Leveraging the Test Effect to Improve Maintenance of the Gains Achieved Through Cognitive Rehabilitation

Rhonda B. Friedman,
Department of Neurology, Center for Aphasia Research and Rehabilitation, Georgetown University Medical Center

Kelli L. Sullivan,
Department of Neurology, Center for Aphasia Research and Rehabilitation, Georgetown University Medical Center

Sarah F. Snider,
Department of Neurology, Center for Aphasia Research and Rehabilitation, Georgetown University Medical Center

George Luta, and
Department of Biostatistics, Bioinformatics, and Biomathematics, Georgetown University Medical Center

Kevin T. Jones
Department of Neurology, Center for Aphasia Research and Rehabilitation, Georgetown University Medical Center

Abstract

Objective—An important aspect of the rehabilitation of cognitive and linguistic function subsequent to brain injury is the maintenance of learning beyond the time of initial treatment. Such maintenance is often not satisfactorily achieved. Additional practice, or overtraining, may play a key role in long-term maintenance. In particular, the literature on learning in cognitively intact persons has suggested that it is testing, and not studying, that contributes to maintenance of learning. The present study investigates the hypothesis that continuing to test relearned words in persons with anomia will lead to significantly greater maintenance compared with continuing to study relearned words.

Method—The current study combines overtraining with the variable of test versus study in examining the effects of overtesting and overstudying on maintenance of word finding in 3 persons with aphasia. First, treatment successfully reestablished the connections between known items and their names. Once the connections were reestablished (i.e., items could be named successfully), each item was placed into 1 of 4 overtraining conditions: test and study, only test, only study, or no longer test or study. Maintenance was probed at 1 month and 4 months following the end of overtraining.

Correspondence concerning this article should be addressed to Rhonda B. Friedman, 4000 Reservoir Road NW, Building D, Suite 207, Washington, DC 20057. friedmar@georgetown.edu.

Results—The results are consistent with an advantage of testing compared with studying. All 3 participants showed significantly greater maintenance for words that were overtested than for words that were overstudied. This testing benefit persisted at 1 month and 4 months after completion of the treatment. In fact, there was no clear evidence for any benefit of overstudying.

Conclusions—The present study demonstrates that overtesting, but not overstudying, leads to lasting maintenance of language rehabilitation gains in patients with anomia. The implications for the design of other treatment protocols are immense.

Keywords

aphasia; anomia; rehabilitation; maintenance; test effect

Studies investigating techniques for restoring language skills in persons with aphasia (PWAs) have been increasing as the population ages and more people are suffering stroke and subsequent language impairment (reviewed in Wisenburn & Mahoney, 2009). Many of these studies have demonstrated success in improving targeted language processes immediately following treatment (e.g., ●●●). Despite these initial gains, long-term maintenance of improvement is not often assessed across studies. A few studies probed maintenance beyond 1 month (Bastiaanse, Hurkmans, & Links, 2006; Drew & Thompson, 1999; Edwards & Tucker, 2006; Furnas & Edmonds, 2014; Thompson, Kearns, & Edmonds, 2006), although none of these examined maintenance of item-specific gain. Many more studies assessed performance only up to 1 month following treatment (Boo & Rose, 2011; Edmonds & Kiran, 2006; Fillingham, Sage, & Lambon Ralph, 2005; Law, Wong, Sung, & Hon, 2006; Leonard, Rochon, & Laird, 2008; Milman, Clendenen, & Vega-Mendoza, 2014; Raymer, Kohen, & Saffell, 2006; Raymer et al., 2012; Robson, Marshall, Pring, Montagu, & Chiat, 2004; Sage, Snell, & Lambon Ralph, 2011; Schneider & Thompson, 2003; van Hees, Angwin, McMahon, & Copland, 2013), and many studies with encouraging language rehabilitation findings in PWA simply did not collect follow-up data that measure the retention of performance gains (Denes, Perazzolo, Piani, & Piccione, 1996; Harnish, Neils-Strunjas, Lamy, & Eliassen, 2008; Hickin, Mehta, & Dipper, 2015; Hinckley & Carr, 2005; Lee, Kaye, & Cherney, 2009; Pulvermüller et al., 2001). Evidence that treatment strategies can lead to improved language abilities is certainly important, but the ultimate goal should be long-term improvement.

Although strategies for improving the recovery of language in PWA have improved of late, little attention has been paid to techniques that might be employed to increase *long-term* retention of learned language skills. Meanwhile, research within the fields of psychology and education has focused on principles of learning that enhance retention of knowledge and information in neurologically intact children and adults. Variables such as massed versus distributed practice, blocked versus mixed practice, and studying versus testing have been investigated and found to be important in the design of learning paradigms. Factors that improve learning in unimpaired adults and children may be good candidates for relearning, or rehabilitation, of cognitive functions subsequent to brain injury. However, with a few exceptions, these variables have not found their way into the rehabilitation research literature. One study (Middleton, Schwartz, Rawson, & Garvey, 2015) demonstrated that retrieval practice facilitated improvement of naming in a population of both fluent and

nonfluent PWA, measured at 1 day and 1 week following the end of treatment. These results provide preliminary evidence of enhanced maintenance of relearned material following retrieval practice, but the length of this benefit remains in question.

Often, treatments for aphasia promote only temporary performance gains; Patients can learn the information initially, but they do not retain it. One factor that may contribute to successful learning in rehabilitation is overlearning. In overlearning paradigms, items continue to be practiced beyond the point at which they are learned to criterion. In the learning and education literature, it has been asserted that in order for learned material to be retained it must be practiced beyond the point of initial competence (Samuels, 1988). This concept has been applied with success in studies of treatment for various aspects of aphasia, including anomia (McNeil et al., 1998), apraxia (Wambaugh, Martinez, McNeil, & Rogers, 1999), and alexia (Lott, Carney, Glezer, & Friedman, 2010; Lott, Sperling, Watson, & Friedman, 2009). Typically, these language rehabilitation paradigms overtrain the entire set of items until most items have been learned. However, this is a time-intensive procedure that is limited in its application. An alternative paradigm would be to monitor performance on each of the individual items and to drop out each item from training as it is "learned," that is, produced correctly on consecutive sessions. This design, which allows the participant to concentrate on the more problematic items, may be a more efficient way to learn. The paradox is that the dropout paradigm eliminates the possibility of overtraining the learned items.

Another factor receiving considerable attention of late is the *test effect*, the finding that repeated testing improves long-term maintenance of learned materials (Karpicke & Roediger, 2008; Rawson & Dunlosky, 2011). Although the benefit of active retrieval, or testing, has been known for some time (Tulving, 1967), recent research has explored in detail the contributions of studying versus testing in long-term maintenance of learning in neurologically intact individuals (Carpenter & DeLosh, 2005; Carrier & Pashler, 1992; Karpicke & Roediger, 2007; Mcdaniel, Roediger, & Mcdermott, 2007; Meyer & Logan, 2013; Rogalski et al., 2014). Results from experimental and classroom studies have suggested that additional studying contributes little to long-term retention, whereas the contribution of additional testing is substantial (Karpicke & Roediger, 2007; Roediger & Karpicke, 2006b).

In paired associate learning, testing can be considered any situation in which the first item in the pair is given and it is incumbent upon the participant to produce the second item in the pair. Indeed, the testing effect can be obtained simply by presenting the first item a few seconds prior to the second, even when no response is required (Metcalfe & Kornell, 2007). This is in contrast to studying, in which the two items (e.g., a picture and a word) are presented simultaneously.

The advantage of testing over studying has been attributed to practice with retrieval (Karpicke & Roediger, 2007; Middleton et al., 2015; Roediger & Karpicke, 2006a, 2006b). That is, the measure of success is how one ultimately performs on a test, so it stands to reason that practicing taking a test would improve performance on a test. This explanation has its roots in notions of transfer-appropriate processing (Morris, Bransford, & Franks,

1977). The benefit of testing might also be attributed to the increased anxiety that testing might induce, leading to heightened arousal and consequently greater attention and better learning. A third possibility is that the testing effect in paired associate learning is the result of spreading activation within the semantic network (Anderson, 1983; Collins & Quillian, 1972), leading to multiple, elaborative associations between the items (Carpenter, 2009). Finally, it has been proposed that the act of retrieving a response requires greater cognitive effort than does that of simply viewing the response and that this is responsible for greater depth of learning (Carpenter, 2009).

Can the test effect be applied to aphasia rehabilitation? The research on the test effect has typically taken place within the context of new learning, whereas aphasia rehabilitation can be viewed as the "relearning" of old information. However, it is not known whether, when PWAs "recover" the ability to name an object, they are learning the name anew (perhaps by a different pathway) or are strengthening that which is known but now exists in a weakened state. In either case, though, it is the association of the name with the object—the retrieval of the target (the name) given the cue (the object)—that is being strengthened. If the advantage of testing over studying is attributable to practice with retrieval, then it should surely apply in situations where the person's difficulty is with word retrieval, as is the case in patients with output anomia. This is the rationale that motivated the Middleton et al. (2015) study.

Retrieval practice is even more likely to be effective in the present study, where we are specifically focusing on the process of retrieving information that has already been successfully stored. All words are being relearned in exactly the same way. Our test versus study variable is applied only once participants have demonstrated retention of the word on two consecutive sessions. That is, we are not applying the principle to relearning but rather to retaining that which has been successfully learned.

Preliminary studies have applied the test effect paradigm to improve memory in patient populations such as those with multiple sclerosis (Sumowski, Chiaravalloti, & DeLuca, 2010; Sumowski et al., 2013) and traumatic brain injuries (Pastötter, Weber, & Bauml, 2013; Sumowski, Wood, et al., 2010). The effect appears to be quite robust and is not limited to verbal learning studies. Despite the breadth of studies applying the test effect, the paradigm has not been directly tested in language rehabilitation settings, where long-term maintenance of relearned items is of great importance. The one exception is Middleton et al. (2015), which did examine the effects of "test-enhanced learning (p. ●●●). In that study, persons with anomia attributed to a failure to retrieve words (output anomia) demonstrated improved naming at 1 day and 1 week posttreatment. However, long-term maintenance beyond 1 week was not assessed. Additionally, the Middleton et al. study focused on the differences between treatment strategies during the learning phase. In the current study, we focus on the factors that contribute to successful *long-term maintenance* following (re)learning of words that are not easily named in persons with output anomia following stroke.

The current study uses an overlearning paradigm to determine the effects that overstudying and overtesting have on long-term maintenance of (re)learned items. We hypothesized that additional testing of relearned words would lead to greater retention of words than would additional studying of relearned words in PWA. We predicted that this benefit would persist

at both 1 month and 4 months following the end of treatment (a length of time rarely probed in rehabilitation studies; Wisenburn & Mahoney, 2009). In addition, we hypothesized that each overtraining condition would be more beneficial than would the no overtraining condition. Last, we hypothesized that the combined overtesting and overstudying condition would lead to the greatest performance at each time point compared to only overtesting or overstudying alone.

Method

Participants

Three individuals with aphasia who met the following criteria were recruited to participate: output anomia, at least 12 months poststroke, native English speaker, and at least 10 years of formal education. All participants gave written informed consent approved by the Georgetown University Institutional Review Board and received financial compensation for their participation in the study.

The first participant, YPR, was a left-handed 52-year-old former teacher and pastor with a master's degree. He suffered a right MCA infarct 3 years prior to his participation in our study. At the time of enrollment in the study, he presented with fluent speech, moderate-to-severe anomia, and an aphasia quotient (AQ) of 58.8 on the Western Aphasia Battery (WAB; ●●●; see Table 1). His auditory comprehension was functional for everyday use. His lesion extended from the anterior pole of the temporal lobe to the occipitoparietal junction, with extension into the parietal lobe and posterior parts of the frontal lobe.

The second participant, ODH, was a 71-year-old right-handed high school graduate who was heavily involved in volunteer activities prior to her stroke. She suffered a left posterior temporal lobe infarct 1 year prior to her enrollment in the study. ODH presented with fluent speech, moderate anomia, relatively good auditory comprehension, and an AQ of 79.3 on the WAB.

The third participant, CLN, was a 65-year-old left-handed high school graduate with 1 year of college education who worked for the government and volunteered as a referee in a local softball league. She suffered a left temporoparietal infarct approximately one year prior to entering the study. She presented with fluent aphasia, moderate anomia, and a WAB AQ of 77.3. Her auditory comprehension was functional for daily activities.

Measures

The participants' performance on semantic tasks across modalities (see Table 2) demonstrates that none had a significant semantic impairment. All participants were consequently classified as having output anomia.

Additionally, standardized and nonstandardized language tests were administered at baseline by a speech–language pathologist to ensure that the participants had adequate auditory comprehension and speech output for participation in the treatment protocol. The Boston Diagnostic Aphasia Examination (3rd ed.; Goodglass, Kaplan, & Barresi, 2001) and the Boston Naming Test (2nd ed.; Kaplan, Goodglass, & Weintraub, 2001) were readministered

at 1 and 4 months posttreatment to assess changes in general language and word-naming abilities.

Cognitive testing—Participants completed a cognitive assessment battery to rule out the presence of any significant cognitive deficits that would impair their ability to learn the treatment items. Tests included the Test of Nonverbal Intelligence (Brown, Sherbenou, & Johnsen, 1997), the Visual Form Discrimination Test (Benton, Sivan, Hamsher, Varney, & Spreen, 1994), and the Biber Figure Learning Test (Glosser, Goodglass, & Biber, 1989).

Stimuli—Prior to treatment, participants were asked to name 384 black-and-white pictures with high name agreement (e.g., cowboy, motorcycle, turtle) presented sequentially on a computer screen. The pictures were named on three separate occasions, 1 week apart. Following the third session of picture naming, we identified 120 items for each participant that were named incorrectly in each of the three sessions.[1] For each of the 120 chosen items, the participants were also asked to name an alternate black-and-white picture depicting the same item. These "Exemplar 2 pictures" were not seen again until posttesting and were used to assess generalization.

For each participant, 24 of the 120 pictures were chosen to serve as control items and were not tested again until the follow-up sessions. The control items were chosen to be representative of the 120 items with regard to the following characteristics: semantic category; frequency; number of syllables, phonemes, and letters; and initial phoneme.

Procedure

Design—Treatment took place twice weekly and lasted approximately 2 hr per session. Each session included at least four alternating blocks of *test* and *study* trials. Participants took short breaks between and during treatment blocks as needed.

The first two sessions were practice sessions, and the order of blocks was Study 1, Test 1, Study 2, Test 2. All subsequent sessions were considered treatment and began with *test* blocks (i.e., the order was Test 1, Study 1, Test 2, Study 2). This enabled us to probe naming accuracy at the beginning of each session (the Test 1 block). An additional pair of blocks (Test 3 and Study 3) was added at the end of the session if time permitted.

Study procedure—During *study* blocks, each of the 96 trained pictures was presented on the screen along with its written name. Simultaneously, the prerecorded spoken name of the picture was played by the computer program through a pair of noise-canceling headphones worn by the participant. The participant was instructed to read or repeat out loud the name of the picture and then to "study" the picture and word until they disappeared from the screen 5 s later (see Figure 1). The researcher controlled the pace of the presented items due to differences in participants' attention.

---

[1]Because of ODH's higher level of naming accuracy, she named only 96 items incorrectly in all three sessions; for her to reach 120 items, we included 24 items that were named incorrectly in two out of the three baseline sessions.

Test procedure—During *test* blocks, each picture was presented alone on a computer screen and the participant was asked to name the picture. After 5 s, the correct word was played through the headphones to provide feedback, regardless of the accuracy of the participant's response. This feedback was given at the end of each testing trial because it has been shown that feedback can enhance testing effects, particularly in older adults (Tse, Balota, & Roediger, 2010). The computerized tasks were programmed in E-Prime 2.0 (●●●). All sessions were recorded by both an audio recorder and a video camera. Sessions were scored by two independent raters to ensure reliability in scoring the participants' responses.

Beginning with Session 3, we tracked performance on the first test block of each session. Items were considered "learned" when they were named correctly on two consecutive Test 1 blocks. Once an item was learned, it was assigned to one of four overtraining conditions in a counterbalanced order (see Table 3). This manner of placing words in the four conditions controls for "difficulty" of the learned word, in that items learned early (for whatever *subject-specific* reasons) are represented in all conditions and likewise for items that are learned later. Consideration was also given to *word-specific* parameters (e.g., frequency, category, length) when placing learned items in the four conditions so that these word characteristics were reasonably well balanced across the conditions. The four overtraining conditions are (1) continuing to test and study (TS) items, (2) continuing to only test items (T), (3) continuing to only study items (S), and dropout, where items are no longer studied or tested (O). All words that were not yet learned continued to be trained in all four blocks (Test 1, Study 1, Test 2, Study 2) of each session. Treatment continued until the participant ceased learning new items.

Follow-up testing—At 1 month and again at 4 months following conclusion of treatment, the participants returned to the lab for follow-up testing. In three separate sessions, with 1 week between consecutive sessions, they were asked to name the treatment and control pictures in three different pseudorandom orders. In a subsequent session, they were asked to name the second set of drawings (Exemplar 2 pictures) corresponding to their treatment items. All picture-naming tests also included 40 items that were consistently named correctly during the initial evaluation.

Statistical Analyses

In order to compare the different overtraining conditions, we conducted group chi-squared analyses between the different treatment conditions at each time point. The chi-squared analyses allowed for direct testing between two treatment conditions, which we conducted for each of our hypotheses. As stated in the introduction, our primary hypothesis was to compare the effects of overtesting and overstudying at each of our follow-up time points (1 and 4 months posttreatment). We did this by comparing set T with set S at each time point. Second, to test the hypothesis that any overtraining is more beneficial than no overtraining, we compared each overtraining condition (TS, T, S) with no overtraining (O). Last, we compared the test and study condition (TS) with continuing to either test (T) or study (S) alone. Answering our hypotheses did not require testing for interaction effects. Therefore, we did not need to conduct logistic regression analyses, whose results would in any case be

expected to be similar, given that the tests from the logistic regression model are asymptotically equivalent to the chi-square tests.

Results

YPR learned 76 of the 96 trained items in 25 treatment sessions. ODH learned 91 of her 96 items in 19 treatment sessions. CLN learned 77 of the 96 trained items in 27 sessions. Only those words that were practiced in an overtrained condition for at least 15 pairs[2] of blocks were included in the analyses. YPR's analyses included 72 learned items (18 items per overtraining condition); ODH's analyses included 85 items (22 items in S; 21 items each in TS, T, and O), and CLN's analyses included 74 items (18 each in TS and S; 19 each in T and O). All of the participants completed treatment within 3 to 4 months.

Effects of Testing and Studying at 1 Month Posttreatment

Results on the naming task at 1 month posttreatment were analyzed to test our main hypothesis that overtesting beyond initial learning is more beneficial than is overstudying beyond initial learning. Overtested items (T) were compared in chi-squared analyses to overstudied items (S) collectively for all participants. There was a significant advantage for overtested items compared to overstudied items, $\chi^2(3, 348) = 19.69$, $p < .0001$, at 1 month following the end of training (see Figure 2).

Effect of Testing and Studying at 4 Months Posttreatment

Figure 3 presents the proportion of items named correctly in each overtraining condition at 4 months posttreatment. For all participants, the relative accuracy across treatment sets was very similar to that seen at 1 month posttreatment. Using chi-squared analysis as was explained earlier, we tested the main hypothesis that at 4 months posttreatment overtesting beyond initial learning is more beneficial than is overstudying beyond initial learning. There was a significant advantage for overtested items compared to overstudied items, $\chi^2(3, 348) = 13.39$, $p = .0002$, at 4 months following the end of training. (see Figure 3).

Second Exemplar

To test whether the training generalized beyond the specific pictures on which the participants were trained, we presented the alternative black-and-white pictures of the trained items (Exemplar 2 pictures) once for naming at both 1- and 4-months posttreatment. At the 1 month mark there was a significant advantage for overtested items compared to overstudied items, $\chi^2(3, 116) = 7.89$, $p = .0086$; see Figure 4A). At 4 months posttreatment a similar pattern emerged, but the effect of overtesting was not significantly different from overstudying, $\chi^2(3, 116) = .79$, $p = .5047$; see Figure 4B).

Overtraining Effects

Chi-square analyses were conducted to test the second set of hypotheses, that each of the overtraining sets (sets TS, T, S) would lead to greater word maintenance at both 1 and 4 months post-treatment compared to the no overtraining set (set O).

---

[2]A pair was one *test* and one *study* block.

At 1 month following the end of treatment, accuracy was significantly greater for both sets that contained overtesting, set TS: $\chi^2(3, 345) = 29.52$, $p < .0001$; set T: $\chi^2(3, 348) = 39.64$, $p < .0001$, but not for the overstudied set, set S: $\chi^2(3, 348) = 3.38$, $p = .0834$. The same analyses conducted at the 4 months posttreatment time point remained significant for the combined overtesting plus overstudying condition, set TS: $\chi^2(3, 345) = 8.19$, $p = .0044$, and the overtesting only condition, set T: $\chi^2(3, 348) = 22.45$, $p < .0001$. There was no significant difference between the overstudying condition and the no overtraining condition, set S: $\chi^2(3, 348) = .99$, $p = .3761$. That is, the additional studying did not lead to improved performance compared with no additional practice.

Exploratory Analyses

Group chi-squared analyses were conducted to test our final hypothesis, that the combined overtesting plus overstudying (set TS) condition would lead to greater performance than would either the overtesting (set T) or overstudying (set S) condition alone. There was no significant difference between the TS set and the T set at either 1 month, $\chi^2(3, 345) = .82$, $p = .4069$, or 4 months, $\chi^2(3, 345) = 3.69$, $p = .0661$, following the end of treatment. However, the combined overtesting condition (set TS) did result in significantly greater word retention compared to the overstudying only condition (set S) at 1 month posttreatment, $\chi^2(3, 345) = 13.48$, $p = .0003$; this difference did not remain significant at the 4 months time point, $\chi^2(3, 345) = 3.54$, $p = .0624$.

It is interesting that although not significant, performance was numerically greater in the overtesting only condition (set T) than in the combined overtesting plus overstudying condition (set TS) at both time points (1 month post: set T: .83, set TS: .79; 4 months post: set T: .82, set T: .74) and for all three participants.

Discussion

The results of this study suggest that overtesting can lead to better maintenance of the positive learning achieved following a program of anomia rehabilitation in persons with aphasia. More important, they demonstrate that continuing to study items learned may be of little or no benefit with regard to maintenance. The benefit of overtesting was maintained at least up to 4 months following the end of treatment. It is important to note that the paradigm resulted in lasting maintenance for all three participants, despite their differing levels of aphasia severity, lesion size and location, age, and handedness. A larger sample size might have revealed differential effects of some of these variables, and it would be important to examine them in the future. The finding of a test effect despite the differences among the participants is encouraging, because it suggests that its applicability could be extensive.

Overtraining and Dropout

In many treatment paradigms, overtraining may be unintentionally built into the procedure. Consider a typical treatment paradigm in which patients are attempting to learn 20 items. Practice continues until they are correct on a certain percentage of the items over two or three consecutive sessions. Note that as items are produced correctly, they continue to be practiced within the list of 20 items while the remaining items are learned. Thus, some items

are actually overtrained. An alternative paradigm, which we employed in the current study, is to monitor performance on each of the individual items and to assign the item to one of the four overtraining conditions as it is learned, that is, produced correctly on two consecutive sessions. This latter design, which allows patients to concentrate on the more problematic items, may be more efficient, but it does not allow for the maintenance benefits of overtraining.

The finding here that periodic testing is more effective than is additional studying supports the use of a hybrid design that incorporates the best of both: Learned items can be dropped out to allow more time to be devoted to not-yet-learned items, but overtraining in the form of testing can be retained, allowing for greater maintenance of learning. This testing effect may actually have contributed to the results of previous rehabilitation studies. Typical studies include regular probe tests. But these probe *tests* might be considered to be a form of treatment that would have effects upon the results of the study. That is, the possibility exists that additional testing is actually a cause of the improvement rather than a mere measure of performance.

An additional feature of the design of this study that bears notice is the manner in which learned items are assigned to the four overtraining conditions. Good experimental design requires that conditions to be compared are equivalent on all relevant variables other than the one being examined (in this case, testing and/or studying). Relevant variables are those that might be expected to have some influence over the outcome. When the stimuli are words and the response is overt naming, relevant variables typically include word frequency, number of syllables, and so forth, that is, variables that are known to correlate with ease of naming. These correlations are based on normative samples and are reasonable estimates of the likelihood of successful naming for any given individual. However, they are indeed only *estimates* of the relative difficulty of items for a given individual. A more precise measure, specific to the individual participant, is the one used in the current study. The current study distributes items into the four training conditions on the basis of the session in which the participant learns the word rather than on normative data, providing the best measure of difficulty for that individual.

One finding of considerable interest is that the *test and study* condition did not result in greater retention than did the *test only* condition. In fact, for all three participants, at both the 1 month and the 4 months time points, the opposite occurred: The *test only* condition yielded better retention than did the *test and study* condition. Although the numbers are too small for statistical significance, the consistency of the finding is remarkable. How might this phenomenon be explained? The explanation that we favor has its roots in a study that compared constructing versus remembering the solution to a problem (Jacoby, 1978). In that study, neurologically intact subjects learned word pairs. On some trials the solution had to be constructed, because only part of the second word was presented. On other trials the solution was provided, that is, both words were presented together and the subject had to simply read the second word. The finding of relevance to the present study is that retention was reduced when the *construction* trials were immediately preceded by the *read only* trials. It was surmised that the second word did not need to be constructed on those trials; it was "remembered" from the previous trial. In our *test and study* condition, *study* trials are

alternated with *test* trials. Thus, the *study* trials may actually interfere with the *test* trials by providing the participants with easier access to the correct response. The implications of this finding would be rather significant for designing cognitive treatment protocols: It may be desirable to avoid studying once an item is learned and concentrate solely on testing.

The present study replicated previous findings on the benefit of retrieval practice (Rawson & Dunlosky, 2011; Roediger & Karpicke, 2006) and extended the paradigm to a neuropsychological population. Previous research (Middleton et al., 2015) demonstrated the efficacy of retrieval practice during the relearning phase of language rehabilitation. The current study demonstrates how its implementation subsequent to relearning can enhance long-term maintenance.

The retention length of treatment benefits in the current study is encouraging for future rehabilitation studies, where performance gains are often temporary. Future treatment studies should aim to maximize the maintenance of positive findings by employing techniques that have been shown to be effective in learning and retention, such as overtraining, dropout, and testing. These techniques could plausibly be applied to the treatment of any linguistic or cognitive deficit—and indeed within the context of any treatment paradigm—where improved function following treatment but unsatisfactory maintenance of that improvement is seen.

Acknowledgments

References

Anderson A. A semantic analysis of spatial descriptions in spontaneous dialog. Bulletin of the British Psychological Society. 1983 Nov.36:A94.

Atkinson, RC., Shiffrin, RM. Human memory: A proposed system and its control processes. Vol. 2. New York, NY: Academic Press; 1968.

Baddeley, A., Eysenck, MW., Anderson, MC. Memory. New York, NY: Psychology Press; 2009.

Bastiaanse R, Hurkmans J, Links P. The training of verb production in Broca's aphasia: A multiple-baseline across-behaviours study. Aphasiology. 2006; 20:298–311. http://dx.doi.org/10.1080/02687030500474922.

Benton, AL., Sivan, AB., de Hamsher, KS., Varney, NR., Spreen, O. Contributions to neuropsychological assessment: A clinical manual. 2nd. New York, NY: Oxford University Press; 1994.

Boo M, Rose ML. The efficacy of repetition, semantic, and gesture treatments for verb retrieval and use in Broca's aphasia. Aphasiology. 2011; 25:154–175. http://dx.doi.org/10.1080/02687031003743789.

Brown, L., Sherbenou, RJ., Johnsen, SK. Test of nonverbal intelligence. 3rd. Austin, TX: Pro-Ed; 1997.

Carpenter SK. Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2009; 35:1563–1569. http://dx.doi.org/10.1037/a0017021.

Carpenter SK, DeLosh EL. Application of the testing and spacing effects to name learning. Applied Cognitive Psychology. 2005; 19:619–636. http://dx.doi.org/10.1002/acp.1101.

Carrier M, Pashler H. The influence of retrieval on retention. Memory & Cognition. 1992; 20:633–642. http://dx.doi.org/10.3758/BF03202713. [PubMed: 1435266]

Collins, AM., Quillian, MR. How to make a language user. New York, NY: Academic Press; 1972.

Denes G, Perazzolo C, Piani A, Piccione F. Intensive versus regular speech therapy in global aphasia: A controlled study. Aphasiology. 1996; 10:385–394. http://dx.doi.org/10.1080/02687039608248418.

Drew RL, Thompson CK. Model-based semantic treatment for naming deficits in aphasia. Journal of Speech, Language, and Hearing Research. 1999; 42:972–989. http://dx.doi.org/10.1044/jslhr.4204.972.

Edmonds LA, Kiran S. Effect of semantic naming treatment on crosslinguistic generalization in bilingual aphasia. Journal of Speech, Language, and Hearing Research. 2006; 49:729–748. http://dx.doi.org/10.1044/1092-4388(2006/053).

Edwards S, Tucker K. Verb retrieval in fluent aphasia: A clinical study. Aphasiology. 2006; 20:644–675. http://dx.doi.org/10.1080/02687030600631827.

Fillingham JK, Sage K, Lambon Ralph MA. Further explorations and an overview of errorless and errorful therapy for aphasic word-finding difficulties: The number of naming attempts during therapy affects outcome. Aphasiology. 2005; 19:597–614. http://dx.doi.org/10.1080/02687030544000272.

Furnas DW, Edmonds LA. The effect of computerised verb network strengthening treatment on lexical retrieval in aphasia. Aphasiology. 2014; 28:401–420. http://dx.doi.org/10.1080/02687038.2013.869304.

Glosser G, Goodglass H, Biber C. Assessing visual memory disorders. Psychological Assessment. 1989; 1:82–91. http://dx.doi.org/10.1037/1040-3590.1.2.82.

Goodglass, H., Kaplan, E., Barresi, B. Boston Diagnostic Aphasia Examination. Austin, TX: Pro-Ed; 2001.

Harnish SM, Neils-Strunjas J, Lamy M, Eliassen JC. Use of fMRI in the study of chronic aphasia recovery after therapy: A case study. Topics in Stroke Rehabilitation. 2008; 15:468–483. http://dx.doi.org/10.1310/tsr1505-468. [PubMed: 19008206]

Hickin J, Mehta B, Dipper L. To the sentence and beyond: A single case therapy report for mild aphasia. Aphasiology. 2015; 29:1038–1061. http://dx.doi.org/10.1080/02687038.2015.1010474.

Hinckley JJ, Carr TH. Comparing the outcomes of intensive and non-intensive context-based aphasia treatment. Aphasiology. 2005; 19:965–974. http://dx.doi.org/10.1080/02687030544000173.

Howard, D., Patterson, K. Pyramids and Palm Trees: A test of semantic access from pictures and words. Bury St Edmunds, Suffolk; United Kingdom: Thames Valley: 1992.

Jacoby LL. On interpreting the effects of repetition: Solving a problem versus remembering a solution. Journal of Verbal Learning and Verbal Behavior. 1978; 17:649–667. http://dx.doi.org/10.1016/S0022-5371(78)90393-6.

Kaplan, E., Goodglass, H., Weintraub, S. Boston Naming Test. Philadelphia, PA: Lippincott Williams & Wilkins; 2001.

Karpicke JD, Roediger HL III. Repeated retrieval during learning is the key to long-term retention. Journal of Memory and Language. 2007; 57:151–162. http://dx.doi.org/10.1016/j.jml.2006.09.004.

Karpicke JD, Roediger HL III. The critical importance of retrieval for learning. Science. 2008 Feb 15.319:966–968. http://dx.doi.org/10.1126/science.1152408. [PubMed: 18276894]

Law SP, Wong W, Sung F, Hon J. A study of semantic treatment of three Chinese anomic patients. Neuropsychological Rehabilitation. 2006; 16:601–629. http://dx.doi.org/10.1080/09602010543000046. [PubMed: 17127568]

Lee JB, Kaye RC, Cherney LR. Conversational script performance in adults with non-fluent aphasia: Treatment intensity and aphasia severity. Aphasiology. 2009; 23:885–897. http://dx.doi.org/10.1080/02687030802669534.

Leonard C, Rochon E, Laird L. Treating naming impairments in aphasia: Findings from a phonological components analysis treatment. Aphasiology. 2008; 22:923–947. http://dx.doi.org/10.1080/02687030701831474.

Lott SN, Carney AS, Glezer LS, Friedman RB. Overt use of a tactile-kinesthetic strategy shifts to covert processing in rehabilitation of letter-by-letter reading. Aphasiology. 2010; 24:1424–1442. http://dx.doi.org/10.1080/02687030903580333. [PubMed: 21170161]

Lott SN, Sperling AJ, Watson NL, Friedman RB. Repetition priming in oral text reading: A therapeutic strategy for phonologic text alexia. Aphasiology. 2009; 23:659–675. http://dx.doi.org/10.1080/02687030801969539. [PubMed: 20664804]

Mcdaniel MA, Roediger HL III, Mcdermott KB. Generalizing test-enhanced learning from the laboratory to the classroom. Psychonomic Bulletin & Review. 2007; 14:200–206. http://dx.doi.org/10.3758/BF03194052. [PubMed: 17694901]

McNeil MR, Doyle PJ, Spencer K, Goda AJ, Flores D, Small S. Effects of training multiple form classes on acquisition, generalization and maintenance of word retrieval in a single subject. Aphasiology. 1998; 12:575–585. http://dx.doi.org/10.1080/02687039808249559.

Metcalfe J, Kornell N. Principles of cognitive science in education: The effects of generation, errors, and feedback. Psychonomic Bulletin & Review. 2007; 14:225–229. http://dx.doi.org/10.3758/BF03194056. [PubMed: 17694905]

Meyer AND, Logan JM. Taking the testing effect beyond the college freshman: Benefits for lifelong learning. Psychology and Aging. 2013; 28:142–147. http://dx.doi.org/10.1037/a0030890. [PubMed: 23437897]

Middleton EL, Schwartz MF, Rawson KA, Garvey K. Test-enhanced learning versus errorless learning in aphasia rehabilitation: Testing competing psychological principles. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2015; 41:1253–1261. http://dx.doi.org/10.1037/xlm0000091.

Milman L, Clendenen D, Vega-Mendoza M. Production and integrated training of adjectives in three individuals with nonfluent aphasia. Aphasiology. 2014; 28:1198–1222. http://dx.doi.org/10.1080/02687038.2014.910590.

Morris CD, Bransford JD, Franks JJ. Levels of processing versus transfer appropriate processing. Journal of Verbal Learning and Verbal Behavior. 1977; 16:519–533. http://dx.doi.org/10.1016/S0022-5371(77)80016-9.

Pastötter B, Weber J, Bäuml KH. Using testing to improve learning after severe traumatic brain injury. Neuropsychology. 2013; 27:280–285. http://dx.doi.org/10.1037/a0031797. [PubMed: 23527656]

Pulvermüller F, Neininger B, Elbert T, Mohr B, Rockstroh B, Koebbel P, Taub E. Constraint-induced therapy of chronic aphasia after stroke. Stroke. 2001; 32:1621–1626. http://dx.doi.org/10.1161/01.STR.32.7.1621. [PubMed: 11441210]

Rawson KA, Dunlosky J. Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? Journal of Experimental Psychology: General. 2011; 140:283–302. http://dx.doi.org/10.1037/a0023956. [PubMed: 21707204]

Raymer AM, Kohen FP, Saffell D. Computerised training for impairments of word comprehension and retrieval in aphasia. Aphasiology. 2006; 20:257–268. http://dx.doi.org/10.1080/02687030500473312.

Raymer AM, McHose B, Smith KG, Iman L, Ambrose A, Casselton C. Contrasting effects of errorless naming treatment and gestural facilitation for word retrieval in aphasia. Neuropsychological Rehabilitation. 2012; 22:235–266. http://dx.doi.org/10.1080/09602011.2011.618306. [PubMed: 22047100]

Robson J, Marshall J, Pring T, Montagu A, Chiat S. Processing proper nouns in aphasia: Evidence from assessment and therapy. Aphasiology. 2004; 18:917–935. http://dx.doi.org/10.1080/02687030444000462.

Roediger HL III, Karpicke JD. The power of testing memory: Basic research and implications for educational practice. Perspectives on Psychological Science. 2006a; 1:181–210. http://dx.doi.org/10.1111/j.1745-6916.2006.00012.x. [PubMed: 26151629]

Roediger HL III, Karpicke JD. Test-enhanced learning: Taking memory tests improves long-term retention. Psychological Science. 2006b; 17:249–255. http://dx.doi.org/10.1111/j.1467-9280.2006.01693.x. [PubMed: 16507066]

Rogalski EJ, Rademaker A, Wieneke C, Bigio EH, Weintraub S, Mesulam MM. Association between the prevalence of learning disabilities and primary progressive aphasia. Journal of the American

Medical Association Neurology. 2014; 71:1576–1577. http://dx.doi.org/10.1001/jamaneurol. 2014.2805. [PubMed: 25486208]

Sage K, Snell C, Lambon Ralph MA. How intensive does anomia therapy for people with aphasia need to be? Neuropsychological Rehabilitation. 2011; 21:26–41. http://dx.doi.org/ 10.1080/09602011.2010.528966. [PubMed: 21181603]

Samuels SJ. Decoding and automaticity: Helping poor readers become automatic at word recognition. Reading Teacher. 1988; 41:756–760.

Schneider SL, Thompson CK. Verb production in agrammatic aphasia: The influence of semantic class and argument structure properties on generalisation. Aphasiology. 2003; 17:213–241. http:// dx.doi.org/10.1080/729255456. [PubMed: 21311720]

Sumowski JF, Chiaravalloti N, Deluca J. Retrieval practice improves memory in multiple sclerosis: Clinical application of the testing effect. Neuropsychology. 2010; 24:267–272. http://dx.doi.org/ 10.1037/a0017533. [PubMed: 20230121]

Sumowski JF, Leavitt VM, Cohen A, Paxton J, Chiaravalloti ND, DeLuca J. Retrieval practice is a robust memory aid for memory-impaired patients with MS. Multiple Sclerosis. 2013; 19:1943– 1946. http://dx.doi.org/10.1177/1352458513485980. [PubMed: 23674564]

Sumowski JF, Wood HG, Chiaravalloti N, Wylie GR, Lengenfelder J, DeLuca J. Retrieval practice: A simple strategy for improving memory after traumatic brain injury. Journal of the International Neuropsychological Society. 2010; 16:1147–1150. http://dx.doi.org/10.1017/S1355617710001128. [PubMed: 20946709]

Thompson CK, Kearns KP, Edmonds LA. An experimental analysis of acquisition, generalisation, and maintenance of naming behaviour in a patient with anomia. Aphasiology. 2006; 20:1226–1244. http://dx.doi.org/10.1080/02687030600875655.

Tse CS, Balota DA, Roediger HL III. The benefits and costs of repeated testing on the learning of face-name pairs in healthy older adults. Psychology and Aging. 2010; 25:833–845. http://dx.doi.org/ 10.1037/a0019933. [PubMed: 20718541]

Tulving E. The effects of presentation and recall of material in free-recall learning. Journal of Verbal Learning and Verbal Behavior. 1967; 6:175–184. http://dx.doi.org/10.1016/ S0022-5371(67)80092-6.

van Hees S, Angwin A, McMahon K, Copland D. A comparison of semantic feature analysis and phonological components analysis for the treatment of naming impairments in aphasia. Neuropsychological Rehabilitation. 2013; 23:102–132. http://dx.doi.org/ 10.1080/09602011.2012.726201. [PubMed: 23098246]

Wambaugh JL, Martinez AL, McNeil MR, Rogers MA. Sound production treatment for apraxia of speech: Overgeneralization and maintenance effects. Aphasiology. 1999; 13:821–837. http:// dx.doi.org/10.1080/026870399401902.

Wisenburn B, Mahoney K. A meta-analysis of word-finding treatments for aphasia. Aphasiology. 2009; 23:1338–1352. http://dx.doi.org/10.1080/02687030902732745.
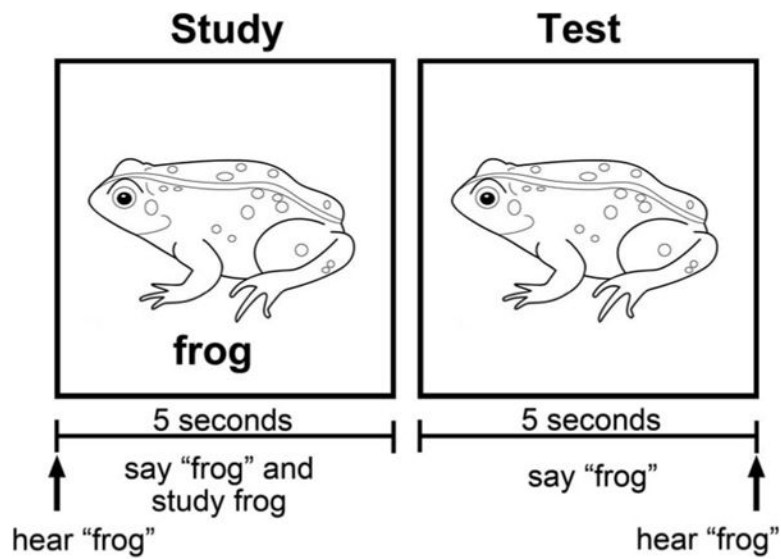
Figure 1.
Paradigm used during treatment for *test* and *study* conditions. *Study* blocks simultaneously display the picture image and the written word and present the auditory name of the picture. The participant must name the image within 5,000 ms. *Test* blocks display the to-be-named picture on the screen for 5,000 ms, during which time the participant names the image out loud. Following the 5,000 ms, participants hear the auditory recording of the name of the image as feedback. Post = posttreatment.
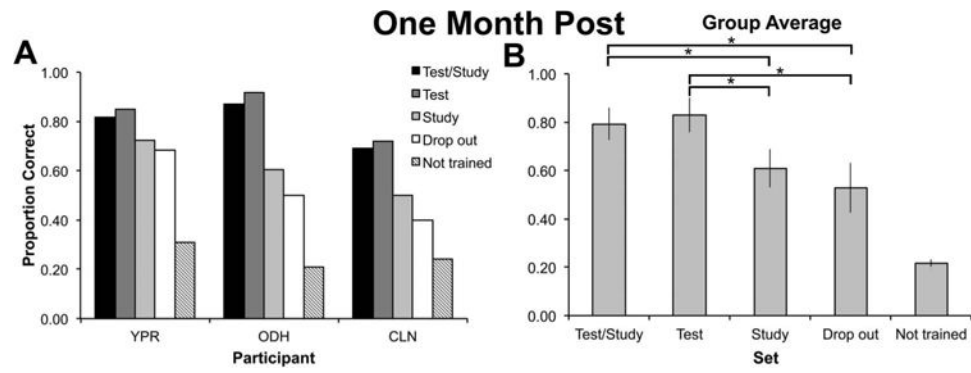
Figure 2.
Bars represent proportion of items correct for each condition, at 1 month posttreatment, for each of the three participants (Panel A; YPR, ODH, and CLN represent the three participants) and for the group as a whole (Panel B; error bars represent standard errors). Post = posttreatment. *p < .05.

Figure 3.
Bars represent proportion of items correct for each condition, at 4 months posttreatment, for each of the three participants (Panel A; YPR, ODH, and CLN represent the three participants) and for the group as a whole (Panel B; error bars represent standard errors). Post = posttreatment. $*p < .05$.
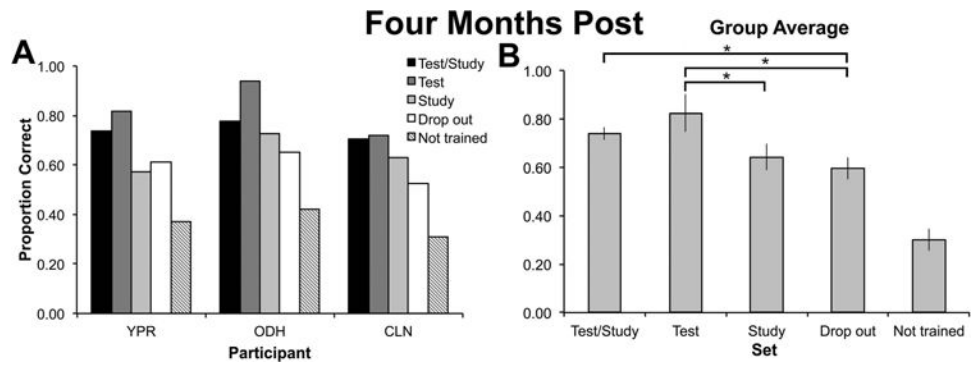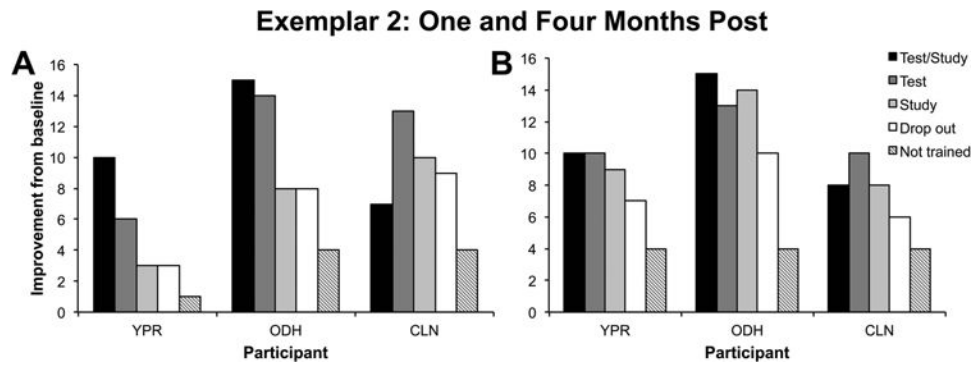
Figure 4.

Bars represent the number of additional items correct compared to baseline for each condition for Exemplar 2 images. Panel A: 1 month posttreatment (post). Panel B: 4 months posttreatment. YPR, ODH, and CLN represent the three participants.

Table 1

Scores on the Western Aphasia Battery and Boston Diagnostic Aphasia Examination

Measure	YPR			ODH			CLN		
	BL	1 MP	4 MP	BL	1 MP	4 MP	BL	1 MP	4 MP
Western Aphasia Battery[a]									
Aphasia Quotient (100)	58.8	59.2	DNT	79.3	82.4	DNT	77.3	74.9	DNT
Auditory Verbal Comprehension (10)	6.5	6.5	DNT	8.65	9.4	DNT	8.15	8.65	DNT
Repetition (10)	4.5	5	DNT	5.6	6.1	DNT	6	4.5	DNT
Naming and Word Finding (10)	5.4	5.1	DNT	7.4	7.7	DNT	6.5	6.3	DNT
Boston Diagnostic Aphasia Examination (BDAE)[b]									
Auditory Comprehension[c] (100)	12	15	15	35	43	47	23	38	40
Repetition									
Words (10)	7	8	7	7	7	7	4	4	3
Sentences (10)	0	1	0	3	1	1	1	1	1
Boston Naming Test[d] (60)	14	14	12	19	18	17	11	13	12
Single Word Reading									
Concrete (30)	23	DNT	DNT	30	DNT	DNT	25	DNT	DNT
Abstract (30)	14	DNT	DNT	29	DNT	DNT	23	DNT	DNT

Note. Numbers in parentheses represent the maximum possible score. YPR, ODH, and CLN represent the three participants. BL = baseline; 1MP = 1 month posttreatment; 4MP = 4 months posttreatment; DNT = did not test.

[a]...

[b]Goodglass, Kaplan, and Barresi (2001).

[c]The BDAE Auditory Comprehension score is the mean of the three subtest percentiles.

[d]Kaplan, Goodglass, and Weintraub (2001).

Table 2

Baseline Scores on Semantic Tasks

Participant	Pyramids and Palm Trees[a] pictures (52)	Spoken word–picture match (48)	Written word–picture match (53)
YPR	51	47	52
ODH	51	46	52
CLN	50	44	52

Note. Numbers in parentheses represent the maximum possible score. YPR, ODH, and CLN represent the three participants.

[a]Howard and Patterson (1992).

Table 3

Overtraining Conditions

Status and condition	Test 1	Study 1	Test 2	Study 2
Not yet learned	✓	✓	✓	✓
Overtraining conditions for learned items				
Test and study (continue to test and study)	✓	✓	✓	✓
Test only (continue to test only)	✓		✓	
Study only (continue to study only)		✓		✓
Dropout (no longer tested or studied)				

Note. Four blocks were run in every session. All items were initially "not yet learned" and appeared in all four blocks. After an item was "learned" (correct on two consecutive sessions), it was assigned to one of the four overtraining conditions.