Mem Cogn (2012) 40:528–539
DOI 10.3758/s13421-011-0168-y

The origin of the interaction between learning method and delay in the testing effect: The roles of processing and conceptual retrieval organization

Adam Congleton · Suparna Rajaram

Published online: 13 December 2011

Abstract Recent research has demonstrated a relationship between retrieval organization and the efficacy of prior repeated retrieval on delayed tests. The present study asked why repeated study engenders higher recall at a short delay despite lower retrieval organization but produces a decline at a long delay, and why repeated retrieval engenders lower recall at a short delay despite higher retrieval organization but produces stable recall over time. This relationship was examined through the inclusion of two successive recall tests—one immediately after learning method and one a week later. Results replicated the interaction in recall between learning method and delay characterizing the testing effect and, critically, revealed the qualitative differences inherent in the retrieval organization of each method. Specifically, stable recall in repeated retrieval was accompanied by strong and sustained conceptual organization, whereas organization for repeated study was tenuous and weakened across tests. These differences quantitatively were assessed through the use of five targeted analyses: specifically, the examination of cumulative recall curves, the accumulation of organization across time (a curve akin to cumulative recall), item gains and losses across time, changes in the size of categories across time, and the fate of specific clusters of recalled items across time. These differences are discussed within the context of differential processes occurring during learning method.

Keywords Memory · Recall · The testing effect · Retrieval organization · Repeated study advantage · Repeated retrieval advantage · Types of processing

A. Congleton · S. Rajaram (✉)
Department of Psychology, Stony Brook University,
Stony Brook, NY 11794, USA
e-mail: suparna.rajaram@sunysb.edu

Interest in the educational application of cognitive principles has increased tremendously in recent years. One promising area of inquiry concerns the influence testing can have on the retention of material and centers on a phenomenon known as the *testing effect*. This phenomenon refers to improved performance from taking a test, and research shows that testing not only assesses knowledge, but also changes it (Roediger & Karpicke, 2006a). Investigations into this phenomenon have demonstrated an interaction between one's learning method, defined as the sequence of repeated study and/or test trials preceding a critical memory assessment, and delay, such that while repeated study confers immediate memory benefits, the act of taking a test can be more beneficial than spending an equivalent amount of time restudying as delay between study and test increases (Hogan & Kintsch, 1971; Roediger & Karpicke, 2006a; Wheeler, Ewers, & Buonanno, 2003). This testing effect is a robust phenomenon and occurs for a wide range of materials, including paired associates (Allen, Mahler, & Estes, 1969) and general knowledge questions (McDaniel & Fisher, 1991), and it also occurs in real-world classroom settings (e.g., McDaniel, Anderson, Derbish, & Morrisette, 2007). The present study was designed to examine why the interaction between learning method (repeated study vs. repeated retrieval) and delay observed in the testing effect occurs, with an emphasis on understanding the basis of the changes in recall occurring over time as a function of learning method.

One of the most important, early findings motivating research on testing was that an equivalent amount of learning takes place during repeated test trials as during repeated study trials (Tulving, 1967). On the basis of such research, and in an effort to more thoroughly understand the relationship between learning methods and delay, Roediger and Karpicke (2006b) had participants take a final test either 5 min or 1 week after their initial study–test cycle. Importantly, results demonstrated

an interaction between learning method and delay. If the final test was taken 5 min after the learning method, repeated study produced significantly better performance than did repeated retrieval. However, if the final test was taken after 1 week, performance was better following repeated retrieval (Karpicke & Roediger, 2007). Thus, while studies have shown that an equivalent amount of learning can take place during test trials as during study trials (Tulving, 1967), the type of learning seems to vary across these methods.

The present study concerns the underlying cause of the testing effect and focuses on one key mechanism implicated by emerging evidence—namely, semantic or conceptual organization in recall (Congleton & Rajaram, in press; Zaromb & Roediger, 2010). Zaromb and Roediger (Experiment 1) had participants engage in a study-only condition (eight study opportunities), a condition containing two recall trials intermixed with six study trials, or a condition containing two study and six recall trials. Two days later, participants performed four final recall tests. Results indicated that the multiple recall condition produced better retention than did the multiple study condition. Although this experiment did not display enhanced category clustering in recall, measured via adjusted ratio of clustering (ARC) scores (a widely used measure of conceptual organization; Roenker, Thompson, & Brown, 1971; described in more detail later) for repeated retrieval, as compared with repeated study, this outcome occurred in the second experiment. In Experiment 2, repeated study (two study opportunities) produced lower recall and less retrieval organization of materials than did repeated retrieval (one study opportunity followed by one testing opportunity) on a final recall test 1 day later, demonstrating a key role for conceptual organization in recall.

Congleton and Rajaram (in press) also demonstrated that the extent of conceptual organization seen in recall seems to differ as a function of the two learning methods. That study additionally included two delays between the learning method and the final test. This inclusion of two delays, 7 min and 2 h, allowed the opportunity to examine the role of conceptual organization as one underlying reason for the *interaction* between learning method and delay observed in the testing effect. The recall patterns replicated the first part of the testing effect. After the short delay, repeated study led to superior recall, as compared with repeated retrieval, on the final test. Strikingly, the patterns were reversed with respect to the conceptual organization of recalled items even at the short delay; repeated study was accompanied by lower retrieval organization in final recall, as compared with repeated retrieval. After the 2-h delay, the recall pattern on the final test changed; repeated study no longer produced higher recall but was at the same level as repeated retrieval. (Because Congleton & Rajaram [in press] used a relatively short delay [2 h], recall following repeated study dropped only to the level of repeated retrieval, and one can predict a further drop as the delay increases.) Repeated retrieval continued to exhibit high levels of conceptual organization in the final recall. These results indicate that repeatedly retrieving material engenders stronger conceptual retrieval organization from the outset and produces stable recall across time, as compared with repeatedly studying material.

Thus, recent findings identify conceptual retrieval organization as an important variable underlying the interaction between learning method and delay seen in the testing effect. In the present study, we investigated the nature of this critical relationship by addressing two key questions that emerge from these findings: (1) why repeated study leads to higher recall at short delay, despite lower retrieval organization, but leads to a decline at long delay, and (2) why repeated retrieval leads to lower recall at short delay, despite higher retrieval organization, but produces stable recall across the same span of delay. To address these questions, we examined the nature of retrieval organization for each participant in terms of changes occurring in their conceptual organization as it unfolds across time.

One potentially useful framework previously used to conceptualize these learning method differences is the processes occurring during the study–test trials of the learning method (Congleton & Rajaram, in press; Zaromb & Roediger, 2010). Early work by Rundus (1971) demonstrated that repeatedly studying words from a particular category increases the likelihood of recalling and rehearsing additional words from the same category. We reason that it is likely that repeatedly retrieving material would produce an even more active form of rehearsal by virtue of the fact that items are explicitly rehearsed via the act of retrieving them. As a consequence of these different types of rehearsal, the two learning methods may engage different degrees of specific forms of processing during encoding—namely, item-specific and relational processing (Hunt & McDaniel, 1993).

Item-specific processing occurs when a person encodes material one item at a time, focusing on the features of items in isolation from one another, creating more distinctive and richer traces (Mulligan, 2001). In contrast, relational processing occurs when a person encodes the associations among items within an overall set of materials. Given that repeated retrieval produces greater conceptual organization than does repeated study, presumably as a result of magnified active rehearsal, it seems likely that repeated retrieval instantiates more relational processing, since the participants would be thinking actively about the relationships among the various stimuli in an effort to maximize retrieval. Similarly, it seems likely that repeated study participants engender relatively less relational processing because the rehearsal here is more covert, while engaging in relatively more item-specific processing as a result of viewing the stimuli three consecutive times, focusing on each item.

The present study investigates whether the differential relationship between learning method, recall, and retrieval organization seen in recent studies (Congleton & Rajaram, in press; Zaromb & Roediger, 2010) is due to such differential

processes occurring due to learning method. Specifically, by viewing each word multiple times during encoding, repeated study participants are expected to engage in more item-specific processing, as compared with repeated retrieval participants, who study the items only once. In addition, since our stimuli consisted of categorized lists, relational processing was also likely to occur for the repeated study participants (Burns & Hebert, 2005; Einstein & Hunt, 1980; Klein, Loftus, Kihlstrom, & Aseron, 1989), thus producing a combination of item-specific and relational processing and, thereby, enhancing immediate recall following repeated study, as compared with repeated retrieval. However, one could predict that the initial advantage afforded repeated study by the influence of item-specific processing might actually interfere with the ability to focus on the relationships among the various study items, resulting in less effective relational processing, indexed by poorer retrieval organization (i.e., ARC), that is qualitatively more fragile, as compared with repeated retrieval. Such inferior and fragile organization for repeated study should prove less effective at preventing items from being forgotten across time, resulting in lower recall at delay.

In contrast, by having to recall the materials twice after only a single exposure, repeated retrieval participants would get to organize their recall repeatedly according to their preferred conceptual organization, and this process is comparable to the category sorting task associated with relational processing (Einstein & Hunt, 1980). Thus, participants would be likely to think about relationships among the various study items in an effort to maximize retrieval. While the limited study exposure to the material, coupled with a lack of rich features necessary to access memory traces of each item, would lead to poorer performance on immediate recall, the opportunity for repeated retrieval participants to repeatedly focus on the associations among study items would enable them to develop a stronger conceptual organization (i.e., higher ARC) that is qualitatively more robust, insulating them against forgetting and, thereby, producing more stable performance on delayed recall (see Masson & McDaniel, 1981). Such differential processing within each learning method may constitute one key basis for the interaction seen between learning method and delay.

We posit that such differential processing due to the nature of the learning methods should give rise to qualitatively different types of organization. Specifically, repeated study, engendering less relational processing, should give rise to a more fragile, transient form of organization that does not sustain across time, while repeated retrieval, engendering more relational processing, should create a more robust and stable organization across time (Hunt & McDaniel, 1993; McDaniel, Moore, & Whiteman, 1998). We present five types of analyses to examine these predictions. One, we compute the ARC in recall to assess the extent of conceptual organization as a function of learning method, and we predict a replication of past patterns, such that repeated retrieval will produce higher ARC scores. Two, we compute cumulative scores across the recall period, assessing increments in both recall and ARC scores to examine the relative contributions of item-specific and relational processing across the learning methods (Burns, 2006; Burns & Hebert, 2005; Burns & Schoff, 1998). Three, we analyze the number of items gained and lost across the 1-week delay to also assess differences in item-specific and relational processing (Burns & Hebert, 2005).

It is important to note that as a measure of organization, ARC is not sensitive to the level of recall produced by a participant, making it useful for comparisons across conditions and experiments. People can recall more or less material across time, but ARC takes into account only the number of times items from the same category are clustered together and can, thus, remain invariant or even improve for lower recall. Despite these advantages, ARC measures conceptual organization only at individual points in time, with no sensitivity to how organization unfolds or changes across time, and it sometimes fails to detect differences in organization between repeated study and repeated retrieval (Zaromb & Roediger, 2010, Experiment 1). Therefore, we present two novel analyses designed to detect qualitative differences between the organizations instantiated by the learning methods, where the fourth analysis indexes changes in the size of categories recalled across delay (*category size* analysis; see also Congleton & Rajaram, in press) and the fifth analysis examines the fate of specific clusters of items recalled by participants across delay (*fate* analysis).

Given our focus on how retrieval organization unfolds across time, resulting in long-term qualitative distinctions, we designed within-subjects conditions to examine the influence of delay. These participants completed both immediate and delayed recall tests so that the changes in retrieval organization and recall levels could be assessed for the same participant across time. This within-subjects design specifically made possible the computations of category size and fate analyses noted above. To also build in a replication, we included two control conditions to provide reference for the basic comparison of the testing effect where the delayed test was unaffected by a prior test. These participants completed only the delayed test. Finally, we ensured that the long delay in this study was 1 week (as opposed to 2 h; Congleton & Rajaram, in press) to enable generalization to past studies that used this length of delay (e.g., Karpicke & Roediger, 2007; Roediger & Karpicke, 2006b).

Method

Participants and design

Eighty college undergraduates participated for experimental credit, with 20 participants assigned to each condition. A

between-subjects, two-factor design was employed, with learning method (repeated study vs. repeated retrieval) and number of recalls (single recall vs. two recalls) as the factors. The four conditions included repeated-study–single-recall, repeated-retrieval–single-recall, repeated-study–two-recalls, and repeated-retrieval–two-recalls. Each condition was divided into two broad segments: the *learning method* segment that consisted of three phases where participants were exposed to the stimuli in one of two formats (repeated study or repeated retrieval) and the *retrieval measures* segment that consisted of the final test(s) where the participants recalled the studied material either once on a delayed test or twice on both immediate and delayed tests. The within-subjects conditions with two successive recalls constituted a key part of the design to test the changes in recall and retrieval organization as a function of delay.

Materials

The stimuli consisted of 180 categorized words, with 15 categories and 12 exemplars per category (Van Overschelde, Rawson, & Dunlosky, 2004). We purposely selected categorized words because they lend themselves to the calculation of conceptual retrieval organization scores via the ARC measure (Roenker et al., 1971). ARC scores represent the strength of retrieval organization in terms of category repetitions and are calculated by quantifying the degree to which participants cluster items from the same taxonomic category at recall. ARC is considered one of the most valid measures of organization because it sets upper and lower bounds on the amount of clustering possible and takes into account the proportion of actual clusters to the total number of clusters possible.

We divided the words into two study lists (each with 90 critical words), which were buffered with four items from separate, nonused categories from Van Overschelde et al. (2004; two buffers each at the beginning and the end). The lists were balanced across two criteria: taxonomic frequency and word length (both ts < 1). The two lists were pseudorandomized, with the constraint that no two items from the same category appeared next to one another. In each condition, half of the 20 participants studied list 1, and the other half studied list 2. Any given participant saw six exemplars from each category. Each study item was presented for 6 s (1 s of an asterisk, 5 s for the exemplar).

Procedure

The experiment began with the *learning method* sequence, which was identical across all conditions. Participants individually viewed the study stimuli via a PowerPoint presentation for the initial study phase (phase 1), with instructions to read the material for a future unspecified memory test and make pleasantness ratings (Craik & Lockhart, 1972), an encoding task commonly used in recall paradigms to boost performance (especially those involving a long study–test delay).

After a single study phase, the procedure began to differ for participants, depending upon the condition. In the repeated study condition, participants viewed the exact list twice more consecutively and again provided pleasantness ratings each time (phases 2 and 3). Participants in the repeated retrieval condition performed two free recall tasks consecutively for 7 min each (phases 2 and 3), writing as many of the items as they could recall in any order from the studied list. Participants also drew a line under the last recalled item every time they heard a computer-emitted tone (once every minute) and then continued (this measure was taken in order to plot cumulative recall curves).

After completing the *learning method* segment, all participants completed a spatial distractor task for 7 min. In the *retrieval measures* segment that followed, participants in the single-recall conditions returned 1 week later to perform the recall task as described above, and those in the two- (successive) recall conditions stayed in the lab to perform the first recall task and returned 1 week later to again perform the same task.

Results

Across all analyses, a corrected recall measure was used, where for each participant the total number of intrusions was subtracted from the total number of items correctly recalled. The intrusions were low in all conditions (from 0.25 to 0.65 items on the immediate test and from 1.40 to 2.09 items on the delayed test). Importantly, all patterns of results were the same in the uncorrected recall measure (the total number of items recalled). Therefore, we present the more conservative corrected recall data. The alpha was set at $p < .05$.

The first set of recall analyses are for the standard between-subjects comparisons of repeated study and repeated retrieval conditions—that is, for the *first recall* that occurred immediately in the two- (successive) recall conditions versus the *first recall* that occurred after the 1-week delay in the single, delayed recall conditions. The second set of analyses focuses on the within-subjects comparisons that are crucial for testing the novel questions of this study, and include comparisons of the *two- (successive) recall* immediate and delayed conditions.

Immediate versus delayed recall—both initial tests

As was predicted, a factorial analysis of variance (ANOVA) between learning method (repeated study vs. repeated retrieval) and delay (short vs. long) demonstrated a significant interaction, $F(1, 76) = 12.00$, $MSE = .01$, $p = .001$.

Immediate recall A one-way completely randomized (between-subjects) ANOVA with learning method (repeated study vs. repeated retrieval) as the factor replicated prior findings, since repeated study (M = .49) led to significantly greater recall than did repeated retrieval (M = .42), $F(1, 38)$ = 5.08, MSE = .01, p = .03, demonstrating the first part of the interaction seen in the testing effect (Roediger & Karpicke, 2006b [Fig. 1, the leftmost bars for recall]; Wheeler et al., 2003).

Delayed recall (single recall) In single delayed recall, a one-way within-subjects ANOVA replicated the second part of the interaction seen in the testing effect, since repeated study recall (M = .09) was significantly lower than repeated retrieval (M = .22), $F(1, 38)$ = 22.58, MSE = .01, p = .001 (Fig. 1, the middle bars for recall).

Retrieval organization A one-way between-subjects ANOVA confirmed that repeated retrieval preferentially enhances conceptual retrieval organization even on immediate recall (Congleton & Rajaram, in press), since ARC scores were significantly higher for participants on the immediate test with a previous history of repeated retrieval (M = .71) than for participants with a history of repeated study (M = .45), $F(1, 38)$ = 25.86, MSE = .03, p = .001 (Fig. 1, leftmost bars for ARC scores). Thus, retrieval organization was lower following repeated study, despite the fact that recall was actually higher on the immediate test.

Mimicking the findings at immediate recall, a one-way ANOVA demonstrated that on the delayed test, ARC scores were significantly higher for single-recall participants with a previous history of repeated retrieval (M = .65) than for participants with a history of repeated study (M = .44), $F(1, 38)$ = 4.20, MSE = .05, p = .05 (Fig. 1, middle bars for ARC scores).

Immediate versus delayed recall—successive tests

Having established the two aspects of the testing effect in a standard, between-subjects design, as well as replicating better retrieval organization for repeated retrieval, as compared with repeated study, at both time points, we turn to the within-subjects findings for both the immediate and delayed tests, since our key questions focused on the relationship between retrieval organization and recall across time. As is elaborated below, the key patterns of findings reported above held true again in the within-subjects comparisons and, thus, extend the delayed recall results of Zaromb and Roediger (2010) and the immediate–delayed interaction in the between-subjects comparison of Congleton and Rajaram (in press) to a within-subjects design.

Delayed recall (second recall) A one-way between-subjects ANOVA with learning method as the factor mirrored the delayed, single-recall conditions, such that recall was significantly higher following repeated retrieval (M = .29) than following repeated study (M = .21), $F(1, 38)$ = 7.50, MSE = .01, p = .01 (Fig. 1, rightmost bars for recall).[1]

Delayed retrieval organization (second recall) A one-way between-subjects ANOVA confirmed that ARC scores mimicked the pattern found on the immediate test, with repeated retrieval participants (M = .66) showing significantly higher ARC scores than did repeated study participants (M = .44), $F(1, 38)$ = 9.31, MSE = .05, p = .004 (Fig. 1, rightmost bars for ARC scores).

Next, we present four analyses targeted specifically for the main goals of the study to specify the process by which this seemingly contradictory pattern takes shape—that is, repeated retrieval improving organization, as compared with repeated study, from the outset, yet producing poorer recall at short delay.

Cumulative recall and cumulative organization curves

Past research has shown that cumulative recall curves are effective in revealing the separate and/or mixed influences of item-specific and relational processing inherent in a participant's recall (Burns & Hebert, 2005; Burns & Schoff, 1998). Specifically, item-specific processing is characterized by curves where recall of studied material is both slower initially and slower to reach asymptotic level of recall, thereby producing a slow, steady rate across the entire recall period. In contrast, relational processing is characterized by curves with high initial recall and a rapid reach to asymptote. Finally, a curve characterized by a combination of item-specific and relational processing exhibits high initial recall but is also slower to reach asymptote.

To examine these characteristics, after the calculations of recall and ARC reported in the previous section, the data were normalized to correct for differences in final performance levels across conditions, and different rates of approach to asymptote between learning methods were assessed in cumulative recall (Fig. 2). For immediate recall, the cumulative recall curve for repeated retrieval participants (black solid line) presents a typical example of relational processing; these participants had a high level of recall initially and appeared to reach their asymptote quickly as well.[2] In comparison, repeated study participants (gray solid line) had slightly lower initial recall (although still

[1] Consistent with past research, recall was higher when the delayed test was preceded by a previous recall than when it was not.

[2] The conclusions from normalized data remain the same for cumulative raw data.

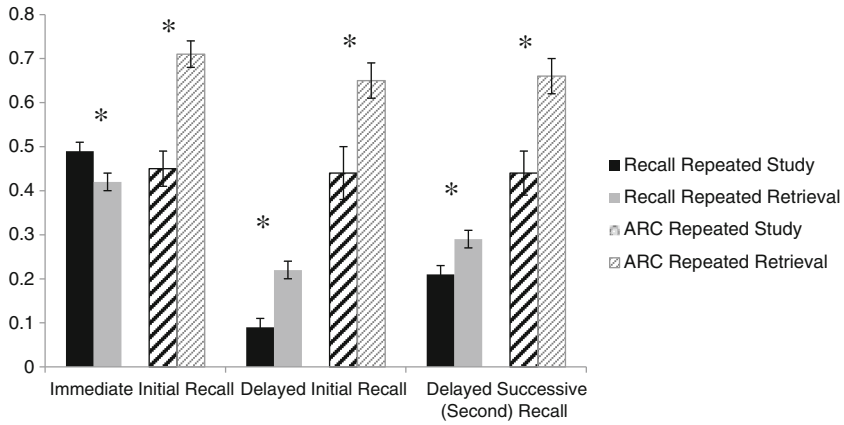

Fig. 1 Means and standard errors for corrected recall and retrieval organization (ARC) scores as a function of retrieval history, delay, and time of recall

high) but had a slower, continued level of recall throughout the rest of the period, indicative of a combination of both item-specific and relational processing. Calculations of lambda (Bousfield & Sedgewick, 1944; Burns & Hebert, 2005) supported these conclusions; repeated retrieval produced a significantly faster rate of reaching asymptote (M = .55) than did repeated study (M = .30), $t(41) = -3.642$, p = .001. Mapping these patterns onto those proposed by Burns and Hebert, these data show that repeated retrieval participants engaged in more relational processing than did repeated study participants, who engaged in a combination of item-specific and relational processing.

Interestingly, at the long delay, the pattern changed for repeated study participants (dashed gray line), such that their pattern of recall now looked like that of the repeated retrieval participants (dashed black line), only with a much lower level of recall. Calculations determined that there was no difference now between repeated study (M = .46) and repeated retrieval (M = .58) in terms of their rate of approaching asymptote, $t(40) = -0.89$, p = .38.

These curves show that the initial advantage afforded repeated study participants by item-specific processing did not last across time. In fact, the influence of item-specific processing during immediate recall could actually work against the processing of relationships among the various list items. This process becomes more visible at delayed recall, where the repeated study participants now relied mainly on a relational strategy to recall materials. However, the material they were able to recall, presumably processed relationally at immediate recall, was noticeably less and/or weaker than that recalled by repeated retrieval participants, given that the repeated study curve is lower. To further quantify these possibilities, we calculated a new type of cumulative curve—namely, *cumulative organization curves*.

Cumulative organization curves are analogous to cumulative recall curves, in that they are designed to track the accumulation of organization across a recall period. To graph such curves, we determined how many clusters of items participants recalled during each of the 7 min allotted during recall. The results support the interpretation of the cumulative recall curves above (Fig. 3). At immediate recall, repeated retrieval produced a large number of clusters very rapidly and, subsequently, reached asymptote quickly as well, while repeated study produced the clusters slowly and steadily across the recall period. Calculations of rate of approaching asymptote showed a clear advantage for repeated retrieval (M = .69), as compared with repeated study (M = .38), $t(35) = -2.92$, p = .006. In contrast, at delayed recall, both repeated retrieval and repeated study

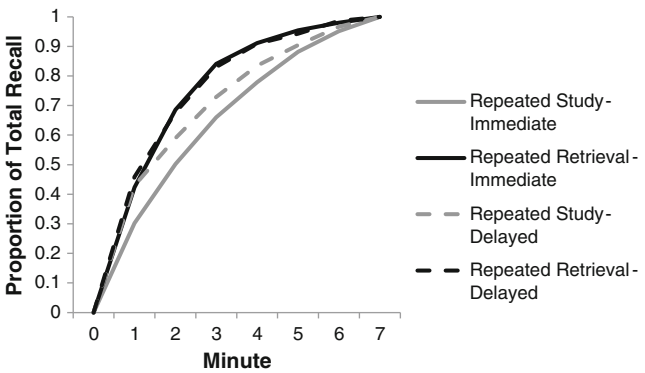

Fig. 2 Normalized cumulative recall curves for repeated study and repeated retrieval at immediate and delayed recalls

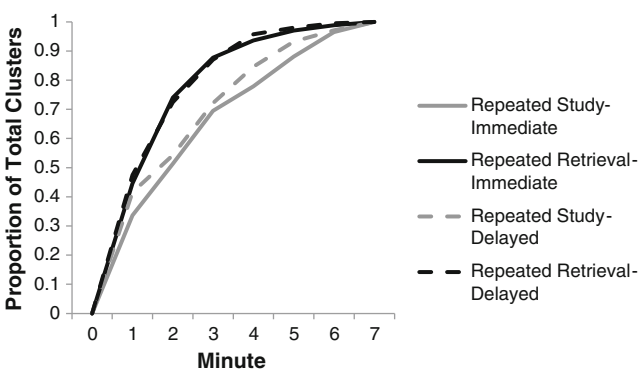

Fig. 3 Normalized cumulative organization curves for repeated study and repeated retrieval at immediate and delayed recalls

participants recalled clusters of items at comparable rates, but repeated study participants did so to a far lesser degree. We unfortunately could not perform calculations of lambda on the delayed organization data, since participants exhausted their recall and organization so quickly as to preclude the calculation of this value (given the requirements of the equation). The results of both cumulative recall and cumulative organization curves demonstrate that repeated study participants rely on an item-specific processing advantage during immediate recall that interferes, to some extent, with their ability to think relationally about the items, leading to poorer recall after delay when the item-specific advantage has dissipated.

To augment the cumulative recall and organization analyses, we performed another set of analyses—namely, *item gains and losses* (Burns, 2006).

Analysis of item gains and item losses

Item gains refers to the newly recalled items on a later test that were not recalled on an earlier test, while *item losses* refers to the number of items forgotten on a later test that were recalled on an earlier test. Manipulations enhancing item-specific processing lead to greater items gained across time, while manipulations enhancing relational processing lead to fewer items lost across time (Klein et al., 1989).

We found no difference between repeated study ($M = 1.86$) and repeated retrieval ($M = 1.80$) for item gains, $F(1, 38) \leq 1$, indicating no difference after delay in the amount of item-specific processing present (see, e.g., Klein et al., 1989). This result is consistent with our cumulative recall data, because the item gains measure indexes changes after delay, where the item-specific advantage for repeated study is expected to already be extinguished (as it is in the cumulative recall data), eliminating quantifiable differences between the two conditions. Furthermore, the item losses analysis showed significantly more items lost for repeated study ($M = .59$) than for repeated retrieval ($M = .34$), $F(1, 38) = 44.74$, $MSE = .02$, $p = .001$. This result once again indicates that there is more relational processing for repeated retrieval, also consistent with our cumulative recall data.

These analyses demonstrate that repeated study and repeated retrieval produce recall and retrieval organizations that accumulate in different ways across time. Next, we assessed whether the retrieval organizations produced by the two learning methods were also qualitatively different. To that end, we first examined how the size of categories recalled changed across time.

Category size analysis

This analysis provided **a** finer-grained examination of exemplar clustering than is provided by ARC. While this analysis is similar to assessing the recall of items within categories (Tulving & Pearlstone, 1966), it further allows us a unique index of the changes in the size of the categories as they unfold across time as a function of learning method. Simply calculating items per category does not capture that aspect of exemplar fluctuations we wanted to uncover. Critically, it allowed us to determine whether the size of the category clusters recalled provides information as to why one particular learning method produces higher performance on immediate recall (e.g., repeated study) and another learning method produces more stable performance across time, as evidenced on delayed recall (e.g., repeated retrieval). We predicted that a less robust retrieval organization present in repeated study would lead to a reduction in the number of exemplars recalled from any given category as the delay between study and test increased. Furthermore, given that having a prior history of repeated retrieval led to a stronger retrieval organization measured via ARC, a similar reduction in exemplar recall would be less likely to occur in this condition, since better organized recall is more likely to survive across delay. In other words, this convergent measure can help uncover the relationship between retrieval organization and learning method, given its greater sensitivity to cluster size, as compared with ARC. To calculate this measure, each protocol was scored for two variables: the number of categories recalled and the number of exemplars per category recalled. The categories recalled were then split into two groups: *small categories*, or categories where participants recalled two or fewer exemplars of the six possible ("2 or fewer"), and *large categories*, or categories where participants recalled three or more exemplars of the six possible ("3 or more").

A 2 × 2 completely randomized, between-subjects ANOVA showed no difference in the number of large categories recalled ("3 or more") on immediate recall between repeated retrieval ($M = 8.95$) and repeated study ($M = 9.50$), $F(1, 38) \leq 1$. Critical for the present hypothesis, for the delayed recall, repeated retrieval ($M = 5.71$) recalled significantly more large categories ("3 or more") than did repeated study ($M = 2.96$), $F(1, 38) = 13.39$, $MSE = 6.36$, $p = .001$ (Fig. 4). Furthermore, the number of smaller categories not only increased over time for both conditions, $F(1, 85) = 65.61$, $MSE = 8.61$, $p = .001$, but also did so to a significantly greater extent following repeated study (immediate $M = 5.39$; delayed $M = 12.04$) than following repeated retrieval (immediate $M = 6.25$; delayed $M = 9.50$), $F(1, 83) = 7.94$, $MSE = 7.87$, $p = .006$, showing that there was not a general decline in the number of categories in recall following repeated study but there was a specific change in the size of the categories recalled. Such differences cannot be discerned simply from the ARC scores, since the ARC measure for repeated study did not change across immediate recall (.45) and delayed recall (.44). Together, these findings show that repeated study produced relatively tenuous organization, such

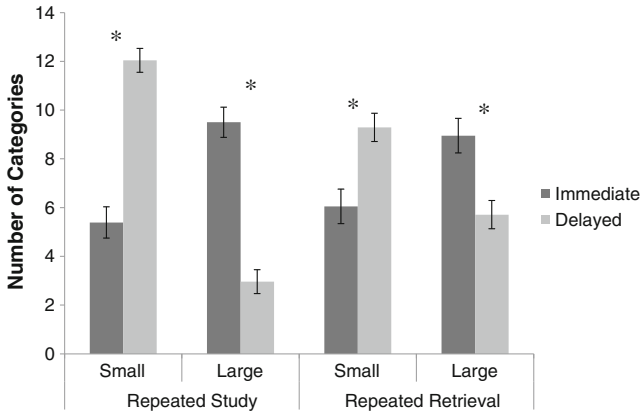

Fig. 4 Category size analysis for changes in retrieval organization from the immediate to the delayed recall test in the two- (successive) recall conditions. Small = when two or fewer exemplars from a category were recalled; large = when three or more exemplars from a category were recalled

that delay reduced the presence of larger categories in recall to a far greater degree than did repeated retrieval.

Fate analysis

We created this novel analysis to examine in greater depth the patterns and implications revealed by the ARC and category size analyses about the relationship between learning method and retrieval organization across time. Specifically, we assessed the "fate" of specific clusters of items across time, from immediate to delayed recall. While part of the calculation of ARC requires the determination of the total number of category repetitions within a recall protocol, it does not take into consideration, or provide information on, *the specific items clustered together during recall*. The fate analysis specifically assessed what happens to these clustered items across time: Do they remain clustered together on the delayed protocol, or are they broken apart or forgotten entirely, and how is this process affected by learning method? In addition, this measure can be thought of as an index of relational processing, since it ascertains, in one particular coding variable, how many clusters of items are maintained in their intact form across a period of delay, something presumed to happen only when one has processed the relationships between the various items composing that cluster. Thus, if repeated study is engaged in more item-specific processing during immediate recall, we should see less emphasis placed on the associations among items within the study list, represented in this analysis as an inability to retain clusters of items across time.

The immediate recall protocol was first examined, and all of the clusters from the same category were identified. The specific clusters identified were then compared with the delayed recall protocol to determine the fate of these clusters and were sorted into one of several categories. *Intact clusters* refer to those clusters that remain completely intact across the 1-week delay, such that if *apple*, *orange*, and *banana* were recalled together at immediate recall, these same three items would appear together at delayed recall as well (although they did not have to be in exactly the same order). *Partial clusters due to fragmented recall* (*partial-fragmented clusters*) refer to those clusters having one or more items from the original cluster (i.e., on the immediate recall protocol) no longer attached to the cluster itself but recalled somewhere else on the delayed protocol, while still retaining at least two items to be classified as a cluster. For example, if *apple*, *orange*, and *banana* were recalled together at immediate recall, at least two of the items—say, *apple* and *orange*—would be recalled together, while the third item would be found somewhere else in the delayed recall protocol. *Partial clusters due to structural loss* (*partial-structural clusters*) refer to those clusters that have lost (i.e., forgotten) one or more items from the original cluster but that retain at least two items to still be considered a cluster on the delayed protocol. *Perished clusters due to structural loss* (*perished-structural clusters*) refer to those clusters that lost (i.e., forgotten) enough items from the original cluster to no longer be considered a cluster at all yet retain at least one item from the original cluster that was not forgotten. *Perished clusters due to fragmented recall* (*perished-fragmented clusters*) refer to those clusters that have all of the items from the original cluster no longer bound together, but recalled at different points throughout the delayed recall protocol. *Perished clusters* refer to those clusters in which all of the items from the original cluster have been lost across delay.

In addition to dividing the clusters up into these six categories, we also lumped all partial clusters into a superordinate category entitled *overall survived* (because at least part of the original cluster survived across delay) and all perished clusters into the superordinate category *overall perished*, allowing us to perform split-half comparisons between the two types of clusters. Finally, we examined how many clusters had added in additional items from the same category not bound to the original cluster at immediate recall. We refer to these new clusters as *augmented clusters.* For interests of economy, and in the spirit of reporting only those analyses that are informative with respect to the hypothesis, we discuss only the *overall survived*, *overall perished*, *intact*, and *perished* clusters. Information concerning all of the various categories can be found in Table 1.

In the process of coding for these fate analyses, we discovered additional variables that could provide us with more evidence of the qualitative differences between the organization instantiated by repeated study and repeated retrieval. We first compared the two learning methods in terms of the *average number of clusters* recalled on both the immediate and delayed protocols. The results were striking. On immediate recall, there was no significant difference between the average number of clusters recalled by repeated

Table 1 Means and standard errors for the fate of clusters from the immediate to the delayed recall test in the two- (successive) recall conditions

Learning method	Repeated Study	Repeated Retrieval
Overall survived clusters	2.58 (0.47)	5.67 (0.63)
Overall perished clusters	7.08 (0.57)	4.62 (0.49)
Intact clusters	0.92 (0.22)	2.90 (0.49)
Partial-fragmented clusters	0.29 (0.13)	0.90 (0.18)
Partial-structural clusters	1.38 (0.31)	1.86 (0.35)
Perished-structural clusters	2.88 (0.42)	1.48 (0.30)
Perished-fragmented clusters	0.67 (0.22)	1.29 (0.27)
Perished clusters	3.54 (0.41)	1.86 (0.29)
Augmented clusters	1.96 (0.36)	2.19 (0.42)

Note. Overall survived = the sum of all the clusters that survived in some form or another across time; overall perished = the sum of all the clusters that did not survive across time; intact = the sum of all the clusters that survived in their original form across time; partial-fragmented = the sum of all clusters having one or more items from the original cluster no longer attached to the cluster itself but recalled somewhere else on the delayed protocol, while still retaining at least two items to be classified as a cluster; partial-structural = the sum of all the clusters that have lost one or more items from the original cluster but retain at least two items to still be considered a cluster on the delayed protocol; perished-structural = the sum of all the clusters that lost enough items from the original cluster to no longer be considered a cluster at all yet retain at least one item from the original cluster that was not forgotten; perished-fragmented = the sum of all the clusters that have all of the items from the original cluster no longer bound together but recalled at different points throughout the delayed recall protocol; perished = the sum of all the clusters where every single item that composed the original cluster did not survive across time; Augmented = the sum of all the clusters had added in additional items from the same category that were not bound to the original cluster at immediate recall

study (M = 9.67) and repeated retrieval (M = 10.29), $F < 1$. However, there was indeed a significant difference on the delayed recall protocol between repeated study (M = 4.38) and repeated retrieval (M = 7.33), $F(1, 38) = 11.96$, MSE = 8.19, $p = .001$. Thus, while both learning methods produced the same number of clusters on immediate recall, demonstrating that they both employed the recall strategy of category clustering, repeated retrieval produced significantly more clusters after the 1-week delay. In addition to this finding, we also looked at the *average size of the clusters* produced on the immediate protocol. Repeated study (M = 2.56) produced smaller clusters, on average, as compared with repeated retrieval (M = 2.83), $F(1, 38) = 5.27$, MSE = .16, $p = .03$.[3] Thus, even though repeated study and repeated retrieval participants recalled the same number of clusters on the immediate protocol, repeated study participants recalled smaller clusters, on average. This difference between the learning methods indicates that while both learning methods are capable of using the same recall strategy, repeated study participants do not use it as effectively as repeated retrieval.

Next, we turn to the fate analysis proper (see Table 1). In terms of *overall survived* clusters, repeated study (M = 2.58) had significantly fewer clusters survive across the delay, as compared with repeated retrieval (M = 5.67), $F(1, 38)$ = 16.09, MSE = 6.62, $p = .001$. In conjunction with this finding, the analysis examining *overall perished* clusters indicated that repeated study (M = 7.08) lost more clusters across the delay than did repeated retrieval (M = 4.62), $F(1, 38) = 10.34$, MSE = 6.58, $p = .002$. Even more striking is our analysis of *intact* clusters, where repeated study (M= 0.92) had significantly fewer clusters that survived in their completely original form across the 1-week delay, as compared with repeated retrieval (M = 2.90), $F(1, 38) = 15.15$, MSE = 2.92, $p = .001$. Converging on this finding is our analysis of *perished* clusters, where repeated study participants forgot significantly more of all of the items from the original clusters (M = 3.54), as compared with repeated retrieval (M = 1.86), $F(1, 38) = 10.47$, MSE = 3.04, p = .002.[4]

These analyses supported our prediction that not only does repeated study produce weaker retrieval organization in general (measured via ARC) and does not allow for the retention of large categories across delay (evidenced by category size analysis), but also it does not retain the same clusters of items across delay, indicating a lack of strength among the items making up the clusters. Thus, this fate analysis provides support for our hypothesis that there are different types of processing engendered by the specific learning method.

Discussion

The aim of this study was to examine the underlying basis of both the immediate and delayed aspects of the testing effect. Past research has reported an interaction between learning method and delay, such that repeatedly studying material

[3] There were more "stray" exemplars (i.e., only one exemplar per category) in repeated study than in repeated retrieval at immediate recall, accounting for higher recall in repeated study on immediate recall.

[4] In order to verify that there was no recall-level confound, we conducted several ANCOVAs to determine whether differences between repeated study and repeated retrieval with regard to category size and fate analyses were influenced by general recall level as a covariate. There was no significant interaction between general recall level and learning method for the selected factors (immediate large categories, delayed large categories, overall survived clusters, overall perished clusters, intact clusters, perished clusters, or augmented clusters). When examining the model with the exclusion of these nonsignificant interaction variables, we obtained the same significant differences between repeated study and repeated retrieval as previously described for every variable tested. Thus, the category size and fate analyses are not confounded with recall levels.

leads to higher recall on immediate memory tests but poor performance across time, while repeatedly retrieving material leads to stable recall across time (Hogan & Kintsch, 1971; Roediger & Karpicke, 2006b). Recent research has demonstrated a relationship between conceptual retrieval organization and the learning methods that give rise to the testing effect (Congleton & Rajaram, in press; Zaromb & Roediger, 2010). Specifically, the superior performance of repeated study on immediate tests is accompanied by poor retrieval organization, while repeated retrieval is accompanied by strong organization (Congleton & Rajaram, in press). The present study built upon these findings, obtained from the typical between-subjects design with respect to delay by using a within-subjects comparison, to better understand the relationship between learning method, retrieval organization, and recall across time, as well as to explain both the immediate and delayed aspects of the testing effect. The critical inclusion of two successive recalls in this design allowed us, for the first time, to examine how one's recall and retrieval organization scores on an immediate memory test influence one's performance on a delayed memory test.

To help us better conceptualize the differences between the two learning methods, we defined them in terms of a framework of differential processes that occur inherently due to the different learning methods. Specifically, it was expected that relational processing would occur for both repeated study and repeated retrieval in the learning of a categorized word list (Burns & Hebert, 2005; Burns & Schoff, 1998; Rundus, 1971), but repeated study would lend itself to more item-specific processing of materials due to increased exposure time and the creation of distinctive memory traces. While this combination of processing was expected to yield superior immediate recall, the operation of some item-specific processing was also deemed likely to interfere with the processing of information relationally. The resultant poorer conceptual organization was expected to lead to lowered recall at delay. In contrast, repeated retrieval was expected to engender more relational processing without such interfering influences of item-specific processing (Hunt & McDaniel, 1993) and, thus, show a more stable level of recall across delay (Masson & McDaniel, 1981). Our findings supported this framework as one important way to explain both the immediate and delayed aspects of the testing effect.

These findings are also consistent with past evidence that pleasantness ratings of items at study promotes more item-specific processing (Klein et al., 1989; see also Burns & Hebert, 2005; Einstein & Hunt, 1980), that pleasantness ratings of categorized words promotes a combination of item-specific and relational processing (Burns & Hebert, 2005) and that category sorting (presumably similar to retrieval organization achieved through repeated retrieval opportunities in the present experimental design) promotes conceptual organization and facilitates relational processing (Einstein & Hunt, 1980). Thus, our study not only replicated the advantages of a combination of item-specific and relational processing in immediate recall, but also produced the novel finding that this combination is less effective for delayed recall, as compared with a selective but superior level of relational processing. As such, our findings showed the expected interaction between delay and learning method inherent in the testing effect and further demonstrated the manner in which recall and retrieval organization interact.

The inclusion of our novel analyses allowed us not only to examine the specific manner by which retrieval organization influences recall across time, but also to provide evidence of the origin of these differences in the learning methods. The results of our cumulative recall and cumulative organization analyses and our item gains and item losses analyses, as well as our category size and fate analyses, demonstrated that the learning methods instantiate two qualitatively distinct types of conceptual organization, a fact ARC alone is not sensitive in detecting. Repeated study leads to a relatively transient, or *tenuous*, form of organization, evidenced and characterized by lower ARC scores on immediate and delayed tests, an inability to retain large categories across time, a lack of the retention of specific category clusters across time, and smaller average clusters on immediate recall tests even though they form the same number of clusters as repeated retrieval. All of these results show some of the key reasons why repeatedly studying material does not protect against forgetting across time, as would be expected by stronger organization. On the other hand, repeatedly retrieving material leads to a stronger, more secure, or *robust* form of organization, characterized by greater ARC scores on both immediate and delayed tests, retention of large categories across time, retention of specific category clusters across time, and larger average clusters on immediate recall tests even though they form the same number of clusters as repeated study.

In addition to the role of conceptual organization considered in the present study, it is possible that repeated study and repeated retrieval may also differ in the extent of temporal organization afforded each method. The repeated study method may afford better temporal order retention than the repeated retrieval method (see, e.g., Karpicke & Zaromb, 2010), and this could account for their superior performance on immediate recall, since research has shown that temporal order information is more useful for immediate recall than for delayed recall (e.g., Burns, Curti, & Lavin, 1993; Nairne, Riegler, & Serra, 1991). Furthermore, the operation of temporal order processing may be of greater force in repeated study, and this might have further blocked the development of conceptual organization. While this possibility could constitute another reason for the differences between the two learning methods, it is important to

note that such competition between temporal order information and conceptual organization can occur even when the order of items is varied from one repeated study presentation to the next, and thus this issue can be applicable to all types of repeated study procedures, and not just to the particulars of our procedure. This is because the experimenter-determined order, even when it varies from session to session, would likely be different from the idiosyncratic order in which each participant would organize his or her information. The latter organization is most readily possible only in the repeated retrieval condition. Clearly, these issues await further research to fully map out the roles multiple mechanisms play in shaping the learning method × delay interaction in the recall performance. The present study was designed to specifically examine one candidate mechanism—namely, how the conceptual organization differs between the two learning methods, and how it evolves across time to produce the intriguing cross-over interaction reported across numerous studies in the literature. It is likely that the factors that explain this interaction may vary as a function of the particulars of the study and test conditions employed across different designs. For example, Karpicke and Zaromb's (2010) experiments also showed a role for item-specific processing in promoting superior recall at delay following retrieval practice, as compared with repeated study, in contrast to the thesis advanced here. The role of item-specific processing in their study could be somewhat specific to the particulars of their design and procedure, developed to compare the relative benefits of incidental generation of targets, as compared with the intentional retrieval of targets in response to word cues. The use of pairs of cue words and fragments of target words for inducing retrieval practice likely promoted more item-specific processing than would occur in the more typical paradigms (those similar to the procedure used in the present study) using repeated recall as the practice task. Nonetheless, it is useful to bear in mind that multiple mechanisms may ultimately account for why repeated retrieval practice promotes long-term retention across a wide variety of learning and testing situations (Karpicke & Zaromb, 2010; Pyc & Rawson, 2010), and future research is needed to identify additional key mechanisms that may provide support for this phenomenon. The present study contributes to this goal by specifying the role of conceptual organization as one important mechanism—in particular, not only in producing the advantages for repeated retrieval in delayed recall, but also in accounting for the learning method × delay interaction noted across the literature.

The results of our various analyses provide strong evidence that in addition to the improvements in retrieval organization that account for the interaction between delay and recall observed in the testing effect (Congleton & Rajaram, in press; Zaromb & Roediger, 2010), it is the *quality or the nature* of conceptual organization engendered by different learning methods that is critical for understanding both the immediate and delayed aspects of the testing effect. In essence, the power of repeated testing lies in its ability to form stable organization of material across time, a benefit simply studying material is unable to provide.

References

Allen, G. A., Mahler, W. A., & Estes, W. K. (1969). Effects of recall tests on long-term retention of paired associates. *Journal of Verbal Learning and Verbal Behavior, 8,* 463–470.

Bousfield, W. A., & Sedgewick, C. H. W. (1944). An analysis of sequences of restricted associative responses. *The Journal of General Psychology, 30,* 149–165.

Burns, D. J. (2006). Assessing distinctiveness: Measures of item-specific and relational processing. In R. R. Hunt & K. Worthen (Eds.), *Distinctiveness and memory* (pp. 109–130). New York: Oxford University Press.

Burns, D. J., Curti, E. T., & Lavin, J. C. (1993). The effects of generation on item and order retention in immediate and delayed recall. *Memory & Cognition, 21,* 846–852.

Burns, D. J., & Hebert, T. (2005). Using cumulative recall curves to assess the extent of relational and item-specific processing. *Memory, 13,* 189–199.

Burns, D. J., & Schoff, K. M. (1998). Slow and steady often ties the race: Effects of item-specific and relational processing on cumulative recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 1041–1051.

Congleton, A. R., & Rajaram, S. (in press). The influence of learning method on collaboration: Prior repeated retrieval enhances retrieval organization, abolishes collaborative inhibition, and promotes post-collaborative memory. *Journal of Experimental Psychology: General.*

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11,* 671–684.

Einstein, G. O., & Hunt, R. R. (1980). Levels of processing and organization: Additive effects of individual-item and relational processing. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 588–598.

Hogan, R. M., & Kintsch, W. (1971). Differential effects of study and test trials on long-term retention and recall. *Journal of Verbal Learning and Verbal Behavior, 10,* 562–567.

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

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

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.

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

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

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

McDaniel, M. A., & Fisher, R. P. (1991). Tests and test feedback as learning sources. *Contemporary Educational Psychology, 16,* 192–201.

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, and Cognition, 24,* 173–185.

Mulligan, N. W. (2001). Generation and hypermnesia. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 436–450.

Nairne, J. S., Riegler, G. L., & Serra, M. (1991). Dissociative effects of generation on item and order retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 702–709.

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science, 330,* 335.

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

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

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.

Rundus, D. (1971). Analysis of rehearsal processes in free recall. *Journal of Experimental Psychology, 89,* 63–77.

Tulving, E. (1967). The effects of presentation and recall of material in free-recall learning. *Journal of Verbal Learning and Verbal Behavior, 6,* 175–184.

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

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.

Wheeler, M. A., Ewers, M., & Buonanno, J. F. (2003). Different rates of forgetting following study versus test trials. *Memory, 11,* 571–580.

Zaromb, F. M., & Roediger, H. L. I. I. I. (2010). The testing effect in free recall is associated with enhanced organizational processes. *Memory & Cognition, 38,* 995–1008.