January 2010

Organizational Processes Contribute to the Testing Effect in Free Recall

Franklin Zaromb
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Department of Psychology

Dissertation Committee:
Henry L. Roediger, III, Chair
Larry L. Jacoby
Mark M. McDaniel
Kathleen B. McDermott
James V. Wertsch
R. Keith Sawyer

ORGANIZATIONAL PROCESSES CONTRIBUTE TO THE TESTING EFFECT IN FREE RECALL

by

Franklin Mendel Zaromb

A dissertation presented to the
Graduate School of Arts and Sciences
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

August 2010

Saint Louis, Missouri

ABSTRACT OF THE DISSERTATION

Organizational Processes Contribute

to the Testing Effect in Free Recall

by

Franklin Mendel Zaromb

Doctor of Philosophy

Washington University in St. Louis, 2010

Professor Henry L. Roediger, III, Chairperson

In educational contexts, tests not only assess what students know, they can also directly improve long-term retention of subject matter relative to restudying it. More importantly, the memorial advantage of testing is not limited to select information that was tested earlier. Research has shown that testing can serve as a versatile learning tool by enhancing the long-term retention of non-tested information that is conceptually related to previously tested information; stimulating the subsequent learning of new information; and permitting better transfer of learning to new knowledge domains. We further investigated the potential benefits of testing on learning by asking whether testing can also improve students' learning and retention of the conceptual organization of study materials, and if so, whether processes involved in mentally organizing information during learning contribute to the memorial advantage of testing.

In three experiments with categorized lists, we asked whether the testing effect in free recall is related to enhancements in organizational processing. In the first experiment, different groups of subjects studied a list either once or twice before a final criterial test or they studied the list once and took an initial recall test before the final test.

Prior testing enhanced total recall of words and reduced false recall of extra-list intrusions relative to restudying. In addition, testing increased the number of categories accessed, the number of items recalled from within those categories, and improved category clustering.

In two additional experiments, manipulating the organizational processing that occurred during initial study and test trials affected delayed recall and measures of output organization. Testing produced superior long-term retention when initial test conditions promoted the use of semantic relational information to guide episodic retrieval, and measures of category clustering and subjective organization were correlated with delayed recall. The results suggest that the benefit of testing in free recall learning arises, at least in part, because testing creates retrieval schemas based upon categorical knowledge and recollections of previous recall attempts that guide and facilitate episodic recall.

Acknowledgments

I would like to thank my graduate advisor and mentor Roddy Roediger for his generous support and professional guidance throughout my graduate school career. I would also like to thank the other members of my dissertation committee: Larry Jacoby, Mark McDaniel, Kathleen McDermott, Keith Sawyer, and Jim Wertsch. In addition to providing helpful suggestions at every stage of the dissertation process, they have been inspiring teachers and research mentors to me.

I have also greatly benefited from the advice and support of current and past members of the Roediger and McDermott labs, and in particular, Pooja Agarwal, Andrew Butler, Jason Chan, Lisa Geraci, Michael Goode, Bridgid Finn, Sean Kang, Jeff Karpicke, Keith Lyle, Dave McCabe, Jane McConnell, Karl Szpunar, and Yana Weinstein. Thanks are also due to Jeff Chiou and Ben Roth for assistance with data collection and scoring. Last but not least, I owe an incredible debt of gratitude to my wife Allison for her patience, understanding, and support in juggling the responsibilities of graduate school and raising four energetic little boys.

Funding for this research was provided by a Collaborative Activity grant from the James S. McDonnell Foundation (#220020041) as well as the Dean's Dissertation Fellowship from the Graduate School of Arts and Sciences at Washington University in St. Louis.

Table of Contents

List of Tables

List of Figures

List of Appendices

Organizational Processes Contribute to the Testing Effect in Free Recall

An established finding in the cognitive psychology literature is that testing a person's memory for previously learned material enhances long-term retention as compared to restudying the material for an equivalent amount of time (e.g., Carrier & Pashler, 1992; for a review see Roediger & Karpicke, 2006b). This finding, known as the testing effect, has been demonstrated using a wide range of study materials; types of tests; in both laboratory and classroom settings; as well as in different subject populations (e.g., Butler & Roediger, 2007; Gates, 1917; Kang, McDermott, & Roediger, 2007; McDaniel, Anderson, Derbish, & Morrisette, 2007; Roediger & Karpicke, 2006a; Spitzer, 1939; Tse, Balota, & Roediger, in press). Recent years have seen renewed interest among researchers investigating the potential benefits of testing as a means to improving learning in educational settings (McDaniel, Roediger, & McDermott, 2007; Pashler, Rohrer, Cepeda, & Carpenter, 2007).

One limitation with this area of research is that testing effects typically report improvements in learners' retention of discrete facts (e.g., foreign vocabulary words) without demonstrating a better understanding of the subject matter through testing (Daniel & Poole, 2009). However, a growing body of research has shown that testing can serve as a versatile learning tool by enhancing the long-term retention of non-tested information that is conceptually related to previously retrieved information (Chan, 2009; Chan, McDermott, & Roediger, 2006); by stimulating the subsequent learning of new information (Izawa, 1970; Karpicke, 2009; Szpunar, McDermott, & Roediger, 2008; Tulving & Watkins, 1974); as well as permitting better transfer to new questions (Butler, in press; Johnson & Mayer, 2009; Rohrer, Taylor, & Sholar, 2010). The present research

further examines the potential benefits of testing by asking whether testing can improve individuals' learning and retention of the conceptual organization of study materials relative to studying the materials alone, a question that has not yet been addressed in the literature.

It also remains unclear what are the underlying mechanisms that determine the presence and magnitude of testing effects. In their recent review, Roediger and Karpicke (2006b) argued that testing has direct benefits on long-term retention. The direct effect of testing is based on the notion that retrieving information from memory leads to a modification of the memory trace that renders it more resistant to forgetting, thereby enhancing the long-term retention of the retrieved information (Bjork, 1975). Indeed, several studies have corroborated the notion that processing that occurs during retrieval can account for testing effects (e.g., Glover, 1989; Karpicke & Roediger, 2008; Pyc & Rawson, 2009). A second aim of the present research was to determine whether the testing effect may be due, at least in part, to cognitive processes involved in mentally organizing information during learning.

Organization in Episodic Recall

The concept of organization is fundamental to the scientific study of human memory. Psychologists have long grappled with questions of how the processes involved in mentally organizing information influence learning and retention (e.g., Ausubel, 1963; Bartlett, 1932; Katona, 1940). One theoretical assumption that has guided much of the cognitive research examining organization and learning was Miller's (1956) conception of recoding, or chunking. Miller observed that the span of immediate memory appeared to be limited to a finite number of items, or units of information—the magical number 7

+/- 2. Thus, he argued that the key to learning and retaining large quantities of information was to mentally repackage, or chunk, the study materials into smaller units. Evidence for chunking has come primarily from studies using both serial recall and free recall paradigms in which subjects often study and attempt to recall verbal materials such as lists of words over multiple alternating study and test trials (e.g., Bower & Springston, 1970; Tulving, 1962), as well as from other techniques (e.g., Mandler, 1967).

In support of chunking, researchers have pointed to the finding that when people study lists of words coming from different conceptual categories in a randomized order, they tend to recall them in an organized fashion by clustering conceptually-related responses together (Bousfield, 1953; Bousfield, Whitmarsh, & Cohen, 1958). Further, response clustering is often associated with greater retention (Mulligan, 2005; Puff, 1979). Similarly, Tulving (1962) found that when asked to learn a list of seemingly unrelated words, individuals tend to recode groups of items into higher-order subjective units, and that this organizing tendency, which is referred to as subjective organization, is predictive of free recall. Subjective organization is presumed to be reflected in the degree to which recall protocols become more consistent over multiple study and test trials even though the sequence of item presentation changes from trial to trial. Mandler (1967) also showed powerful effects of organization on recall; after subjects sorted unrelated words into consistent groupings, they remembered them better than subjects in other conditions exposed to the words the same number of times.

One question that was never addressed in this line of research is whether organizational phenomena such as category clustering and subjective organization are determined by processes that occur during study trials, test trials, or both. The present

research investigated the effects of testing on organization by comparing learning conditions in which subjects performed both study trials and test trials of free recall in learning lists of categorized words with learning conditions in which subjects only performed study trials. The conditions of studying and testing were equated by allotting the same amount of time for study and test trials, and by equating the total number of study and test trials in each learning condition. In addition, the present experiments examined how subjects mentally organize words from the lists by varying study and test instructions to manipulate the manner and degree to which subjects processed and utilized organizational information. Of interest was whether varying the number of times subjects studied or attempted to recall lists of categorized words and types of study and test instructions affected both number of words recalled and organization in both initial and delayed tests of free and cued recall.

Theoretical Explanations for How Retrieval Affects Organization

Theories and models of memory have staked out a variety of positions on the question of how testing affects organization and whether organizational phenomena such as category clustering and subjective organization are determined by encoding or retrieval processes, or both. In one of the first studies demonstrating the positive effects of recall testing, or recitation, on retention relative to rereading, Kühn (1914, p. 440) argued that, "By learning with recitation the construction of groups can be carried on more readily than through reading. Many persons say, in fact, that in really pure reading such a construction of groups is impossible." In his classic large-scale study comparing the effects of recitation and rereading on retention among different groups of children, Gates (1917, pp. 96-97) made a similar point that "recitation was of great service in

assisting the subject to organize the material into some sort of compact and connected whole, such an organization being essential to a thorough mastery of it." He further argued that recitation fosters this organization, because as subjects attempt to reproduce the subject material they will analyze it more carefully, pick out striking information, and employ a better schema of reconstruction than by rereading (Gates, 1917, p. 9). According to this view, processes that occur during recall testing directly enhance organization relative to restudying.

A few decades later, Bousfield (1953) argued that the tendency to cluster categorically-related items during recall is due to processes that occur at retrieval. When subjects retrieve an item from a given category, an increment of memory strength is added to other list items from the same category, which Bousfield called the "relatedness increment," and as a result, the probability of then recalling an item from the same category increases relative to other lists items from different categories. Bousfield and Cohen (1953) further developed the concept of the relatedness increment into a hierarchical theory of mental organization based on the ideas of Hebb (1949). When subjects attempt to recall lists of words that represent instances from different categories, associative bonds between superordinate (e.g., category names) and subordinate (e.g., category instances) mental representations of words are strengthened. Thus, improvements in category clustering or output organization, in general, occur primarily during test trials whereby retrieving previously learned items strengthens their representation in memory and their capacity to evoke semantically-related items.

Slamecka's (1968) independent trace storage hypothesis is just as strong as Bousfield's (1953) concept of the relatedness increment in its emphasis on the notion that

organization occurs at retrieval. According to this view, information stored in long-term memory is disorganized; however, the process of retrieval is organized in the sense that during study, subjects formulate and adopt an organized retrieval plan for the future that might rely upon the use of higher order concepts to guide retrieval (for a similar theory of how information storage is disorganized see Landauer, 1975). In other words, during study trials, subjects store information in memory in a random fashion, but might notice relations among to-be-learned items that aid in the formation of a subsequent retrieval strategy. Free recall test trials then serve as an opportunity for subjects to implement their prepared retrieval strategies.

More recent associative theories of memory also propose mechanisms whereby the retrieval of information activates or strengthens the memorial representation of related concepts. For instance, computational models such as Free Recall by an Associative Net (FRAN; Anderson, 1972), Human Associative Memory (HAM; Anderson & Bower, 1973), Adaptive Control of Though-Rational (ACT-R; Anderson, 1996), Context Maintenance and Retrieval (CMR; Polyn, Norman, & Kahana, 2009), Search of Associative Memory (SAM; Raaijmakers & Shiffrin, 1980, 1981) along with its recent extensions eSAM (Sirotin, Kimball, & Kahana, 2005) and fSAM (Kimball, Smith, & Kahana, 2007), and the Temporal Context Model (TCM, Howard & Kahana, 2002) have demonstrated success in accounting for a variety of organizational phenomena observed in free recall. Although these models differ in many fundamental respects, such as in the ways verbal information is represented in the mind and what mental operations are performed at various stages of cognition, one key feature shared by all these models is that the processing of relational information, or organizational

learning, does occur during retrieval. On the other hand, these same models either explicitly state (or in the very least do not deny) the possibility that the same degree of processing or activation of relational information can also occur during study. This means that, all things being equal, study trials may be just as effective as test trials in promoting output organization in free recall.

Theories such as the transfer-appropriate processing framework (Morris, Bransford, & Franks, 1977) and encoding specificity principle (Tulving & Thomson, 1973) suggest alternative approaches to explain how retrieval might enhance organization in episodic recall. According to both views, performance on a test of memory benefits to the extent that conditions at retrieval match encoding conditions during prior learning. To the extent that subjects retrieve and utilize relational information such as higher-order taxonomic category or semantic associative information during free recall to guide episodic retrieval of previously learned items, prior testing should facilitate subsequent recall performance and promote a greater degree of output organization than studying. This is because the cognitive operations and conditions required to retrieve and organize information on an initial recall test more closely match those required to perform later recall tests.

Conversely, Bjork and Bjork (1992) argued in their "New Theory of Disuse" that although the act of retrieving previously learned information strengthens its memorial representation and increases the likelihood that the information may be retrieved in the future, related information may be weakened, thereby impairing its subsequent retrieval. Specifically, Bjork and Bjork argued that the learning and retention of information depends upon two properties: its storage strength and retrieval strength. Storage strength

describes how well information has been learned, and retrieval strength describes how easy it is to access the information in memory. One critical difference between these two properties is that while there is presumably no limit to the human mind's capacity to store information, there is a limit to how much information can be retrieved at any given time.

Both studying and retrieving information can result in increments to its storage strength and retrieval strength, but retrieval is a more potent event. The assumption is that the successful retrieval of previously learned information produces greater increments to its storage strength and retrieval strength relative to the act of restudying that information. Due to limitations in retrieval capacity, increasing the retrieval strength of certain information incurs the cost of rendering other information more difficult to retrieve. Bjork and Bjork (1992, p. 44) further argue that "such competitive effects will tend to be governed by similarity or category relationships defined semantically or episodically." In other words, the retrieval of previously learned items may weaken the retrieval strength of related items, which can explain the phenomenon of retrieval-induced forgetting (Anderson, 2003; Anderson, Bjork, & Bjork, 1994).

Another negative consequence of this latter property of retrieval is that testing might not enhance organization, and may even lead to worse output organization than repeated studying, because the successful retrieval of some items from a previously learned list of items may impair subsequent recall of semantically related list items. Chan (2009) has recently shown that these concerns may be especially warranted either when the final test is administered at a short retention interval (i.e., 20 minutes) following an earlier test or under conditions of "poor integration" in which study items are presented in

a disorganized (i.e. random) order and subjects are discouraged from forming inter-item conceptual relations (see also Anderson, 2003; Anderson & McCulloch, 1999).

Empirical Evidence that Retrieval Influences Organization

Although a variety of theories offer explanations for how retrieval affects organization, there is surprisingly little evidence that this is so. Two studies of hypermnesia have shown that taking multiple successive recall tests (without intervening study episodes) can enhance organization relative to taking a single test of equal total duration (Mulligan, 2002; Mulligan, 2005). For example, Mulligan (2005, Experiment 2) found that taking 4 successive 5-minute recall tests produced greater clustering two days later than taking a single 20 minute recall test. These findings are consistent with the view that repeated testing promotes the development of increasingly stable retrieval strategies (e.g., Hunt & McDaniel, 1993; McDaniel, Moore, & Whiteman, 1998). However, these data do not directly address the question of whether there are differential effects of studying and testing on recall organization.

One set of data that does suggest that testing may improve recall organization relative to studying alone comes from an experiment conducted by Masson and McDaniel (1981). In their first experiment, they presented subjects with a list of words representing several taxonomic categories in a random order. Half of the subjects were given intentional, and the other half given incidental, learning instructions. All subjects performed several different encoding tasks for the study of individual words. The encoding tasks required subjects to write on a sheet of paper either a category name, adjective, or rhyme word associated to a list item during its presentation. Last, half of the

subjects were given a free recall test immediately following the initial study period, and all of the subjects were given delayed recall and recognition tests a day later.

Not surprisingly, subjects who were given the immediate recall test demonstrated superior recall of the word lists one day later relative to subjects who were not previously tested (see Table 1). More importantly, Masson and McDaniel (1981) measured the degree of output organization in the recall protocols by computing adjusted ratio of clustering (ARC; Roenker, Thompson, & Brown, 1971) scores. ARC quantifies the extent to which subjects tend to cluster responses according to taxonomic categories (or other pre-defined types of categories). ARC scores range in value from -1 to 1, where 0 indicates that the amount of clustering reflected in subjects' response protocols is no greater than that expected by chance alone and 1 indicates perfect clustering. More importantly, ARC is considered to be a relatively pure measure of output organization, because it controls for differences in level of recall. As shown in Table 1, subjects who were initially tested on the word list produced higher ARC scores (.40 and .47) than subjects who did not receive a recall test during the first session (.20 and .11). In other words, the initially tested subjects tended to cluster their responses according to taxonomic categories in delayed free recall to a greater extent than non-tested subjects. These data suggest that testing can improve the organization of episodic retrieval.

Testing also eliminated differences in the effects of study instructions on long-term retention. When no immediate recall test was provided, intentional encoding instructions promoted better long-term retention of the word list than incidental encoding instructions. However, the advantage of intentional encoding disappeared with the administration of an immediate recall test. Masson and McDaniel (1981) argued that

Table 1. *Mean proportion of words recalled and adjusted ratio of clustering (ARC) scores as a function of study instructions and testing schedule in Experiment 1of Masson and McDaniel (1981).*

		Immediate Recall		Delayed Recall	
Prior Recall	Instructions	Words	ARC	Words	ARC
Yes	Intentional	.39	.06	.37	.40
Yes	Incidental	.35	.23	.30	.47
No	Intentional			.28	.20
No	Incidental			.16	.11

Note. The two prior recall groups received a free recall test in the initial session, whereas the remaining two groups did not take a free recall test until the second session 24 hours later.

testing promoted the additional processing of relational information among study items that, in turn, was utilized to aid retrieval. In other words, even under incidental learning conditions, testing appears to have stimulated the kind of processing associated with intentional learning.

While Masson and McDaniel's (1981) results are suggestive, they are not conclusive, because the higher organization scores for the prior recall condition may be attributed to the fact that subjects had an additional opportunity to learn the material; an additional study trial during the first session might have been just as effective in promoting additional processing of relational information among list items. Alternatively, one could argue that during study subjects performed encoding tasks that may have promoted greater processing of semantic and/or phonological features unique to each item while diminishing the processing of inter-item relational information. Output organization might have been greater had subjects simply been given the opportunity to study the list items as they saw fit, in which case they might have been more likely to notice and better process inter-item semantic relations.

To the extent that organization may be important for learning and retention, it is also worth pointing out that there is some evidence for the role of organizational processing in determining the presence and magnitude of testing effects. Wheeler and Roediger (1992; Experiment 1) conducted a study in which subjects studied a series of 60 pictures under one of two conditions. In one condition, the pictures were presented within the context of an orally narrated story, and in the second condition, the pictures were shown as a list, and subjects heard the name of each picture as it appeared. Afterwards, subjects filled out a brief questionnaire and then completed either one or three successive

free recall tests on the studied pictures, whereas another group did not take a free recall test. Then all groups were tested a week later.

One of the key findings was that when subjects attempted to recall the pictures a week later, pictures embedded in the story were generally remembered better than pictures presented only with their names (see Table 2). However, this benefit of meaningfully embedding the pictures in a story only occurred in the groups initially tested a week earlier. That is, testing itself appears to have improved the retention of picture materials organized in a more meaningful way. More importantly, the recall advantage of learning the pictures in the context of a story improved even further as the number of prior retrieval attempts increased from one to three. Thus, testing may facilitate the recall of previously learned meaningful materials to a greater degree than materials that are poorly understood or less-well organized. Consistent with these ideas, Chan and colleagues demonstrated that taking an initial recall test for previously studied prose passages can enhance long-term retention of related information that was not initially tested relative to studying the passages alone (Chan, 2009; Chan, McDermott, & Roediger, 2006; but see Gates, 1917, who reported larger testing effects for nonsense syllables as compared to meaningful prose).

Table 2. *Mean number of pictures recalled as a function of presentation context and testing schedule in Experiment 1 of Wheeler and Roediger (1992).*

	Initial Tests			Delayed Tests		
Group	Test 1	Test 2	Test 3	Test 1	Test 2	Test 3
			Pictures + Names			
3-3	26.6	27.2	28.4	25.2	26.3	26.0
1-3	25.7			20.2	21.7	23.0
0-3				16.7	17.5	17.5
			Pictures + Story			
3-3	32.7	35.0	36.4	31.8	33.0	33.4
1-3	31.8			23.3	25.0	25.6
0-3				17.4	17.2	18.4

Note—All groups took three tests in the delayed session. Group 3-3 received three tests in the initial session, Group 1-3 received one test in the initial session, and Group 0-3 took no memory tests until the delayed session.

Measures of Output Organization

The present research compared the effects of studying and testing during the acquisition of lists of words representing several conceptual categories on long-term retention and organization. We focused on several different measures to examine recall performance and organization. Total recall was measured by the proportion of all words recalled from each list. Recall of the categorized lists in Experiments 1 and 2 was also decomposed into two components which multiplied together give total recall: category recall (Rc) and recall of items within categories (Rw/c; Tulving & Pearlstone, 1966). Rc is defined as the number of times at least one member of a taxonomic category represented in the original study list is recollected, and Rw/c is the average number of items recalled from each of the list categories represented in a subject's output protocol (Cohen, 1963). The measures index how many categories can be recalled and the completeness of the recall from the categories once accessed.

The organization of recall was measured using the adjusted ratio of clustering (ARC; Roenker, Thompson, & Brown, 1971). As mentioned earlier, ARC assesses the degree to which subjects' recall patterns correspond to the conceptual structure of the study materials and is also considered a relatively pure measure of organization, because it controls for differences in recall level across subjects or learning conditions [for reviews of ARC and other clustering measures, see Kahana, Howard, & Polyn (2008); Murphy (1979); Murphy & Puff (1982); Pellegrino & Hubert (1982)].

Another form of organization that may be directly influenced by retrieval practice is subjective organization (e.g., Mulligan, 2002). Even with the use of categorized lists, subjects may tend to adopt idiosyncratic forms of conceptual organization to chunk list

items into higher order subjective units, or they may adopt uniform organization within category recall. The measure of subjective organization that we used in Experiments 2 and 3 is bi-directional intertrial repetition (B-ITR; Boufield & Bousfield, 1966; Bousfield, Puff, & Cowan, 1964), also referred to as pair frequency (PF; Sternberg and Tulving, 1977). Pair frequency represents the number of pairs of items recalled on adjacent test trials in adjacent output positions in either forward or reverse order. Moreover, pair frequency takes into account the baseline level of subjective organization that might be expected by chance alone in a given recall protocol. The measure can go from 0 (chance organization) to much higher levels (depending on the number of items recalled).

Of course, there are other measures of organization, and debates surrounding the issue of which is the best measure have not been resolved (Murphy, 1979). The measures we employed are commonly accepted in the literature and when used in combination provide a comprehensive picture of how testing affects the learning and utilization of organizational information to aid episodic retrieval relative to studying alone.

Overview of the Experiments

At present there is hardly any evidence that testing affects memory organization. Therefore, the current experiments were designed to investigate the potential effects of testing on organization as well as the potential contribution of organizational processes to the testing effect in free recall. First, do operations that occur during retrieval promote the additional processing of relational information, or does a second study trial produce a similar or perhaps even greater degree of output organization in delayed recall than a test trial? Experiment 1 addressed this question by using an experimental design similar to

that of Masson and McDaniel (1981, Experiment 1), but with some changes. In addition to comparing long-term retention and organization for subjects who either received one study trial followed by an immediate recall test with groups that received one study trial alone, there was an additional control group that performed two consecutive study trials. The additional control condition permitted answering the question of whether Masson and McDaniel's original finding that testing improved output organization was merely due to subjects having additional exposure to list items.

Experiments 2 and 3 were designed to examine whether organizational processes directly contribute to the testing effect in free recall. Experiment 2 asked whether varying the organizational processing that occurs during initial tests of free recall influences long-term retention and output organization of categorized word lists following a one-day retention interval. Four groups of subjects initially studied a categorized word list by performing one of several different semantic judgment tasks on each item. Immediately following the study trial, one group took a standard free recall test on the word list. A second group was given a two-dimensional chart at the start of free recall and asked to record items that belong to the same taxonomic category in the same columns and items that do not belong together in different columns. This condition was designed to enhance the overt retrieval and utilization of inter-item semantic relational information during recall relative to standard free recall testing.

A third group was also given a chart at the start of free recall, but was instructed to record items previously studied using the same judgment task in the same columns and items studied with different judgment tasks in different columns. This condition was designed to minimize the overt retrieval and utilization of inter-item semantic relational

information by focusing subjects' recollections on seemingly arbitrary inter-item relations based upon the type of judgment task assigned to each word. Last, there was an additional control group that performed two consecutive study trials.

Experiment 3 further examined whether organizational processes contribute to the testing effect in free recall by asking whether varying the perceived organization of the study materials mediates the benefits of testing on long-term retention and recall organization. We manipulated the organization of the study materials using lists of words representing ad-hoc categories (e.g., Barsalou, 1983, 1985), such as "things dogs chase" or "weekend entertainment", under conditions in which subjects were either aware or unaware of the categorical structure of the lists during learning.

In contrast to taxonomic categories whose knowledge structures are presumably well-established in long-term memory and may be automatically activated and brought to mind when particular category instances are encoded and/or retrieved, ad-hoc categories represent disparate knowledge that becomes organized into coherent categories in particular situations to achieve goal-relevant tasks. We also manipulated testing conditions by assigning different groups of subjects into conditions in which they either studied a word list for two consecutive study trials or took a recall test following an initial study trial. Final recall performance and output organization were assessed a day later in order to determine whether subjects given prior tests achieved higher levels of recall and organization than those who only studied the lists. The control condition permitted examination of recall and organization for what subjects perceived as an unrelated word list.

Experiment 1

The purpose of Experiment 1 was to examine the effects of testing on the learning and retention of lists of words representing different taxonomic categories. Of interest was whether the retrieval processes that occur during a recall test stimulate organizational processing to a greater extent than does a study trial of equal duration. Using an experimental design adapted from Masson and McDaniel (1981), we compared delayed recall performance, as measured by total word recall, category recall (Rc), and words per category recall (Rw/c), and organization, as measured by response output organization (ARC), for subjects who either received one study trial followed by an immediate recall test with groups that received one or two study trials alone. All groups were given a delayed test 24 hours later.

In one study-only condition, a group of subjects studied several lists of words for one study trial each with instructions to rate the pleasantness of each word. A second study-only group studied each list once with intentional learning instructions. A third repeated-study group rated the pleasantness of each word during an initial study trial, and then they studied each list a second time under intentional learning instructions. Last, a fourth prior-testing group initially studied each list of words with instructions to make pleasantness judgments, and then they attempted to recall each list immediately following list presentation.

The logic underlying these comparisons is as follows. The comparison of the pleasantness rating study phase by itself with the same kind of study phase plus an initial test conceptually replicates the design of Masson and McDaniel (1981, Experiment 1). The condition with two study conditions (pleasantness rating and intentional learning) equates exposure to that of the study + test condition. The addition of the single

intentional study control condition asks what effect studying under intentional learning has on later performance and permits comparison to the pleasantness-rating single-study condition. A day later, subjects in all four conditions took final tests of free and category cued recall.

Method

Subjects. 64 Washington University undergraduates participated for either payment or for course credit.

Design. There were four learning conditions distributed between subjects. In the S_p condition, 16 subjects studied 3 lists of words only once with instructions to rate the pleasantness of each list item on a 5 point scale. In the S_i condition, 16 subjects studied all 3 lists of words only once with intentional learning instructions to learn each of the list items as well as possible during list presentation. In the S_pS_i condition, another group of 16 subjects rated the pleasantness of each list item during an initial study trial, and then they studied the list a second time with standard intentional learning instructions before proceeding to the next list. Last, in the prior-testing condition (S_pT), 16 subjects first studied the list of words with instructions to make pleasantness judgments for each item, and then they attempted to recall the list immediately afterwards before proceeding to the next list. Words were presented in a different randomized order on each study trial in the one condition that involved two study trials. The critical tests took take place a day later when subjects in all four conditions attempted to recall the word lists using tests of free and category cued recall.

Materials. Ninety words sampled from 18 categories (5 words per category) in the expanded and updated version of the Battig and Montague word norms (Van

Overschelde, Rawson, & Dunlosky, 2004) were used to create 3, 30-word study lists. The 30 words in each list included 5 medium frequency nouns belonging to each of 6 taxonomic categories.

Procedure. Subjects participated in two sessions scheduled 1 day apart. In the first session, subjects were informed that they would study several lists of words presented by a computer in preparation for a memory test the next day. During the study trials, the computer displayed each word in the center of the monitor display one at a time for 4.5 seconds, followed by a 500 ms inter-stimulus interval. Words were presented in randomized order on each study trial. For the S_p study trials, subjects were informed that they had 5 seconds during the presentation of each word to type a number between 1 and 5 indicating their pleasantness judgment for the current item. For the S_i study trials, subjects were only instructed to learn each word as best as possible as it was presented. The total time for each study trial was 2.5 minutes.

During the test trials in the S_pT condition, subjects were given 2.5 minutes to write down on a blank sheet of paper as many words as they could remember from the most recently studied list in any order that the words come to mind (free recall). In order to keep the spacing between each of the 3 study lists constant across the 4 learning conditions, subjects in the S_p and S_i conditions played Tetris for an additional 2.5 minutes in between study trials. E-Prime experimental software (Psychology Software Tools, Inc.) was used for stimulus presentation and recording subjects' keyboard responses. The first session lasted about 30 minutes.

Following a 1-day retention interval, subjects were given tests of final free and cued recall. During the free recall test, subjects had 10 minutes to write down on a blank

sheet of paper as many words as they could remember from all 3 lists in any order that the words came to mind. Last, subjects had 10 minutes to recall words from all three lists; however, in contrast to the previous test, subjects were also provided a list of all of the category names to aid recall of the words. Of course, because cued recall followed free recall, effects in cued recall may be partly due to the prior free recall test. The second session lasted 20 minutes.

Results

All results, unless otherwise stated, were significant at the .05 level. For all sets of individual comparisons, we controlled the Type I error rate using the False Discovery Rate procedure (Benjamini & Hochberg, 1995; Benjamini & Yekutieli, 2001). We only report analyses for the delayed tests of free and cued recall, because only one learning condition (S_pT) included tests during the initial learning phase, and it was only possible to compare recall performance and organization across all conditions in the delayed tests. On the initial test trial, subjects in the S_pT condition recalled, on average, 20.31 (*SD* = 3.69) words or .68 (*SD* = .12) of the list from 5.48 (*SD* = .36) categories (Rc) and 3.71 (*SD* = .56) items per category (Rw/c) of each 30-item list. Recall was also highly organized, as indicated by a mean ARC score of .79 (*SD* = .12).

Recall of Words. The top row of Table 3 shows that testing during the initial learning phase improved recall performance in delayed tests of free and cued recall. We conducted a 2 (Test Type: Free Recall vs. Cued Recall) X 4 (Learning Condition: S_p vs. S_i vs. S_pS_i vs. S_pT) ANOVA, which revealed superior performance in cued recall relative to free recall, (.40 vs. .26), $F(1,60) = 511.39$, $MSE = .00$, $\eta_p^2 = .90$. There was a

significant effect of learning condition, $F(3,60) = 23.59$, $MSE = .03$, $\eta_p^2 = .54$, as well as a significant interaction between the two factors, $F(3,60) = 3.95$, $MSE = .00$, $\eta_p^2 = .17$.

These effects were due to enhanced free recall in the prior testing condition (S_pT) relative to the study-only S_p (.45 vs. .19), $t(30) = 6.48$, $SEM = .04$, $d = 2.35$, S_i (.45 vs. .18), $t(30) = 7.84$, $SEM = .03$, $d = 2.84$, and S_pS_i (.45 vs. .21), $t(30) = 5.99$, $SEM = .04$, $d = 2.17$, conditions. Cued recall was also enhanced in the S_pT condition as compared to the S_p (.61 vs. .34), $t(30) = 6.46$, $SEM = .04$, $d = 2.24$, S_i (.61 vs. .29), $t(30) = 7.60$, $SEM = .04$, $d = 2.66$, and S_pS_i (.61 vs. .37), $t(30) = 5.27$, $SEM = .05$, $d = 1.85$, conditions. No other comparisons among the study-only conditions were statistically significant.

In general, the pattern of results is the same in cued recall as that for free recall and similar patterns of statistical significance obtained for these and subsequent analyses across all three experiments. It is important to keep in mind that cued recall followed free recall, so the parallel trends may be carryover effects from free recall. Thus, analyses for free and cued recall were reported separately, even when there were no significant interactions between the two measures.

In sum, testing improved long-term free and cued recall relative to studying alone, and neither varying the encoding instructions (pleasantness ratings vs. standard

Table 3. *Mean proportion of words recalled, number of categories recalled (Rc), number of words per category recalled (Rw/c), adjusted ratio of clustering (ARC) scores, and proportion of recalled words that were extra-list intrusions (XLIs) as a function of the study with pleasantness ratings (S_p), study with intentional learning instructions (S_i),repeated study with pleasantness ratings on the first trial and intentional learning instructions on the second trial (S_pS_i), and study with pleasantness ratings followed by a recall test (S_pT) initial learning conditions in delayed tests of free and cued recall in Experiment 1.*

		Free Recall				Cued Recall			
Measure		S_p	S_i	S_pS_i	S_pT	S_p	S_i	S_pS_i	S_pT
Recall	*Prop.*	.19	.18	.21	.45	.34	.29	.37	.61
	CI	(.06)	(.04)	(.06)	(.06)	(.05)	(.06)	(.06)	(.06)
Rc	*M*	8.31	7.56	8.19	12.56	14.69	13.06	15.69	17.25
	CI	(1.68)	(1.50)	(1.32)	(.74)	(1.23)	(1.70)	(1.09)	(.67)
Rw/c	*M*	1.99	2.04	2.16	3.17	2.07	1.93	2.09	3.17
	CI	(.23)	(.25)	(.35)	(.28)	(.22)	(.18)	(.26)	(.27)
ARC	*M*	.60	.48	.60	.85				
	CI	(.20)	(.17)	(.17)	(.04)				
XLIs	*Prop.*	.23	.36	.21	.06	.39	.52	.41	.12
	CI	(.10)	(.12)	(.11)	(.04)	(.10)	(.13)	(.12)	(.07)

Note—Values in parentheses are 95% confidence intervals (CI).

intentional learning) nor the number of study opportunities (1 vs. 2 study trials) affected delayed recall performance.

Recall of Categories. The second row of Table 3 shows that testing during the initial learning phase improved Rc in delayed tests of free and cued recall. An ANOVA revealed a significant effect of test type, with more categories accessed in cued recall than in free recall, (15.17 vs. 9.16), $F(1,60) = 494.19$, $MSE = 2.34$, $\eta_p^2 = .89$. There was a significant effect of learning condition, $F(3,60) = 11.06$, $MSE = 11.03$, $\eta_p^2 = .36$, as well as a significant interaction between the two factors, $F(3,60) = 4.96$, $MSE = 2.34$, $\eta_p^2 = .20$.

These effects were due to enhanced Rc in the prior testing condition (S_pT) relative to study-only S_p (12.56 vs. 8.31), $t(30) = 4.64$, $SEM = .92$, $d = 1.64$, S_i (12.56 vs. 7.56), $t(30) = 6.01$, $SEM = .83$, $d = 2.13$, and S_pS_i (12.56 vs. 8.19), $t(30) = 5.80$, $SEM = .75$, $d = 2.05$, conditions in free recall. Rc was enhanced, but to a lesser extent, in cued recall in the S_pT condition relative to the S_p (17.25 vs. 14.69), $t(30) = 3.65$, $SEM = .70$, $d = 1.29$, S_i (17.25 vs. 13.06), $t(30) = 4.59$, $SEM = .91$, $d = 1.62$, and S_pS_i (17.25 vs. 15.69), $t(30) = 2.44$, $SEM = .64$, $d = .86$, conditions. No other comparisons were statistically significant. In sum, testing during the initial learning phase improved Rc relative to studying alone, and neither varying the encoding instructions nor the number of study trials affected category recall.

Recall of Items Within Categories. As shown in the third row of Table 3, testing during the initial learning phase improved Rw/c in delayed tests of free and cued recall. An ANOVA revealed that there was no effect of test type, $F < 1$. There was a significant effect of learning condition, $F(3,60) = 21.16$, $MSE = .49$, $\eta_p^2 = .51$, but no interaction

between the two factors, $F(3,60) = 1.08$, $MSE = .06$, $\eta_p^2 = .05$, *ns*. The effect of learning condition was due to enhanced Rw/c in free recall in the prior testing condition (S_pT) relative to the S_p (3.17 vs. 1.99), $t(30) = 6.53$, $SEM = .18$, $d = 2.30$, S_i (3.17 vs. 2.04), $t(30) = 6.04$, $SEM = .19$, $d = 2.13$, and S_pS_i (3.17 vs. 2.09), $t(30) = 4.49$, $SEM = .22$, $d = 1.59$, conditions. In cued recall, Rw/c was similarly enhanced in the S_pT condition relative to S_p (3.17 vs. 2.07), $t(30) = 6.36$, $SEM = .17$, $d = 2.23$, S_i (3.17 vs. 1.93), $t(30) = 7.61$, $SEM = .16$, $d = 2.68$, and S_pS_i (3.17 vs. 2.09), $t(30) = 5.83$, $SEM = .19$, $d = 2.04$, conditions. No other comparisons among the study-only conditions were significant. In sum, long-term free and cued recall of words within categories were superior in the prior testing condition relative to the study-only conditions, and neither varying the encoding instructions nor the number of study trials affected Rw/c.

Category Clustering. As shown in the fourth row of Table 3, testing during the initial learning phase improved category clustering in delayed free recall. An ANOVA confirmed a significant effect of learning condition on category clustering, $F(3,58) = 3.93$, $MSE = .10$, $\eta_p^2 = .16$, which was due to enhanced ARC scores in the prior testing condition (S_pT) relative to study-only S_p (.85 vs. .60), $t(30) = 2.50$, $SEM = .10$, $d = .87$, S_i (.85 vs. .48), $t(29) = 4.41$, $SEM = .08$, $d = 1.59$, and S_pS_i (.85 vs. .61), $t(29) = 2.78$, $SEM = .09$, $d = .97$, conditions. No other comparisons among the study-only conditions were significant. In addition, ARC scores were positively correlated with delayed recall ($r = .51$). Thus, testing improved the organization of recall and organization was correlated with the number of words recalled. Furthermore, neither varying the encoding instructions nor the number of study trials affected output organization.

Intrusions. We further examined recall accuracy by measuring the proportion of all words recalled in delayed tests of free and cued recall that were words not presented during the course of the experiment (extra-list intrusions). The bottom row of Table 3 shows that category cueing increased the commission of extra-list intrusions relative to free recall across all learning conditions, but testing during the learning phase reduced false recall on the delayed test. An ANOVA confirmed that a greater proportion of extra-list intrusions were committed in cued recall than in free recall, (.36 vs. .21), $F(1,60) = 99.16$, $MSE = .01$, $\eta_p^2 = .62$. There was a significant effect of learning condition $F(3,60) = 8.14$, $MSE = .08$, $\eta_p^2 = .29$, as well an interaction between the two factors, $F(3,60) = 3.79$, $MSE = .01$, $\eta_p^2 = .16$.

These effects were due to a lower proportion of extra-list intrusions committed in free recall in the prior testing condition (S_pT) relative to the S_p (.06 vs. .23), $t(30) = 3.07$, $SEM = .06$, $d = 1.09$, S_i (.06 vs. .36), $t(30) = 4.64$, $SEM = .07$, $d = 1.63$, and S_pS_i (.06 vs. .21), $t(30) = 2.71$, $SEM = .06$, $d = .92$, conditions. Even fewer extra-list intrusions occurred in cued recall in the S_pT condition relative to S_p (.12 vs. .39), $t(30) = 4.21$, $SEM = .06$, $d = 1.27$, S_i (.12 vs. .52), $t(30) = 5.43$, $SEM = .07$, $d = 1.93$, and S_pS_i (.12 vs. .41), $t(30) = 4.03$, $SEM = .07$, $d = 1.25$, conditions. No other comparisons among the study-only conditions were significant. Thus, testing during the initial learning phase reduced false recall as compared to studying alone following a long delay.

Discussion

This experiment confirmed a powerful effect of testing (relative to restudying) on delayed retention tests of free and cued recall. Consistent with prior research, studying a list and taking an immediate recall test produced greater recall a day later compared to

conditions in which subjects only studied a list one or two times (Masson & McDaniel, 1981). Somewhat surprisingly, neither varying the conditions of encoding nor increasing the number of study trials affected recall after 24 hours. Although it is reasonable to expect that repeatedly studying information should improve recall relative to a single study opportunity, repetition does not always boost retention (e.g, Callendar & McDaniel, 2009) especially after long delays (Karpicke & Roediger, 2008).

Keep in mind that if sheer exposure were the primary factor determining performance, the repeated study condition should have outperformed the prior testing condition. When subjects were given a test in the initial learning phase, they only recalled (on average) about 70% of the items, whereas subjects in the repeated study condition were of course re-exposed to 100% of the items on each study trial. In addition, prior testing improved overall accuracy by minimizing false recall of extra-list intrusions relative to repeated studying alone. These results extend previous findings that testing reduces the commission of prior-list intrusions in recall (Szpunar, McDermott, & Roediger, 2008). Taken together, these findings provide further striking evidence for the power of testing (Roediger & Karpicke, 2006b).

The purpose of this experiment was to determine what components of recall were improved by testing relative to studying alone – access to higher order units, access to items within units, or both. The last option was confirmed because testing benefited both measures of category access (Rc) and recall of items within each accessed category (Rw/c) in delayed tests of free and cued recall. These results are surprising, because many prior studies have shown that these two factors contribute independently to recall. That is,

variables that influence Rc usually have no influence on Rw/c, and vice versa (e.g., Burns & Brown, 2000; Cohen, 1963, 1966; Hunt & Seta, 1984; Tulving & Pearlstone, 1966).

If individuals learn categorized word lists by chunking items into category-based units then, presumably, once they can access the units during retrieval, their contents (the individual items) will be accessed as well to some degree. In their classic work supporting the distinction between item availability and accessibility, Tulving and Pearlstone (1966) showed that Rc and Rw/c were largely independent of each other, because variables that affected Rc (such as category cuing and list length) had little influence on Rw/c. Hunt and Seta (1984) argued that Rc and Rw/c measure the extent to which relational and item-specific information, respectively, is used to guide episodic retrieval. While Rc measures the extent to which individuals can retrieve higher order units or chunks, Rw/c reflects the degree to which individuals can retrieve category members.

Indeed, experimental conditions designed to promote organizational processing (e.g., instructing subjects to organize study items, providing category names during study) have been found to selectively increase Rc, and those designed to enhance item-specific processing (e.g., generating study items) have been shown to increase Rw/c (e.g., Cohen, 1963, 1966; McDaniel, Waddill, & Einstein, 1988; Schmidt & Cherry, 1989). To the extent that these measures assess the extent to which relational (Rc) and item-specific (Rw/c) information is used to guide episodic retrieval (e.g., Hunt & Seta, 1984), then our findings show that testing may promote both relational and item-specific processing relative to studying alone.

In addition, our results provide definitive confirmation that testing can improve organization of recall, or category clustering, in delayed free recall relative to restudying material (Masson & McDaniel, 1981). That organization was positively correlated with delayed recall further suggests that the testing effect in free recall may be due in part to enhanced organization during retrieval. Of course, a positive correlation does not establish a causal relationship. In order to more directly examine whether processes involved in mentally organizing information during learning contribute to the testing effect in free recall, we asked in Experiments 2 and 3 whether manipulating the organizational processing that occurs as subjects study and attempt to recall categorized word lists affects long-term retention and recall organization.

Experiment 2

The purpose of Experiment 2 was to investigate whether organizational processing that occurs during retrieval contributes to the testing effect in free recall. Specifically, we asked whether varying the retrieval instructions designed to either enhance or reduce organizational processing during initial tests of free recall influences long-term retention and output organization of categorized word lists following a one-day retention interval. Similar to Experiment 1, following either one study trial and one test trial or two study trials, we compared delayed recall performance, as measured by total word recall, category recall (Rc), and words per category recall (Rw/c), and organization, as indexed by response output organization measures (ARC and PF). All groups were given delayed tests of free and cued recall 24 hours later.

In an initial study trial for all conditions, subjects performed one of a variety of encoding tasks on each item of a categorized word list. Specifically, subjects provided

one of 6 types of judgments (e.g., pleasantness, imagery, survival processing) during the presentation of each word using a 1-5 scale. A repeated study (S_jS, where "S_j" refers to the initial study trial performed with various judgment tasks and "S" denotes the subsequent study trial performed without making judgments) condition in which subjects studied the categorized word list for a second time under standard intentional learning conditions and without making judgments served as a control condition.

In a standard free recall condition (S_jT, where "S_j" refers to the initial study trial performed with various judgment tasks and "T" denotes the subsequent free recall test trial), subjects were asked to recall previously studied items in any order that the words came to mind. In two additional testing conditions, subjects were given a two-dimensional (6 rows X 5 columns) chart at the start of each test of free recall and asked to write down list items starting from the upper left hand corner of the chart and then to record items that belong together conceptually in the same columns and items that do not belong together in different columns.

In the free recall by category (S_jT_c) condition, subjects were instructed to record previously studied items that belong to the same taxonomic category in the same columns and items that belong to different categories in different columns. This condition was designed to enhance the overt retrieval and utilization of inter-item relational information during recall relative to standard free recall testing. In the free recall by judgments (S_jT_j) condition, subjects were instructed to record items previously studied using the same judgment task in the same columns. In contrast to the standard free recall (S_jT) and free recall by categories (S_jT_c) conditions, the S_jT_j condition was designed to minimize the

overt retrieval and utilization of inter-item semantic relational information by focusing subjects' recollections on the type of judgment performed on each word.

Method

Subjects. 96 Washington University undergraduates participated for either payment or for course credit.

Design. There were four learning conditions distributed between subjects with 24 subjects assigned to each condition. In the study-only (S_jS) condition, subjects studied each of 3 categorized word lists for two consecutive trials. On the first study trial, subjects provided 1 of 6 different types of judgments (described below) for each list item, and then had the opportunity to study the list a second time with standard intentional learning instructions. In the standard free recall (S_jT) condition, subjects performed one study trial followed by a free recall test trial for each list. In the free recall by judgments (S_jT_j) condition, subjects performed one study trial with judgments followed by a test trial that required subjects to write down words from the list in any order that they came to mind in a two-dimensional chart such that items that were given the same type of judgment were to be written in the same column, and items given different judgments were to be written in different columns. Last, in the free recall by categories (S_jT_c) condition, subjects performed one study trial with judgments followed by a test trial that required subjects to write down words from the list in any order that they came to mind starting from the upper left hand corner of the two-dimensional chart and recording items that belonged to the same taxonomic category within the same column, and recording items that belonged to different categories in separate columns.

Materials. 90 words sampled from 18 categories (5 words per category) in the expanded and updated version of the Battig and Montague word norms (Van Overschelde et al., 2004) were used to create 3, 30-word study lists. The 30 words in each list included 5 medium frequency nouns belonging to each of 6 taxonomic categories.

Procedure. Subjects participated in two sessions scheduled 1 day apart. In the first session, subjects were informed that they would study and attempt to recall several lists of words presented by a computer. During the study trials, the computer displayed each word one at a time for 4.5 seconds, followed by a 500 millisecond inter-stimulus interval. Words were presented in randomized order on each study trial. Just as each of the lists included words representing 6 different taxonomic categories, subjects were instructed to provide 1 of 6 different types of judgments for each list item using a 5 pt. scale. Specifically, subjects rated the pleasantness, concreteness, survival value, activity (passive to active), potency (weak to strong), or valence (negative to positive) of each item. However, subjects were not informed about the specific categories represented in each list. The assignment of judgment task was counterbalanced such that no two words within a category were assigned the same judgment task, and each judgment task was assigned to every list item an equal number of times across subjects.

Subjects were informed that they have up to 5 seconds during the presentation of each list item to type a number between 1 and 5 indicating their judgment for the current word. A label appeared at the top of the computer screen indicating which type of judgment was to be made for a given item. The second study trial in the study-only condition did not require subjects to make judgments. Rather, the list of words was shown again at the same rate of presentation in a new random order and subjects were

given standard intentional learning instructions. The total study time was 2.5 minutes per trial.

During the test trials, subjects in the S_jT condition had 5 minutes to write down on a blank sheet of paper as many words as they could remember from the most recently studied list in any order that the words came to mind. In the S_jT_c condition, subjects were provided with a 6 column X 5 row chart and asked to write down words from the list just presented in any order that they came to mind starting from the upper left hand corner of the grid and record items that belong to the same taxonomic category within the same column and items that belong to different categories in different columns. In addition, subjects were instructed to write a label representing each category recalled at the top of each column and to number each recalled word in the order in which it was written, thereby permitting the computation of output organization scores (ARC and PF) for the output protocols.

In the S_jT_j condition, subjects were provided with a 6 column X 5 row chart and asked to write down words from the just-presented list in any order that they came to mind starting from the upper left hand corner of the grid and to record items that were given the same type of judgment within the same column and items given different judgments in different columns (see Appendix 1 for the chart administered in the S_jT_c and S_jT_j conditions). In addition, subjects were instructed to write a label representing each judgment type recalled at the top of each column and to number each recalled word in the order in which it was written. The same procedure for the study and test trials was repeated for the remaining two categorized word lists. This first session lasted about 30 minutes.

Following a 1-day retention interval, subjects took final tests of free and cued recall. During the free recall test, subjects had 10 minutes to write down on a blank sheet of paper as many words as they could remember from all 3 lists in any order that the words came to mind. Last, subjects had 10 minutes to again recall words from all three lists; however, in contrast to the previous test, subjects were also provided a list of all of the category names to aid recall of the words. The second session lasted 20 minutes.

Results

Recall of Words. Figure 1 shows that the mean proportion of words recalled from each list was similar across recall tests in the learning phase. When subjects attempted to organize words during retrieval according to their assigned judgment tasks (S_jT_j condition), recall was poor relative to the standard free recall (S_jT) and free recall by categories (S_jT_c) conditions. We conducted a 3 (Test Trial: Test 1 vs. Test 2 vs. Test 3) X 3 (Learning Condition: S_jT vs. S_jT_j vs. S_jT_c) ANOVA, which confirmed a significant effect of learning condition, $F(2,69) = 19.11$, $MSE = .03$, $\eta_p^2 = .36$, with enhanced recall in the S_jT and S_jT_c conditions relative to the S_jT_j condition (.52 vs. .35), $t(46) = 5.76$, $SEM = .03$, $d = 1.69$, and (.51 vs. .35), $t(46) = 5.34$, $SEM = .03$, $d = 1.51$, respectively. However, there was neither a significant effect of test trial, $F < 1$, nor a significant interaction between test trial and learning condition, $F(4, 138) = 1.23$, $MSE = .01$, $\eta_p^2 = .03$, *ns*.

A day later, the top row of Table 4 shows that delayed recall was superior in the S_jT_c condition, followed by the S_jT condition, and poorest in the S_jS and S_jT_j conditions. We conducted a 2 (Test Type: Free Recall vs. Cued Recall) X 4 (Learning Condition: S_jS vs. S_jT vs. S_jT_j vs. S_jT_c) ANOVA, which revealed enhanced performance in cued relative

Figure 1.

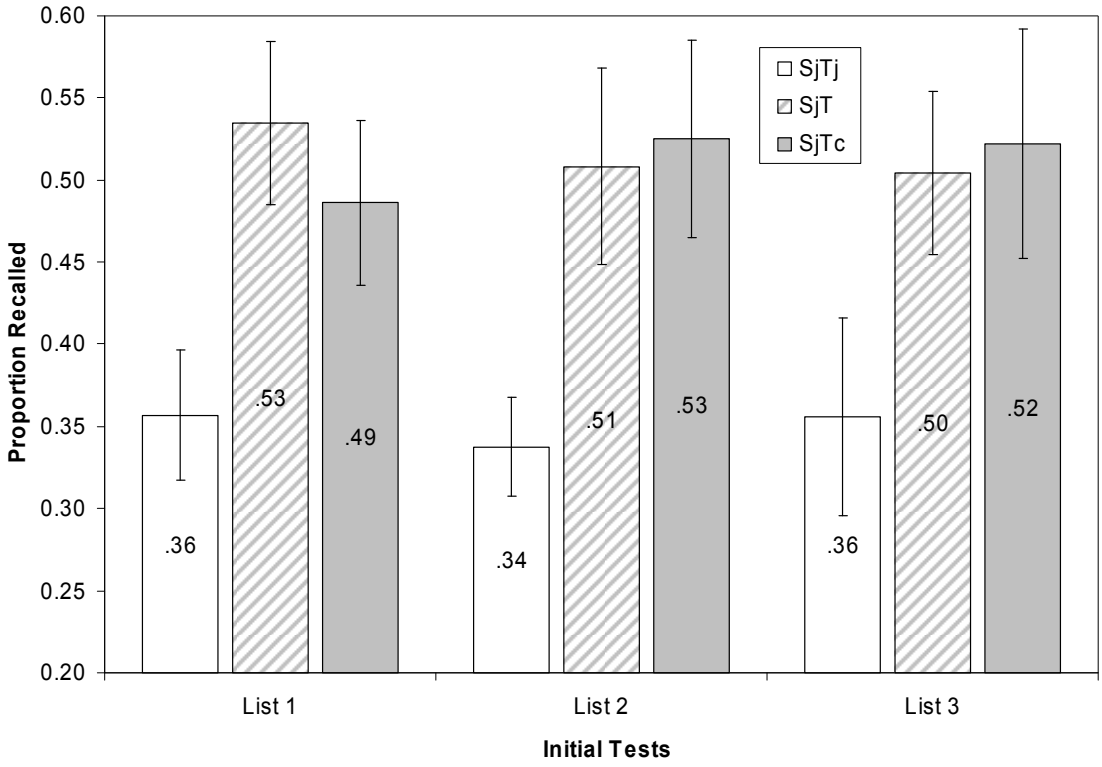

0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
Proportion Recalled
SjTj
SjT
SjTc
.36
.53
.49
.34
.51
.53
.36
.50
.52
List 1
List 2
List 3
Initial Tests

Table 4. *Mean proportion of words recalled, number of categories recalled (Rc), number of words per category recalled (Rw/c), adjusted ratio of clustering (ARC) scores, pair frequency (PF) scores, and proportion of recalled words that were extra-list intrusions (XLIs) as a function of the repeated study (S_jS), free recall by judgment tasks (S_jT_j), standard free recall (S_jT), and free recall by categories (S_jT_c) conditions in delayed tests of free and cued recall in Experiment 2.*

		Free Recall				Cued Recall			
Measure		S_jS	S_jT_j	S_jT	S_jT_c	S_jS	S_jT_j	S_jT	S_jT_c
Recall	*Prop.*	.23	.23	.28	.36	.38	.33	.43	.48
	CI	(.06)	(.05)	(.04)	(.05)	(.06)	(.04)	(.05)	(.05)
Rc	*M*	8.83	9.92	10.46	12.00	15.33	14.92	16.33	16.96
	CI	(1.41)	(1.25)	(1.22)	(1.25)	(.98)	(.98)	(.73)	(.53)
Rw/c	*M*	2.17	2.02	2.34	2.65	2.14	1.93	2.35	2.54
	CI	(.29)	(.16)	(.20)	(.18)	(.25)	(.20)	(.20)	(.18)
ARC	*M*	.62	.62	.66	.71				
	CI	(.12)	(.10)	(.08)	(.08)				
PF	*M*		2.29	4.17	6.25				
	CI		(.82)	(1.43)	(1.57)				
XLIs	*Prop.*	.23	.11	.10	.14	.25	.20	.16	.21
	CI	(.08)	(.05)	(.05)	(.05)	(.05)	(.06)	(.04)	(.05)

Note—Values in parentheses are 95% confidence intervals (CI).

to free recall (.40 vs. .27), $F(1, 92) = 432.77$, $MSE = .00$, $\eta_p^2 = .83$. There was a significant effect of learning condition, $F(3, 92) = 6.29$, $MSE = .03$, $\eta_p^2 = .17$, as a well as a significant interaction between the two factors, $F(3, 92) = 3.73$, $MSE = .00$, $\eta_p^2 = .11$. These effects were due to enhanced free recall in the S_jT_c condition relative to the S_jS condition (.36 vs. .23), $t(46) = 3.27$, $SEM = .04$, $d = 1.00$, S_jT_j condition (.36 vs. .23), $t(46) = 3.94$, $SEM = .03$, $d = 1.23$, and S_jT condition (.36 vs. .28), $t(46) = 2.29$, $SEM = .03$, $d = .67$. Although performance in the S_jT condition was higher than in the S_jS and S_jT_j conditions, the differences were not statistically significant (.28 vs. .23), $t(46) = 1.22$, $SEM = .04$, $d = .38$, *ns*, and (.28 vs. .23), $t(46) = 1.54$, $SEM = .03$, $d = .43$, *ns*, respectively, and recall was identical in the S_jS and S_jT_j conditions, $t < 1$.

The fact that recall in the S_jT condition was not significantly greater than in the S_jS condition is somewhat surprising, because numerous studies have demonstrated significant positive effects of prior testing on long-term retention relative to restudying (see Roediger & Karpicke, 2006b), and yet testing only enhanced long-term retention when subjects were explicitly instructed to semantically organize their responses during retrieval on the initial free recall tests. The absence of testing effects in the S_jT condition and S_jT_j condition is likely due in part to low initial recall performance permitting subjects to have re-exposure to only 52% and 35% of the words they recalled (averaged over Lists 1-3) during the test trial in these respective conditions as compared to the S_jS condition where subjects were re-exposed to 100% of the words during the second study trial. Nevertheless, a robust testing effect did occur in the S_jT_c condition where subjects demonstrated a similarly low level of initial recall performance (.51), which further indicates that enhanced organizational processing in the S_jT_c condition contributed to the

testing effect, and that the absence of testing effects of in the S_jT and S_jT_j conditions may be due to poorer or sub-optimal organizational processing.

Last, cued recall was also enhanced in the S_jT_c relative to the S_jS (.48 vs. .38), $t(46) = 2.68$, $SEM = .04$, $d = .75$, and S_jT_j (.48 vs. .33), $t(46) = 4.64$, $SEM = .03$, $d = 1.36$, conditions. In addition, recall was superior in the S_jT as compared to the S_jT_j condition (.43 vs. .33), $t(46) = 2.97$, $SEM = .03$, $d = .83$. No other individual pair-wise comparisons were statistically significant.

In sum, the benefit of testing on long-term free recall was greatest when subjects were explicitly instructed to semantically organize their responses during initial free recall testing. However, when subjects were initially tested with standard free recall instructions or with instructions to organize responses according to their assigned encoding tasks, delayed recall performance was not significantly better than that obtained from studying alone.

Recall of Categories. Although total word recall remained constant, Figure 2 shows that the mean number of categories recalled (Rc) from each list declined across the initial recall tests performed in the learning phase. Rc also varied as a function of the retrieval conditions during testing, with greatest recall of semantic categories in the S_jT condition, followed by the S_jT_c and S_jT_j conditions. An ANOVA revealed a significant effect of list, $F(2,69) = 24.18$, $MSE = .72$, $\eta_p^2 = .26$, due to higher Rc in the first list recalled relative to recall of the second and third lists (5.07 vs. 4.63), $t(71) = 2.93$, $SEM = .15$, $d = .48$, and (5.07 vs. 4.38), $t(71) = 4.98$, $SEM = .14$, $d = .73$, respectively. The difference in Rc between the second and third recall trials was not significant (4.63 vs. 4.38), $t(71) = 1.65$, $SEM = .15$, $d = .25$, *ns*.

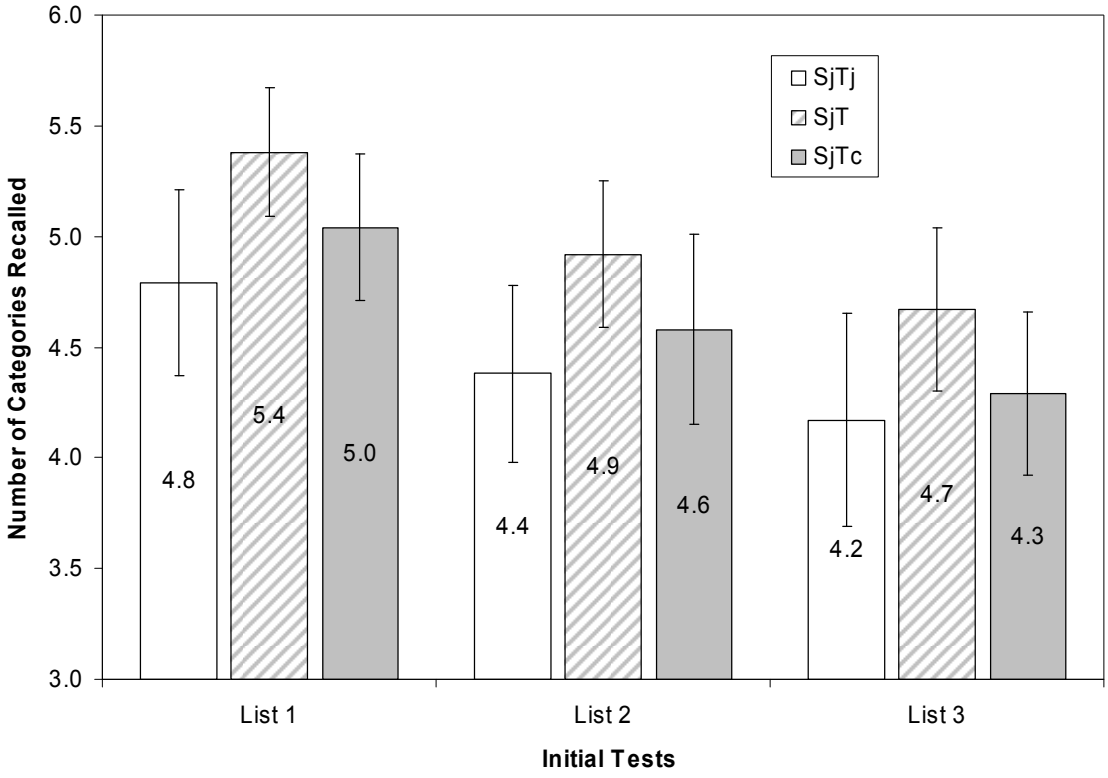
In addition, there was a significant effect of learning condition, $F(2,138) = 5.19$, $MSE = 1.04$, $\eta_p^2 = .13$, due to enhanced Rc in the S_jT relative to the S_jT_j and S_jT_c conditions (4.99 vs. 4.44), $t(46) = 3.40$, $SEM = .16$, $d = 1.00$. Neither the difference in Rc between the S_jT and S_jT_c (4.99 vs. 4.64), $t(46) = 2.19$, $SEM = .16$, $d = .64$, *ns*, nor between the S_jT_j and S_jT_c was significant (4.44 vs. 4.64), $t(46) = 1.02$, $SEM = .19$, $d = .30$, *ns*. Moreover, there was non-significant interaction between test trial and learning condition, $F < 1$. Apparently, telling subjects to organize recall by categories actually led to their recalling fewer categories than in standard free recall.

There was a shift in the pattern of results following the 24-hour retention interval. The second row of Table 4 shows that for delayed tests of free and cued recall Rc was superior in the S_jT_c condition, followed by the S_jT condition, and poorest in the S_jT_j and S_jS conditions. An ANOVA confirmed that Rc was greater in cued relative to free recall (15.89 vs. 10.30), $F(1, 92) = 519.89$, $MSE = 2.88$, $\eta_p^2 = .85$. There was a significant effect of learning condition, $F(3, 92) = 4.74$, $MSE = 11.78$, $\eta_p^2 = .13$, but a non-significant interaction between the two factors, $F(3, 92) = 2.30$, $MSE = 2.88$, $\eta_p^2 = .07$, *ns*.

In free recall, these effects were due to enhanced Rc in the S_jT_c condition relative to the S_jS condition (12.00 vs. 8.83), $t(46) = 3.27$, $SEM = .97$, $d = .94$, S_jT_j condition (12.00 vs. 9.92), $t(46) = 2.29$, $SEM = .91$, $d = .66$, and S_jT condition; however, the latter difference was not statistically significant (12.00 vs. 10.46), $t(46) = 1.73$, $SEM = .89$, $d = .50$, *ns*. Although Rc was higher in the S_jT condition as compared to the S_jS and S_jT_j conditions, their differences were not statistically significant (10.46 vs. 8.83), $t(46) = 1.71$, $SEM = .95$, $d = .49$, *ns*, and (10.46 vs. 9.92), $t < 1$, respectively.

Similarly, in cued recall, Rc was also enhanced in the S_jT_c condition relative to

Figure 2.


6.0
5.5
5.0
4.5
4.0
3.5
3.0
Number of Categories Recalled
SjTj
SjT
SjTc
4.8
5.4
5.0
4.4
4.9
4.6
4.2
4.7
4.3
List 1
List 2
List 3
Initial Tests

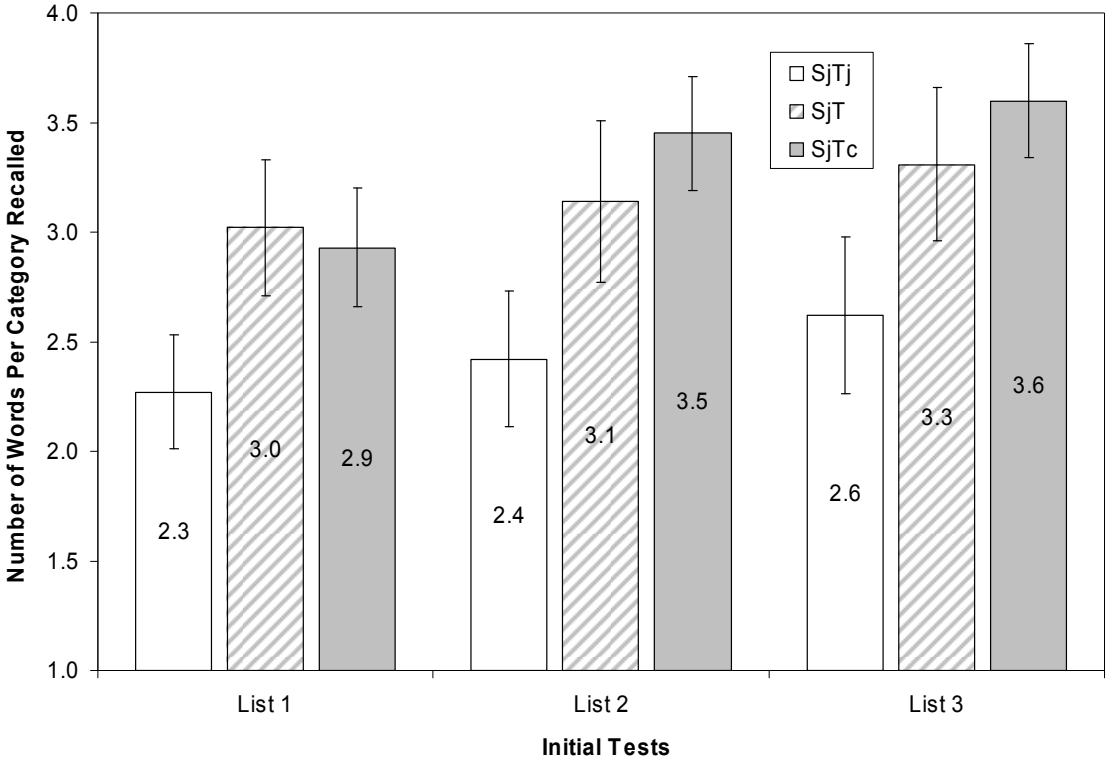
the S_jS condition (16.96 vs. 15.33), $t(46) = 2.86$, $SEM = .57$, $d = .83$, and S_jT_j condition (16.96 vs. 14.92), $t(46) = 3.63$, $SEM = .56$, $d = 1.30$. In addition, Rc was superior in the S_jT condition as compared to the S_jT_j condition (16.33 vs. 14.92), $t(46) = 2.28$, $SEM = .62$, $d = .66$. No other individual pair-wise comparisons were statistically significant.

Similar to the pattern of results obtained in word recall performance, testing during the initial learning phase only enhanced Rc in delayed recall when subjects were explicitly instructed to semantically organize their responses during initial free recall tests. When subjects were initially tested with standard free recall instructions or with instructions to organize responses according to their assigned judgment tasks, Rc following a long delay was not significantly better than in the repeated study condition.

Recall of Items Within Categories. Figure 3 shows that the mean number of words recalled within accessed categories (Rw/c) from each list increased across the lists during the initial phase. Rw/c also varied as a function of the retrieval conditions during testing, with greatest Rw/c in the S_jT_c condition, followed by the S_jT and S_jT_j conditions. An ANOVA confirmed a significant effect of list, $F(2,138) = 6.62$, $MSE = .53$, $\eta_p^2 = .09$, due to higher Rw/c in the third list as compared to the first list (3.18 vs. 2.74), $t(71) = 3.51$, $SEM = .12$, $d = .53$. Neither the difference in Rw/c between the second and first lists (3.00 vs. 2.74), $t(71) = 2.13$, $SEM = .12$, $d = .32$, *ns*, nor between the second and third lists was statistically significant (3.00 vs. 3.18), $t(71) = 1.55$, $SEM = .11$, $d = .20$, *ns*.

In addition, there was a significant effect of learning condition, $F(2,69) = 24.09$, $MSE = .67$, $\eta_p^2 = .41$, due to enhanced Rw/c in the S_jT_c and S_jT conditions relative to the S_jT_j condition (3.33 vs. 2.44), $t(46) = 6.91$, $SEM = .13$, $d = 2.00$, and (3.16 vs. 2.44), $t(46) = 5.05$, $SEM = .14$, $d = 1.45$, respectively. The difference in Rw/c between the S_jT_j and

Figure 3.


4.0
3.5
3.0
2.5
2.0
1.5
1.0
Number of Words Per Category Recalled
SjTj
SjT
SjTc
2.3
3.0
2.9
2.4
3.1
3.5
2.6
3.3
3.6
List 1
List 2
List 3
Initial Tests

S_jT_c conditions was not significant (3.16 vs. 3.33), $t(46) = 1.22$, $SEM = .14$, $d = .36$, *ns*. There was no interaction between test trial and learning condition, $F < 1$.

The third row of Table 4 shows that for delayed tests of free and cued recall Rw/c was superior in the S_jT_c condition, followed by the S_jT condition, and poorest in the S_jS and S_jT_j conditions. An ANOVA confirmed a significant effect of learning condition, $F(3, 92) = 6.48$, $MSE = .52$, $\eta_p^2 = .17$. However, there was neither a significant effect of test type, $F(1, 92) = 3.30$, $MSE = .05$, $\eta_p^2 = .04$, *ns*, nor a significant interaction between the two factors, $F < 1$.

Individual pair-wise comparisons revealed that in free recall, Rw/c was greater in the S_jT_c condition relative to the S_jS condition (2.65 vs. 2.17), $t(46) = 2.78$, $SEM = .17$, $d = .80$, S_jT_j condition (2.65 vs. 2.02), $t(46) = 5.18$, $SEM = .12$, $d = 1.50$, and S_jT condition (2.65 vs. 2.34), $t(46) = 2.29$, $SEM = .13$, $d = .66$, however, the latter difference was not statistically significant. In addition, Rw/c was superior in the S_jT condition as compared to the S_jT_j condition (2.34 vs. 2.02), $t(46) = 2.42$, $SEM = .13$, $d = .70$.

Similarly, in cued recall, Rw/c was enhanced in the S_jT_c condition relative to the S_jS condition (2.54 vs. 2.14), $t(46) = 2.44$, $SEM = .16$, $d = .71$, and S_jT_j condition (2.54 vs. 1.93), $t(46) = 4.39$, $SEM = .14$, $d = 1.28$. Rw/c was also greater in the S_jT condition as compared to the S_jT_j condition (2.35 vs. 1.93), $t(46) = 2.94$, $SEM = .14$, $d = .87$. No other individual pair-wise comparisons were statistically significant.

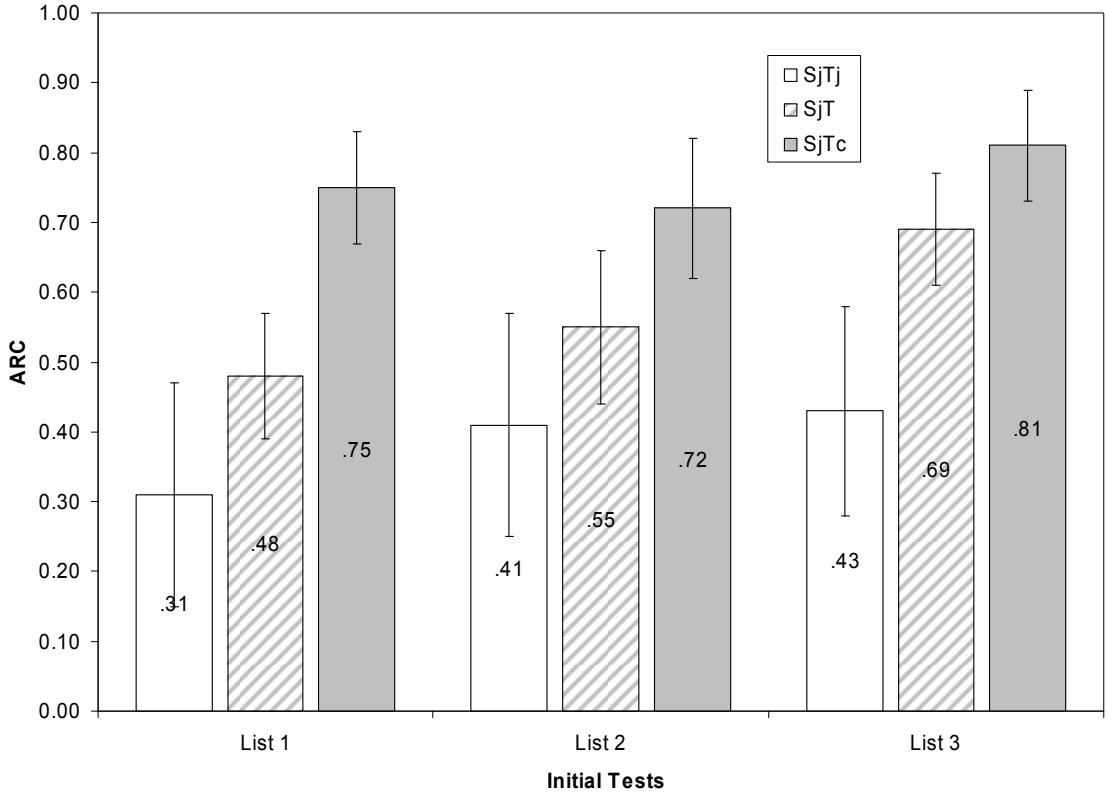
Again, testing during the initial learning phase only enhanced Rw/c in delayed free and cued recall when subjects were explicitly instructed to semantically organize their responses during initial free recall testing. When subjects were initially tested with standard free recall instructions or with instructions to organize responses according to

their assigned encoding tasks, Rw/c following a long delay was similar to the repeated study condition.

Category Clustering. Figure 4 shows that category clustering, as measured by ARC, increased across the lists in the learning phase. Category clustering also varied as a function of the retrieval conditions during testing, with greatest clustering in the S_jT_c condition, followed by the S_jT and S_jT_j conditions. An ANOVA confirmed a significant effect of list, $F(2,138) = 3.95$, $MSE = .08$, $\eta_p^2 = .05$, which was due to greater ARC scores in the third as compared to the first list (.64 vs. .51), $t(71) = 3.16$, $SEM = .04$, $d = .41$. Thus, consistent with previous research, organization improved across lists despite constant recall performance (Thompson, & Roenker, 1971). However, neither the difference in category clustering between the first and second, nor between the second and third lists was significant, (.51 vs. .56), $t < 1$, and (.56 vs. .64), $t(71) = 1.70$, $SEM = .05$, $d = .25$, *ns*.

The ANOVA further revealed a significant effect of learning condition, $F(2,69) = 29.23$, $MSE = .09$, $\eta_p^2 = .46$, which was due to enhanced clustering in the S_jT_c and S_jT conditions relative to the S_jT_j condition (.76 vs. .39), $t(46) = 7.92$, $SEM = .05$, $d = 2.27$, and (.57 vs. .39), $t(46) = 3.50$, $SEM = .05$, $d = .97$, respectively. Category clustering was also greater in the S_jT_c as compared to the S_jT condition (.76 vs. .57), $t(46) = 4.08$, $SEM = .05$, $d = 1.21$. There was a non-significant interaction between test trial and learning condition, $F < 1$. In delayed free recall, the fourth row of Table 4 shows that category clustering was highest in the S_jT_c condition, followed by the S_jT condition, and poorest in the S_jS and S_jT_j conditions. However, an ANOVA revealed that ARC scores did not significantly vary as function of learning condition, $F < 1$.

Figure 4.


1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
ARC
SjTj
SjT
SjTc
.31
.48
.75
.41
.55
.72
.43
.69
.81
List 1
List 2
List 3
Initial Tests

As expected, whereas instructing subjects to semantically organize responses during initial tests of free recall produced greater category clustering than standard free recall testing, instructing subjects to organize responses according to their assigned judgment tasks reduced category clustering in immediate free recall. Nevertheless, manipulating the retrieval conditions during initial testing did not reliably affect clustering in delayed free recall.

Subjective Organization. Another form of recall organization that may be affected by testing is subjective organization (Mulligan, 2002). Even with the use of categorized lists, subjects may tend to adopt idiosyncratic forms of conceptual organization to chunk list items into higher order subjective units, or they may adopt uniform organization within category recall. Subjective organization was measured using pair frequency (PF; Sternberg and Tulving, 1977). Again, PF represents the number of pairs of items commonly recalled on adjacent test trials in adjacent output positions in either forward or reverse order.

Keep in mind that in the current experiment subjects took recall tests on three separate categorized word lists during the initial learning phase and a day later took a final, delayed free recall test on all three lists together. It was therefore necessary to combine the output protocols from the initial tests into one output protocol representing free recall during the initial learning phase and then to calculate a PF score for each subject based upon the number of pairs of items commonly recalled across the combined initial tests and single delayed test of final free recall. As a minimum of two recall trials are required to compute PF, it was not possible to measure subjective organization in the Study-only condition.

The fifth row of Table 4 shows that mean PF scores measured between the combined initial tests of free recall and the delayed final free recall test were highest in the S_jT_c condition, followed by the S_jT condition, and poorest in the S_jT_j condition. An ANOVA revealed a significant effect of learning condition, $F(2,69) = 9.16$, $MSE = 10.32$, $\eta_p^2 = .21$, which was due to enhanced PF scores in the S_jT_c condition relative to the S_jT_j condition (6.25 vs. 2.29), $t(46) = 4.49$, $SEM = .88$, $d = 1.29$. Although PF scores were greater in the S_jT_c as compared to the S_jT condition (6.25 vs. 4.17), and higher as well in S_jT relative to the S_jT_j condition (4.17 vs. 2.29), these differences were not statistically significant, $t(46) = 1.97$, $SEM = 1.06$, $d = .57$, *ns*, and $t(46) = 2.29$, $SEM = .82$, $d = .66$, respectively.

In addition, whereas ARC scores were moderately correlated with delayed recall performance ($r = .37$), PF scores were highly correlated with delayed recall ($r = .79$). The PF measure captures a form of organization that is more highly correlated with delayed recall than the ARC measure, despite the fact we used categorized lists. This outcome supports the hypothesis that even though category clustering was high for all groups of subjects, differences in later recall among the S_jT_j, S_jT_c, and S_jT conditions were more highly correlated with consistent responding in recall of items within and across categories, as measured by PF. Enhanced organization may responsible, at least in part, for the testing effect in free recall.

Intrusions. We further examined recall accuracy by measuring the proportion of all words recalled in delayed tests of free and cued recall that were words not presented during the course of the experiment (extra-list intrusions). Extra-list intrusions (XLIs)

were seldom committed during the initial learning phase—on average, subjects only committed .30 XLIs per recall trial—and were, therefore, not included in the analyses.

The bottom row of Table 4 shows that category cueing increased the commission of extra-list intrusions relative to free recall across all learning conditions, while testing during the learning phase reduced false recall in the delayed tests. An ANOVA confirmed that a greater proportion of extra-list intrusions were committed in cued relative to free recall (.21 vs. .15), $F(1,92) = 28.00$, $MSE = .01$, $\eta_p^2 = .23$. There was a significant effect of learning condition $F(3,92) = 3.32$, $MSE = .03$, $\eta_p^2 = .10$, due to a lower proportion of extra-list intrusions committed in the S_jT as compared to the Study-only condition in both free recall (.10 vs. .23), $t(46) = 2.81$, $SEM = .05$, $d = .77$, and cued recall (.16 vs. .25), $t(46) = 2.50$, $SEM = .03$, $d = .78$. However, the interaction between the two factors did not reach the conventional level of statistical significance, $F(3,92) = 2.19$, $MSE = .01$, $\eta_p^2 = .07$, $p = .10$. No other comparisons were significant. Consistent with the findings of Experiment 1 and prior work (Szpunar et al., 2008), standard free recall testing during the initial learning phase reduced false recall as compared to studying alone following a long delay.

Discussion

The purpose of this experiment was to investigate whether organizational processes contribute to the testing effect in free recall. One key finding was that, after study of a categorized word list, manipulating organizational processing during an immediate test of free recall affected retention of the list a day later. Relative to a standard free recall condition, semantically organizing responses by taxonomic categories

produced greater category clustering, and organizing responses by judgment tasks produced poorer category clustering in initial recall tests taken during the learning phase.

Testing only enhanced long-term retention, however, when subjects semantically organized their initial recall responses by categories. Under these conditions, testing enhanced performance in delayed tests of free and cued recall by improving both category access (Rc) and recall of items within accessed categories (Rw/c). By contrast, when subjects attempted to organize responses during the initial free recall tests according to previously assigned judgment tasks, long-term retention and category clustering were no better than restudying. Taken together, these findings suggest that the positive effect of testing on long-term retention depends upon the organizational processing that occurs during testing. When test conditions during the initial learning phase fostered the use of semantic relational information to guide episodic recall, testing enhanced long-term retention. However, when initial test conditions interfered with semantic organizational processing by requiring subjects to organize information according to arbitrary associations among list items, the testing effect disappeared.

These findings also raise a puzzle. When subjects were initially tested with standard free recall instructions, long-term retention was not significantly better than restudying. This outcome is surprising, because numerous studies have demonstrated significant positive effects of prior recall testing on long-term retention relative to restudying (see Roediger & Karpicke, 2006b). Moreover, the standard free recall (S_jT) condition was nearly identical to the study with pleasantness rating + testing (S_pT) condition in Experiment 1, and yet a testing effect only obtained in the latter condition.

One reason for the divergent findings may have to do with the fact that the two standard free recall testing conditions in the current and first experiment only differed in the judgment tasks performed during the initial study trial. Whereas in the current experiment subjects performed 6 different judgment tasks in a randomized order during list presentation, subjects in Experiment 1 only made pleasantness ratings. In performing the six different judgment tasks during list presentation, subjects had to exert considerable attention and cognitive control to frequently switch among judgment tasks and focus on the particular semantic attributes of each list item relevant to its assigned judgment task. Such tasks require extensive item-specific processing, and as a result, subjects may have had more difficulty than in Experiment 1 processing inter-item relational information to facilitate list recall.

Consistent with this hypothesis, recall performance and mean ARC scores for the initial recall tests were lower in the standard free recall condition of the current experiment (52% of words were recalled from each list with a mean ARC score of .57) than in the S_pT condition of Experiment 1 (68% of words were recalled from each list with a mean ARC score of .79). However, a testing effect did occur in the recall by categories (S_jT_j) condition with similarly low initial test performance but greater category clustering than the standard free recall condition (51% of words were recalled from each list with a mean ARC score of .76). The absence of a testing effect in the standard free recall condition may still be partly due to the fact of lower initial recall performance permitting subjects to have re-exposure to only 52% of the words they recalled during the test trial as compared to the repeated study condition which re-exposed subjects to 100% of the words during the second study trial. Nevertheless, it appears that the primary factor

determining the presence of a testing effect was the degree of organization achieved during initial recall.

As in Experiment 1, our primary interest was in examining the effects of testing conditions on long-term recall organization. Although varying organizational processing during initial testing influenced recall organization in the initial learning phase, category clustering in delayed free recall did not differ across the learning conditions. However, when we used the more subtle pair frequency measure of subjective organization, we found significant differences among conditions. That is, increased organization during initial testing produced greater consistency in recall across initial and delayed recall tests (measured by PF) without affecting category clustering (measured by ARC). Moreover, ARC scores were only moderately correlated with delayed recall (as in Experiment 1), while PF scores were highly correlated with delayed recall. Thus, even with the use of categorized lists, subjects may have adopted idiosyncratic forms of conceptual organization to chunk list items into higher order subjective units, or they may have adopted uniform organization within category recall.

The strong correlation between PF scores and delayed recall suggests that the processes underlying subjective organization may also contribute to the positive effects of testing on long-term retention (see also Zaromb & Roediger, in press). Moreover, the finding that delayed recall and PF scores were also correlated with ARC scores indicates that subjects may have adopted complementary retrieval schemas based upon their categorical knowledge (ARC) and recollection of previous recall attempts (PF) to guide episodic recall.

Experiment 3

Experiments 1 and 2 demonstrated that the testing effect in free recall may be due in part to enhanced organizational processes, as reflected in measures of category clustering (ARC) and subjective organization (PF). While these findings may hold true for study materials that are conceptually structured such as categorized word lists, it is unclear whether testing affects recall organization and, if so, whether organizational processes mediate the benefits of testing on long-term retention using study materials that lack a coherent conceptual structure, such as unrelated word lists.

Experiment 3 further examined the effects of testing on long-term retention and organization in free recall using both unrelated and categorized word lists. As mentioned earlier, previous studies have demonstrated that even when asked to learn a list of seemingly unrelated words, individuals tend to recode groups of items into higher-order subjective units, and that this organizing tendency, which is referred to as subjective organization, is predictive of free recall (Mandler, 1967; Tulving, 1962).

If the benefits of testing on long-term retention are associated with subjective organizational processes, then testing individuals' recall of seemingly unrelated words during learning should still produce superior recall following a long delay relative to restudying, and measures of subjective organization (PF) should be correlated with recall. To the extent that individuals also utilize categorical knowledge to guide episodic recall, the testing effect should be further enhanced for previously categorized word lists, and measures of category clustering (ARC) should also be correlated with recall performance.

In order to simultaneously test these predictions, we manipulated the organization of the study materials using lists of words representing ad-hoc categories (e.g., Barsalou, 1983, 1985), such as "things dogs chase" or "weekend entertainment", under conditions

in which subjects were either aware or unaware of the categorical structure of the lists during learning. In contrast to taxonomic categories whose knowledge structures are presumably well-established in long-term memory and may be automatically activated and brought to mind when particular category instances are encoded and/or retrieved, ad-hoc categories represent disparate knowledge that becomes organized into coherent categories in particular situations to achieve goal-relevant tasks.

When individuals are presented with a list of words representing ad-hoc categories without being informed of the list's categorical structure, the words may appear to be unrelated. However, when individuals are informed about the ad-hoc categories, they can readily organize the list items according to these categories. Similar to Experiments 1 and 2, we also manipulated testing conditions by assigning different groups of subjects to conditions in which they either studied a word list for two consecutive study trials or took a recall test following an initial study trial. All groups took final delayed tests of free and cued recall a day later.

Method

Subjects. 80 Washington University undergraduates participated for either payment or for course credit.

Design. The experiment followed a 2 Learning condition (Study-only vs. Study-Test) X 2 Conceptual Awareness (Aware vs. Unaware) between-subjects design with 20 subjects assigned to each of the four conditions: Study-only Aware (SS_A), Study-only Unaware (SS_U), Study-Test Aware (ST_A), and Study-Test Unaware (ST_U). Half of the subjects were (SS_A and ST_A) and the other half were not (SS_U and ST_U) presented with the names of ad-hoc categories corresponding to each list during the initial study trial. In

the Study-only conditions, subjects studied a list on two consecutive trials. In the Study-Test conditions, subjects took a free recall test following an initial study trial. The assignment of the 4 learning conditions to the 2 study lists and the order of list presentation were counterbalanced such that each study list was assigned to each of the 4 conditions and presented as either the first or second list an equal number of times across subjects.

Materials. 40 words were sampled from 8 ad-hoc categories (5 words per category) reported in Barsalou (1985), Little, Lewandowsky, and Heit (2006), and Vallée-Tourangeau, Anthony, and Austin (1998) to create 2, 20-word study lists (see Appendix 2 for the lists of words and ad-hoc categories). The 20 words in each list included 5 medium frequency nouns belonging to each of 4 ad-hoc categories.

Procedure. Subjects participated in two sessions scheduled 1 day apart. In the first session, subjects were informed that they would study and attempt to recall several lists of words presented by a computer. During an initial study trial, the computer displayed each word one at a time for 8 seconds, followed by a 1 second inter-stimulus interval. Words were presented in randomized order on each study trial. In the Study-only Aware and Study-Test Aware conditions, the computer also displayed the names of the 4 ad-hoc categories represented in the list at the bottom of the computer screen and numbered 1 through 4, and subjects were informed that each word in the study list belonged to one of the categories. As each item was displayed on the computer screen, subjects were instructed to type a number between 1 and 4 indicating to which ad-hoc category the item belonged. In the Study-only and Study-Test Unaware conditions, subjects were instructed

to study the list in preparation for a later memory test without being shown the names of the ad-hoc categories during list presentation.

In the second study trial of the Study-only (SS_A and SS_U) conditions, the list of words was presented again (without the ad-hoc category names) in a new random order and subjects were given standard intentional learning instructions. The total study time was 3 minutes per trial. In the Study-Test (ST_A and ST_U) conditions, the initial study trial was followed by a test of free recall in which subjects had 3 minutes to write down as many words on a blank sheet of paper as they could remember from the most recently studied list in any order that the words came to mind. The same procedure for the study and test trials was repeated for the second list of words. This first session lasted approximately 30 minutes.

Following a 1-day retention interval, subjects took final tests of free and cued recall. During the free recall test, subjects had 10 minutes to write down on a blank sheet of paper as many words as they could remember from the two lists in any order that the words came to mind. Last, subjects had 10 minutes to again recall words from both lists; however, in contrast to the previous test, subjects were also provided a list of all of the ad-hoc category names to aid recall of the words. The second session lasted 20 minutes.

Results

Recall of Words. The top two rows of Table 5 show that recall performance during the initial learning phase, measured as the proportion of words recalled from each study list, was similar for the first and second lists (.78 vs. .81), $F(1,38) = 2.08$, $MSE = .01$, $\eta_p^2 = .05$, *ns*, and was not affected by subjects' awareness of the ad-hoc categories during learning, $F < 1$.

Table 5. *Mean proportion of words recalled and adjusted ratio of clustering (ARC) scores for the Aware and Unaware learning conditions in initial tests of free recall in Experiment 3.*

		Initial Tests			
		List 1		List 2	
Measure	Condition	*M*	CI	*M*	CI
Recall	Unaware	.77	(.06)	.79	(.06)
	Aware	.80	(.06)	.83	(.05)
ARC	Unaware	.02	(.10)	.02	(.06)
	Aware	.29	(.14)	.51	(.16)

Note—Values in parentheses are 95% confidence intervals (CI).

For delayed tests of free and cued recall taken a day later, the top row of Table 6 shows that both providing organizational information (ad-hoc category names) and testing during the learning phase improved long-term retention. Performance was highest in the ST_A condition, followed by the ST_U and SS_A conditions, and poorest in the SS_U condition. We conducted a 2 (Test Type: Free vs. Cued Recall) X 2 (Learning condition: Study-only vs. Study-Test) X 2 (Conceptual Awareness: Aware vs. Unaware) ANOVA, which revealed superior retention in cued relative to free recall, (.57 vs. .48), $F(1, 76) = 68.70$, $MSE = .00$, $\eta_p^2 = .48$. In addition, there was a significant benefit of testing as compared to repeated study (.62 vs. .42), $F(1, 76) = 27.27$, $MSE = .06$, $\eta_p^2 = .26$. Although providing subjects with the names of the ad-hoc categories during study did not affect initial recall performance, providing this organizational information during learning significantly improved delayed recall relative to withholding this organizational information (.62 vs. .43), $F(1, 76) = 25.08$, $MSE = .06$, $\eta_p^2 = .25$.

In addition, there was a significant interaction between test type and learning condition, $F(1, 76) = 11.67$, $MSE = .00$, $\eta_p^2 = .13$, due to a larger testing effect in cued relative to free recall. Similar to Experiments 1 and 2, the cued recall test followed free recall, which raises the possibility that this interaction may be complicated by carryover effects from free recall. Last, there was neither a significant interaction between test type and conceptual awareness, $F < 1$, between learning condition and conceptual awareness, $F(1, 76) = 1.75$, $MSE = .06$, $\eta_p^2 = .02$, *ns*, nor was there a significant interaction among the three factors, $F(1, 76) = 1.95$, $MSE =$, .00, $\eta_p^2 = .03$, *ns*. Thus, testing improved long-term free and cued recall relative to restudying, and recall was further enhanced when subjects organized list items into their corresponding ad-hoc categories during learning.

Table 6. *Mean proportion of words recalled, adjusted ratio of clustering (ARC) scores, pair frequency (PF) scores, and proportion of recalled words that were extra-list intrusions (XLIs) for the Study-only Aware (SS_A), Study-only Unaware (SS_U), Study-Test Aware (ST_A), and Study-Test Unaware (ST_U) conditions in delayed tests of free and cued recall in Experiment 3.*

		Free Recall				Cued Recall			
Measure		SS_U	SS_A	ST_U	ST_A	SS_U	SS_A	ST_U	ST_A
Recall	*Prop.*	.25	.47	.52	.67	.35	.62	.58	.71
	CI	(.09)	(.10)	(.07)	(.09)	(.07)	(.05)	(.06)	(.08)
ARC	*M*	.01	.32	.21	.63				
	CI	(.29)	(.12)	(.06)	(.09)				
PF	*M*			3.43	5.11				
	CI			(1.32)	(1.74)				
XLIs	*Prop.*	.28	.18	.07	.02	.29	.13	.09	.04
	CI	(.11)	(.13)	(.04)	(.02)	(.09)	(.06)	(.06)	(.03)

Note—Values in parentheses are 95% confidence intervals (CI).

Category Clustering. As shown in the bottom two rows of Table 5, mean category clustering (ARC) scores for recall tests performed in the learning phase were greater than chance for subjects in the Aware (ST_A) condition who organized list items into their corresponding ad-hoc categories during the initial study trial. Further, ARC scores increased across the two lists. An ANOVA confirmed that category clustering was enhanced in the ST_A relative to the ST_U condition (.40 vs. .02), $F(1,38) = 28.37$, $MSE = .10$, $\eta_p^2 = .43$. ARC scores were also greater in second list as compared to the first list (.26 vs. .16), $F(1,38) = 4.37$, $MSE = .10$, $\eta_p^2 = .43$. Moreover, there was a significant interaction between the two factors, $F(1,38) = 4.77$, $MSE = .05$, $\eta_p^2 = .11$, which was due to an increase in ARC scores across test trials for the ST_A (.22), but not in the ST_U condition (.00). This finding is consistent with results of Experiment 2 and previous research using categorized word lists showing a "learning to cluster" effect in which organization improves across tests of immediate free recall despite constant recall performance (Thompson, & Roenker, 1971).

Following a 24-hour retention interval, the second row of Table 6 shows that prior testing during the initial learning phase improved category clustering in delayed free recall relative to restudying, and clustering was further enhanced when subjects were made aware of the categorical structure of the study lists. Category clustering was greatest in the ST_A condition, followed by the SS_A and ST_U conditions, and poorest in the SS_U condition. An ANOVA confirmed that testing enhanced output organization relative to restudying (.42 vs. .16), $F(1,72) = 9.58$, $MSE = .13$, $\eta_p^2 = .12$. Providing organizational information during study also improved category clustering relative to withholding this information during study (.48 vs. .11), $F(1, 72) = 20.48$, $MSE = .13$, $\eta_p^2 = .22$. There was

no interaction between the two factors, $F < 1$. In addition, ARC scores across all four conditions were positively correlated with delayed recall ($r = .46$). Thus, testing improved the organization of recall, and organization was correlated with the number of words recalled. Recall organization was further enhanced when subjects organized list items during study into their corresponding ad-hoc categories.

Subjective Organization. If the testing effect in free recall is associated with enhanced organization, then why did testing improve delayed recall when subjects were not made aware of the categorical structure of the lists and category clustering was near chance and uncorrelated with recall ($r = .17$, *ns*)? The answer is probably that subjects may have adopted idiosyncratic forms of organization, or subjective organization, to learn and remember list items.

To examine this possibility, we measured subjective organization using pair frequency (PF; Sternberg and Tulving, 1977). Similar to Experiments 1 and 2, subjects took separate recall tests on each categorized word list during the initial learning phase and a day later took a final, delayed free recall test on all three lists together. It was therefore necessary to combine the output protocols from the initial tests into one output protocol representing free recall during the initial learning phase and then to calculate a PF score for each subject based upon the number of pairs of items commonly recalled across the combined initial tests and single delayed test of final free recall. As a minimum of two recall trials are required to compute PF, it was not possible to measure subjective organization in the Study-only condition.

The third row of Table 6 shows that mean PF scores measured between the combined initial tests of free recall and the delayed final free recall test in the ST_U and

ST_A conditions were much higher than chance (which is zero). Although PF scores were numerically greater in the ST_A relative to the ST_U condition, the difference was not statistically significant (5.11 vs. 3.43), $t(38) = 1.54$, $SEM = 1.09$, $d = .49$, *ns*. Consistent with Experiment 2, whereas ARC scores in the ST_U and ST_A conditions were moderately correlated with delayed recall ($r = .39$), PF scores were highly correlated with delayed recall ($r = .68$). The PF measure captures a form of organization that is more highly correlated with delayed recall than the ARC measure, especially for seemingly unrelated materials.

Intrusions. We further examined recall accuracy by measuring the proportion of all words recalled in delayed tests of free and cued recall that were words not presented during the course of the experiment (extra-list intrusions). Extra-list intrusions (XLIs) were seldom committed during the initial learning phase—on average, subjects only committed .08 XLIs per recall trial—and these were, therefore, not included in the analyses.

The bottom row of Table 6 shows that testing during the learning phase reduced false recall in the delayed tests of free and cued recall, and that making subjects aware of the categorical structure of the study lists also reduced false recall. An ANOVA confirmed that testing significantly reduced false recall of XLIs relative to restudying (.06 vs. .22), $F(1, 76) = 20.37$, $MSE = .05$, $\eta_p^2 = .21$. Providing organizational information also reduced false recall relative to withholding this information during study (.10 vs. .18), $F(1, 76) = 5.85$, $MSE = .05$, $\eta_p^2 = .07$. However, there was a non-significant effect of test type (free vs. cued recall), $F < 1$. No interaction effects were significant.

In sum, consistent with the findings of Experiments 1 and 2 and prior work (Szpunar et al., 2008; Zaromb & Roediger, in press) free recall testing during the initial learning phase reduced false recall after a delay. False recall was also reduced when subjects were made aware of the categorical structure of the word lists before initial study.

Discussion

This experiment demonstrated powerful effects of testing (relative to restudying) on long-term retention and recall organization. Consistent with the first experiment, studying a list and taking an immediate recall test produced greater recall and reduced false recall of extra-list intrusions a day later compared to restudying the list. Recall was further improved when subjects were informed of the conceptual structure of the lists and required to organize list items according to their corresponding categories during study. Under these learning conditions, testing also enhanced category clustering, measured by ARC, just as it did in Experiment 1 and in prior work (Masson & McDaniel, 1981).

Not surprisingly, when subjects were uninformed of the ad-hoc categorical structure of the word lists, the lists appeared as sets of unrelated words, and long-term retention was poorer than in conditions where subjects were informed of the categorical structure. This finding is consistent with prior work and serves as a powerful demonstration of the benefits of organization or meaningful learning on long-term retention (e.g., Asch, 1969; Katona, 1940; Mandler, 1967). More importantly, in the SS_U and ST_U conditions, category clustering was near chance levels and uncorrelated with recall, and yet testing still enhanced recall following a long delay.

This testing effect arose in part because subjects adopted personal idiosyncratic forms of conceptual organization, or subjective organization, to facilitate learning and episodic recall. Indeed, using the pair frequency measure of subjective organization, a high degree of consistency in recall was observed across initial and delayed tests. Further, recall was strongly correlated with PF scores regardless of whether or not subjects were initially informed of the categorical structure of the word lists. Replicating one of the outcomes of Experiment 2, even when subjects were initially informed of the ad-hoc categories, and category clustering was above chance levels, delayed recall was more highly correlated with PF scores than with ARC scores. Taken together, these findings provide further evidence that the processes underlying subjective organization contribute to the positive effects of testing on long-term retention.

General Discussion

Three experiments confirmed the positive effects of testing to enhance long-term retention relative to restudying categorized word lists. Studying a list and taking an immediate recall test produced greater recall and reduced the false recall of extra-list intrusions a day later compared to conditions in which subjects repeatedly studied the list. The main novel finding of our experiments is that the benefits of testing were also associated with enhanced recall organization, as reflected in measures of category clustering (Experiments 1 and 3) and subjective organization (Experiments 2 and 3). Moreover, manipulating the organizational processing that occurred during initial study (Experiment 3) and test trials (Experiment 2) was found to modulate the effects of testing on long-term retention and recall organization. Taken together, these findings provide

further striking evidence for the power of testing (Roediger & Karpicke, 2006b) and help to provide an understanding of why testing effects occur, at least in free recall.

Testing Enhances Organizational and Item-Specific Processing

Our primary objective was to investigate whether the benefits of testing extended to individuals' learning of conceptual organization relative to studying alone, a question that had not yet been addressed in the literature. First, we asked what components of recall were improved by testing relative to studying alone – access to higher order units, access to items within units, or both. In Experiment 1, the last option was confirmed because testing benefited both measures of category access (Rc) and recall of items within each accessed category (Rw/c) in delayed tests of free and cued recall.

If individuals learn categorized word lists by chunking items into category-based units, then once they can access the units during retrieval, their contents (the individual items) will be accessed as well to some degree. Moreover, many researchers have demonstrated that Rc and Rw/c are largely independent of each other—whereas experimental conditions designed to promote organizational processing (e.g., instructing subjects to organize study items, providing category names during study) have been found to selectively increase Rc, those designed to enhance item-specific processing (e.g., generating study items) have been shown to increase Rw/c (e.g., Cohen, 1963, 1966; McDaniel et al., 1988; Schmidt & Cherry, 1989). To the extent that these measures assess the extent to which relational (Rc) and item-specific (Rw/c) information is used to guide episodic retrieval (e.g., Hunt & Seta, 1984), then our findings show that testing may promote both relational and item-specific processing relative to studying alone.

It is worth noting that several other studies have corroborated the notion that testing enhances item-specific processing. Karpicke and Zaromb (2010) recently found that testing enhances memory for previously read list items on final tests of recall and recognition relative to passively re-reading or actively generating the items. They also showed that these effects are robust in both within- and between-subjects experimental designs (unlike the generation effect). They argued that testing may enhance item-specific processing that constrains retrieval to the set of list items to be remembered on a later test.

This explanation is consistent with our finding in all three experiments that testing reduced the false recall of extra-list intrusions relative to restudying. Moreover, when subjects in Experiment 1 falsely recalled extra-list intrusions, over 80% of these intrusions were other category exemplars, which suggests that testing may reduce false recall by constraining retrieval to the target category exemplars. Gallo and Roediger (2002) showed a similar effect in that recall testing of previously studied associate (DRM) lists reduced later false recognition. They argued that testing enhanced the recollective distinctiveness of list items, which, in turn, reduced false recognition on a later final test (see also Brewer, Marsh, Meeks, & Clark-Foos, 2010). Taken together, one might argue that it is the combination of these two types of processing—relational and item-specific—that produces superior retention and underlies the positive effects of testing on long-term retention (Hunt, 2006; Matthews, Smith, Hunt, & Pivetta, 1999; see also Kühn, 1914, p. 443).

One criticism with interpreting Rc and Rw/c as measures of organizational and item-specific processing is that they do not adjust for differences in recall performance

across individuals or learning conditions (Burns & Brown, 2000; Murphy, 1979). For instance, Burns and Brown (2000) have argued for the use of the adjusted category access ratio (ACA) and adjusted items per category recalled ratio (AIPC) in conjunction with Rc and Rw/c, because these measures do correct for recall-level differences (see Burns & Brown, 2000, for details). ACA and AIPC scores of zero indicate chance-level Rc and Rw/c scores, respectively, and scores above zero indicate that Rc and Rw/c scores are greater than expected by chance alone.

We applied Burns and Brown's (2000) measures to our data and obtained curious outcomes. In Experiment 1, access to categories (ACA, the corrected version of Rc) was well below chance in final recall in both the non-tested and tested conditions. Further, corrected access of items within categories (AIPC, the corrected version of Rw/c) was near chance levels in the non-tested conditions and above chance in the tested condition.

These findings raise questions, one of which is the interpretation of "below chance" access of categories during free recall of categorized lists. Burns and Brown (2000) argued that negative ACA scores indicate that during free recall subjects attempt to exhaustively recall items within a category before transitioning to items from another category. As a result, subjects are likely to access fewer categories but recall more words per accessed category than that expected by chance alone given their recall level. This interpretation may also help to explain the pattern of results obtained in Experiment 2. Although recall performance remained constant for the three tests taken during the learning phase of Experiment 2, ARC scores and Rw/c increased, while Rc declined. In other words, as recall became more organized, subjects accessed fewer categories and recalled more words per accessed category.

Nevertheless, the finding of negative ACA scores gives one pause about the assumptions being used in the measure. If subjects are obviously using organized recall, then perhaps the estimate of "chance" is too high in these measures (hence leading the data to appear to be below chance). Our preferred use of Rc and Rw/c measures is the same as that of Tulving and Pearlstone (1966) and many others, as descriptive measures: Total recall of categorized lists can be decomposed into two components that bear a multiplicative relationship (i.e., recall of words or Rw = Rc x Rw/c). The Rc and Rw/c measures are, by definition, components of overall recall and do not need to be corrected for descriptive purposes. On the other hand, future research may indeed show that Hunt and Seta's (1984) interpretation of Rc and Rw/c as reflecting relational and item-specific processing may be in need of re-examination, as Burns and Brown (2000) claim.

A second question we asked was whether testing improves recall organization. In Experiment 1, we found that testing produced greater category clustering relative to restudying, and organization was correlated with delayed recall. These effects were replicated in Experiment 3 under conditions in which subjects were informed of the categorical structure of the study lists during the initial learning phase and utilized this categorical knowledge to guide recall a day later. These findings provide evidence that testing enhances organizational processes, and they further suggest that organizational processes may directly contribute to the testing effect in free recall.

Organizational Processes Modulate the Testing Effect

Experiments 2 and 3 further examined whether processes involved in mentally organizing information during study and test trials contribute to the testing effect in free recall. In both experiments, we found that manipulating organizational processing during

the initial phase modulated the effects of testing on long-term retention and recall organization. In Experiment 2, testing only significantly enhanced long-term retention when subjects semantically organized their initial recall responses. By contrast, when subjects attempted to organize responses during the initial free recall tests according to previously assigned judgment tasks, long-term retention and category clustering were not appreciably better than restudying. In Experiment 3, studying a list of words from ad-hoc categories and taking an immediate test of free recall enhanced long-term retention compared to restudying the list. More importantly, delayed recall was further improved when subjects were informed of the conceptual structure of the list and required to organize list items according to their corresponding categories during initial study.

Taken together, these findings suggest that the positive effects of testing on long-term free recall depend in part upon the organizational processing that occurs during prior study episodes and recall tests. Testing produced superior long-term retention when study and/or test conditions during the initial learning phase fostered the use of semantic relational information to guide episodic recall. However, the testing effect was either reduced (Experiment 3) or eliminated (Experiment 2) when initial learning conditions were designed to attenuate processing of inter-item semantic relational information based on taxonomic categories by requiring subjects to organize information according to arbitrary associations among list items (Experiment 2), or by having subjects study and attempt to recall a list of seemingly unrelated words (Experiment 3).

Somewhat surprisingly, a testing effect did not occur under standard free recall test conditions in Experiment 2, a finding that stands in stark contrast to the testing effects observed in Experiments 1 and 3 under similar conditions. As mentioned earlier,

the only difference between standard free recall testing conditions in Experiments 1 and 2 lay in the types of judgment tasks performed during the initial study trial. One possible explanation is that performing six different judgment tasks during the initial study trial (as opposed to one judgment task in Experiment 1) required extensive item-specific processing and made it more difficult for subjects to process and utilize inter-item semantic relational information in the subsequent recall test trial.

Returning to Experiment 3, when subjects were uninformed of the categorical structure of the word lists of "ad hoc" items, delayed recall was poor relative to conditions in which the organizational information was provided. This finding underscores the benefits of organizational processing or meaningful learning on long-term retention (e.g., Asch, 1969; Katona, 1940; Mandler, 1967). Critically, we found that when organizational information (ad-hoc category names) were withheld from subjects, category clustering was near chance levels and uncorrelated with delayed recall, and yet testing still enhanced long-term retention of the seemingly unrelated word lists relative to restudying.

A likely explanation for this finding is that instead of utilizing categorical knowledge, subjects adopted personal idiosyncratic forms of organization, or subjective organization, to facilitate learning and episodic recall. When we used the pair frequency measure of subjective organization, we found a high degree of consistency in recall across initial and delayed tests. Further, recall was strongly correlated with PF scores regardless of whether or not subjects were initially informed of the categorical structure of the word lists. Yet, even when subjects recalled categorized word lists and category clustering was evident, delayed recall was still more highly correlated with PF scores

than with ARC scores. These findings provide further evidence that the processes underlying categorical clustering and subjective organization may independently contribute to the positive effects of testing on long-term retention. Put another way, testing appears to stimulate the development of retrieval schemas based upon both categorical knowledge (ARC) and previous recall attempts (PF) to guide and facilitate episodic recall.

Theoretical Implications

Although a growing body of research has corroborated the notion that retrieval processes in testing enhance later recall, the specific underlying mechanisms responsible for the testing effect remain unclear (e.g., Glover, 1989; Karpicke & Roediger, 2008; Karpicke & Zaromb, 2010; Pyc & Rawson, 2009). The results of our experiments advance theoretical understanding of the testing effect, at least in free recall, in showing that organizational and retrieval processes bear a reciprocal relationship.

Recall testing can stimulate organizational processing, as measured by increased category access (Rc) and output organization (ARC, PF). Testing may also enhance item-specific processing, as measured indirectly by increased recall of items within accessed categories (Rw/c) and reduced false recall of items not presented during the earlier study episode. Matthews and colleagues (1999) have argued that the benefits of testing arise through this confluence of superior organizational and item-specific processing relative to restudying. Acts of retrieval utilize relational information to organize the memory search, and item-specific information is utilized to specify target items within that search.

This interpretation may account for why recall tests tend to promote greater retention than recognition tests (Butler & Roediger, 2007; Glover, 1989; Kang et al.,

2007; McDaniel et al., 2007). Whereas recall tests require organizational and item-specific processing to guide and facilitate episodic recall, tests of item recognition rely more on item-specific processing to aid in the discrimination of target items from non-target lures. Thus, recall tests promote greater retention than recognition tests, because recall tests improve both the organizational and item-specific processing of study materials, while recognition tests primarily contribute to item-specific processing. If this view is correct, one implication is that taking tests of item recognition during a learning phase should have little or no impact on organization in delayed free recall.

Our results go a step further in showing that testing effects in free recall may be due in large part to processes involved in mentally organizing to-be-learned information. First, manipulating organizational processing during initial study episodes (Experiment 3) and test trials (Experiment 2) directly influenced the effects of testing on long-term retention and recall organization. Moreover, in all three experiments, the benefits of testing were associated with measures of recall organization (ARC and/or PF), and recall organization was predictive of recall performance.

As discussed earlier, theories and models of human memory have staked out a variety of positions on the questions of whether retrieval processes can affect recall organization, and conversely, whether organizational processes mediate the effects of retrieval on long-term retention. Our findings are generally consistent with the notion that testing fosters the development of retrieval schemas (Gates, 1917; Kühn, 1914), or retrieval plans (Slamecka, 1968), that guide and facilitate episodic recall. Depending upon the conceptual structure of the study materials and learning (study and/or test) conditions, such retrieval schemas may be based upon categorical knowledge, temporal

associations among list items, other types of semantic or non-semantic associative information, or a combination thereof.

Our main finding that testing stimulates organization also places constraints on associative theories of memory. Computational models such as FRAN (Anderson, 1972), HAM (Anderson & Bower, 1973), ACT-R (Anderson, 1996), CMR (Polyn, Norman, & Kahana, 2009), SAM (Raaijmakers & Shiffrin, 1980, 1981) along with its recent extensions eSAM (Sirotin, Kimball, & Kahana, 2005) and fSAM (Kimball, Smith, & Kahana, 2007), and TCM (Howard & Kahana, 2002) have demonstrated success in accounting for a variety of organizational phenomena observed in free recall. Although these models differ in many fundamental respects, one key feature shared by all these models is that the processing of relational information does occur during retrieval. On the other hand, these same models either explicitly state (or in the very least do not deny) the possibility that the same degree of processing or activation of relational information can also occur during study. In order to account for our findings, models of associative memory need to better specify how retrieval processes may differentially affect the processing and utilization of organizational information in episodic recall.

The results of the current experiments also highlight a limitation in Bjork and Bjork's (1992) "New Theory of Disuse." Their theory proposes that the act of retrieving previously learned information may weaken the memorial representation of conceptually related information, thereby impairing its subsequent retrieval. Bjork and Bjork argued that due to limitations in the human mind's capacity to retrieve information at any given time, increasing the retrieval strength of certain information through testing incurs the cost of rendering other conceptually related information more difficult to retrieve. Thus,

testing might not enhance organization, and may even lead to worse output organization than repeated studying, because the successful retrieval of some items from a previously learned list of items may impair subsequent recall of semantically related list items. While the New Theory of Disuse may shed light on conditions that produce retrieval-induced forgetting (Chan, 2009; Anderson, 2003; Anderson et al., 1994), this theory cannot account for our findings of retrieval-induced facilitation—that testing enhances the retrieval of relational information.

On the other hand, theories such as the transfer-appropriate processing framework (Morris, Bransford, & Franks, 1977) and encoding specificity principle (Tulving & Thomson, 1973) can help to explain how retrieval might enhance organization in episodic recall. According to both views, performance on a test of memory benefits to the extent that conditions at retrieval match encoding conditions during prior learning. To the extent that tests of free recall require the use of relational information such as higher-order taxonomic category, temporal, and semantic associative information to guide episodic retrieval of previously learned items (e.g., Kahana, Howard, & Polyn, 2008), prior testing should facilitate subsequent recall performance and promote a greater degree of output organization than restudying. This is because the cognitive operations and conditions required to retrieve and organize information on an initial recall test more closely match those required to perform later recall tests.

Consistent with this prediction, we found that testing enhanced long-term retention and recall organization the most when initial test conditions promoted the use of semantic relational information in episodic recall. Nevertheless, our findings still do not provide strong evidence for the transfer appropriate processing framework or encoding

specificity principle, because the current experiments only used final tests of free and category cued recall. It is possible that initial test conditions that promote semantic organizational processing promote greater retention in delayed item recognition and other tests. Future research should be aimed at testing further predictions of these theoretical frameworks by, for instance, varying the types of final tests (recall vs. recognition) or retrieval cues made available in cued recall (semantic vs. episodic).

There are also several limitations with the measures we employed to assess recall organization that leave some questions unanswered. ARC, PF, and other measures of category clustering and subjective organization, are limited in the sense that they focus on single dimensions of semantic organization. On the other hand, most theories and computational models of memory presume that knowledge is organized in a multi-dimensional mental space (Landauer & Dumais, 1997; Osgood, Suci, & Tannenbaum, 1957; Steyvers, Shiffrin, & Nelson, 2005; Tulving & Bower, 1974; Voss, 1979). ARC and PF are also limited by the fact that they only measure chunking in groups of two items at a time and cannot directly measure chunking that might occur among three or more items. It is, therefore, clear that ARC and PF do not reveal the rich and complex modes of how knowledge is mentally organized; the measures are a first step in a more complex understanding. A better theoretical understanding of the relationship between measures of recall organization and the structure of semantic memory awaits future research.

Educational Implications

One criticism with studies of the testing effect research is that testing effects typically report improvements in learners' retention of discrete facts (e.g., foreign

vocabulary words) without demonstrating a better understanding of the subject matter through testing (Daniel & Poole, 2009). Our finding that tests can enhance students' learning of the conceptual organization of study materials relative to restudying contributes to a steadily growing body of research demonstrating that testing holds promise as a versatile learning tool.

Testing has already been shown to enhance the long-term retention of non-tested information that is conceptually related to previously retrieved information (Chan, 2009; Chan et al., 2006); to stimulate the subsequent learning of new information (Izawa, 1970; Karpicke, 2009; Szpunar et al., 2008; Tulving & Watkins, 1974); as well as to permit better transfer to new questions (Butler, in press; Johnson & Mayer, 2009; Rohrer et al., 2010). It is also worth noting that many education researchers have found that having students answer questions while reading textbook material can improve both their retention and comprehension of the material (e.g., Hamaker, 1986; Rothkopf, 1966; but see Agarwal & Roediger, submitted). Although answering such adjunct questions is not the same as taking a formal test independent of studying the text, it may still be considered a "test-like event," especially when the questions are placed at the end of text (Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008; Rothkopf, 1966).

Of relevance to the current focus on recall testing, when educators use recall tests such as short-answer or essay exams to assess students learning of course materials, they are not only interested in assessing how much information students remember, but they may be just as, if not more, interested in assessing how well students understand the subject matter. Just as measures of output organization in free recall list learning experiments help memory researchers assess how subjects mentally organize list items,

students' understanding of course materials may be best reflected in the organization or coherence of their responses to short answer or essay exam questions. One educational implication of our findings is that the regular use of recall testing in the classroom may help educators improve their students' understanding of the subject matter (see also McDaniel, Howard, & Einstein, 2009). Such tests may include short essay questions that explicitly encourage students to practice organizing their recollections of the subject material in a well-structured manner.

Short-answer and multiple-choice tests may also harbor the potential to improve students' conceptual understanding of subject matter provided the questions challenge students to adopt retrieval strategies that approximate those of free recall learning situations. For instance, Chan and colleagues (2006; Experiment 3) demonstrated that conscious retrieval strategies may be necessary for testing to enhance the retention of semantic associative information. They observed that when students were asked to study and take initial short-answer tests on prose passages, memory for facts that were not initially tested, but were conceptually related to the previously tested facts, was enhanced on a final test relative to a condition in which the passages were re-studied. However, this retrieval-induced facilitation only occurred when subjects adopted a broad retrieval strategy on the initial test in which they attempted to recollect all of the information in the passages that might serve as potential responses to the target questions. When students adopted a narrow retrieval strategy of only trying to think of the correct answers to initial short-answer test questions without thinking of anything else, testing did not facilitate later recall of semantically-related information that was not previously tested.

Conclusion

In sum, the main findings from our experiments are that testing enhances three different measures of categorized list recall: access to higher order units (Rc), access to their contents (Rw/c), and organization of the lists (ARC and/or PF). We conclude that testing stimulates the development of both categorized knowledge (assessed by ARC) and personal idiosyncratic organization (measured by PF). Put another way, testing appears to permit subjects to develop schemas of reconstruction (Gates, 1917; Kühn, 1914) or retrieval plans (Slamecka, 1968) based on both their categorical knowledge and recollection of previous recall attempts. These complementary retrieval schemas that arise through testing may be largely responsible for the testing effect obtained in delayed free recall. These findings contribute to the theoretical understanding that organizational and retrieval processes can enhance learning through a reciprocal relationship. Just as testing can enhance organizational processes, so too do organizational processes contribute to the positive effects of testing on learning.

References

Agarwal, P.K., & Roediger, H.L. (submitted). *Expectancy of an open-book test decreases final retention.*

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger, H. L., & McDermott, K. B. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology*, *22*, 861-876.

Anderson, J.R. (1972). FRAN: A simulation model of free recall. In G. Bower (Ed.) *The Psychology of Learning and Motivation* (Vol. 5, pp. 315-378). New York: Academic Press.

Anderson, J.R. (1996). ACT: A simply theory of complex cognition. *American Psychologist, 51,* 355-365.

Anderson, M.C. (2003). Rethinking interference theory: Executive control and the mechanisms of forgetting. *Journal of Memory & Language, 49,* 415-445.

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, & Cognition, 20,* 1063-1087.

Anderson, J.R., & Bower, G.H. (1973). *Human associative memory*. Washington, D.C.: Winston.

Anderson, M.C., & McCulloch, K.C. (1999). Integration as a general boundary condition on retrieval-induced forgetting. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 25,* 608-629.

Asch, S. E. (1969). A reformulation of the problem of associations. *The American Psychologist, 24,* 92-102.

Ausubel, D.P. (1963). *The psychology of meaningful verbal learning.* New York: Grune & Stratton.

Bangert-Drowns, R.L., Kulik, J.A., & Kulik, C.C. (1991). Effects of frequent classroom testing. *Journal of Educational Research, 85,* 89-99.

Barsalou, L.W. (1983). Ad hoc categories. *Memory & Cognition, 11,* 211-227.

Barsalou, L.W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 11,* 629-654.

Bartlett, F. C. (1932). *Remembering: A Study in Experimental and Social Psychology.* Cambridge: Cambridge University Press.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B, 57,* 289-300.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics, 29,* 1165-1188.

Bjork, R.A. (1975). Retrieval as a memory modifier. In R. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 123-144). Hillsdale, NJ: Erlbaum.

Bjork, R.A., & Bjork, E.L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S.Kosslyn, & R. Shiffrin (Eds.), *From Learning Processes to Cognitive Processes: Essays in Honor of William K. Estes* (Vol. 2, pp. 35-67). Hillsdale, NJ: Erlbaum.

Bousfield, W. (1953). The occurence of clustering in the recall of randomly arranged associates. *The Journal of General Psychology*, *49*, 229-240.

Bousfield, A.K., & Bousfield, W.A. (1966). Measurement of clustering and of sequential constancies in repeated free recall. *Psychological Reports, 19,* 935-942.

Bousfield, W., Cohen, B., & Whitmarsh, G. (1958). Associative clustering in the recall of words of different taxonimic frequencies of occurrence. *Psychological Reports*, *4*, 39-44.

Bousfield, W.A., Puff, C.R., & Cowan, T.M. (1964). The development of constancies in sequential organization during repeated free recall. *Journal of Verbal Learning & Verbal Behavior, 3,* 489-495.

Bower, G.H., & Springston, F. (1970). Pauses as recoding points in letter series. *Journal of Experimental Psychology, 83,* 421-430.

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.

Burns, D.J., & Brown, C.A. (2000). The category access measure of relational processing. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 26,* 1057-1062.

Butler, A. C. (in press). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, & Cognition.*

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

Callender, A.A., & McDaniel, M.A. (2009). The limited benefits of rereading educational texts. *Contemporary Educational Psychology, 34,* 30-41.

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition, 20,* 633-642.

Chan, J.C.K. (2009). When does retrieval induce forgetting and when does it induce facilitation? Implications for retrieval inhibition, testing effect, and text processing. *Journal of Memory and Language, 61,* 153-170.

Chan, J.C.K., McDermott, K.B., & Roediger, H.L. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General, 135,* 553-571.

Cohen, B.H. (1963). Recall of categorized word lists. *Journal of Experimental Psychology, 65,* 368-376.

Cohen, B.H. (1966). Some or none characteristics of coding behavior. *Journal of Verbal Learning & Verbal Behavior, 5,* 182-187.

Daniel, D. B., & Poole, D. A. (2009). Learning for life: An ecological approach to pedagogical research. *Perspectives on Psychological Science, 4,* 91–96.

Gallo, D.A., & Roediger, H.L. (2002). Variability among word lists in eliciting memory illusions: Evidence for associative activation and monitoring. *Journal of Memory & Language, 47,* 469-497.

Gates, A.I. (1917). Recitation as a factor in memorizing. *Archives of Psychology, 6,* 1-104.

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

Hamaker, C. (1986). The effects of adjunct questions on prose learning. *Review of Educational Research, 56,* 212-242.

Hebb, D.O. (1949). *The organization of behavior: A neuropsychological theory*. New York: Wiley.

Howard, M. W. and Kahana, M. J. (2002a). A distributed representation of temporal context. *Journal of Mathematical Psychology*, **46**, 269-299.

Hunt, R.R. (2006). The concept of distinctiveness in memory research. In R.R. Hunt and J.B. Worthen (Eds.), *Distinctiveness and memory* (pp. 3-26). New York: Oxford University Press.

Hunt, R.R., & McDaniel, M.A. (1993). The enigma of organization and distinctiveness. *Journal of Memory and Language, 32,* 421-445,

Hunt, R.R., & Seta, C.E. (1984). Category size effects in recall: The roles of relational and individual item information. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 10,* 454-464.

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

Jacoby, L.L. (1973). Test appropriate strategies in categorized lists. *Journal of Verbal Learning and Verbal Behavior, 12,* 675-682.

Johnson, C.I., & Mayer, R.E. (2009). The testing effect with multimedia learning. *Journal of Educational Psychology, 101,* 621-629.

Kahana, M.J., Howard, M.W., & Polyn, S.M. (2008). Associative processes in episodic memory (pp. 467-490). In H.L. Roediger III (Ed.), Cognitive psychology of

memory. Vol. 2 of *Learning and memory: A comprehensive reference*, 4 vols. (J. Byrne, Editor). Oxford: Elsevier.

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

Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General, 138*, 469-486.

Karpicke, J. D., & Roediger, H. L. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language, 57*, 151-162.

Karpicke, J.D., & Roediger, H.L. (2008). The critical importance of retrieval for learning. *Science, 319,* 966-968.

Karpicke, J.D., & Zaromb, F.M. (2010). Retrieval mode distinguishes the testing effect from the generation effect. *Journal of Memory and Language, 62,* 227-239.

Katona, G. (1940). *Organizing and memorizing.* New York: Columbia University Press.

Kimball, D.R., Smith, T.A., & Kahana, M.J. (2007). The fSAM model of false recall. *Psychological Review, 114,* 954-993.

Klein, S.B., Kihlstrom, J.F., Loftus, J., & Aseron, R. (1989). Effects of item-specific and relational information on hypermnesic recall. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 15,* 1192-1197.

Klein, S.B., Loftus, J., & Schell, T. (1994). Repeated testing: A technique for assessing the roles of elaborative and organizational processing in the representation of social knowledge. *Journal of Personality and Social Psychology, 66,* 830-839.

Kühn, A. (1914). Über einprägung durch lesen und durch rezitieren. *Zeitschrift für Psychologie, 68,* 396-481.

Landauer, T.K. (1975). Memory without organization: Properties of a model with random storage and undirected retrieval. *Cognitive Psychology, 7,* 495-531.

Landauer, T.K., & Dumais, S.T. (1997). A solution to Plato's Problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104,* 211-240.

Lewis-Smith, M.Q. (1976). Expectancy and cued recall. *Bulletin of the Psychonomic Society, 7,* 145-147.

Little, D.R., Lewandowsky, S., & Heit, E. (2006). Ad hoc restructuring. *Memory & Cognition, 34,* 1398-1431.

Mandler, G. (1967). Organization and memory. In K.W. Spence & J.T. Spence (Eds.), *The psychology of learning and motivation* (Vol. 1, pp. 327-372). New York: Academic Press.

Masson, M.E.J., & McDaniel, M.A. (1981). The roles of organizational processes in long-term retention. *Journal of Experimental Psychology: Human Learning and Memory, 7,* 100-110.

Matthews, T.D., Smith., R.E., Hunt, R.R., & Pivetta, C.E. (1999). Role of distinctive processing during retrieval. *Psychological Reports, 84,* 904-916.

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

McDaniel, M. A., Howard, D., & Einstein, G. O. (2009). The read-recite-review study strategy: Effective and portable. *Psychological Science*, *20*, 516-522.

McDaniel, M.A., Moore, B.A., & Whiteman, H.L. (1998). Dynamic changes in hypermnesia across early and late tests: A relational/item-specific account. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 24,* 173-185.

McDaniel, M. A., Roediger, H. L., & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review, 14,* 200-206.

McDaniel, M.A., Waddill, P.J., & Einstein, G.O. (1988). A contextual account of the generation effect: A three factor theory. *Journal of Memory & Language, 27,* 521-536.

Miller, G. (1956). The magical number seven, plus or minus two. *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 & Verbal Behavior, 16,* 519-533.

Mulligan, N.W. (2002). The emergence of item-specific encoding effects in between-subjects designs: Perceptual interference and multiple recall tests. *Psychonomic Bulletin & Review, 9,* 375-382.

Mulligan, N.W. (2005). Total retrieval time and hypermnesia: Investigating the benefits of multiple recall tests. *Psychological Research, 69,* 272-284.

Murphy, M.D. (1979). Measurement of category clustering in free recall. In C.R. Puff (Ed.), *Memory organization and structure* (pp. 51-83). New York: Academic Press.

Murphy, M.D., & Puff, C.R. (1982). Free recall: Basic methodology and analyses. In C.R. Puff (Ed.), *Handbook of research methods in human memory and cognition* (pp. 99-128). New York: Academic Press.

Osgood, C.E., Suci, G.J., & Tannenbaum, P.H. (1957). *The measurement of meaning.* Urbana, IL: University of Illinois Press.

Pashler, H., Rohrer, D., Cepeda, N. J., & Carpenter, S. K. (2007). Enhancing learning and retarding forgetting: Choices and consequences. *Psychonomic Bulletin & Review, 14,* 187-193.

Pellegrino, J.W., & Hubert, L.G. (1982). The analysis of organization and structure in free recall. In C.R. Puff (Ed.), *Handbook of research methods in human memory and cognition* (pp. 129-172). New York: Academic Press.

Polyn, S.M., Norman, K.A., & Kahana, M.J. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review, 116,* 129-156.

Puff, C.R. (1979). Memory organization research and theory: The state of the art. In C. R. Puff (Ed.), *Memory organization and structure* (pp. 3-17). New York: Academic Press.

Pyc, M. A., & Rawson, K. A. (2009). Testing the Retrieval Effort Hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language, 60,* 437-447.

Raaijmakers, J.G.W., & Shiffrin, R.M. (1980). SAM: A theory of probabilistic search of 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: Academic Press.

Raaijmakers, J.G.W., & Shiffrin, R.M. (1981). Search of associative memory. *Psychological Review, 88*, 93-134.

Roediger, H.L., & Karpicke, J.D., (2006a). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17,* 249-255.

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

Roenker, D.L., Thompson, C.P., & Brown, S.C. (1971). Comparison of measures for the estimation of clustering in free recall. *Psychological Bulletin, 76,* 45-48.

Rohrer, D., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 233-239.

Rothkopf, E.Z. (1966). Learning from written instructive materials: An exploration of the control of inspection behavior by test-like events. *American Educational Research Journal, 3,* 241-249.

Schmidt, S.R., & Cherry, K. (1989). The negative generation effect: Delineation of a phenomenon. *Memory & Cognition, 17,* 359-369.

Sirotin, Y.B., Kimball, D.R., & Kahana, M.J. (2005). Going beyond a single list: Modeling the effects of prior experience on episodic recall. *Psychonomic Bulletin & Review, 12,* 787-807.

Slamecka, N.J. (1968). An examination of trace storage in free recall. *Journal of Experimental Psychology, 76,* 504-513.

Spitzer, H.F. (1939). Studies in retention. *Journal of Educational Psychology, 30,* 641-656.

Sternberg, R.J., & Tulving, E. (1977). The measurement of subjective organization in free recall. *Psychological Bulletin, 84,* 539-556.

Steyvers, M., Shiffrin, R.M., & Nelson, D.L. (2005). Word Association Spaces for predicting semantic similarity effects in episodic memory. In A.F. Healy (Ed.), *Experimental cognitive psychology and its applications* (pp. 237-249). Washington, D.C.: American Psychology Association.

Szpunar, K. K., McDermott, K. B., & Roediger, H. L. (2008). Testing during study insulates against the build-up of proactive interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 1392–1399.

Thompson, C.P., & Roenker, D.L. (1971). Learning to cluster. *Journal of Experimental Psychology, 91,* 136-139.

Tse, C. –S., Balota, D.A., & Roediger, H.L. (in press.). The benefits and costs of repeated testing on the retention of face-name pairs across healthy aging. *Psychology and Aging.*

Tulving, E. (1962). Subjective organization in free recall of "unrelated" words. *Psychological Review*, *69*, 344-354.

Tulving, E., & Bower, G.H. (1974). The logic of memory representations. In G.H. Bower (Ed.), *The Psychology of Learning and Motivation, Vol. 8* (pp. 265-301). New York: Academic Press.

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

Tulving, E., & Thomson, D.M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review, 80,* 352-373.

Tulving, E., & Watkins, M. J. (1974). On negative transfer: Effects of testing one list on the recall of another. *Journal of Verbal Learning and Verbal Behavior, 13,* 181–193.

Wheeler, M.A., & Roediger, H.L. (1992). Disparate effects of repeated testing: Reconciling Ballard's (1913) and Bartlett's (1932) results. *Psychological Science, 3,* 240-245.

Vallée-Tourangeau, F., Anthony, S.H. & Austin, N.G. (1998). Strategies for generating multiple instances of common and ad hoc categories. *Memory, 6,* 555-592.

Van Overschelde, J.P., Rawson, K.A., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the Battig and Montague (1969) norms. *Journal of Memory and Language, 50,* 289-335.

Voss J.F. (1979). Organization, structure, and memory: Three perspectives. In C.R. Puff (Ed.), *Memory organization and structure* (pp. 375-400). New York: Academic Press.

Zaromb, F.M., & Roediger, H.L. (in press). The testing effect in free recall is associated with enhanced organizational processes. *Memory & Cognition.*

Appendix 1

Chart used for the initial recall tests in the free recall by categories (S_jT_c) and free recall by judgment tasks (S_jT_j) conditions in Experiment 2.

A	B	C	D	E	F

Appendix 2

Ad-hoc categories and corresponding words used to construct the two study lists in Experiment 3.

List 1	List 2
Things dogs chase	Things that you see at a police station
Cats	Cells
Sticks	Computers
Bones	Donuts
Postmen	Fingerprints
Bicycles	Uniforms
Weekend Entertainment	Things that people hate when they are ill
Drinking	Medicine
Concerts	Vomiting
Dancing	Noise
Picnics	Pain
Movies	Hospitals
Camping Equipment	Things that people keep in their pockets
Tent	Pens
Lantern	Tissues
Canteen	Coins
Fuel	Keys
Pots	Cards
Things that can fall on your head	Things to take out of a fire
Apples	Children
Confetti	Documents
Leaves	Pets
Sleet	Pictures
Water	Memorabilia