Increases in the number of items studied on a list generally result in only small impairments, at least when one uses controls to reduce the impact of such factors as serial position effects and lag effects. Dennis and Humphreys (2001) pointed this out and suggested that the primary (perhaps only) cause of forgetting of words in recognition was *context noise*, interference caused not by activation of similar list items, but rather caused by activation of traces of the test item itself that were stored prior to the list study. Context noise surely contributes to recognition difficulty, but the relative amount of interference due to *item noise* (e.g., from similar items on the list) and due to context noise remains in dispute (Criss & Shiffrin, 2004; Criss et al., 2011; Dennis & Humphreys, 2001).

It is possible that small list length effects in recognition are in part the result of context changes from study to test: If the probe cue uses context significantly different from the context stored during list study, then the similarity of a test probe to stored traces of other words would be quite low because the traces would be dissimilar in both content and context. If so, only the trace of the test word, if it exists, would tend to be activated.

However, this hypothesis implies that storage of the test traces themselves would be quite similar to subsequent tests. If so, test traces would tend to be activated (depending also on content similarity) and would cause interference for subsequent tests. Such interference has been found in many studies and can be quite strong. Recognition accuracy has been found to decrease over tests in forced choice testing (Criss et al., 2011; Murdock & Anderson, 1975) and in old–new testing (Criss et al., 2011). The natural interpretation is interference caused by storage of test traces that are similar enough in context and content to be activated when a different item is subsequently tested.

The role of item interference from the storage of item information is further revealed by experiments that demonstrate how output interference may be reduced. Wickens, Born, and Allen (1963; see also Wickens, 1970) reported the results of several experiments demonstrating what Watkins and Watkins (1975) referred to as the release from proactive interference or *release from PI*. In a typical study, trigrams of stimuli from some category are studied and followed by a period of distraction activity that serves to empty short-term memory. The test performance then measures retrieval from long-term memory. Different trials use different stimuli. If successive trials use stimuli from the same category, then performance drops, but a switch to a new category causes performance to revert to the initially higher level. The typical interpretation is that there is increasing competition as traces of items in the same category accumulate, competition that abates when the switch to a new category occurs. In a recent experiment, we found a form of release from PI within successive recognition tests following list study. The study list contained items from several categories, presented in random order. The sequence of forced choice tests was blocked by category. Performance dropped over tests within category but reverted to the initial level when testing of a new category began (Malmberg, Criss, Gangwani, & Shiffrin, 2012).

4. THE INFLUENCE OF ONE TEST ON THE NEXT: SEQUENTIAL DEPENDENCIES

Responses on successive test trials often interact, for many different reasons. In recall, for example, what is recalled on one trial may be used as a probe or source of additional information on the next trial. In recognition, the response (or stimulus) on trial n predicts the nature and speed of the response on trial $n+j$, even when items tested were not in fact studied

(Kachergis, Cox, & Shiffrin, 2013; Malmberg & Annis, 2012; Ratcliff & McKoon, 1978). For instance, when recognition is tested using a yes–no procedure and the order of target and foil test trials is determined randomly, a "yes" response is more to follow a "yes" response than a "no" response, and "yes" responses are made more quickly when they follow a "yes" response than when they follow a "no" response. These correlations are positive and known as *assimilation*. In other instances, the current response is negatively correlated with prior responses (or stimuli) and this is known as *contrast*. For instance, when recognition memory is tested using a judgment of frequency (JOF) procedure, the JOF on trial $n+1$ is often greater if the prior stimulus was an infrequently studied item than it was a frequently studied item (J. Annis & K. J. Malmberg, in press; Malmberg & Annis, 2012). In contrast to the decline in accuracy associated with output interference, overall accuracy of yes–no recognition is unaffected since the magnitude of the assimilation is similar for both target and foil test trials (Malmberg & Annis, 2012). In this sense, the decreases in accuracy with increases in testing and sequential dependencies are distinct phenomena that require separate accounts.

Taken together, assimilation and contrast comprise a set of *sequential dependencies*, and several recent studies have documented them in recognition memory testing. The cogent reader may not be surprised by this finding since sequential dependencies have long been known to exist in sequences of perceptual decision trials (Collier, 1954a, 1954b; Collier & Verplanck, 1958; Verplanck, Collier, & Cotton, 1952). Indeed, the sequential dependencies found in a series of absolute identification trials were a basis for linking memory and perception in Miller (1956). However, the patterns of sequential dependencies observed in memory testing can be quite different than those observed in perceptual testing, like absolute identification, and organizing these observations in a theoretical framework may go a long way in developing a better general understanding of cognition, as well as the individual systems that support it.

A key question, borrowed from the perception literature, concerns whether assimilation is the result of shifts in response bias and/or shifts in the nature of the information on which the recognition decision is made (Lockhead, 2004; Treisman & Williams, 1984). According to the former hypothesis, response biases are based on the prior probabilities of target versus foil test trials. If one is in midst of a long series of mostly target trials and bias reflects recent experience, then one should be more biased to respond "yes" than if one is in the midst of a long series of foil trials. However,

sequential dependencies are found in recognition testing in the absence of direct knowledge about the nature of the memory tests, and therefore prior responses must be the basis for determining the bias. According to the later hypothesis, assimilation is the result of a positive correlation between the sources of information on which consecutive decisions are made. We refer to this as carryover, and we have developed a model of carryover that shares some assumptions of compound cue models (Annis & Malmberg, 2013; cf. Ratcliff & McKoon, 1988). Nevertheless, it is quite possible that sequential dependencies arise from both transient fluctuations in response bias and carryover.

To tease apart the contributions of bias and carryover, one must make some assumptions. When interpreting the data from several recent experiments, Malmberg and Annis worked within a signal detection framework by assuming that response bias is influenced only by the prior probabilities of the classes of stimuli and the costs and rewards associated with various outcomes; so long as the evidence on which the decision is made is a continuous random variable, what is represented by the information used to make the decision is not important (Green & Swets, 1966). In other words, sequential dependencies produced by bias should be similar regardless of whether memory or perception is being tested (Malmberg & Annis, 2012).

In one experiment in which recognition memory and absolute identification were directly compared, subjects were presented with words that varied at six levels on two dimensions: the frequency with which they were encountered and the font size in which the words were displayed. During this phase of the experiment, subjects performed absolute identification based on the font size of the words, and assimilation was observed, with contrast observed at lags >1. Following the study list, recognition memory for the words was tested with the JOF procedure. Assimilation was again observed in the JOFs, but contrast was not observed at lags >1. Rather, contrast was observed in adjacent memory tests between the prior stimulus and the current response. In another experiment, feedback was manipulated. Feedback is often provided in perception experiments and is thought to influence the subject's knowledge about the prior probabilities of the stimulus classes, and contrast is only observed at lags >1 and only when feedback is provided (Holland & Lockhead, 1968). For JOFs, on the other hand, contrast occurs in the absence but not in the presence of feedback. As a package, this pattern of sequential dependencies observed in JOFs is different from those commonly reported for absolute identification.

Other research has observed that assimilation diminishes with increases in the ISI during absolute identification tests (Matthews & Stewart, 2009) but not during recognition testing (Malmberg & Annis, 2012). Thus, there are reports of differences in the patterns of sequential dependencies in recognition and perception testing, and therefore, they are most likely not caused by simple changes in response bias within the standard signal detection framework. It is possible that these different patterns of sequential dependencies are related to differences in memory and perception systems and/or subtle differences in the procedures used for examining them.

Modeling the sequential dependencies observed in recognition testing may help us better understand the nature of memory traces and evidence used to make memory decisions. For instance, sequential dependencies are also found in recognition memory testing when the confidence ratings procedure is used (Malmberg & Annis, 2012); a "yes" response is more likely following a highly confident "yes" response than following a low-confidence "yes" response. In addition, a "yes" response was more likely following a low-confidence "yes" than following a miss (i.e., "no" response). Hence, the amount of assimilation was correlated with the amount of evidence used to make the yes–no decision on the prior trial. This suggests that the nature of the evidence is either a continuous random variable or a discrete random variable with at least three categories: not retrieved, weakly retrieved, or strongly retrieved.

Schwartz, Howard, Jing, and Kahana (2005) examined the relationship between the order in which items were studied and the order in which items were tested. When items were studied in adjacent positions, but not distant serial positions, and tested in adjacent positions, they found that hits were more likely following a high-confidence response on the immediately prior test trial, which they stated provided evidence that the high confidence associated with first response is due to a recollection of the co-occurrence of the two items on the study list. However, there was a boost in hit rates even when items were studied in distant serial positions and therefore did not co-occur and were unlikely to be rehearsed together (Kachergis et al., 2013; Malmberg & Annis, 2012). In addition, similar boosts in false alarm rates are observed even when the first item in the test sequence was not studied (Malmberg & Annis, 2012). Since recollection can be ruled out as the cause of the increase in the tendency to recognize an unstudied item, sequential dependencies either do not depend on the recollection of an item on a prior test trial or may also reflect a combined strength of evidence obtained on adjacent test trials.

5. PAST DECISIONS INFLUENCE FUTURE DECISIONS: SHIFTS IN BIAS

On one hand, it is only sensible to acknowledge that the criterion for responding old or new may change during the course of a recognition test (cf. Treisman & Williams, 1984). However, there exists very little evidence to suggest this is so. The classic paradigm for eliciting changes in response bias involves manipulating the contents of the test list. Specifically, changing the proportion of targets on the test list typically elicits a change in the location of the criterion exactly as expected in a signal detection framework, that is, participants become more conservative as the proportion of foils increases if participants are informed of the relative proportion of targets (or foils) on the test list (e.g., Criss, 2009, 2010). When this information is withheld, there is little to no evidence of a change in response bias (Healy & Kubovy, 1978). In a striking example, Cox and Dobbins (2011) presented target-free or distractor-free lists for recognition decisions without informing participants, but the tendency to respond "old" was nearly identical for a target-free and distractor-free list as for standard lists containing half targets and half foils (see also G. Koop, A. H. Criss, & K. J. Malmberg, in preparation).

Although subjects in typical recognition experiments do not seem very sensitive to the proportion targets and foils test trials, feedback appears to be a critical factor that modulates changes in response bias during the course of recognition testing (Estes & Maddox, 1995). In fact, one experiment conducted by K. J. Malmberg and J. Xu (unpublished) found that providing feedback concerning the proportion of correct responses only after all memory testing was complete influenced response bias on subsequent lists using a continuous recognition procedure. Since feedback is almost never provided in memory testing experiments, the upshot is that there is little evidence indicating systematic shifts in response bias occur.

Another common scenario where the criterion is assumed to change during the course of testing is when the expected memorability of the test item varies. For example, early theories of the word-frequency mirror effect assumed that participants adjusted the amount of evidence required to endorse an item as studied depending on the expected accuracy for each class of items (e.g., Glanzer & Adams, 1990). Low-frequency targets are easy to remember, and therefore, a higher level of evidence is required to endorse a LF word, reducing the false alarm rate. The WFE is now attributed to

stimulus attributes such as feature frequency (Malmberg, Steyvers, Stephens, & Shiffrin, 2002; Shiffrin & Steyvers, 1997); however, the idea that expected familiarity of the test stimuli is a driving force behind the criterion placement remains (e.g., Hirshman, 1995).

Direct attempts to elicit changes in criterion on this basis in a trial-by-trial manner have not been successful. For example, in one experiment, Stretch and Wixted (1998) had participants study some items one time and other items five times. They emphasized the differences in expected memory by color-coding items such that strongly encoded targets and a subset of foils were presented in red font, and weakly encoded targets and the remaining foils were presented in green font even when fully informed about the experimental design. There are many similar examples, showing that participants do not change their criteria during the test list in response to changes in difficulty. The one apparent exception seems to be when the difficulty of the test is altered, not by changing the nature of the targets but by changing the nature of the foils.

Brown, Steyvers, and Hemmer (2007) changed the nature of target items in a blocked fashion going from either easy (randomly chosen) or difficult (mirror image of studied items) foils. Not surprisingly, the false alarm rate was substantially higher for the difficult foils. The critical finding was that the hit rate also changed with block—increasing when the foils became easier and decreasing when foils become more difficult to reject (see Benjamin & Bawa, 2004 for similar data). Brown et al. interpreted this as strong evidence of a change in criterion during the test. An alternative explanation comes from Turner, Van Zandt, and Brown (2011) who describe a model where stimulus representations develop over the course of the experiment and these changes in stimulus representation result in data that are typically interpreted as a criterion change. In other words, even cases that appear to result from changes in the criteria may in fact result from changes in the distributions of memory evidence. The Turner et al. model provides important insight about the role of feedback in memory, which we will return to. However, the model is a signal detection model and has nothing to say about the encoding and retrieval processes that underlie memory or the processes that resulting updated representations. Within the REM framework, representations are updated during test by updating the best matching memory trace if the test item is judged to be old and storing a new trace if the test item is judged to be new (Criss et al., 2011).

A critical feature of the Turner et al. model is that feedback (externally provided by the experimenter or internally generated from a subject's own response when feedback is not provided) is used to establish accurate

representations. Interestingly, providing feedback during recognition is the single manipulation that produces compelling criterion shifts, at least under some conditions. Han and Dobbins (2008, 2009) provided accurate feedback for correct responses and biased feedback for incorrect responses. In one condition, they provided feedback indicating that all false alarms were actually correct, and in another, they provided feedback indicating that all misses were actually correct. Performance showed that participants did change their criterion in response to this biased feedback. These studies used standard recognition lists with half targets and half foils. In contrast, Koop et al. used distractor-free and target-free lists. Recall that when no feedback is provided, performance in these "pure" lists is identical to lists with half targets and half foils. In all cases, presenting feedback causes the criterion setting to be closer to optimal. That is, for pure lists, the presence of feedback causes the probability of calling a test item old to move toward one for the distractor-free list and zero for the target-free list.

6. CONCLUSIONS

Recent interest in the positive consequences of memory testing has spawned a number of systematic empirical investigations of the circumstances in which one would expect to observe them. Here, we presented some of these results in a broad context that reflects the variety of influences of memory testing on subsequent testing. These consequences are sometimes positive, but often they are negative, depending on the manner in which memory is tested. A comprehensive understanding will explain both types of outcomes within a framework that views memories as the outcome of a continuous parallel process of encoding and retrieval. The extant literature suggests to us that such an account will also necessarily require both a description of the nature of the information encoded during memory testing and a description of the control processes invoked to carry out the testing. In addition, a comprehensive account of the consequences of memory testing will describe how prior tests of memory affect future decisions one makes.

REFERENCES

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 1063–1087.

Anderson, M. C., Bjork, E. L., & Bjork, R. A. (2000). Retrieval-induced forgetting: Evidence for a recall-specific mechanism. *Psychonomic Bulletin & Review, 7*, 522–530.

Annis, J., & Malmberg, K. J. (2013). A model of positive sequential dependencies in judgments of frequency. *Journal of Mathematical Psychology*, *57*, 225–236.
Benjamin, A. S. (2003). Predicting and postdicting the effects of word frequency on memory. *Memory & Cognition*, *31*, 297–305.
Benjamin, A. S., & Bawa, S. (2004). Distractor plausibility and criterion placement in recognition. *Journal of Memory and Language*, *51*, 159–172.
Brewer, G. A., Marsh, R. L., Meeks, J. T., Clark-Foos, A., & Hicks, J. L. (2010). The effects of free recall testing on subsequent source memory. *Memory*, *18*, 385–393.
Brown, J. (1958). Some tests of the decay theory of immediate memory. *Quarterly Journal of Experimental Psychology*, *10*, 12–21.
Brown, S. D., Steyvers, M., & Hemmer, P. (2007). Modeling experimentally induced strategy shifts. *Psychological Science*, *18*, 40–45.
Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition*, *34*, 268–276.
Chan, J. C. K., & McDermott, K. B. (2007). The testing effect in recognition memory: A dual process account. *Journal of Experimental Psychology. Learning Memory and Cognition*, *33*, 431–437.
Collier, G. (1954a). Intertrial association at the visual threshold as a function of intertribal Interval. *Journal of Experimental Psychology*, *48*, 330–334.
Collier, G. (1954b). Probability of response and intertrial association as functions of monocular and binocular stimulation. *Journal of Experimental Psychology*, *47*, 75–83.
Collier, G., & Verplanck, W. S. (1958). Nonindependence of successive responses at threshold as a function of interpolated stimuli. *Journal of Experimental Psychology*, *55*, 429–437.
Cox, J. C., & Dobbins, I. G. (2011). The striking similarities between standard, distractor-free, and target-free recognition. *Memory and Cognition*, *39*, 925–940.
Criss, A. H., & Shiffrin, R. M. (2004). Context noise and item noise jointly determine recognition memory: A comment on Dennis & Humphreys (2001). *Psychological Review*, *111*(3), 800–807.
Criss, A. H. (2009). The distribution of subjective memory strength: List strength and response bias. *Cognitive Psychology*, *59*, 297–319.
Criss, A. H. (2010). Differentiation and response bias in episodic memory: Evidence from reaction time distributions. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *36*, 484–499.
Criss, A. H., Malmberg, K. J., & Shiffrin, R. M. (2011). Output interference in recognition memory. *Journal of Memory and Language*, *64*, 316–326.
Crowder, R. G. (1976). *Principles of learning and memory*. Hillsdale, NJ: Erlbaum.
Darley, C. F., & Murdock, B. B. (1971). Effects of prior free recall testing on final recall and recognition. *Journal of Experimental Psychology*, *91*, 66–73.
Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic recognition memory. *Psychological Review*, *108*, 452–478.
Dunn, J. A. (2004). RK: A matter of confidence. *Psychological Review*, *111*, 524–542.
Ebbinghaus, H. (1885). *Memory: A contribution to experimental psychology*. New York, NY: Teachers College, Columbia University.
Estes, W. K., & Maddox, W. T. (1995). Interactions of stimulus attributes, base rates, and feedback in recognition. *Journal of Experimental Psychology. Learning Memory and Cognition*, *21*, 1075–1095.
Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, *91*, 1–67.
Glanzer, M., & Adams, J. K. (1990). The mirror effect in recognition memory: Data and theory. *Journal of Experimental Psychology. Learning Memory and Cognition*, *16*, 5–16.

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology*, *81*, 392–399.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York, NY: Wiley.

Han, S., & Dobbins, I. G. (2008). Examining recognition criterion rigidity during testing using a biased-feedback technique: Evidence for adaptive criterion learning. *Memory and Cognition*, *36*, 703–715.

Han, S., & Dobbins, I. G. (2009). Regulating recognition decisions through incremental reinforcement learning. *Psychonomic Bulletin & Review*, *16*, 469–474.

Healy, A. F., & Kubovy, M. (1978). Probability matching and the formation of conservative decision rules in a numerical analog of signal detection. *Journal of Experimental Psychology. Human Learning and Memory*, *7*, 344–354.

Hirshman, E. (1995). Decision processes in recognition memory: Criterion shifts and the list strength paradigm. *Journal of Experimental Psychology. Learning Memory and Cognition*, *21*, 302–313.

Holland, M. K., & Lockhead, G. R. (1968). Sequential effects in absolute judgments of loudness. *Perception & Psychophysics*, *3*, 409–414.

Izawa, C. (1970). Optimal potentiating effects and forgetting-prevention effects of tests in paired-associate learning. *Journal of Experimental Psychology*, *83*(2, Pt.1), 340–344.

Izawa, C. (1971). The test trial potentiating model. *Journal of Mathematical Psychology*, *8*(2), 200–224.

Jakab, E., & Raaijmakers, J. G. W. (2009). The role of item strength in retrieval-induced forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 607–617.

Jones, T. C., & Roediger, H. L. (1995). The experiential basis of serial position effects. *European Journal of Cognitive Psychology*, *7*, 65–80.

Kachergis, G., Cox, G. E., & Shiffrin, R. M. (2013). The effects of repeated sequential context on recognition memory. In *Proceedings of the 35th annual conference of the cognitive science society*.

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: An episodic context account. In B. H. Ross (Ed.), *The Psychology of Learning and Motivation*: *Vol. 61*. (pp. 237–284). San Diego, CA: Elsevier Academic Press.

Keppel, G., & Underwood, B. J. (1962). Proactive inhibition in short term retention of single items. *Journal of Verbal Learning and Verbal Behavior*, *1*, 153–161.

Koriat, A., & Bjork, R. A. (2005). Illusions of competence in monitoring one's knowledge during study. *Journal of Experimental Psychology. Learning Memory and Cognition*, *31*, 187–194.

Koriat, A., & Bjork, R. A. (2006). Illusions of competence during study can be remedied by manipulations that enhance learners' sensitivity to retrieval conditions at test. *Memory & Cognition*, *34*, 959–972.

Lattal, K. M., & Wood, M. A. (2013). Epigenetics and persistent memory: Implications for reconsolidation and silent extinction beyond the zero. *Nature Neuroscience*, *16*, 124–129.

Lehman, M., & Malmberg, K. J. (2009). A global theory of remembering and forgetting from multiple lists. *Journal of Experimental Psychology. Learning Memory and Cognition*, *35*, 970–988.

Lehman, M., Smith, M. A., & Karpicke, J. D. (in press). Toward an episodic context account of retrieval-based learning: Dissociating retrieval practice and elaboration. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Lehman, M., & Malmberg, K. J. (2013). A buffer model of encoding and temporal correlations in retrieval. *Psychological Review*, *120*(1), 155–189.

Lockhead, G. R. (2004). Absolute judgments are relative: A reinterpretation of some psychophysical ideas. *Review of General Psychology*, *8*, 265–272.
Malmberg, K. J., & Annis, J. (2012). On the relationship between memory and perception: Sequential dependencies in recognition testing. *Journal of Experimental Psychology. General*, *141*(2), 233–259.
Malmberg, K. J., Criss, A. H., Gangwani, T. H., & Shiffrin, R. M. (2012). Overcoming the negative consequences of interference that results from recognition memory testing. *Psychological Science*, *23*(2), 115–119.
Malmberg, K. J., Holden, J. E., & Shiffrin, R. M. (2004). Modeling the effects of repetitions, similarity, and normative word frequency on judgments of frequency and recognition memory. *Journal of Experimental Psychology. Learning Memory and Cognition*, *30*, 319–331.
Malmberg, K. J., & Shiffrin, R. M. (2005). The "one-shot" hypothesis for context storage. *Journal of Experimental Psychology. Learning Memory and Cognition*, *31*, 322–336.
Malmberg, K. J., Steyvers, M., Stephens, J. D., & Shiffrin, R. M. (2002). Feature-frequency effects in recognition memory. *Memory & Cognition*, *30*(4), 607–613.
Malmberg, K. J., & Xu, J. (2007). On the flexibility and on the fallibility of associative memory. *Memory & Cognition*, *35*(3), 545–556.
Malmberg, K. J. (2008). Recognition Memory: A Review of the Critical Findings and an Integrated Theory for Relating Them. *Cognitive Psychology*, *57*, 335–384.
Malmberg, K. J., Zeelenberg, R., & Shiffrin, R. M. (2004). Turning up the noise or turning down the volume? on the nature of the impairment of episodic recognition memory by midazolam. *Journal of Experimental Psychology. Learning Memory and Cognition*, *30*(2), 540–549.
Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review*, *87*, 252–271.
Matthews, W. J., & Stewart, N. (2009). The effect of inter-stimulus interval on sequential effects in absolute identification. *The Quarterly Journal of Experimental Psychology*, *62*, 2014–2029.
McGeoch, J. A. (1942). *The psychology of human learning*. New York: Longmans, Green, and Co.
Mensink, G. J. M., & Raaijmakers, J. G. W. (1989). A model for contextual fluctuation. *Journal of Mathematical Psychology*, *33*, 172–186.
Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*, 81–97.
Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*, 519–533.
Murdock, B. B., & Anderson, R. E. (1975). Encoding, storage and retrieval of item information. In R. L. Solso (Ed.), *Theories in cognitive psychology: The Loyola symposium* (pp. 145–194). Hillsdale, NJ: Erlbaum.
Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect". *Psychological Science*, *2*, 267–270.
Nelson, T. O., & Leonesio, R. J. (1988). Allocation of self-paced study time and the "labor-in-vain effect". *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *14*, 676–686.
Nelson, A. B., & Shiffrin, R. M. (2013). The co-evolution of knowledge and event memory. *Psychological Review*, *120*(2), 356–394.
Palmeri, T. J., & Flanery, M. A. (2002). Memory systems and perceptual categorization. In B. Ross (Ed.), *The psychology of learning and motivation: Vol. 41.* (pp. 141–189). USA: Elsevier.
Peterson, L. R., & Peterson, M. J. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology*, *58*, 193–198.

Raaijmakers, J. G. W., & Jakab, E. (2013a). Rethinking inhibition theory: On the problematic status of the inhibition theory for forgetting. *Journal of Memory and Language*, *68*, 98–122.
Raaijmakers, J. G. W., & Jakab, E. (2013b). Is forgetting caused by inhibition? *Current Directions in Psychological Science*, *22*, 205–209.
Raaijmakers, J. G. W., & Shiffrin, R. M. (1980). SAM: A theory of probabilistic search in associative memory. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory*: *Vol. 14*. (pp. 207–262). New York, NY: Academic Press.
Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review*, *88*, 93–134.
Ratcliff, R., Clark, S. E., & Shiffrin, R. M. (1990). List-strength effect: I. Data and discussion. *Journal of Experimental Psychology. Learning Memory and Cognition*, *16*, 163–178.
Ratcliff, R., & McKoon, G. (1978). Priming in item recognition, evidence for propositional structure of sentences. *Journal of Verbal Learning and Verbal Behavior*, *17*, 403–417.
Ratcliff, R., & McKoon, G. (1988). A retrieval theory of priming in memory. *Psychological Review*, *95*, 385–408.
Rodriguez-Ortiz, C. J., & Bermúdez-Rattoni, F. (2007). Memory reconsolidation or updating consolidation? In F. Bermúdez-Rattoni (Ed.), *Neural plasticity and memory: From genes to brain imaging*. Boca Raton, FL: CRC Press, (Chapter 11).
Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*, 181–210.
Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology. Learning Memory and Cognition*, *21*, 803–814.
Schiller, D., & Phelps, E. A. (2011). Does reconsolidation occur in humans? *Frontiers in Behavioral Neuroscience*, *5*(24), 1–12.
Schwartz, G., Howard, M. W., Jing, B., & Kahana, M. J. (2005). Shadows of the past: Temporal retrieval effects in recognition memory. *Psychological Science*, *16*, 898–904.
Shiffrin, R. M., Ratcliff, R., & Clark, S. E. (1990). List-strength effect: II. Theoretical mechanisms. *Journal of Experimental Psychology. Learning Memory and Cognition*, *16*, 179–195.
Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—Retrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*, 145–166.
Shiffrin, R. M., & Steyvers, M. (1998). The effectiveness of retrieval from memory. In M. Oaksford, & N. Chater (Eds.), *Rational models of cognition* (pp. 73–95).London: Oxford University Press.
Slamecka, N. J. (1968). An examination of trace storage in free recall. *Journal of Experimental Psychology*, *76*, 504–513.
Slamecka, N. J. (1969). Testing for associative storage in multitrial free recall. *Journal of Experimental Psychology*, *81*, 557–560.
Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *26*, 204–221.
Stretch, V., & Wixted, J. T. (1998). On the difference between strength-based and frequency-based mirror effects in recognition memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *24*, 1379–1396.
Treisman, M., & Williams, T. C. (1984). A theory of criterion setting with an application to sequential dependencies. *Psychological Review*, *91*, 68–111.
Turner, B. M., Van Zandt, T., & Brown, S. (2011). A dynamic stimulus-driven model of signal detection. *Psychological Review*, *118*, 583–613.
Verplanck, W. S., Collier, G. H., & Cotton, J. W. (1952). Nonindependence of successive responses in measurements of the visual threshold. *Journal of Experimental Psychology*, *44*, 273–282.

Watkins, O. C., & Watkins, M. J. (1975). Buildup of proactive inhibition as a cue-overload effect. *Journal of Experimental Psychology. Human Learning and Memory*, *1*(4), 442–452.
Wickens, D. D. (1970). Encoding categories of words: An empirical approach to meaning. *Psychological Review*, *77*, 1–15.
Wickens, D. D., Born, D. G., & Allen, C. K. (1963). Proactive inhibition and item similarity in short-term memory. *Journal of Verbal Learning and Verbal Behavior*, *2*, 440–445.
Xu, J., & Malmberg, K. J. (2007). Modeling the effects of verbal- and nonverbal-pair strength on associative recognition. *Memory & Cognition*, *35*(3), 526–544.