

Science, in press

The Critical Importance of Retrieval for Learning

Jeffrey D. Karpicke¹ and Henry L. Roediger, III²

¹Department of Psychological Sciences, Purdue University, West Lafayette, IN 47907, USA.

²Department of Psychology, Washington University in St. Louis, St. Louis, MO 63130, USA.

Category: Report

Summary: Tests are widely considered neutral assessments of learning, but here we show that the act of retrieval during testing produces much more learning than studying.

Address correspondence to:

Jeffrey D. Karpicke
Department of Psychological Sciences
Purdue University
703 Third Street
West Lafayette, IN 47907-2081

Email: karpicke@purdue.edu

Phone: (765) 494-0273

Fax: (765) 496-1264

This is the author's version of the work. It is posted here by permission of the AAAS for personal use, not for redistribution. The definitive version was published in *Science*, 319, 966 (2008), doi:10.1126/science.1152408, and may be found at <http://www.sciencemag.org/cgi/content/full/319/5865/966>

Abstract

Learning is often considered complete when a student can produce the correct answer to a question. In our research, students in one condition learned foreign language vocabulary words in the standard paradigm of repeated study-test trials. In three other conditions, once a student had correctly produced the vocabulary item, it was repeatedly studied but dropped from further testing; repeatedly tested but dropped from further study; or dropped from both study and test. Repeated studying after learning had no effect on delayed recall, but repeated testing produced a large positive effect. In addition, students' predictions of their performance were uncorrelated with actual performance. The results demonstrate the critical role of retrieval practice in consolidating learning and show that even university students seem unaware of this fact.

Ever since the pioneering work of Ebbinghaus (*1*), scientists have generally studied human learning and memory by presenting people with information to be learned in a study period and testing them on it in a test period to see what they retained. When this procedure occurs over many trials, an exponential learning curve is produced. The standard assumption in virtually all research is that learning occurs while people study and encode material. Therefore, additional study should increase learning. Retrieving information on a test, however, is sometimes considered a relatively neutral event that measures the learning that occurred during study but does not by itself produce learning. Over the years, researchers have occasionally argued that learning can occur during testing (*2-6*). However, the assumptions that repeated studying promotes learning, and that testing represents a neutral event that merely measures learning, still permeate contemporary memory research as well as contemporary educational practice, where tests are also considered purely as assessments of knowledge.

Our goal in the present research was to examine these long-standing assumptions regarding the effects of repeated studying and repeated testing on learning. Specifically, once information can be recalled from memory, what are the effects of repeated encoding (during study trials) or repeated retrieval (during test trials) on learning and long-term retention, assessed after a week delay? A second purpose of this research was to examine students' assessments of their own learning. After learning a set of materials under repeated study or repeated test conditions, we asked students to predict their future recall on the week-delayed final test. Our question was: Would students show any insight into their own learning?

A final purpose of the experiment was to address another venerable issue in learning and memory, concerning the relation between the speed with which something is learned and the rate at which it is forgotten. Is speed of learning correlated with long-term retention, and if so, is the correlation positive (processes that promote fast learning also slow forgetting and promote good retention) or negative (quick learning may be superficial and produce rapid forgetting)? Early research led to the conclusion that quick learning reduced the rate of forgetting and improved long-term retention (7), but later critics argued that when forgetting is assessed more properly than in the early studies, no differences exist between forgetting rates for fast and slow learning conditions (8-9). By any account, conditions that exhibit equivalent learning curves should produce equivalent retention after a delay (9).

Using the learning of foreign language vocabulary word pairs, we examined the contributions of repeated study and repeated testing to learning by comparing a standard learning condition to three novel dropout conditions. The standard method of measuring learning, used since Ebbinghaus's research (1), involves presenting subjects with information in a study period, then testing them on it in a test period, then presenting it again, testing on it again, and so on. The dropout learning conditions of the present experiment differed from the standard learning condition in that, once an item was successfully recalled once on a test, it was either 1) dropped from study periods but still tested in one condition, or 2) dropped from test periods but still repeatedly studied in a second condition, or 3) dropped altogether from both study and test periods in a third condition (see Table 1).

Surprisingly, standard learning conditions and dropout conditions have seldom been compared in memory research, despite their critical importance to theories of learning and their practical importance to students (in using flashcards and other study methods). Dropout conditions were originally developed to remedy methodological problems that arise from repeated practice in the standard learning condition (10), but they can also be used to examine the effect of repeated practice in its own right, as we did in the present experiment. If learning happens exclusively during study periods and if tests are neutral assessments, then additional study trials should have a strong positive effect on learning, whereas additional test trials should produce no effect. Further, if repeated study or test practice after an item has been learned does indeed benefit long-term retention, this would contradict the conventional wisdom that students should drop material that they have learned from further practice in order to focus their effort on material they have not yet learned. Dropping learned facts may create the same long-term retention as occurs in standard conditions, but in a shorter amount of time, or it may improve learning by allowing students to focus on items they have not yet recalled. This strategy is implicitly endorsed by contemporary theories of study-time allocation (11-12) and is explicitly encouraged in many popular study guides (13).

In the experiment, we had college students learn a list of foreign language vocabulary word pairs and manipulated whether pairs remained in the list (and were repeatedly practiced) or were dropped after the first time they were recalled, as shown in Table 1. All students began by studying a list of 40 Swahili-English word pairs (e.g., *mashua* – *boat*) in a study period and then testing over the entire list in a test period (e.g.,

mashua – ?). All conditions were treated the same in the initial study and test periods. Once a word pair was recalled correctly, it was treated differently in the four conditions. In the standard condition, subjects studied and were tested over the entire list in each study and test period (denoted ST). In a second condition, once a pair was recalled, it was dropped from further study but tested in each subsequent test period (denoted $S_N T$, where S_N indicates that only non-recalled pairs were restudied). In a third condition, recalled pairs were dropped from further testing but studied in each subsequent study period (denoted ST_N , where T_N indicates that only non-recalled pairs were kept in the list during test periods). Finally, in a fourth condition, recalled pairs were dropped entirely from both study and test periods ($S_N T_N$). The final condition represents what conventional wisdom and many educators instruct students to do: Study something until it is learned (i.e., can be recalled) and then drop it from further practice.

At the end of the learning phase, students in all four conditions were asked to predict how many of the 40 pairs they would recall on a final test in 1 week. They were then dismissed and returned for the final test a week later. Of key importance were the effects of the four learning conditions on the speed with which the vocabulary words were learned, on students' predictions of their future performance, and on long-term retention assessed after a week delay (14).

Figure 1 shows the cumulative proportion of word pairs recalled during the learning phase, which gives credit the first time a student recalled a pair. We also analyzed traditional learning curves (the proportion of the total list recalled in each test period) for the two conditions that required recall of the entire list (ST and $S_N T$) and the

results by the two measurement methods were virtually identical. Thus we restrict our discussion to the cumulative learning curves on which all four conditions can be compared. Figure 1 shows that performance was virtually perfect by the end of learning (i.e., all 40 English target words were recalled by nearly all subjects). More importantly, there were no differences in the learning curves of the 4 conditions.

Given the similarity of acquisition performance, it is not too surprising that students in the 4 conditions did not differ in their aggregate judgments of learning (their predictions of their future performance). On average, the subjects in all conditions predicted they would recall approximately 50% of the pairs in 1 week. The mean number of words predicted to be recalled in each condition were as follows: ST = 20.8; S_NT = 20.4; ST_N = 22.0; S_NT_N = 20.3. An analysis of variance did not reveal significant differences among the conditions ($F < 1$).

Although subjects' cumulative learning performance was equivalent in the 4 conditions, and predicted final recall was also equivalent, actual recall on the final delayed test differed widely across conditions, as shown in Figure 2. The results clearly show that testing (and not studying) is the critical factor for promoting long-term recall. In fact, repeated study after one successful recall did not produce any measurable learning a week later. In the learning conditions that required repeated retrieval practice (ST and S_NT), subjects correctly recalled approximately 80% of the pairs on the final test. In the other conditions in which items were dropped from repeated testing (ST_N and S_NT_N) subjects recalled just 36% and 33% of the pairs. It is worth emphasizing that despite the fact that students repeatedly studied all of the word pairs in every study period

in the ST_N condition, their long-term recall was much worse than students who were repeatedly tested on the entire list. Combining the two conditions that involved repeated testing (ST and $S_N T$) and combining the two conditions that involved dropping items from testing after they were recalled once (ST_N and $S_N T_N$), repeated retrieval increased final recall by 4 standard deviations ($d = 4.03$). The distributions of scores in these two groups were non-overlapping: Final recall in the drop-from-testing conditions ranged from 10% to 60% across subjects, whereas final recall in the repeated test conditions ranged from 63% to 95%. Whether students repeatedly studied the entire set or whether they restudied only pairs they had not yet recalled produced virtually no effect on long-term retention. The dramatic difference shown in Figure 2 was caused by whether or not the pairs were repeatedly tested.

Even though cumulative learning performance was identical in the 4 conditions, the total number of trials (study or test) in each condition varied greatly. Table 1 shows the mean number of trials in each study and test period and the total number of trials in each condition. The standard condition (ST) involved the most trials (320) because all 40 items were presented in each study and test period. The $S_N T_N$ condition involved the fewest trials (154.8, on average) because the number of trials in each period grew smaller as items were recalled and dropped from further practice. The other two conditions ($S_N T$ and ST_N) involved approximately the same number of trials (236.8 and 243.0, respectively) but because they differed in terms of whether items were dropped from study or test periods, they produced dramatically different effects on long-term retention. In other words, about 80 more study trials occurred in the ST_N condition than in the $S_N T_N$

condition, but this produced practically no gain in retention. Likewise, about 80 more study trials occurred in the ST condition than in the S_NT condition and this produced no gain whatsoever in retention. However, when approximately 80 more test trials occurred in the learning phase (in the ST condition vs. the ST_N condition, and in the S_NT condition vs. the S_NT_N condition), repeated retrieval practice led to greater than 150% improvements in long-term retention.

The present research shows the powerful effect of testing on learning: Repeated retrieval practice enhanced long-term retention, whereas repeated studying produced virtually no benefit. Although educators and psychologists often consider testing a neutral process that merely assesses the contents of memory, practicing retrieval during tests clearly produces more learning than additional encoding or study once an item has been recalled (*15-17*). Dropout methods such as the ones used in the present experiment have seldom been used to investigate effects of repeated practice in their own right, but in the present research, comparison of the dropout conditions to the repeated practice conditions revealed dramatic effects of retrieval practice on learning.

The experiment also shows a striking absence of any benefit of repeated studying once an item could be recalled from memory. A basic tenet of human learning and memory research is that repetition of material improves its retention. This is often true in standard learning situations, yet our research demonstrates a situation that stands in stark contrast to this principle. The benefits of repetition for learning and long-term retention clearly depend on the processes learners engage in during repetition. Once information can be recalled, repeated encoding in study trials produced no benefit, whereas repeated

retrieval in test trials generated large benefits for long-term retention. Further research is necessary to generalize these findings to other materials. However, the basic effects of testing on retention have been shown with many kinds of materials (16), so we have confidence that the present results will generalize, too.

Our experiment also speaks to an old debate in the science of memory, concerning the relation between speed of learning and rate of forgetting (7-9). Our study shows that the forgetting rate for information is not necessarily determined by speed of learning but, instead, is greatly determined by the type of practice involved. Even though the four conditions in the experiment produced equivalent learning curves, repeated recall slowed forgetting relative to recalling each word pair just one time.

Importantly, students exhibited no awareness of the mnemonic effects of retrieval practice, as evidenced by the fact that they did not predict they would recall more if they had repeatedly recalled the list of vocabulary words than if they only recalled each word one time. Indeed, questionnaires asking students to report on the strategies they use to study for exams in education also indicate that practicing recall (or self-testing) is a seldom-used strategy (18). If students do test themselves while studying, they likely do it to assess what they have or have not learned (19), rather than to enhance their long-term retention by practicing retrieval. In fact, the conventional wisdom shared among students and educators is that if information can be recalled from memory, it has been “learned” and can be dropped from further practice, so students can focus their effort on other material. Research on students’ use of self-testing as a learning strategy shows that students do tend to drop facts from further practice once they can recall them (20).

However, the present research shows that the conventional wisdom existing in education and expressed in many study guides is wrong. Even after items can be recalled from memory, eliminating those items from repeated retrieval practice greatly reduces long-term retention. Repeated retrieval induced through testing (and not repeated encoding during additional study) produces large positive effects on long-term retention.

References and Notes

1. H. Ebbinghaus, *Memory: A Contribution to Experimental Psychology*, H. A. Ruger, C. E. Bussenius, Trans. (Dover, New York, 1964, original work published 1885).
2. R. A. Bjork, in *Information Processing and Cognition: The Loyola Symposium*, R. L. Solso, Ed. (Erlbaum, Hillsdale NJ, 1975), pp. 123-144.
3. M. Carrier, H. Pashler, *Mem. Cognit.*, **20**, 633 (1992).
4. A. I. Gates, *Arch. Psychol.*, **6** (1917).
5. C. Izawa, *J. Math. Psychol.*, **8**, 200, (1971).
6. E. Tulving, *J. Verb. Learn. Verb. Behav.*, **6**, 175 (1967).
7. J. A. McGeoch, *The Psychology of Human Learning* (Longmans, Green and Co., New York, 1942).
8. N. J. Slamecka, B. McElree, *J. Exp. Psychol. Learn. Mem. Cognit.*, **9**, 384 (1983)
9. B. J. Underwood, *J. Verb. Learn. Verb. Behav.*, **3**, 112 (1964).
10. W. F. Battig, *Psychon. Sci. Monogr. Supp.*, **1**, 1 (1965).
11. J. Metcalfe, N. Kornell, *J. Exp. Psychol. Gen.*, **132**, 530 (2003)
12. K. W. Thiede, J. Dunlosky, *J. Exp. Psychol. Learn. Mem. Cognit.*, **25**, 1024 (1999)
13. S. Frank, *The Everything Study Book* (Adams Media Company, Avon, MA, 1996).
14. Materials and methods are available as supporting material on *Science* online.
15. J. D. Karpicke, H. L. Roediger, *J. Mem. Lang.*, **57**, 151 (2007).

16. H. L. Roediger, J. D. Karpicke, *Perspectives Psychol. Sci.*, **1**, 181 (2006).
17. H. L. Roediger, J. D. Karpicke, *Psychol. Sci.*, **17**, 249 (2006).
18. N. Kornell, R. A. Bjork, *Psychon. Bull. Rev.*, **14**, 219 (2007).
19. J. Dunlosky, K. Rawson, S. McDonald, in *Applied Metacognition*, T. Perfect, B. Schwartz, Eds. (Cambridge, Cambridge University Press, 2002) pp. 68-92.
20. J. D. Karpicke, thesis, Washington University in St. Louis (2007)
21. We thank J. S. Nairne for helpful comments on the manuscript. This research was supported by a Collaborative Activity Grant of the James S. McDonnell Foundation to the second author.

Supporting Online Material

www.sciencemag.org

Materials and Methods

Table S1

References

Table 1

Conditions used in the experiment, average number of trials within each study or test period, and total number of trials in the learning phase in each condition. S_N indicates that only pairs not recalled in the previous test period were studied in the current study period. T_N indicates that only pairs not recalled in the previous test period were tested in the current test period. Subjects in all conditions performed a 30 sec distracter task that involved verifying multiplication problems after each study period.

Condition	Study or Test Period and Number of Trials Per Period								Total Number of Trials
	1	2	3	4	5	6	7	8	
ST	S	T	S	T	S	T	S	T	
	40	40	40	40	40	40	40	40	320
$S_N T$	S	T	S_N	T	S_N	T	S_N	T	
	40	40	26.8	40	8.0	40	2.0	40	236.8
ST_N	S	T	S	T_N	S	T_N	S	T_N	
	40	40	40	27.9	40	11.8	40	3.3	243.0
$S_N T_N$	S	T	S_N	T_N	S_N	T_N	S_N	T_N	
	40	40	27.1	27.1	8.8	8.8	1.5	1.5	154.8

Figure Captions

Fig. 1. Cumulative performance during the learning phase.

Fig. 2. Proportion recalled on the final test 1 week after learning. Error bars represent standard errors of the mean.

Fig. 1

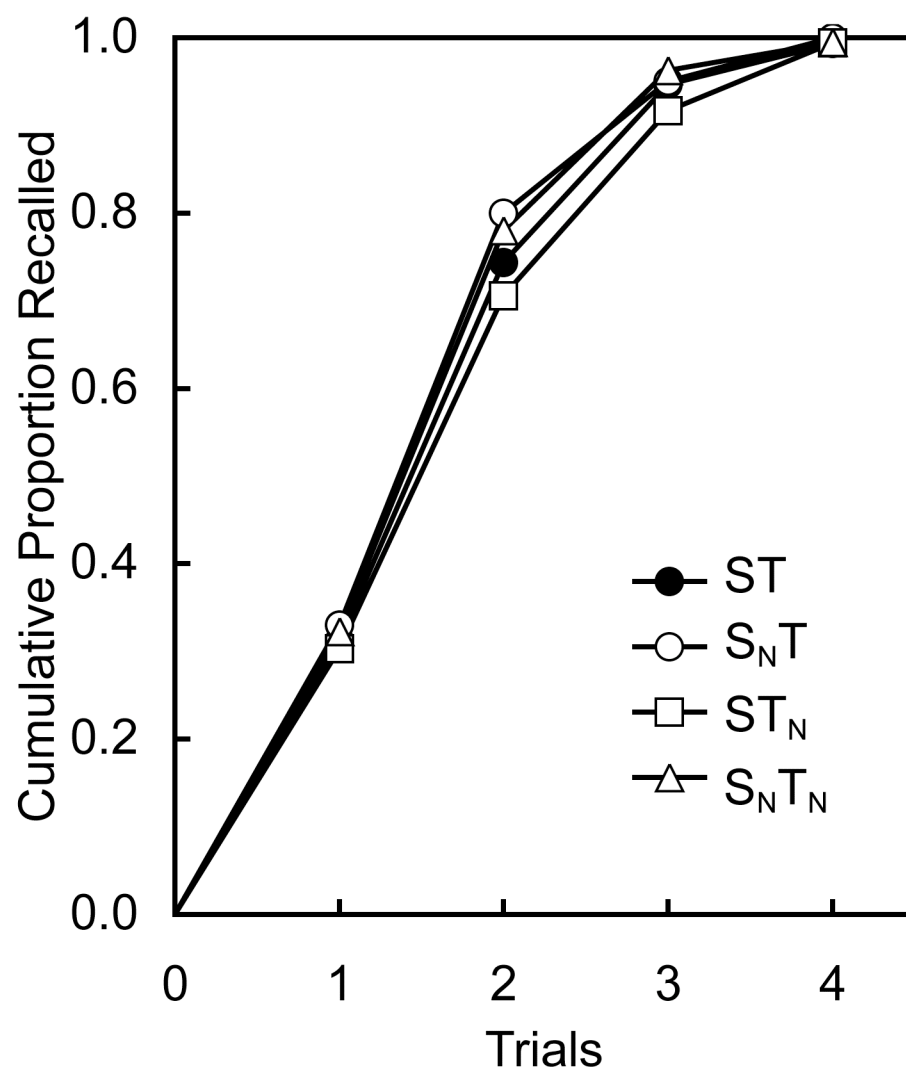
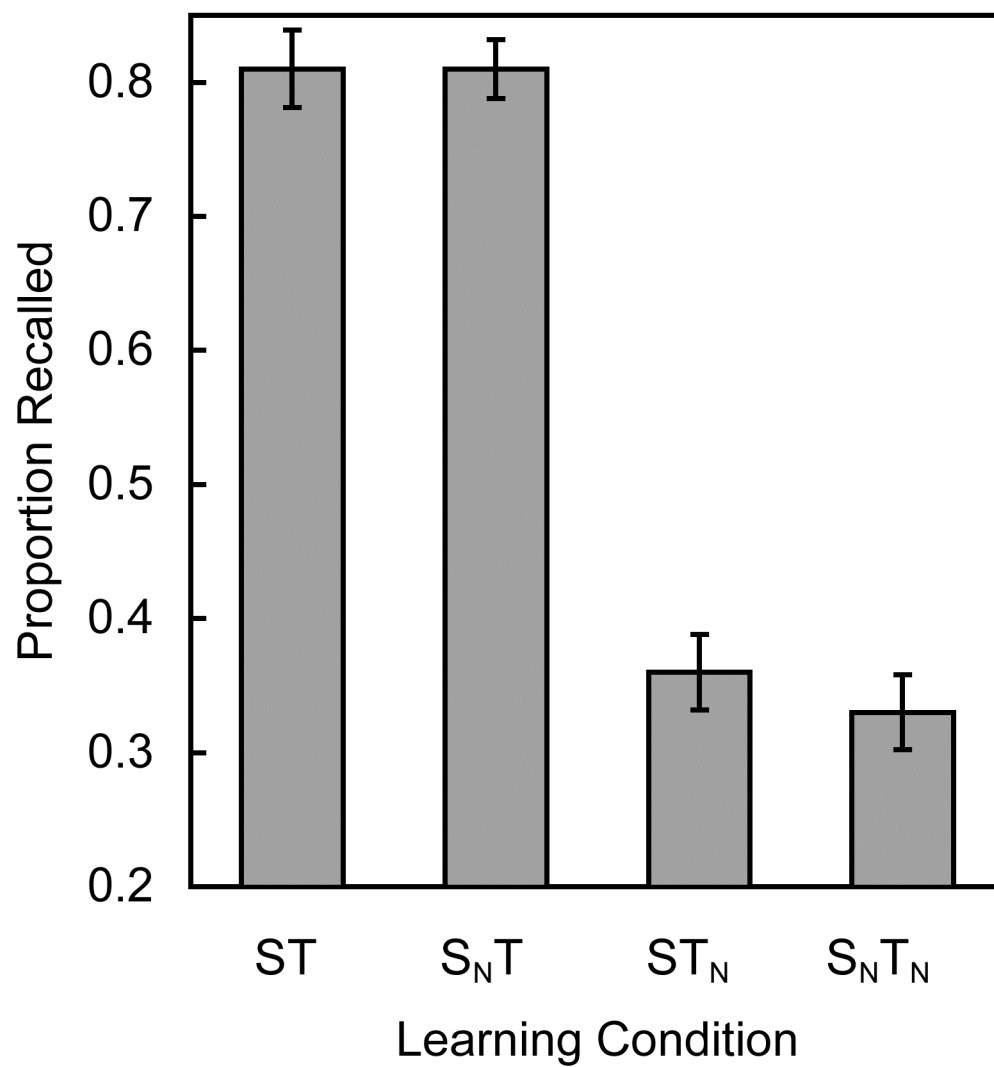


Fig. 2



Supporting Online Material

Materials and Methods

In the experiment, 40 Washington University undergraduates learned a list of 40 Swahili-English word pairs (e.g., *mashua* – *boat*) selected from previously published norms (*SI*). The list of word pairs is shown in Table S1. Students learned the list across a total of 8 alternating study (S) and test (T) periods. The first study period consisted of 40 study trials followed by 40 test trials. After that, the number of study and test trials varied according to the condition. During study trials, students saw each Swahili word and its English translation on a computer screen for 5 sec and were told to study the pair so they could recall the English word given the Swahili word. After every study period, subjects performed a 30 sec distracter task that involved verifying multiplication problems. Each test period consisted of 40 or fewer test trials (depending on the condition). During test trials, students saw each Swahili word and a cursor and their task was to type the correct English translation. Each test trial lasted 8 sec after which the computer program automatically advanced to the next item regardless of whether the student had entered a response. If subjects failed to recall an item during testing, they were not given feedback.

The students learned the vocabulary words in one of four conditions. In the standard condition, students studied and were tested over the entire list of 40 pairs in each study and test period (denoted ST). In a second condition, students studied the entire list in the first period, were tested over the entire list in the second period, but then they restudied only the pairs they had not recalled on the previous test (denoted $S_N T$, where S_N indicates that only non-recalled pairs were restudied). The entire list was tested in each

test period in the $S_N T$ condition. In a third condition, students studied the entire list in each study period, but only items that they had not yet recalled were tested in test period (denoted ST_N , where T_N indicates that only non-recalled pairs were repeatedly tested). All 40 pairs were studied each time in this condition. Finally, in a fourth condition, students restudied only non-recalled pairs and were tested only over non-recalled pairs ($S_N T_N$). Therefore, both the number of pairs studied and tested diminished across periods in this condition. This $S_N T_N$ condition is the adjusted learning or dropout technique used in prior research (*S2-S4*). The condition also represents what students are often told to do: Study something until it is learned (or recallable) and then drop it from further practice. In the ST_N and $S_N T_N$ conditions, students recalled each pair one time in the learning phase, whereas in the $S_N T$ and ST conditions, students repeatedly recalled the pairs in every test period.

At the end of the learning phase, we asked students to predict how many of the 40 pairs they would recall on a final test in 1 week (an aggregate judgment of learning). Students were then dismissed and returned for the final test a week later. On the final test, subjects were shown each Swahili word for 15 sec and were told to type the correct English translation. After completing the test, subjects were debriefed and thanked for their participation.

Table S1

Forty Swahili-English word pairs used in the experiment

Number	Swahili	English
1	adhama	honor
2	adui	enemy
3	bustani	garden
4	buu	maggot
5	chakula	food
6	dafina	treasure
7	elimu	science
8	embe	mango
9	fagio	broom
10	farasi	horse
11	fununu	rumour
12	godoro	mattress
13	goti	knee
14	hariri	silk
15	kaa	crab
16	kaburi	grave
17	kaputula	shorts
18	leso	scarf
19	maiti	corpse
20	malkia	queen
21	mashua	boat
22	ndoo	bucket
23	nyanya	tomato
24	pazia	curtain
25	pipa	barrel
26	pombe	beer
27	punda	donkey
28	rembo	ornament
29	roho	soul
30	sala	prayer
31	sumu	poison
32	tabibu	doctor
33	theluji	snow
34	tumbili	monkey
35	usingizi	sleep
36	vuke	steam
37	yai	egg
38	zeituni	olives
39	ziwa	lake
40	zulia	carpet

References

- S1. T. O. Nelson, J. Dunlosky, *Memory*, **2**, 325 (1994).
- S2. R. S. Woodworth, *Psychol. Bull.*, **11**, 58 (1914).
- S3. W. F. Battig, *Psychon. Sci. Monogr. Supp.*, **1**, 1 (1965).
- S4. A. L. Gillette, *Arch. Psychol.*, **28** (1936).