Do Testing Effects Change Over Time?

Insights from Immediate and Delayed Retrieval Speed

Abbreviated Title: Do Testing Effects Change Over Time

Abstract

Retrieving information from memory improves recall accuracy more than continued studying, but this testing effect often only becomes visible over time. In contrast, the present study documents testing effects on recall *speed* both immediately after practice and after a delay. Forty participants learned the translation of 100 Swahili words and then further restudied the words with translations or retrieved the translations from memory during testing. As in previous experiments, recall accuracy was higher for restudied words than for tested words immediately after practice, but higher for tested words after seven days. Response times for correct answers, however, showed a different result: Learners were faster to recall tested words than restudied words both immediately after practice and after seven days. These results are interpreted in light of recent suggestions that testing selectively strengthens cue-response associations. An additional outcome was that testing effects on recall accuracy were related to perceived retrieval success during practice. When several practice retrievals were successful, testing effects on recall accuracy were significant already immediately after practice. Together with the reaction time data, this supports recent models that attribute changes in testing effects over time to limited item retrievability during practice.

Keywords: testing effects; retrieval speed; response times; retrieval success; word learning

Do Testing Effects Change Over Time?

Insights from Immediate and Delayed Retrieval Speed

Numerous studies have documented *testing effects*, i.e., the phenomenon that retrieving information from memory improves the long-term retention of that information more than continued studying (review in Roediger & Butler, 2011). For example, learners benefit less from *restudying* a foreign vocabulary and its translation than from retrieving the translation from memory like on a *test* (e.g., Carrier & Pashler, 1992; Metcalfe & Kornell, 2007). However, these benefits of testing are often only visible after a delay and not immediately after practice, when outcomes may even be better for restudied materials than for tested materials (for reviews, see Kornell, Bjork, & Garcia, 2011; Roediger & Karpicke, 2006; Toppino & Cohen, 2009). In the present study, we investigated why this is the case by analyzing response times after restudy and testing practice, and by relating later recall to judgments of retrieval success during practice.

Although there is a growing literature on the cognitive mechanisms that might underlie testing effects, it is not yet clear why testing effects change over time. For example, a prominent account is that testing improves the efficiency of later recall processes (Karpicke & Smith, 2012), such that relevant information comes to mind earlier and less irrelevant associations are activated (Thomas & McDaniel, 2012). The exact mechanisms of this process have not been established yet, but they could involve increased suppression of competing irrelevant information after repeated selection of target-information during testing (M. C. Anderson, Bjork, & Bjork, 1994, 2000). Also, the search set of items treated as candidates in response to retrieval cues could be reduced (Karpicke & Smith, 2012), for example, due to refined mnemonic associations (Pyc & Rawson, 2010) or improved recapitulation of the encoding context (Jacoby, Shimizu, Daniels, & Rhodes, 2005).

Such mechanistic accounts, however, do not readily explain why benefits of testing practice are typically only visible after a delay and not immediately after learning. In the literature, this timing of testing effects is usually discussed in terms of reduced forgetting after testing in comparison to restudying (Carpenter, Pashler, Wixted, & Vul, 2008; Wheeler, Ewers, & Buonanno, 2003) and it has been suggested that the accessibility of items in memory decreases faster for weak (restudied) than for strong (tested) memories (Bjork & Bjork, 1992). The present study investigates an alternative explanation which was recently presented by Halamish and Bjork (2011) and Kornell, Bjork and Garcia (2011), who suggested that the timing of testing effects can be explained without assuming differences in forgetting rates, referring only to limited retrieval success during testing practice.

Limited retrieval success during testing practice could explain the timing of testing effects because it leads to a “bifurcation” of items (Kornell, et al., 2011, p. 85) into some tested items with high memory strength (those that were successfully retrieved during practice) and some with low memory strength (those that were not retrieved during practice). In contrast, restudying should lead to a comparably large number of items with moderate memory strength, assuming that restudying is not as effective as successful testing but more effective than unsuccessful testing (Kornell, et al., 2011). Assuming further that the high memory strength of successfully tested items and the moderate memory strength of restudied items but not the low memory strength of unsuccessfully tested items is sufficient for recall on a later test, the situation can arise that more restudied items than tested items are recalled although the average memory strength of the restudied items is lower than the average memory strength of the successfully tested items (Kornell, et al., 2011). However, memory strength decays over time and due to their initially higher memory strength, (successfully) tested items are more likely than restudied items to remain accessible enough for recall over time, leading to higher recall on delayed tests. Note that this explanation of changes in testing effects over time does not require that the memory decay over time is different for tested and restudied items.

The bifurcation model was based on previous studies of testing effects on recall accuracy: The strongest support for the model comes from experiments showing that increasing the difficulty of performance measures can make testing effects visible already immediately after learning, arguing that only (successfully tested) items with high memory strength but not restudied items with moderate memory strength can be recalled on such relatively difficult tests (Halamish & Bjork, 2011). However, in order to directly test the bifurcation model, measures of recall accuracy do not suffice because they only provide information on the outcome of the recall (recalled or not recalled), and not on the difficulty of the recall. In the present study, we therefore measured response times to collect additional information on the difficulty of the retrieval act and on the accessibility of the target information among competing representations in memory, assuming that longer reaction times reflect more difficulty in retrieving information (cf. J. R. Anderson, 1981; MacLeod & Nelson, 1984; Wixted & Rohrer, 1993).

The first purpose of this study was to investigate whether testing practice (in comparison to restudying) influences later retrieval speed at all. The facilitation of later retrieval processes as described by mechanistic accounts of testing effects has so far almost always been measured in terms of the amount of information which the learners recalled but it is likely that recalls also become *faster* if more efficient retrieval routes become available. Although there has been some interest in changes of response times over the course of repeated retrieval practice (e.g., Karpicke & Roediger, 2007), very few studies measured response times *after* restudy and testing practice. The first study that we found dates back to the 1980s, when MacLeod and Nelson (1984) reported shorter response times but lower recall success immediately after four testing cycles in comparison to three study cycles and one testing cycle. Testing effects on response times did not reach statistical significance in their study, but the authors concluded that accuracy and response times reflect different dimensions of memory, with accuracy depending on whether an item is sufficiently encoded to be retrieved at all, and response times depending on processing steps necessary during retrieval (MacLeod and Nelson, 1984). More support for the relevance of testing effects on response times comes from recent Neuroimaging studies of testing effects focusing at its neural correlates, in which significant response time effects were reported as a side result (Keresztes, Kaiser, Kovács, & Racsmány, in press; van den Broek, Takashima, Segers, Fernández, & Verhoeven, 2013). Therefore, the present study was set up to more systematically investigate whether testing not only improves recall accuracy but also recall speed indicating that testing practice reduces the amount of processing needed for later memory retrieval.

The second purpose of this study was to test the bifurcation explanation of changes in testing effects over time. This was done in two ways. First, we investigated if and how testing effects on response times change over time. The bifurcation model predicts that testing effects on response times should, unlike testing effects on recall accuracy, be visible already immediately after learning and remain visible over time. The reason for this is that the memory strength of those tested items that are successfully retrieved during practice and thus remembered over time should be higher than that of restudied items, both immediately after learning (even when at that moment overall less tested items than restudied items are recalled) and on delayed tests. From this, we derived the hypothesis that both immediately after learning as well as on a delayed test, response times for correctly remembered items would be shorter for tested than for restudied items. A different possible outcome would testing effects on response times change over time similar to testing effects on recall accuracy, such that testing only leads to shorter response times after a delay but not immediately after learning. In that case, the data would directly contradict the bifurcation model but be in line with the idea that testing effects only appear after a delay because memory representations of tested items are more resistant to forgetting over time than representations of restudied items (Carpenter, et al., 2008; Wheeler, et al., 2003).

As a second test of the bifurcation model, we collected judgments of retrieval success during practice to investigate the prediction that testing effects are restricted to items that are successfully retrieved during practice. So far, there has been limited direct research on this topic. In one recent study, Jang and colleagues used an initial test to establish retrievability of items, and then exposed participants to further restudy and testing practice (Jang, Wixted, Pecher, Zeelenberg, & Huber, 2012). By dividing the data into retrievable and nonretrievable items, they showed that immediate benefits of restudy over testing were almost completely explained by effects on initially nonretrievable items, whereas delayed benefits of testing over restudy were explained fully by testing effects on initially retrievable items. In the present study, we further explored the relation between item retrievability and the timing of testing effects.

Methods

Participants. Forty female university students (M_{age}= 19.5 years, SD_{age} = 2.1) from a Psychology Participant Pool took part in the experiment for course credits or a monetary compensation (10 Euro per hour). To increase their motivation, there was an additional bonus of 10 Euro for the 10% best performing participants. Participants reported investing a high amount of mental effort during practice, with an average score of 15.9 (*SD* = 2.6) on a 20-point rating scale (0 = *very low effort*, 20 = *very high effort*). All participants spoke Dutch fluently (88% native speakers), and none of them had prior knowledge of Swahili.

Stimuli. The stimuli were 100 Swahili nouns with Dutch translations, which were pronounceable for Dutch speakers, such as *bustani* (garden), *kaza* (work), *anga* (sky), *samaki* (fish), *jiwe* (stone), *tofaa* (apple).

Overview of the Experiment. There were two sessions: The first session comprised an initial encoding phase, a practice phase with testing and restudy trials, and an immediate test. The second session seven days later comprised a second test. Session 1 took about 1 hour and 40 minutes; Session 2 took about 20 minutes.

Encoding phase. The purpose of the initial encoding phase was to ensure that participants learned the meaning of the majority of the words and to control for item-selection differences between testing and restudy condition (cf. Karpicke & Smith, 2012). For this purpose, we used an adaptive study program that presented the word-pairs one at a time, in a randomized order, and let participants indicate after each presentation whether they thought they knew the word-pair or not. The presentation of each pair continued until the participants had indicated in two consecutive encoding rounds that they knew the pair. In addition, all word-pairs were presented one more time at the end of the encoding phase to control for recency effects. The presentation durations for the word-pairs were reduced in steps of 500 ms for every encoding round from 4000 ms in the first round to a minimum duration of 2000 ms. To minimize opportunities for retrieval during the encoding phase, Swahili words were always presented simultaneously with their translation. At the end of encoding, words were randomly assigned to the testing, restudy, or control condition for every participant in such a way that the mean number of presentations during the encoding phase was equal in all conditions ($M_{\mathrm{T}} = 4.6$, $SD_{\mathrm{T}} = 3.1$; $M_{\mathrm{RS}} = 4.6$, $SD_{\mathrm{RS}} = 3.1$).

Practice phase. The critical experimental manipulation took place in the practice phase, when the participants practiced 40 of the 100 previously encoded words in a *testing* condition and 40 of the words in a *restudy* condition. The remaining 20 words served as controls that were not presented during practice. The difference between the conditions was that the complete word-pair was visible on the screen in the *restudy* condition (e.g., *roho* - *soul*), whereas only the Swahili word was visible in the *testing* condition (e.g., *roho – xxx*).

The words were presented for 800 ms before they were replaced by a prompt to make a retrieval success judgment. There were three practice blocks, in which trials were presented in a randomized order. Each block lasted about 9 minutes.

Analysis of perceived retrieval success during practice. To obtain a measure of perceived retrieval success during practice, the participants answered the question “Did you already know the translation?” with “Yes” or “No” after each practice trial. Responses for the three practice rounds were then summarized in five categories: No/No/No (*NNN*), No/No/Yes (*NNY*), No/Yes/Yes (*NYY*), Yes/Yes/Yes (*YYY*), and any other combination in a rest category. For example, NYY indicates words to which participants responded “No” in the first practice block, and “Yes” in the second and third practice block. The words in the “Rest” category (4.8 % of all words) were not included in the analysis reported here, as they form a less interpretable category. However, including them did not change the overall picture of results.

Immediate and delayed test. Every participant was tested on a random selection of 10 words from each condition immediately after practice and on the remaining words on a delayed test after seven days. During both tests, the participants saw the Swahili words (one at a time) on a computer screen and entered the Dutch translation with the keyboard. Responses were later categorized as either correct or incorrect. The test program (Inquisit 3.0.4.0 (2009). Seattle, WA: Millisecond Software LLC) recorded how long it took the students to fill in the translation and to click on a button to proceed to the next word, after the Swahili word had appeared on screen. The students received no instruction to respond fast. Only response times for correct responses were included in the following data analyses, in order to avoid confounding by performance differences between the conditions (correct and incorrect responses often differ in terms of response times (e.g., J. R. Anderson, 1981)). Individual response times that deviated more than three standard deviations from the participant's average response time (these were 1.3 % of all correct responses) were excluded before response times were summarized per participant for further statistical analysis.

Data Analysis. Data on participant level were subjected to two 3 x 2 repeated measures analyses of variance (ANOVA) with Practice Condition (Test, Restudy, Control) and Testing Moment (immediate, delayed) as within subject factors and Later Recall (i.e., the mean proportion of correctly translated words) or Response Times for correct responses as dependent variables. In a second step, the word-specific data were subjected to a repeated measures logistic regression analysis with SPSS Generalized Estimating Equations function to account for the hierarchical structure of the data (words in participants) (cf. Hanley, Negassa, & Forrester, 2003). We entered Practice Condition (Test, Restudy), Testing Moment (immediate, delayed), and Retrieval Success during Practice (NNN, NNY, NYY, YYY) as predictors and Later Recall Success (correct = 1, not correct = 0) as dependent variable. Note that we could not analyze the relation between retrieval success during practice and response times for correct answers in the same way because there were not enough correct answers for some categories of retrieval success (in particular, very few words were later correctly recalled on the test if they could not be retrieved during practice before). All analyses were performed using SPSS version 15.01.

Results

Table 1 contains summary statistics for later recall success (the proportion of correctly translated words) and response times for correct answers.

Recall Success. There were significant main effects of Time, $F(1, 39) = 87.94$, $p < .001$, $\eta_p^2 = .69$ and Practice Condition, $F(2, 78) = 15.67$, $p < .001$, $\eta_p^2 = .29$ on Recall Success*;* as well as an interaction between the two factors, $F(2, 78) = 17.21$, $p <.001$, $\eta_p^2 = .31$. Further investigation of this interaction with t-tests for paired samples revealed a classic testing effect: On the immediate test, performance was significantly better for restudied words than for tested words, $t(39) = -3.58$, $p = .001$, $d = 0.57$, and control words, $t(39) = 3.85$, $p < .001$, $d = 0.61$, whereas the tested and control words did not differ from each other, $t(39) = 0.64$, $p = .53$, $d = 0.10$. On the delayed test after seven days, the effect was reversed: Performance was significantly better for tested words than for restudied words, $t(39) = 5.57$, $p < .001$, $d = 0.88$, and control words, $t(39) = 7.38$, $p < .001$, $d = 1.17$. Performance was marginally better for restudied words than for control words, $t(39) = 2.015$, $p = .051$, $d = 0.32$.

(Table 1. Summary statistics Recall Success and Response Times)

Response Times. There were significant main effects of Time, $F(1, 34)$[1] = 16.33, $p < .001$, $\eta_p^2 = .32$, and Practice Condition, $F(2, 68) = 6.70$, $p = .002$, $\eta_p^2 = .17$ on Response Times for correct responses, but no interaction between the two factors, $F(2, 68) = 1.08$, $p = .35$, $\eta_p^2 = .031$. The main effect of Time was caused by shorter response times immediately after practice than on the test after seven days. The main effect of Practice Condition was caused by shorter overall response times for tested words (estimated marginal mean: 4690 ms) than for restudied words (estimated marginal mean: 5066 ms), $F(1, 34) = 10.95$, $p = .002$, $\eta_p^2 = .24$, and shorter response times for tested words than for control words (estimated marginal mean: 5173 ms), $F(1, 34) = 11.39$, $p = .002$, $\eta_p^2 = .25$. The difference in response times between control and restudied words was not significant, $F(1, 34) = .47$, $p = .50$, $\eta_p^2 = .01$.

Perceived Retrieval Success During Practice. To check the reliability of participants' judgments of retrieval success we compared the retrieval success judgments of test items during practice with the recall accuracy of the same items on the immediate test. We found that in case participants indicated during the last practice round that they knew the translation of a word, they correctly recalled the translation on the immediate test a few minutes later in 85.6% (*n* = 268, retrieval condition) or 81.4% (*n* = 301, restudy condition) of the cases. This indicates that the retrieval judgments were quite reliable.

To test the research questions related to the bifurcation model, we further investigated the relation between perceived retrieval success during practice and later recall on the two testing moments (see Figure 1). The number of words (percentage of all words in parentheses) per retrieval success category were as follows: tested words 221 NNN (15%), 43 NNY (2.9%), 77 NYY (5.2%), and 1133 YYY (76.9%); restudied words 42 NNN words (2.7%), 38 NNY (2.5%), 57 NYY (3.7%), and 1396 YYY words (91.1%). We further investigated these data with repeated measures logistic regression analyses with words within participants as units of analysis. First, we tested a simple model with main effects of Practice Condition, Retrieval Success during Practice, and Time. All main effects were significant, due to, respectively, higher later recall success for tested words than for restudied words, $\chi^2(1) = 64.02$, $p < .001$, higher later recall success when words were retrieved more often during practice $\chi^2(3) = 125.15$, $p < .001$, and higher recall success on the immediate test than on the test after seven days $\chi^2(1) = 93.31$, $p < .001$ (a complete overview of B values, SE_B and confidence intervals of the odds ratios can be found in Appendix 1). In a second step, we added the interaction between Practice Condition and Retrieval Success to the model, which was significant, $\chi^2(3) = 12.03$, $p = .007$ due to the fact that testing effects on Later Recall were significant for the YYY, $\chi^2(1) = 52.70$, $p < .001$, and the NYY words $\chi^2(1) = 22.44$, $p < .001$, but not significant[2] for the NNY, $\chi^2(1) =0.001$, $p = .97$, or the NNN words, $\chi^2(1) = 0.76$, $p = .38$. The graphs in Figure 1 suggest that this effect was more pronounced on the delayed test than on the immediate test, but in subsequent analyses, the 3-way interaction between Practice Condition, Retrieval Success during Practice, and Time was not significant, $\chi^2(1) = 0.49$, $p = .92$.

(Figure 1. Later Recall against Practice Condition and Perceived Retrieval Success)

Discussion

In the present study, we investigated immediate and delayed effects of successful retrieval during testing practice on later recall accuracy and response times. There were three main results. First, testing improved not only later recall accuracy but also response times in comparison to restudying. Second, the timing of these effects differed: As in previous studies, testing effects on recall accuracy only became visible over time (overviews in Kornell, et al., 2011; Roediger & Karpicke, 2006). In contrast, testing effects on response times were visible already immediately after practice as well as after seven days. Third, testing effects on later recall accuracy were related to retrieval success during practice: For those words for which participants indicated that they successfully retrieved the translation in at least two practice rounds, there were testing effects already immediately after practice as well as after seven days. Together, these results indicate that testing improves memory both in terms of later recall success and recall speed but affects only those items that are retrieved successfully during practice. Such limited item retrievability could explain why overall testing effects on recall success only became visible on the delayed test, whereas testing effects on response times for correct answers were already visible immediately after practice.

First, the fact that learners not only recalled *more* tested words than restudied words on the delayed test, but also recalled the tested words *faster*, suggests that successful retrieval practice increases both the chance that information can later be recalled and the accessibility of that information in memory in terms of processing steps (i.e., time) needed for recall. This interpretation of reaction time results fits well with recent accounts that testing effects could partly be due to increased efficiency of practiced recall processes (Karpicke & Smith, 2012). In terms of the present study, testing may have facilitated the activation of the correct translation in response to the Swahili cue or increased the suppression of incorrect translations. Importantly, this facilitation of later retrieval processes has so far only been measured in terms of the *amount* of recalled information but if testing works by "narrowing the scope of the memory search [during later retrieval] to hone in on targeted information" (Thomas & McDaniel, 2012, p.1), a straight forward prediction is that retrieval should also become *faster*. Therefore, the reduced response times that we found after testing than restudying support mechanistic accounts that explain testing with the (selective) strengthening of cue-response associations.

The present results converge with the few previous studies that reported reaction time outcomes after testing and restudy practice (Keresztes, et al., in press; MacLeod & Nelson, 1984; van den Broek, et al., 2013). Note that Keresztes et al. (in press) reported significant testing effects on reaction times both immediately after learning and after a delay of one week, similar to the results reported here, but used onset latencies to measure reaction times whereas submission latencies were used in the present study. The fact that the pattern of results was the same in both studies, suggests that testing effects on response times generalize across different measurements (i.e., onset and submission latencies).

Second, the reported results support the bifurcation model in two ways. First, the pattern of changes over time that we found for response times and recall accuracy support the bifurcation idea that items that are remembered after testing practice may have a higher memory strength than restudied items, even at a moment when the number of recalled tested items is smaller than the number of recalled restudied items (Halamish & Bjork, 2011; Jang, et al., 2012; Kornell, et al., 2011). Reaction times were shorter for tested words than for restudied words already immediately after learning, although at that moment overall recall was higher for restudied than for tested words. There was, however, no difference in the rate with which response speed decreased over time for restudied and tested materials. Therefore, the present results do not support theories that changes in testing effects over time are due to differences in forgetting rates (Carpenter, et al., 2008; Wheeler, et al., 2003). Further research is needed with more measurement moments to determine how exactly reaction times change after repeated testing and restudy practice. However, as far as the present study goes, the timing of effects on reaction times can be explained just by referring to limited item retrievability during practice.

The present results also support this bifurcation idea in a second way because when participants indicated that two or three practice retrievals were successful, recall success was better for tested than for restudied items, and this was the case already immediately after learning as well as after seven days. These results are in line with the bifurcation model (Halamish & Bjork, 2011; Kornell, et al., 2011). However, a replication of the reported results with a more objective measure of retrieval success is desirable because in the present study, the accuracy of judgments could differ for testing and restudy trials (cf. Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008), which could partly explain differences in recall success. To control for this potential confound, we repeated our analyses with a measure of word difficulty as covariate (the average recall for each specific word when used as control word) to correct for differences between word difficulty of tested and restudied words within the categories of retrieval success judgments. This analysis again showed that testing led to higher later recall success than restudying at both measurement moments (only) when two or three retrievals were successful during practice. Hence, the conclusion seems warranted that testing without feedback can indeed improve recall success already immediately after learning if several practice retrievals are successful. This is in line with previous studies showing strong benefits of repeated retrieval over a single retrieval opportunity (e.g., Karpicke & Roediger, 2008). However, more research is needed to establish how many successful retrievals are necessary to produce such immediate testing effects.

To conclude, the present study showed that successful retrieval during testing increases not only the amount of information that is remembered over time but also the speed with which that information is accessed. We documented these testing effects on response times at a moment when testing effects on recall success were not yet visible, which supports the idea that limited item retrievability could explain why overall testing effects on recall success only became visible over time. These results open up interesting new possibilities to investigate changes in the accessibility of memories after repeated testing practice even when recall accuracy is at a ceiling level. The reported results further improve insight into the powerful memory-enhancing effects of testing as a tool for learning by measuring not only response accuracy but also response times.

References

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger, H. L., & McDermott, K. B. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology, 22*(7), 861-876. doi: 10.1002/acp.1391

Anderson, J. R. (1981). Interference: The relationship between response latency and response accuracy. *Journal of Experimental Psychology: Human Learning and Memory, 7*(5), 326. doi: 10.1037/0278-7393.7.5.326

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(5), 1063. doi: 10.1037/0278-7393.20.5.1063

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (2000). Retrieval-induced forgetting: Evidence for a recall-specific mechanism. *Psychonomic Bulletin & Review, 7*(3), 522-530. doi: 10.3758/BF03214366

Bjork, R. A. (1994). Memory and meta-memory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185-205). Cambridge, MA: MIT Press.

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

Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory & Cognition, 36*(2), 438-448. doi: 10.3758/mc.36.2.438

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition, 20*(6). doi: 10.3758/BF03202713

Halamish, V., & Bjork, R. A. (2011). When Does Testing Enhance Retention? A Distribution-Based Interpretation of Retrieval as a Memory Modifier. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(4), 801-812. doi: 10.1037/a0023219

Hanley, J. A., Negassa, A., & Forrester, J. E. (2003). Statistical analysis of correlated data using generalized estimating equations: an orientation. *American journal of epidemiology, 157*(4), 364-375. doi: 10.1093/aje/kwf215

Jacoby, L. L., Shimizu, Y., Daniels, K. A., & Rhodes, M. G. (2005). Modes of cognitive control in recognition and source memory: Depth of retrieval. *Psychonomic Bulletin & Review, 12*(5), 852-857. doi: 10.3758/BF03196776

Jang, Y., Wixted, J. T., Pecher, D., Zeelenberg, R., & Huber, D. E. (2012). Decomposing the interaction between retention interval and study/test practice: The role of retrievability. *The Quaterly Journal of Experimental Psychology., 65*(5), 962-975. doi: 10.1080/17470218.2011.638079

Karpicke, J. D., & Roediger, H. L. (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(4), 704. doi: 10.1037/0278-7393.33.4.704

Karpicke, J. D., & Roediger, H. L. (2008). The Critical Importance of Retrieval for Learning. *Science, 319*(5865), 966-968. doi: 10.1126/science.1152408

Karpicke, J. D., & Smith, M. A. (2012). Separate mnemonic effects of retrieval practice and elaborative encoding. *Journal of Memory and Language, 67*(1), 17-29. doi: 10.1016/j.jml.2012.02.004

Keresztes, A., Kaiser, D., Kovács, G., & Racsmány, M. (in press). Testing Promotes Long-Term Learning via Stabilizing Activation Patterns in a Large Network of Brain Areas. *Cerebral Cortex*.

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language, 65*(2), 85-97. doi: 10.1016/j.jml.2011.04.002

MacLeod, C. M., & Nelson, T. O. (1984). Response latency and response accuracy as measures of memory. *Acta Psychologica, 57*(3), 215-235. doi: 10.1016/0001-6918(84)90032-5

Metcalfe, J., & Kornell, N. (2007). Principles of cognitive science in education: The effects of generation, errors, and feedback. *Psychonomic Bulletin & Review, 14*(2), 225-229. doi: 10.3758/BF03194056

Pyc, M. A., & Rawson, K. A. (2010). Why Testing Improves Memory: Mediator Effectiveness Hypothesis. *Science, 330*(6002), 335. doi: 10.1126/science.1191465

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15*(1), 20-27. doi: 10.1016/j.tics.2010.09.003

Roediger, H. L., & Karpicke, J. D. (2006). Test-Enhanced Learning: Taking memory tests improves long-term memory. *Psychological Science, 17*(3), 249-255. doi: 10.1111/j.1467-9280.2006.01693.x

Thomas, R. C., & McDaniel, M. A. (2012). Testing and Feedback Effects on Front-End Control Over Later Retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Advance online publication. doi: 10.1037/a0028886

Toppino, T. C., & Cohen, M. S. (2009). The Testing Effect and the Retention Interval: Questions and Answers. *Experimental Psychology, 56*(4), 252-257.

van den Broek, G. S., Takashima, A., Segers, E., Fernández, G., & Verhoeven, L. (2013). Neural Correlates of Testing Effects in Vocabulary Learning. *Neuroimage, 78*, 94-102.

Wheeler, M., Ewers, M., & Buonanno, J. (2003). Different rates of forgetting following study versus test trials. *Memory, 11*(6), 571 - 580. doi: 10.1080/09658210244000414

Wixted, J. T., & Rohrer, D. (1993). Proactive interference and the dynamics of free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*(5), 1024-1039. doi: 10.1037/0278-7393.19.5.1024

Footnotes

[1] Three participants were not included in this analysis because they did not correctly recall any words from the restudy condition on the second test, and therefore had a missing value for response times for correct recalls. Two more participants were excluded because their score on at least one variable was a univariate outlier (z-score > 3.29). Excluding these outlier cases did not change the direction or significance of results.

[2] Statistical power to investigate testing effects on the 81 NNY and 263 NNN words was limited due to the relatively small number of words in relation to the very small observed differences in recall success (0.03 and 0.02 respectively). However, performance in these conditions was actually slightly higher for the restudied words than for the tested words. Therefore, it is unlikely that the absence of significant benefits of testing is simply due to a lack of power.

Table 1.

Average Proportion of Swahili Words Translated Correctly (short: Recall Success) and Average Response Times for Correct Responses (ms), per Practice Condition, as Measured Immediately and Seven Days After Practice.

Testing moment	**Dependent variable**	**Retrieval**		**Restudy**		**Control**	
		M	*SD*	***M***	*SD*	***M***	*SD*
Immediate	Recall Success	0.69	0.24	0.77	0.22	0.67	0.29
	Response time	4105	897	4654	1313	4826	1332
After 7 days	Recall Success	0.56	0.25	0.45	0.25	0.40	0.24
	Response time	5275	1104	5478	1567	5520	2066

Note. Response times are based on correct responses only.

Appendix 1.

Test statistics for the logistic regression analysis of word-level data of Later Recall (1 = correct, 0 = incorrect) against Practice Condition, Perceived Retrieval Success during Practice, and Testing Moment.

	Model with main effects		**Main effects and Interaction PC x PRSP**	
	B (SE_B)	**OR [95%CI]**	**B (SE_B)**	**OR [95%CI]**
Intercept				
Intercept	-2.51(0.36)	0.08 [0.04-0.16]	-1.34 (0.48)	0.26 [0.10 – 0.67]
Practice Condition (PC)				
Restudy [a]	0	1	0	1
Testing	0.76*** (0.09)	2.13 [1.77 – 2.56]	-0.49 (0.50)	0.61[0.23 – 1.62]
Perceived Retrieval Success during Practice (PRSP)	Wald $\chi^2(3)$ = 125.15, p < .001		Wald $\chi^2(3)$ = 102.99, p < .001	
No-No-No [a]	0	1	0	1
No-No-Yes	1.86*** (0.31)	6.43 [3.52 - 11.75]	1.15** (0.41)	3.15 [1.40 - 7.07]
No-Yes-Yes	2.67 *** (0.34)	14.50 [7.43 - 28.30]	1.04 $^{p\,=.068}$ (0.57)	2.827 [0.93 - 8.62]
Yes-Yes-Yes	3.60 *** (0.33)	36.56 [19.21 - 69.57]	2.47*** (0.48)	11.79 [4.60 - 30.26]
Testing Moment (TM)				
Immediate [a]	0	1	0	1
Delayed	-1.29*** (0.13)	0.276 [0.21 – 0.36]	-1.29*** (0.13)	0.27 [0.21 – 0.36]
Interaction PC x PRSP	NI		Wald $\chi^2(3)$ = 12.029, p = .007	
Restudy x NNN	NI	NI	0[a]	1
Restudy x NNY	NI	NI	0[a]	1
Restudy x NYY	NI	NI	0[a]	1
Restudy x YYY	NI	NI	0[a]	1
Retrieval x NNN	NI	NI	0[a]	1
Retrieval x NNY	NI	NI	0.51(0.57)	1.67 [0.54 - 5.11]
Retrieval x NYY	NI	NI	2.04** (0.61)	7.71 [2.31 - 25.72]
Retrieval x YYY	NI	NI	1.25* (0.50)	3.48 [1.31 – 9.24]
Interaction PC x TM	NI		NI	
Interaction TM x PRSP	NI		NI	
Immediate x NNN	NI	NI	NI	NI
Immediate x	NI	NI	NI	NI

	NNY				
	Immediate x NYY	NI	NI	NI	NI
	Immediate x YYY	NI	NI	NI	NI
	Delayed x NNN	NI	NI	NI	NI
	Delayed x NNY	NI	NI	NI	NI
	Delayed x NYY	NI	NI	NI	NI
	Delayed x YYY	NI	NI	NI	NI
3-way interaction PC x TM x PRSP		NI	NI	NI	NI

Note. OR = odds ratio; CI = confidence interval; NI = not included in model; PC = Practice Condition; PRSP = Perceived retrieval success during the three practice rounds (NNN = No/No/No, NNY = No/No/Yes, NYY = No/Yes/Yes, and YYY = Yes/Yes/Yes); TM = Testing Moment.

[a] set to zero because parameter is redundant.

*** *p* < .001, ** *p* < .01, * *p* < .05.