Why wait if you can switch? A short term testing effect in cross-language recognition.

Document Version:
Peer reviewed version

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

Open Universiteit
www.ou.nl

Psychological Science

A Short-Term Testing Effect in Cross-Language Recognition

>> Version of Record - Jun 11, 2012

OnlineFirst Version of Record - Apr 26, 2012

What is This?

A Short-Term Testing Effect in Cross-Language Recognition

Peter P. J. L. Verkoeijen, Samantha Bouwmeester, and Gino Camp
Erasmus University Rotterdam

Abstract

Taking a memory test after an initial study phase produces better long-term retention than restudying the items, a phenomenon known as the testing effect. We propose that this effect emerges because testing strengthens semantic features of items' memory traces, whereas restudying strengthens surface features of items' memory traces. This novel account predicts that a testing effect should be observed even after a short retention interval when a language switch occurs between the learning phase and the final test phase. We assessed this prediction with Dutch-English bilinguals who learned Dutch Deese-Roediger-McDermott word lists through restudying or through testing (retrieval practice). Five minutes after this learning phase, they took a recognition test in Dutch (within-language condition) or in English (across-language condition). We observed a testing effect in the across-language condition, but not in the within-language condition. These findings corroborate our novel account of the testing effect.

In recent years, there has been an explosion of research on the beneficial effect of retrieval practice on memory. A robust finding that has emerged from a large number of studies is that taking one or more intervening tests after an initial encoding (study) episode produces better retention of the to-be-remembered material than does restudying the same material for an equivalent amount of time. This phenomenon is known as the testing effect (for recent reviews, see Delaney, Verkoeijen, & Spirgel, 2010; Roediger & Butler, 2011; and Roediger & Karpicke, 2006a).

A typical experiment on this effect compares retention of restudied information and tested information at multiple retention intervals (RIs).[1] For short RIs of several minutes, these experiments usually show either that restudying has a memory advantage over testing or that restudied and previously tested items are remembered equally well. However, after an RI of multiple days, previously tested items are remembered better than restudied items. Furthermore, a significant interaction between RI and learning procedure (restudy vs. testing) emerges because forgetting is slower for tested than for restudied items. This interaction has been replicated many times with different stimulus materials and with a variety of test types (e.g., Carpenter & Pashler, 2007; Coppens, Verkoeijen, & Rikers, 2011; Hogan & Kintsch, 1971; Roediger & Karpicke, 2006b;

Thompson, Wenger, & Bartling, 1978; Toppino & Cohen, 2009; Wheeler, Ewers, & Buonanno, 2003).

Although the testing effect has been replicated repeatedly, the majority of the studies in the literature have had an empirical rather than a theoretical focus. Only in the past few years have researchers started to systematically investigate the mechanism (or mechanisms) underlying the testing effect (e.g., Carpenter, 2009; Carpenter, 2011; Pyc & Rawson, 2010). However, despite these efforts, researchers still know little about why testing (retrieval practice) enhances retention. To contribute to the development of a theoretical explanation of the testing effect, we (Bouwmeester & Verkoeijen, 2011) proposed that testing and restudying are operations that tap into different types of memory traces. Our account is rooted in fuzzy-trace theory (see Brainerd & Reyna, 2004; Brainerd, Reyna, & Ceci, 2008; Reyna & Brainerd, 1995), but it also fits well within other memory models. The goal of the present study was to assess an important prediction of our novel account of the testing effect.

Corresponding Author:
Peter P. J. L. Verkoeijen, Department of Psychology, Erasmus University Rotterdam, P. O. Box 1738, NL-3000 DR Rotterdam, The Netherlands
E-mail: verkoeijen@fsw.eur.nl

In memory models such as Search of Associative Memory (SAM; Raaijmakers & Shiffrin, 1981), Retrieving Effectively From Memory (REM; Shiffrin, 2003), Processing Implicit and Explicit Representations (PIER; Nelson, Schreiber, & McEvoy, 1992), and fuzzy-trace theory, episodic memory traces are represented as collections of both surface and semantic features. Consider a participant who is presented with the word *tree* during a verbal learning task. The surface features of this word are its visual (orthographic) and phonological characteristics. By contrast, its semantic features are representations of its meaning. These semantic features include the concept "tree," to which the word *tree* refers, but also the concepts that are semantically related to the concept "tree" (see Griffiths, Steyvers, & Tenenbaum, 2007, for an overview of different types of semantic representations). We propose that retrieval practice strengthens semantic features of memory traces because participants use semantic cues, such as word meanings or semantic associates, to recover information from memory. By contrast, participants receive more exposure to the surface features of stimuli during restudy, and as a result, restudy strengthens surface features of memory traces more than retrieval practice does.

Our account predicts that testing will produce better memory performance than restudying whenever participants have to rely primarily on semantic cues during a final test. This occurs when the final test is administered after a multiday RI. Indeed, a typical finding from the testing literature is that the testing effect emerges only after a long delay between the learning phase and the final test. However, our account predicts that under certain conditions, a testing effect will also be found after a short RI of several minutes. Specifically, a short-term testing effect should emerge if the final test prevents participants from using surface retrieval cues, and instead requires them to rely on semantic retrieval cues. We conducted the present study to test this prediction.

In our study, Dutch-English bilingual adults learned a number of Dutch Deese-Roediger-McDermott (DRM; Deese, 1959; Roediger & McDermott, 1995) lists through restudying or through retrieval practice. A DRM list contains target words that are all semantically related to a particular nonpresented distractor word (lure). Subsequently, after a 5-min RI, half of the participants received a final yes/no recognition test in English (across-language condition), whereas the other half of the participants received the test in Dutch (within-language condition). In the across-language condition, the final test provided only partial information about the targets, because participants were cued with word meanings (note that we assume Dutch words and their English translations have a common semantic representation; e.g., Sahlin, Harding, & Seamon, 2005; Zeelenberg & Pecher, 2003) but not with the surface features of the studied words (cf. Howe, Gagnon, & Thouas, 2008). By contrast, in the within-language condition, the final test cued participants with both surface information (the visual presentation of the target words) and semantic information.

We expected to find a testing effect in the across-language condition, because participants in this condition had to rely on semantic cues, and according to our account, retrieval practice strengthens semantic features of memory traces more than restudying does. In the within-language condition, we expected to obtain one of the results commonly obtained in studies on the testing effect, namely, either an advantage for restudying or no difference between restudying and testing. Furthermore, if semantic cues drive the recognition of items learned through testing, we would find the same final-test recognition performance for these items in the across-language condition and in the within-language condition. And if recognition of restudied items depends more strongly on surface cues, performance on restudied items would be worse in the across-language condition than in the within-language condition.

Finally, according to our account of the testing effect, retrieval practice should also strengthen concepts semantically related to the studied items more than restudying does. Hence, compared with restudy, retrieval practice should yield a higher false alarm rate to related distractors on the final recognition test. However, we expected that it might be difficult to detect this predicted difference because with adult participants, DRM lists tend to produce very high levels of false alarms to related distractors.

Method

Participants and design

Participants were 64 individually tested Dutch psychology undergraduates at the Erasmus University Rotterdam. They took part in this experiment to fulfill a course requirement. These participants can be considered bilinguals with respect to the words presented at the final test (see Zeelenberg & Pecher, 2003, for relevant arguments). We used a 2 (learning procedure: restudy vs. testing) × 2 (final test: across language vs. within language) mixed design with repeated measures on the first factor. Participants were randomly assigned to the final-test condition.

Materials

For the learning phase of this study, we selected 12 DRM study lists from Zeelenberg and Pecher's (2002) Dutch translations of Stadler, Roediger, and McDermott's (1999) word lists. Each list comprised eight words, each of which had a strong backward association with the same semantically related distractor. The 12 lists were randomly split in two sets of 6 lists. We counterbalanced order of learning procedure (i.e., restudy first vs. testing first) and presentation order of the two lists (Set 1 first vs. Set 2 first), so that there were four study sequences for both the across-language condition and the within-language condition.

For the final recognition test, we created a list of 72 words: 36 target words (3 words from each list), the 12 semantically

related distractors, and 24 unrelated distractors. In the across-language condition, the final test was administered in English, whereas in the within-language condition, the final test was administered in Dutch. We made sure that the surface features of the English test items and their Dutch counterparts did not overlap.

Procedure

At the start of the experiment, participants were informed that they would be presented with word lists of eight words each, and that they should try to remember as many words as possible for an upcoming, unspecified memory test. The learning phase for each list began with an initial study phase in which the words were presented in the center of a computer screen at a rate of 4 s per word, with a 1-s interstimulus interval. Participants were instructed to type each word and to memorize it. After this initial study phase, either the list was presented again using the same procedure (restudy) or a free-recall test was administered (testing). During the free-recall test, participants were asked to type as many words as they could remember from the previously studied list. Participants were given 4 s to type each word, and there was a 1-s interval between each of the eight response opportunities. Participants first completed either a block of six restudy lists or a block of six tested lists. After a short break, they completed a block in which they learned the other six lists using the other learning procedure.

Following the second block, participants engaged in a 2-min distractor task in which they counted backward by 3s from a given number. After the distractor task, the final yes/no recognition task was administered either in Dutch or in English. Words were presented one by one on the computer screen, in random order, and participants had to indicate for each word whether it had been presented during the learning phase (i.e., an "old" judgment) or was new. The final test was self-paced, and a new test item appeared after a participant made a response.

Results and Discussion

A significance level (p) of .05 was used as the threshold for statistical significance. The reported t-tests were two-tailed.

Test performance during the learning phase

The mean proportion of correctly retrieved words from the tested lists during the learning phase did not differ between the across-language condition ($M = .80$, $SD = .13$) and the within-language condition ($M = .83$, $SD = .10$), $t(62) = 0.83$, $p = .41$, $d = 0.18$.

Final-test performance

For each participant, we calculated the proportion of unrelated distractors correctly classified as "new," the proportion of

Table 1. Mean Proportion of Correct Recognition

	Final test	
Learning procedure	Across language	Within language
Restudy	.67 (.04)	.81 (.03)
Testing	.78 (.03)	.78 (.03)

Note: Standard errors are given in parentheses.

targets correctly classified as "old," and the proportion of related distractors incorrectly classified as "old" (i.e., the proportion of false memories). An independent t test revealed that the mean proportion of correctly classified unrelated distractors was comparable for the across-language condition ($M = .91$, $SD = .10$) and the within-language condition ($M = .94$, $SD = .11$), $t(62) = 1.22$, $p = .22$, $d = 0.32$. This finding suggests that participants in the two conditions employed a similar response criterion.

A 2 (learning procedure: restudy vs. testing) × 2 (final test: across language vs. within language) mixed-model analysis of variance (ANOVA) on the proportion of correctly classified targets (see Table 1) did not reveal significant main effects of learning procedure, $F(1, 62) = 2.73$, $MSE = 0.02$, $p = .10$, $\eta_p^2 = .04$, or final test, $F(1, 62) = 3.15$, $MSE = 0.04$, $p = .08$, $\eta_p^2 = .05$. However, and more important, the Learning Procedure × Final Test interaction was significant, $F(1, 62) = 8.18$, $MSE = 0.02$, $p < .05$, $\eta_p^2 = .12$. To examine this interaction effect, we conducted four follow-up t tests using a Bonferroni-corrected alpha level of .013. As predicted, we found that testing produced better recognition than restudying did in the across-language condition, $t(30) = 3.20$, $p < .013$, $d = 0.55$, but not in the within-language condition, $t(32) = 0.85$, $p = .40$, $d = 0.15$. Also as predicted, the language switch between the learning phase and the final-test phase decreased recognition performance for restudied items (i.e., across-language vs. within-language comparison), $t(62) = 2.80$, $p < .013$, $d = 0.68$, but not for items learned through retrieval practice, $t(62) = 0.19$, $p = .84$, $d = 0.05$.

A 2 (learning procedure) × 2 (final test) mixed-model ANOVA on the proportion of incorrectly classified related distractors (i.e., the proportion of false memories; see Table 2) did not reveal significant main effects or a significant interaction (all $Fs < 1$, all $\eta_p^2s < .01$).

Table 2. Mean Proportion of False Memories

	Final test	
Learning procedure	Across language	Within language
Restudy	.64 (.05)	.61 (.05)
Testing	.66 (.04)	.60 (.04)

Note: Standard errors are given in parentheses.

Discussion

The most important finding in this study was that a short-term testing effect emerged in the across-language condition, in which participants were prevented from using surface cues and instead had to rely exclusively on semantic cues provided by the English test words. By contrast, in the within-language condition, in which participants could base their recognition decisions on both surface and semantic cues, recognition was comparable for previously tested items and restudied items. The latter finding replicates the results typically reported for short RIs. Moreover, recognition memory for restudied items declined when the final-test language was different from the learning-phase language. However, this language switch did not influence recognition performance for items learned through testing. Taken together, these findings suggest that there are qualitative differences in how restudying and testing strengthen memory. Specifically, restudying strengthens surface features of items' memory traces more than testing does, whereas testing strengthens semantic features of items' memory traces more than restudying does.

Our findings for the target words seem to be inconsistent with the recently proposed distribution-based framework (e.g., Halamish & Bjork, 2011). According to this framework, a short-term testing effect will emerge when the final test is sufficiently difficult, that is, when the minimum memory strength required for retrieval of the items is sufficiently high. In the present study, final-test performance for previously tested items was similar in the across-language condition and the within-language condition. This indicates that the final tests in the two conditions were of comparable difficulty. Nevertheless, the language-switch manipulation gave rise to a testing effect. These results suggest that the type of final-test cue, rather than final-test difficulty, is a crucial factor in the emergence of the testing effect.

We also predicted that, compared with restudying, testing would increase the proportion of falsely recognized related distractors (i.e., the proportion of false memories). Contrary to this prediction, the proportion of false memories was similar for restudied and previously tested DRM lists. Perhaps this unexpected outcome arose because we employed relatively short, eight-word lists. Hence, during the immediate free-recall test, participants might have relied primarily on word meanings for item retrieval and much less on semantic associates, such as related distractors. Consequently, relative to restudying, retrieval practice might have resulted in only a small increase in the related distractors' memory strength (see Sugrue, Strange, & Hayne, 2009, for a study showing that false recall increases with DRM list length). This, in turn, might have made it hard for us to detect differences in false memory between restudied and tested lists on the final recognition test. However, this account of our false-memory findings is post hoc, and further research is required to determine the conditions under which testing strengthens memory traces of related distractors more than restudying does.

To conclude, we note that our findings might inform classroom practice because they suggest that testing enhances meaningful, semantic processing of to-be-learned materials.

Acknowledgments

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Note

1. It should be noted that *RI* refers to the interval between the learning phase (i.e., initial study plus either restudy or testing) and the final test.

References

Bouwmeester, S., & Verkoeijen, P. P. J. L. (2011). Why do some children benefit more from testing than others? Gist trace processing to explain the testing effect. *Journal of Memory and Language*, *65*, 32–41.

Brainerd, C. J., & Reyna, V. F. (2004). Fuzzy-trace theory and memory development. *Developmental Review*, *24*, 396–439.

Brainerd, C. J., Reyna, V. F., & Ceci, S. J. (2008). Developmental reversals in false memory: A review of data and theory. *Psychological Bulletin*, *134*, 343–382.

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 1563–1569.

Carpenter, S. K. (2011). Semantic information activated during retrieval contributes to later retention: Support for the mediator effectiveness hypothesis of the testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 1547–1552.

Carpenter, S. K., & Pashler, H. (2007). Testing beyond words: Using tests to enhance visuospatial map learning. *Psychonomic Bulletin & Review*, *14*, 474–478.

Coppens, L. C., Verkoeijen, P. P. J. L., & Rikers, R. M. J. P. (2011). Learning Adinkra symbols: The effect of testing. *Journal of Cognitive Psychology*, *23*, 351–357.

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology*, *58*, 17–22.

Delaney, P. F., Verkoeijen, P. P. J. L., & Spirgel, A. S. (2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. In B. H. Ross (Ed.), *Psychology of learning and motivation* (Vol. 53, pp. 63–147). Burlington, MA: Academic Press.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*, 211–244.

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*, 801–812.

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

Howe, M. L., Gagnon, N., & Thouas, L. (2008). Development of false memories in bilingual children and adults. *Journal of Memory and Language*, *58*, 669–681.

Nelson, D. L., Schreiber, T. A., & McEvoy, C. L. (1992). Processing implicit and explicit associations. *Psychological Review*, *99*, 322–348.

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

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

Reyna, V. F., & Brainerd, C. J. (1995). Fuzzy-trace theory: An interim synthesis. *Learning and Individual Differences*, *7*, 1–75.

Roediger, H. L., III, & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*, 20–27.

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

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

Roediger, H. L., III, & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 803–814.

Sahlin, B. H., Harding, M. G., & Seamon, J. G. (2005). When do false memories cross language boundaries in English-Spanish bilinguals? *Memory & Cognition*, *33*, 1414–1421.

Shiffrin, R. M. (2003). Modeling memory and perception. *Cognitive Science*, *27*, 341–378.

Stadler, M. A., Roediger, H. L., & McDermott, K. B. (1999). Norms for words that create false memories. *Memory & Cognition*, *27*, 494–500.

Sugrue, K., Strange, D., & Hayne, H. (2009). False memories in the DRM paradigm: Age-related differences in lure activation and source monitoring. *Experimental Psychology*, *56*, 354–360.

Thompson, C. P., Wenger, S. K., & Bartling, C. A. (1978). How recall facilitates subsequent recall: A reappraisal. *Journal of Experimental Psychology: Human Learning and Memory*, *4*, 210–221.

Toppino, T. C., & Cohen, M. S. (2009). The testing effect and the retention interval. *Experimental Psychology*, *56*, 252–257.

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

Zeelenberg, R., & Pecher, D. (2002). False memories and lexical decision: Even twelve primes do not cause long-term semantic priming. *Acta Psychologica*, *109*, 269–284.

Zeelenberg, R., & Pecher, D. (2003). Evidence for cross-language repetition priming in conceptual implicit memory tasks. *Journal of Memory and Language*, *49*, 80–94.