Journal of Cognitive Psychology

ISSN: 2044-5911 (Print) 2044-592X (Online) Journal homepage: https://www.tandfonline.com/loi/pecp21

Does retrieval practice depend on semantic cues? Assessing the fuzzy trace account of the testing effect

Gerdien G. van Eersel, Peter P. J. L. Verkoeijen, Samantha Bouwmeester, Huib K. Tabbers & Remy M. J. P. Rikers

To cite this article: Gerdien G. van Eersel, Peter P. J. L. Verkoeijen, Samantha Bouwmeester, Huib K. Tabbers & Remy M. J. P. Rikers (2017) Does retrieval practice depend on semantic cues? Assessing the fuzzy trace account of the testing effect, Journal of Cognitive Psychology, 29:5, 583-598, DOI: 10.1080/20445911.2017.1300156

To link to this article: https://doi.org/10.1080/20445911.2017.1300156

© 2017 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

Published online: 09 Mar 2017.

Submit your article to this journal

Article views: 833

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=pecp21

JOURNAL OF COGNITIVE PSYCHOLOGY, 2017
VOL. 29, NO. 5, 583–598
http://dx.doi.org/10.1080/20445911.2017.1300156

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

Does retrieval practice depend on semantic cues? Assessing the fuzzy trace account of the testing effect

Gerdien G. van Eersel[a], Peter P. J. L. Verkoeijen[a,b], Samantha Bouwmeester[a], Huib K. Tabbers[a] and Remy M. J. P. Rikers[a,c]

[a]Department of Psychology, Education & Child Studies, Faculty of Social Sciences, Erasmus University Rotterdam, Rotterdam, Netherlands; [b]Learning and Innovation Center, Avans University of Applied Sciences, Breda, Netherlands; [c]Roosevelt Center for Excellence in Education, University College Roosevelt, Utrecht University, Middelburg, Netherlands

ABSTRACT

Retrieval practice enhances long-term retention more than restudying; a phenomenon called the testing effect. The fuzzy trace explanation predicts that a testing effect will already emerge after a short interval when participants are solely provided with semantic cues in the final test. In the present study, we assessed this explanation by gradually reducing the surface features overlap between cues in the learning phase and the final recognition test. In all five experiments, participants in the control/word condition received as final test cues the same words as in the learning phase. The experimental final test cues consisted of scrambled words, words in a new context, scrambled words in a new context (Experiment 1), synonyms (Experiment 2), or images (Experiments 3, 4a, 4b). A short-term testing effect was only observed for the image final test cues. These results do not provide strong support for the fuzzy trace explanation of the testing effect.

ARTICLE HISTORY
Received 25 August 2016
Accepted 17 February 2017

KEYWORDS
Testing effect; semantic processing; recognition memory; retrieval practice; fuzzy trace theory

The testing effect occurs when retrieving information from memory after an initial study phase enhances long-term retention more than restudying does (for reviews, see Delaney, Verkoeijen, & Spirgel, 2010; Karpicke, Lehman, & Aue, 2014; Roediger & Butler, 2011; Rowland, 2014). In a typical testing effect experiment, participants learn a set of words during an initial study phase either by restudying or by testing (i.e. retrieval practice). After a certain retention interval, they receive a final test. When no feedback on their memory performance is provided during the intervening test, performance for restudied stimuli is generally better than, or comparable to, performance for tested stimuli after a short interval of five minutes (exceptions can be found in, e.g. Carpenter, 2009; Halamish & Bjork, 2011). However, after a long interval (generally one week), retrieval practice is more effective than restudy, giving rise to an interaction effect of study method and retention interval on memory performance (e.g. Hogan & Kintsch, 1971; Roediger & Karpicke, 2006; Wheeler, Ewers, & Buonanno, 2003). This testing effect has been demonstrated under a variety of practice tests, such as cued-recall, recognition, free recall, fill-in-the-blank, and short answer questions (Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013).

Several theories (for overviews, see Delaney et al., 2010; Karpicke et al., 2014; Rowland, 2014) have been proposed to explain the mechanism underlying testing effect. One category of explanations can be classified as elaboration theories (e.g. Carpenter, 2009, 2011; Pyc & Rawson, 2010), which propose that retrieval practice induces more semantic elaboration of a memory trace than restudy. When retrieving a target, information that is semantically related to the cue is activated, and becomes linked to the target. As a result, the number of retrieval routes is larger for tested items than for restudied items, which in turn leads to a testing benefit on a final memory test administered after a long retention interval.

CONTACT Gerdien G. van Eersel gerdienveersel@gmail.com

© 2017 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

Verkoeijen, Bouwmeester, and Camp (2012; see also Bouwmeester & Verkoeijen, 2011) postulated another explanation of the testing effect within this category, based on the fuzzy trace theory (Brainerd & Reyna, 2004). The central idea of the fuzzy trace theory is that information is stored on two different types of memory traces: verbatim/surface and gist traces. Verbatim traces are representations of a memory target's literal, contextual, and item-specific surface features. Gist traces, in contrast, are representations of semantic, relational, and other elaborative information about a target. The empirical support for the distinction between verbatim and gist traces is extensive (for an overview, see Brainerd & Reyna, 2004). According to the fuzzy trace explanation of the testing effect (Verkoeijen et al., 2012), restudying an item strengthens its verbatim memory traces more than retrieval practice does. By contrast, retrieval practice is assumed to activate the gist memory traces of the item, because people mainly use semantic cues to retrieve information from memory.

The fuzzy trace account can explain an important boundary condition of the testing effect, namely that the superior memory performance for tested items typically emerges after a long retention interval. Several studies (e.g. Anderson, 1974; Kintsch, Welsch, Schmalhofer, & Zimny, 1990; Sachs, 1967) have shown that information on verbatim traces decays more rapidly from memory than information on gist traces. The fuzzy trace explanation for this observation is that verbatim traces are more sensitive to sources of interference than gist traces, and therefore do not consolidate as much as gist traces (Brainerd & Reyna, 2004). Hence, after a short retention interval of several minutes, people can retrieve information from memory based on surface traces, gist traces, or a combination of both. After a retention interval of multiple days, though, they need to rely almost exclusively on gist traces. Because, according to the fuzzy trace explanation, retrieval practice is assumed to strengthen gist traces more than restudy, the testing effect is stronger after a multi-day retention interval than after a short retention interval of several minutes.

An interesting prediction that follows from the fuzzy trace account of the testing effect is that a testing effect will be obtained after a short retention interval (i.e. "short-term testing effect") when people cannot use verbatim/surface cues in a final test, and have to rely exclusively on semantic cues instead. Verkoeijen et al. (2012) tested this prediction. A group of 64 Dutch psychology undergraduates was asked to study 12 Dutch Deese–Roediger–McDermott word lists (DRM: Deese, 1959; Roediger & McDermott, 1995) either by restudying or by testing. Each list consisted of eight words, each of which had a strong backward association with one semantically related distractor. Immediately after the learning phase, participants took a final yes–no recognition test in Dutch (within-language condition) or in English (across-language condition). The participants were bilingual with respect to the English final test words in the across-language condition. It was assumed that participants in the within-language condition were cued with both semantic and verbatim/surface information (i.e. the visual appearance of a word) of the studied words. By contrast, in the across-language condition, surface features of the previously studied words were unavailable, so participants were only cued with semantic information. As indicated, according to the fuzzy trace account of the testing effect, recognition of restudied items depends more strongly on surface cues than the recognition of tested items. The recognition of tested items, on the other hand, hinges more strongly on semantic cues. For that reason, the fuzzy trace theory predicts a testing effect to arise in the across-language but not in the within-language condition. The results of Verkoeijen and colleagues' experiment (2012) were in line with these predictions. The proportion of correctly recognised items was higher for tested items (.78) than for restudied items (.67) in the across-language condition, but did not differ between tested items (.78) and restudied items (.81) in the within-language condition. In other words, there was a short-term testing effect in the across-language condition but not in the within-language condition.

As outlined above, the fuzzy trace account proposes that the short-term testing effect found by Verkoeijen and colleagues (2012)[1] emerged because the final test recognition of restudied items, but not of tested items, suffered from the lack of surface features overlap between the items in the learning phase and the cues in the final test. In the present study, we assessed this fuzzy trace account by

[1]Carpenter (2011) and Rawson, Vaughn, and Carpenter (2015) have also used semantic final test cues yet within a cued-recall setting, which differs from the recognition memory framework that is of interest in the current study.

gradually reducing the degree of surface features overlap between the items in the learning phase and the items in final test over five experiments. In this way it was possible to examine whether a testing effect would occur when there was small surface features overlap. In line with the fuzzy trace account, we expected that the smaller the surface features overlap, the larger the benefit of testing over restudy.

In all five experiments, participants studied a list of unrelated words through restudying or through testing. The crucial manipulation took place at the final yes/no recognition test, which was administered five minutes after the learning phase. In the first study, the factor "surface features overlap" differed in the extent to which there was a surface features overlap between the words in the learning phase and the cues in the final test. The factor had four levels: word, scrambled, background, and background scrambled. In the word condition, the final test cues were the studied targets (plus distractors) presented in exactly the same manner as in the learning phase, thereby guaranteeing a maximum overlap of surface cues between the learning phase and test phase. This condition was similar to the within-language condition of Verkoeijen and colleagues (2012). In the scrambled condition, scrambled versions of the words were presented, and here the surface features overlap between the words from the learning phase and the items in the final test was still considerable. In the background condition, words were vertically presented in a different font, at the top right of the computer screen on a colourful, flowered background. In the background scrambled condition, scrambled versions of the words were vertically presented in a different font, at the top right of the screen, on a colourful and flowered background. The surface features overlap in the latter condition was smaller than in the other three conditions. In the second experiment, there were two versions of the final recognition test: words and synonyms. In the synonym condition, synonyms of the target words were shown. In the last three experiments, the final recognition test consisted of either the target words or images of the target words, with the latter being purely semantic cues.

In general, we predicted that the smaller the surface features overlap between the words in the learning phase and the cues in the final test, the larger the benefit of testing compared to restudying. For this reason, we expected that no testing effect would emerge in any of the five word conditions. Furthermore, in the first experiment, there was still some surface features overlap between the targets in the learning phase and the cues in the final tests. We therefore predicted the advantage of testing in the background condition and the scrambled condition to be small or absent, possibly just like in the background scrambled condition, where the manipulation was a bit stronger but still subtle. In the final tests of the synonym condition (Experiment 2) and the image conditions (Experiments 3, 4a, and 4b), the surface features of the studied targets were absent, so we predicted a large benefit of testing to actuate in these two conditions. In addition, we expected that the smaller the surface features overlap, the more difficult the final test, and the lower the overall performance on the targets as well as on the unrelated distractors.

Experiment 1

Method

Participants and design

One hundred eighty-three participants were recruited online through Amazon's Mechanical Turk (AMT) (http://www.mturk.com). Twelve participants were excluded on the basis of one of the three following criteria, resulting in a total of 171 participants. Firstly, a score lower than zero on the equation "percentage correct on the targets minus percentage incorrect on the distractors". In this case, participants choose the option "old" (i.e. presented during the learning phase) more often when a word was new than when it was old. This indicates that they did not pay full attention to the task or that they (coincidentally) switched the response buttons. A second criterion for exclusion was a score of less than 30% correct on the distractors in the final test, since such a score of at least 20% points lower than chance level is also an indication of not giving full attention to the task or switching response buttons. Thirdly, participants were excluded when the log files showed that they had performed retrieval practice during the 2-minute distractor task, because practicing when they were supposed not to study would have an undesired effect on the final test scores. If a log file contained one or more words from the last learning phase instead of numbers counted backwards, all data from that participant were discarded. For more information on the demographic characteristics of the AMT population, see Paolacci, Chandler,

and Ipeirotis (2010), and Ross, Irani, Six Silberman, Zaldivar, and Tomlinson (2010). All participants were native English speakers and residents of the USA. They were paid $0.80 for their participation, which took about 25 minutes.

A 2 Study Method (restudy vs. testing) × 4 Surface Features Overlap (word vs. scrambled vs. background vs. background scrambled) mixed design was used with repeated measures on the first factor. Participants were randomly assigned to the levels of the between-subjects factor.

Materials

For the learning phase of Experiment 1, we selected 80 concrete, simple English nouns. Thirty-six words were used as "targets" (i.e. they would later appear in the final recognition test), while the other 44 words in the learning phase were used as fillers (i.e. not appearing in the final test). Mean word frequency was determined using the SUBTLEXus database, and did not differ statistically between targets and fillers (1.38 ± 1.62 and 1.49 ± 0.46 lnLog per million, respectively). Also, mean word length did not differ statistically between targets (4.67 ± 1.16 letters) and fillers (4.68 ± 1.29 letters). There were ten lists of eight words, and these ten lists were randomly split into two sets of five lists. Then four study sequences were created by counterbalancing across presentation order of sets (set 1 first vs. set 2 first) and study method (testing first vs. restudy first), such that participants first received a set of five restudy lists and then a set of five test lists, or vice versa. The order of words within a list and the order of lists within a set were fixed.

The final recognition test consisted of 73 words: 36 target words and 37 unrelated distractors,[2] the latter also being concrete English nouns. The words in the final test were randomly assigned to the serial positions, and the resulting test sequence was administered to all participants. In the word condition of the final test, words were presented in the same way (i.e. same font, letter size, letter type, and screen position) as they were during the learning phase. Hence, in this condition there was complete surface features overlap between the words in the learning phase and the words in the final test. In the scrambled condition, scrambled versions of the words were displayed. For example, the word "black" was presented as "cklba". Participants were asked to mentally unscramble the word and then to indicate whether the word had been presented during the learning phase. In the background condition, words were vertically presented in a different font at the top right of the computer screen on a colourful, flowered background. In the background scrambled condition, scrambled versions of the words were vertically presented in a different font, at the top right of the screen on a colourful and flowered background (see Appendix A for an example in black and white). As a result, the surface features overlap in the latter condition was minimal.

Procedure

The experiment was programmed and presented in the Qualtrics survey research suite (http://www.qualtrics.com). Participants were first informed that they would be presented with ten lists of eight words to memorise. They then started with an initial study phase in which words were presented in the centre of the computer screen at a 3.75-s rate. After each list, a free recall test was conducted (testing), or the list was presented again using the same procedure (restudy). In the free recall test, participants were asked to type in all the words that they could remember from the preceding study list. This free recall test took 30 seconds in total, which was equal to the duration of the restudy condition. The learning phase was followed by a 2-minute distractor task in which participants counted backwards on a sheet of paper in steps of three from a given number. Subsequently, participants completed the final recognition test. This task was varied according to the levels of the factor "surface features overlap": word, scrambled, background, and background scrambled. In all conditions, the final test required participants to indicate whether the word was old or new (i.e. presented in the previous learning phase or not). Words were presented one by one on the computer screen. In the scrambled condition and the background scrambled condition, participants were asked to first mentally unscramble the word and then to indicate whether the word was old or new. The final test was self-paced, and a new test item appeared after the participant had clicked on the next item button.

[2]Due to a few programming errors, the numbers of targets and distractors were not always equal, and also slightly differed across test sessions and experiments.

Results

The three outcome variables are the responses to the immediate free recall test, the unrelated distractors in the final recognition test and the targets in the final recognition test. Because these are all binomial count variables (correct = 1 / incorrect = 0), they were entered into a logistic regression analysis with a random intercept to deal with the dependence of the repeated measures. With this type of outcome variable, a regular ANOVA on the proportion of correct responses can lead to spurious findings because it might attribute probability mass to impossible values (i.e. values below 0 or above 1) and because the assumption of homogeneity is easily violated (Jaeger, 2008). The level of significance was set at $a = .05$.

Immediate test

There were no statistical[3] differences in the mean proportion correctly retrieved tested items in the learning phase between the word condition ($M = .73$, $SD = .15$), the scrambled condition ($M = .74$, $SD = .15$), the background condition ($M = .72$, $SD = .16$), and the background scrambled condition ($M = .73$, $SD = .18$), $Wald\ (156) = 1.76$, $p = .620$.

Final test performance

A 2 Study Method (restudy vs. testing) × 4 Surface Features Overlap (word vs. scrambled vs. background vs. background scrambled) logistic regression on the binomial count targets (see Table 1) did not reveal a statistical study method × surface features overlap interaction, $Wald\ (162) = 4.63$, $p = .200$. We did not find a statistical main effect of study method, $Wald\ (162) = 1.33$, $p = .250$. The regression coefficient (b) was −0.16, which is the log (ln) of the odds ratio between the restudy and the testing condition. The corresponding 95% confidence interval was [−0.42, 0.11], and the odds ratio for a correct answer was 0.85. This is a small effect size. An odds ratio of 0.85 means that the odds of a correct answer in the restudy condition is 0.85 times the odds of a correct answer in the test condition. The closer an odds ratio is to 1, the smaller the effect. Furthermore, there was a main effect of surface features overlap, $Wald\ (162) = 46.41$, $p < .001$. Recognition performance was $M = .80$ ($SD = .13$) in the word condition, it was $M = .72$ ($SD = .12$) in the scrambled condition (odds ratio = 0.55), it was $M = .80$ ($SD = .13$) in the background condition (odds ratio = 0.86), and $M = .61$ ($SD = .11$) in the background scrambled condition (odds ratio = 0.34), with the word condition acting as the baseline category for the odds ratios, which are all small. Additionally, the mean proportion of correctly classified distractors differed between the word condition ($M = .79$, $SD = .16$), the scrambled condition ($M = .75$, $SD = .15$, odds ratio = 0.79), the background condition ($M = .79$, $SD = .16$, odds ratio = 0.96), and the background scrambled condition ($M = .70$, $SD = .14$, odds ratio = 0.61), $Wald\ (144) = 46.08$, $p < .001$, with the word condition taken as the baseline category for the odds ratios, which are all small.

Table 1. Mean proportion of correctly recognised targets in Experiment 1 by surface features overlap and study method.

	Surface features overlap			
Study method	Words	Scrambled	Background	Background scrambled
Restudy	.79 (.02)	.72 (.02)	.82 (.02)	.60 (.02)
Testing	.81 (.02)	.72 (.02)	.79 (.03)	.62 (.02)

Note: Standard errors are between brackets.

Discussion

In our first experiment, we performed a subtle manipulation of the final test cues in order to assess the fuzzy trace explanation of the testing effect. According to this explanation, testing strengthens the gist traces of stimuli in memory, while restudying strengthens the surface traces. In Experiment 1, we varied the overlap between the surface features of presented words in the learning phase and the targets in the final test. By doing so, we sought to examine to what extent the surface features overlap had to fade in order for a short-term testing effect to emerge. We expected to find no advantage of testing in the word condition, a small or no advantage of testing in the background condition and the scrambled condition, and a relatively larger advantage of testing in the background scrambled condition. However, we did not observe an interaction effect of study method and surface features overlap, that is, the difference between the recognition of tested and restudied items was very small in all four conditions. Apparently, there was still sufficient surface features overlap between the learning phase and the final tests of all four conditions. We therefore conducted a

[3]Following Kline (2004, Chapter 3) and Cumming (2014), we use the term "statistically" instead of "'significantly", because the latter is often erroneously understood as meaning "important".

second experiment in which the surface manipulation of the final test cues was stronger, namely synonyms, resulting in a small surface features overlap between the learning phase and the final test words.

Experiment 2

Method

Participants and design

A total of 96 native English-speaking participants were recruited online through AMT. They were paid $0.80 for participating, which required about 25 minutes. Five participants were excluded from this experiment on the basis of one of the criteria mentioned in the method section of Experiment 1, resulting in a total number of 91 participants.

A 2 Study Method (restudy vs. testing) × 2 Surface Features Overlap (words vs. synonyms) mixed design was used with repeated measures on the first factor. Participants were randomly assigned to the levels of the between-subjects factor.

Materials

For the learning phase of this experiment, we selected 80 new English nouns and adjectives. Thirty-six words were used as targets and the other 44 were used as fillers. The synonyms in the final test were selected on the basis of the Edinburgh Associative Thesaurus word association norms (http://www.eat.rl.ac.uk/), for example movie/film and pants/trousers. After initial selection, we verified on Merriam-Webster's Learner's dictionary whether the words were indeed regarded as synonyms. Mean word frequency was determined using the SUBTLEXus database, and did not differ statistically between targets and fillers (1.49 ± 0.72 and 1.64 ± 0.69 lnLog per million, respectively). Also, mean word length did not differ between targets (5.11 ± 1.43 letters) and fillers (4.61 ± 1.21 letters). The counterbalancing method was the same as in Experiment 1. The final recognition test consisted of 36 target words and 36 unrelated distractors. In the word condition, words were presented identical to the way they were presented in the learning phase. In the synonym condition, synonyms of the words were shown.

Procedure

The procedure was identical to that in Experiment 1, except that the scrambled condition, the background condition, and background scrambled condition were replaced by one synonym condition. In this condition, participants were asked to indicate whether a synonym of the word on the screen had been in one of the studied lists ("old or new"). Participants in the synonym condition were given the following instruction:

> Next you will receive a test that consists of 72 words. For each word, you have to indicate whether a **synonym** of the word on the screen was in one of the lists you have just studied (yes) or not (no). For example, in the following test you see the word "act". If you have seen a synonym of "act" in one of the lists you've studied, for example the word "play", you answer yes. If you have not just studied a synonym of the word "act", you answer no.

We had ensured that none of the distractors in the final task was a synonym of one of the studied words.

Results

The three outcome variables and their analyses are the same as in Experiment 1.

Immediate test performance

The mean proportion correctly retrieved tested items during the learning phase did not statistically differ between the word condition ($M = .70$, $SD = .21$) and the synonym condition ($M = .72$, $SD = .14$), *Wald* (78) = 1.39, $p = .240$, regression coefficient −0.08, 95% confidence interval of the regression coefficient [−0.23, 0.06], odds ratio = 0.92. This is a small effect size.

Final test performance

A 2 Study Method (restudy vs. testing) × 2 Surface Features Overlap (words vs. synonyms) logistic regression on the binomial count targets (see Table 2) did not reveal a statistical interaction effect, *Wald* (86) = 0.62, $p = .430$, regression coefficient = −0.13, 95% confidence interval [−0.47, 0.20], odds ratio = 0.88. This is small effect size. This odds ratio of 0.88 means that in the word

Table 2. Mean proportion of correctly recognised targets in Experiment 2 by surface features overlap and study method.

	Surface features overlap	
Study method	Words	Synonyms
Restudy	.78 (.03)	.69 (.03)
Testing	.77 (.03)	.65 (.03)

Note: Standard errors are between brackets.

condition, the difference in odds of a correct answer between the tested and restudied words is 0.88 times this difference in the synonym condition. We did not find a statistical main effect of study method, *Wald* (86) = 3.22, $p = .073$, regression coefficient = 0.20, 95% confidence interval [−0.02, 0.43], odds ratio = 1.22. This is a small effect size. There was a statistical main effect of surface features overlap, *Wald* (86) = 12.88, $p < .001$, regression coefficient = 0.78, 95% confidence interval [0.35, 1.22], odds ratio = 2.18, with the proportion of recognised words being higher in the word condition ($M = .77$, $SD = .17$) than in the synonym condition ($M = .67$, $SD = .16$). This effect size is small. In addition, the mean proportion of correctly classified distractors was higher in the word condition ($M = .81$, $SD = .17$) than in the synonym condition ($M = .73$, $SD = .14$), *Wald* (70) = 28.55, $p < .001$, regression coefficient 0.45, 95% confidence interval [0.28, 0.62], odds ratio = 1.57. This effect size is small.

Discussion

In the second experiment, we expected an advantage of tested words compared to restudied words in the synonym condition but not in the word condition, since only in the former the surface cues of the studied words were unavailable. The results of the experiment were incongruent with these expectations. We did not find an interaction effect between the factors surface features overlap and study method, and even numerically there was no tendency in the hypothesised direction. These results were surprising, because the synonym condition was conceptually identical to the across-language condition in Verkoeijen et al. (2012). In the latter an interaction effect did occur, thereby substantiating the fuzzy trace account of the testing effect.

There is a possibility that the final test synonym cues did activate the surface representations of the studied words after all. Support for this idea might be found in studies using the lexical decision task (LDT). This task measures how fast participants can classify letter strings as words or non-words, which can be used to show a *priming effect*—the implicit memory effect that exposure to one word influences the response time to another word. Several authors have claimed that the LDT mainly relies on *orthographic* or lexical processes (e.g. De Groot, 2002; Zeelenberg & Pecher, 2003). Now some studies (e.g. Perea & Rosa, 2002) have shown a masked priming effect for related synonym pairs on the LDT. When a word was presented between 66 and 166 ms (i.e. the prime) and then followed by its synonym (i.e. the target) in the LDT, participants respond faster to the target than when the prime and the target were not related (Perea & Rosa, 2002). However, studies have failed to find *across-language* repetition priming effects on the LDT, that is, when the targets are translations of the primes (e.g. Gerard & Scarborough, 1989; Kirsner, Brown, Abrol, Chadna, & Sharma, 1980; Scarborough, Gerard, & Cortese, 1984; Zeelenberg & Pecher, 2003). This distinction might be due to the LDT primarily depending on orthographical or lexical processes. In support of this claim, Zeelenberg and Pecher (2003) showed that when a *semantic* classification task was used instead of the LDT, cross-language priming did occur. Together these studies indicate that cross-language cues directly active their semantic representations, while synonyms activate their orthographic representations. This hypothesis would explain the discrepancy between the results of our second experiment and the cross-language testing effect observed by Verkoeijen et al. (2012).

In sum, it is possible that the final test words of Experiment 2 did activate the orthographic representations of their studied synonyms after all. We therefore conducted another experiment with *non-verbal* cues as final test cues, namely images. Research has shown that images primarily activate their semantic representations (e.g. Johnson, Paivio, & Clark, 1996). We consider the image cues to be the strongest of all manipulations, probably even stronger than the across-language condition in Verkoeijen et al. (2012), since the latter condition is still of a verbal nature. In the image final test condition, there is no surface features overlap between the studied words and the final test cues.

Experiment 3

Method

Participants and design

A total of 152 native English-speaking participants were recruited online through AMT. They were paid $0.40 for their participation, which required approximately 25 minutes. Twelve participants were excluded on the basis of one of the criteria mentioned in the method section of Experiment 1, resulting in a number of 140 participants.

A 2 Study Method (restudy vs. testing) × 2 Surface Features Overlap (word vs. image) mixed design was

used with repeated measures on the first factor. Participants were randomly assigned to the levels of the between-subjects factor.

Materials

For the learning phase we used the same words as in Experiment 1, except for six words that could not easily be translated into images. We replaced these six words by six other concrete nouns, resulting in a total number of 36 targets and 44 fillers. Mean word frequency was determined using the SUBTLE-Xus database, and did not differ statistically between targets and fillers (1.37 ± 0.50 and 1.47 ± 1.62 lnLog per million, respectively). Moreover, mean word length did not differ statistically between targets (4.66 ± 1.26 letters) and fillers (4.73 ± 1.20 letters). The counterbalancing method was the same as in Experiment 1. The final recognition test consisted of 35 or 36 target words—depending on the test session—and 38 unrelated distractors. The word–image combinations were validated by asking six PhD candidates what the 73 images depicted (by free association, so without offering them any possible alternatives). Only if all six candidates mentioned the same object, the image was used. The images were obtained from the following website: http://users.skynet.be/taal/pictos/Page.html (see Appendix B for an example).

Procedure

The procedure of Experiment 3 was identical to that in Experiment 1, except that the scrambled condition, the background condition, and background scrambled condition were replaced by one image condition. In the image condition, images of the words were shown, and participants were asked whether the word that was represented by the image was old or new.

Results

The three outcome variables and their analyses are the same as in Experiment 1.

Immediate test performance

The mean proportion correctly retrieved tested items during the learning phase did not statistically differ between the word condition ($M = .74$, $SD = .15$) and the image condition ($M = .72$, $SD = .17$), *Wald* (78) = 2.16, $p = .140$, regression coefficient = −.09, 95% confidence interval [−0.21, 0.03], odds ratio = .91. This is a small effect size.

Final test performance

A 2 Study Method (restudy vs. testing) × 2 Surface Features Overlap (images vs. words) logistic regression on the binomial count targets (see Table 3) did not yield a statistical study method × surface features overlap interaction effect, *Wald* (135**)** = 0.62, $p = .430$, regression coefficient = −0.11, 95% confidence interval [−0.41, 0.18], odds ratio = 0.89. This odds ratio is of a small size. The analysis further showed a statistical main effect of surface features overlap, *Wald* (135) = 15.38, $p < .001$, regression coefficient = 0.61, 95% confidence interval [0.30, 0.92], odds ratio = 1.84. This effect size is small. The mean proportion of recognised words was higher in the word condition ($M = .83$, $SD = .11$) than in the image condition ($M = .74$, $SD = .14$). We did not find a statistical main effect of study method, *Wald* (135) = 2.63, $p = .11$, regression coefficient = 0.17, 95% confidence interval [−0.04, 0.39], odds ratio = 1.19. This effect size is small. In addition, the mean proportion of correctly classified distractors was higher in the word condition ($M = .80$, $SD = .16$) than in the image condition ($M = .71$, $SD = .18$), *Wald* (73) = 59.47, $p < .001$, regression coefficient = 0.50, 95% confidence interval [0.38, 0.63], odds ratio = 1.66. This is a small effect size.

Discussion

In Experiment 3, a study method × surface features overlap interaction effect did not occur: there was a numerical advantage of testing compared to restudying in the image condition and also in the word condition. However, both simple effects were too small to be statistically significant. Accordingly, the overall final test performance did not differ between the restudy condition and the testing condition. These outcomes are not in line with the findings by Verkoeijen and colleagues (2012). They observed a benefit of testing over restudy with purely semantic final test cues in their across-language condition, but no difference between

Table 3. Mean proportion of correctly recognised targets in Experiment 3 by surface features overlap and study method.

	Surface features overlap	
Study method	Words	Images
Restudy	.82 (.01)	.72 (.02)
Testing	.83 (.01)	.75 (.01)

Note: Standard errors are between brackets.

testing and restudy in their within-language condition.

The question then is what could be underlying the discrepancy between the results from our Experiment 3 and those from the across-language condition in Verkoeijen and colleagues' study (2012). One possibility is that the findings differed because the participant pools, settings, and procedures differed as well. That is, the study by Verkoeijen et al. (2012) was performed by Dutch psychology undergraduates in the laboratory at the Erasmus University Rotterdam, rather than by AMT workers anonymously at home, as in our Experiment 3 (as well as in our Experiments 1 and 2). Furthermore, there were three small differences between the procedure by Verkoeijen and colleagues (2012) and the procedure on AMT in our first three experiments (see the procedure below). However, it should be noted that there are no theoretical reasons as to why any of these differences—or a combination of some of these differences—should influence the testing effect. Nevertheless, we decided to repeat Experiment 3 in the laboratory with Dutch psychology undergraduates and Dutch materials, using exactly the same procedure as Verkoeijen et al. (2012). The only difference was in the final test cues, which were images instead of noncognate translations. Since the outcomes of the first laboratory experiment (Experiment 4a) supported the fuzzy trace theory of the testing effect, we repeated this experiment (Experiment 4b) to see if its results were robust.

Experiments 4a and 4b

Method

Participants and design

The participants were 60 (Experiment 4a) and 61 (Experiment 4b) Dutch undergraduates from the Erasmus University Rotterdam, the Netherlands, who were rewarded with course credits or €5.

A 2 Study Method (restudy vs. testing) × 2 Surface Features Overlap (words vs. images) mixed design was used with repeated measures on the first factor. Participants were randomly assigned to the levels of the between-subjects factor.

Material

We used the same words as in Experiment 3, except that we replaced six words that were not easily translatable into Dutch. There were 36 targets and 44 fillers. Mean word frequency was determined using the Dutch CELEX database, and did not differ statistically between targets and fillers (1.41 ± 0.62 and 1.41 ± 0.62 lnLog per million respectively). Also, mean word length did not differ statistically between targets (4.3 ± 1.03 letters) and fillers (4.2 ± 1.00 letters). The counterbalancing method was the same as in Experiment 1. The final recognition test consisted of 36 target words, and 37 (Experiment 4a) or 36 (Experiment 4b) unrelated distractors.

Procedure

Experiments 4a and 4b were programmed and presented in E-Prime software and conducted at the Erasmus University Rotterdam. The procedures were identical to that of Verkoeijen et al. (2012). Participants were first informed that they would be presented with ten lists of eight words, and they were asked to memorise these words. They then started with an initial study phase in which words were presented in the centre of the computer screen at a 4-s rate with a 1-s interstimulus interval (cf. the 3.75-s rate of Experiment 1, 2, and 3). In this initial study phase, participants were instructed to type in each word and memorise it (cf. Experiments 1, 2, and 3, where typing in the words was not required). After each list, participants engaged in free recall or restudy. The restudy phase was identical to the initial study phase. In the free recall phase, participants were asked to type in all words that they could remember from the preceding study list. Free recall time was divided into eight periods of four seconds, with a 1-second interval between periods (cf. Experiments 1, 2, and 3, where all remembered words were typed during one uninterrupted period). In this way, free recall time was equally distributed over the words, to make the procedure more similar to the restudy condition. The total test time added up to 40 seconds in total, which was equal to the time-on-task in the restudy condition. Participants completed the 2-minute distractor task and afterwards the final recognition test, which was similar to the final task in the previous experiments. Participants were asked whether the word was old or new. A new test item appeared after the participant gave a response (instead of after clicking on a button, as in Experiments 1, 2, and 3).

Results Experiment 4a

The three outcome variables and their analyses are the same as in Experiment 1.

Immediate test performance

The mean proportion of correctly retrieved tested items during the learning phase did not statistically differ between the word condition ($M = .74$, $SD = .15$) and the image condition ($M = .76$, $SD = .10$), *Wald* (54) = 1.36, $p = .240$, regression coefficient = −0.11, 95% confidence interval [−0.31, 0.08], odds ratio = 0.90. This is a small effect size.

Final test performance

A 2 Study Method (restudy vs. testing) × 2 Surface Features Overlap (images vs. words) logistic regression on the binomial count targets (see Table 4) showed a statistical study method × surface features overlap interaction effect, *Wald* (55) = 7.09, $p = .008$, regression coefficient = 0.62, 95% confidence interval [0.16, 1.08], odds ratio = 1.86. This is a small effect size. Specifically, there was a recognition advantage of testing over restudying in the image condition, *Wald* (27) = 9.82, $p = .002$, regression coefficient = 0.50, 95% confidence interval [0.19, 0.82], odds ratio = 1.65 (small effect size), but not in the word condition, *Wald* (27) = 0.45, $p = .500$, regression coefficient = −0.11, 95% confidence interval [−0.44, 0.21], odds ratio = 0.90 (small effect size). In addition, we found a main effect of study method, *Wald* (55) = 9.92, $p = .002$, regression coefficient −0.51, 95% confidence interval [−0.82, −0.19], odds ratio = 0.60. This is a small effect size. The mean proportion of recognised words was higher after testing ($M = .83$, $SD = .12$) than after restudying ($M = .80$, $SD = .15$). The main effect of surface features overlap did not reach statistical significance, *Wald* (55) = 0.21, $p = .650$, regression coefficient = −0.11, 95% confidence interval [−0.59, 0.37], odds ratio = 0.89. This is a small effect size. Additionally, the mean proportion of correctly classified distractors was higher in the word condition ($M = .85$, $SD = .10$) than in the image condition ($M = .80$, $SD = .13$), *Wald* (56) = 8.29, $p = .004$, regression coefficient = 0.33, 95% confidence interval [0.11, 0.55], odds ratio 1.39. This is a small effect size.

Table 4. Mean proportion of correctly recognised targets in Experiment 4a by surface features overlap and study method.

	Surface features overlap	
Study method	Words	Images
Restudy	.84 (.02)	.77 (.03)
Testing	.82 (.03)	.84 (.02)

Note: Standard errors are between brackets.

Results Experiment 4b

The three outcome variables and their analyses are the same as in Experiment 1.

Immediate test performance

The mean proportion correctly retrieved tested items during the learning phase differed between the word condition ($M = .68$, $SD = .12$) and the image condition ($M = .74$, $SD = .12$), *Wald* (59) = 9.94, $p = .002$, regression coefficient −0.28, 95% confidence interval [−0.46, −0.11], odds ratio = 0.76. This is a small effect size. However, the Pearson correlation coefficient between immediate test performance and the final test difference scores (correctly classified tested words—correctly classified restudied words) was $r = 0.02$, $p = .860$, which indicates that the differences on the immediate test did not confound the final test results.

Final test performance

A 2 Study Method (restudy vs. testing) × 2 Surface Features Overlap (images vs. words) logistic regression on the binomial count targets (see Table 5) only showed a trend toward a statistical study method × surface features overlap interaction effect, *Wald* (56) = 3.18, $p = .075$, regression coefficient 0.42, 95% confidence interval [−0.04, 0.88], odds ratio = 1.52. This is a small effect size. The odds ratio of 1.52 means that in the word condition, the difference in odds of a correct answer between the tested and restudied words is 1.52 times this difference in the image condition. There was no main effect of study method, *Wald* (56) = 1.09, $p = .300$, regression coefficient = −0.16, 95% confidence interval [−0.45, 0.14], odds ratio = 0.85. This is a small effect size. The main effect of surface features overlap did not reach statistical significance, *Wald* (56) = 2.61, $p = 0.11$, regression coefficient = 0.38, 95% confidence interval [−0.09, 0.84], odds ratio = 1.45. This is a small effect size. Additionally, the mean proportion of correctly classified distractors was higher in the word condition ($M = .83$, SD

Table 5. Mean proportion of correctly recognised targets in Experiment 4b by surface features overlap and study method.

	Surface features overlap	
Study method	Words	Images
Restudy	.88 (.02)	.77 (.03)
Testing	.85 (.02)	.79 (.03)

Note: Standard errors are between brackets.

= .10) than in the image condition (M = .79, SD = .16), *Wald* (59) = 6.38, p = .012, regression coefficient = 0.26, 95% confidence interval [0.06, 0.46], odds ratio = 1.30. This is a small effect size.

Discussion

In Experiment 4a, we found an interaction effect of study method × surface features overlap on recognition performance. In the word condition, there was no statistical difference between tested items and restudied items, while in the image condition a testing effect did emerge. In Experiment 4b, although there was no statistical interaction effect, numerically there was a tendency in the expected direction. In addition, there was a main effect of study method in Experiment 4a but not in Experiment 4b. When zooming in on the results of Experiment 4a, it appears that this main effect of study method resulted from the relatively high testing score in the image condition, which at the same time gave rise to the interaction effect in Experiment 4a.

What can we conclude from these outcomes? In Experiment 4a the p-value of the critical interaction effect was smaller than .05, while in Experiment 4b it was larger than .05 (i.e. p = .075). However, identical replication studies are likely to produce different outcomes as a result of random sampling fluctuation, especially for sample sizes that are typically used in psychological research (e.g. Coursey, Hovis, & Schulze, 1987; Gámez, Diaz, & Marrero, 2011; Lakens & Etz, 2017; Morey & Lakens, 2016). It is therefore best to evaluate the results of replication studies on more than just the criterion of statistical significance. That is, when the replication estimation is imprecise, the conclusion based on statistical significance might be opposite to what the evidence warrants (Simonsohn, 2015). It is possible that a replication study obtains an effect size similar to that of the original study, but still produces a non-significant finding because the replication estimation isnoisy or underpowered. On the other hand, two effect sizes can differ to a large extent but both lead to statistically significant outcomes. In the latter case, the replication attempt cannot be said to be successful. When evaluating the findings of replication studies, it is therefore important to also check whether the effect sizes and the confidence intervals are similar (Cumming, 2014). Now, the effect sizes of the interaction effects in Experiments 4a and 4b are comparable (i.e. odds ratios of 1.86 and 1.52, resp.), and the 95% confidence intervals of their regression coefficients largely overlap. This indicates that the interactions effects in these experiments were consistent with one another.

Furthermore, we conducted a 2 Experiment (4a vs. 4b) × 2 Study Method (restudy vs. testing) × 2 Surface Features Overlap (images vs. words) logistic regression on the binomial count targets, thus combining the data of Experiments 4a and 4b. This analysis yielded a statistical interaction effect between study method and surface features overlap, *Wald* (114) = 9.27, p = .002, odds ratio = 1.88, which is a small effect. Specifically, there was a recognition advantage of testing over restudying in the image condition, *Wald* (58) = 8.56, p = .003, regression coefficient = 0.32, 95% confidence interval [0.10, 0.54], odds ratio = 1.37 (small effect size), but not in the word condition, *Wald* (57) = 2.18, p = .140, regression coefficient = −0.18, 95% confidence interval [−0.42, 0.06], odds ratio = 0.83 (small effect size). Moreover, we did not observe a statistical three-way interaction effect, *Wald* (114) = 1.05, p = .300, odds ratio = 0.79 (small effect), which means that there is no indication that the interaction effect between surface features overlap and study method statistically differed between Experiments 4a and 4b. Again, this suggests that the outcomes of Experiments 4a and 4b are comparable. Both experiments clearly reinforce each other and together they provide evidence that the short-term testing effect is larger for images than for words.

Summary of all findings

In Figure 1, the regression coefficients are plotted, together with their confidence intervals, corresponding to the log(ln) of the odds ratio between restudy and testing within the surface features overlap conditions of the five experiments. In this plot, a positive regression coefficient signifies an advantage of testing on the final recognition test (i.e. a testing effect), while a negative coefficient denotes an advantage of restudying. In general, the overlap between the 12 confidence intervals is large, suggesting that the differences between conditions are small (and/or that the parameter estimations are imprecise). At the top of the figure the regression coefficients of the five *word* conditions are presented, which were predicted to be close to zero. Figure 1 shows that they are indeed centred around zero, signifying the absence of a testing effect in these conditions. Below are the regression

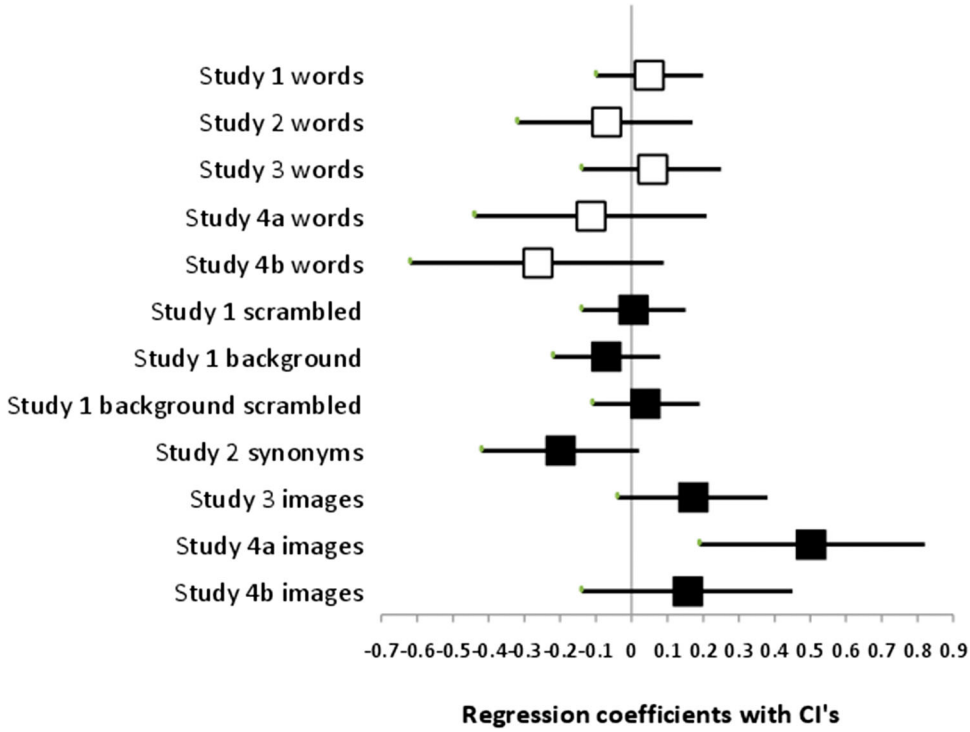

Figure 1. The regression coefficients, with their 95% confidence intervals, corresponding to the log of the odds ratio between restudy and testing with in the twelve different surface features overlap conditions of the five experiments. The white squares correspond to the five word-cue conditions, and the black squares correspond to the conditions where the surface features of the final test cues were altered as compared to the learning phase.

coefficients of the seven *non-word* conditions. Because there was still some surface features overlap in the final test conditions of Experiment 1 (scrambled, background, background scrambled), we expected a (very) small advantage of testing in the scrambled condition and the background condition, and possibly also in the background scrambled condition. We predicted larger advantages of testing to occur in the synonym and the image conditions, since these cues were purely semantic. It turned out that the subtle manipulations in Experiment 1 did not yield a testing effect in any of the surface features overlap conditions. In the synonym condition of Experiment 2, we did not find a testing benefit either. However, Figure 1 clearly shows that the results from the three image conditions in Experiments 3, 4a, and 4b stand out. Contrary to all other surface features overlap conditions, the image conditions consistently produced a mean recognition benefit of tested items over restudied items in the studied samples. Although the 95% confidence intervals indicate that there was only a two-tailed statistically significant testing effect in Experiment 4a, the results as a whole provide evidence that using image cues in the final test can give rise to a short-term testing effect in recognition.

General discussion

The fuzzy trace account of the testing effect predicts that a short-term testing effect will emerge when there is a low degree of surface features overlap between the items in the learning phase and the final test. In the present study, we assessed the fuzzy trace account by gradually reducing the availability of surface cues in the final test. In Experiment 1, the four surface features overlap conditions consisted of words, scrambled words, words vertically presented in a different font on a colourful flowered background ("background"), or a combination of the latter two conditions. In Experiment 2, the surface features overlap conditions contained either the same words or synonyms of the studied words. Experiments 3, 4a, and 4b had two surface features overlap conditions: words and images. In Experiments 1 and 2, the reduction of the availability of surface cues in the final tests did not result in a benefit of testing over restudying, which is not congruent with the fuzzy trace theory. The findings in

Experiments 3, 4a, and 4b, however, differed markedly from the findings in Experiments 1 and 2. These experiments showed an (numerical) advantage of testing as compared to restudying in the image conditions, which is in keeping with the fuzzy trace theory. Moreover, in the word conditions of Experiments 4a and 4b, no benefit of testing occurred. Experiments 4a and 4b were identical in their methods and subject pools, and overall produced corresponding results. However, although these testing benefits show that image cues can produce short-term testing effects in recognition memory, we hasten to add that more research is needed to examine the robustness of the short-term testing effect with image cues, because the results in the image conditions were small and quite variable. All in all, the present study provides only weak evidence in support of the fuzzy trace theory of the testing effect.

How can we explain the difference in results between the image studies on the one hand and Experiments 1 and 2 on the other? Apparently, the images trigger a distinctive response as for the effect of testing versus restudying on recognition. In Experiment 1, the manipulation of surface features overlap was more subtle than the manipulations in the other four experiments. Possibly, a considerable number of surface cues were still present in the final test of Experiment 1, such that the recognition of restudied words was not sufficiently impaired. The results in the synonym condition of Experiment 2, however, were surprising, because it was conceptually identical to the across-language condition in Verkoeijen et al. (2012). A potential explanation for these deviating outcomes might be that the synonyms in the final test did in fact activate the surface features of their intermediate test equivalents. Evidence for this idea comes from lexical decision studies that have demonstrated cross-synonyms priming (e.g. Perea & Rosa, 2002), but no cross-language repetition priming (e.g. Gerard & Scarborough, 1989; Kirsner et al., 1980; Scarborough et al., 1984; Zeelenberg & Pecher, 2003). This difference might be due to LDTs depending primarily on orthographic processes (e.g. De Groot, 2002; Zeelenberg & Pecher, 2003). Now if it is true that the surface features of the synonyms were in fact activated, then this would explain the difference between the findings in the synonym and the image conditions. However, this idea is speculative, and future research could focus on the differences between the memory representations of translation equivalents and synonyms.

In our last two experiments, as well as in the Verkoeijen, Bouwmeester, and Camp study (2012), the tasks were performed by Dutch college undergraduates at our laboratory. In the first three studies, on the other hand, AMT workers performed the task anonymously. This complicates the comparison between Experiments 1/2/3 versus 4a/4b. It might be possible that the AMT population has some distinctive characteristics that the Dutch undergraduates population lacks, which in turn interacted with the study method × surface features overlap effect in the present experiments. However, although there are known differences between the ATM population and a typical undergraduate pool (e.g. Paolacci et al., 2010), there are no theoretical reasons as to why these differences should produce a three-way interaction in present study. In addition, when looking at task performance measures, the ATM participants were very similar to the psychology undergraduates. That is, on the immediate test scores and the scores on the distractors we obtained highly comparable results across our ATM experiments and our experiments with psychology undergraduates. Also, the final test scores and the standard deviations were fairly similar across experiments. Moreover, many replication studies have shown that the behaviour of the AMT population resembles the behaviour of laboratory participants (e.g. Buhrmester, Kwang, & Gosling, 2011; Casler, Bickel, & Hackett, 2013; Horton, Rand, & Zeckhauser, 2011; Paolacci et al., 2010; Rand, 2012). Furthermore, Klein et al. (2014) assessed the replicability of a number of studies and found that very little of the variability in effect sizes could be attributed to whether the data collection occurred online or in the laboratory. All things considered, it seems unlikely that there were *relevant* differences in this study between the lab population and the AMT population.

However, perhaps procedural differences between Experiments 1/2/3 and Experiments 4a/4b might underlie the deviating results. Specifically, it might be possible that typing the words during the initial learning phase (Experiments 4a and 4b), and/or typing the words in separate periods during the free recall phase (Experiments 4a and 4b) versus typing them all in one go (Experiments 1, 2, and 3), made a difference to the outcomes. For example, typing responses during the initial learning phase (Experiments 4a and 4b) could

have strengthened the verbatim traces to a higher extent than not typing. However, if this were true, one would not have expected any testing effects in these last two experiments. Moreover, Verkoeijen et al. (2012) and Coppens, Verkoeijen, and Rikers (2011) also asked participants to type their responses during the initial study phase, as well as to type the words in separate periods during free recall. In the latter study, an advantage of testing over restudy did emerge after seven days, again suggesting that in the testing condition, the gist traces had been strengthened more than the verbatim traces. Taken together, we think it is unlikely that the procedural differences between Experiments 1/2/3 and Experiments 4a/4b led to variability in outcome patterns.

A different kind of explanation for the opposing outcomes evidently concerns the fuzzy trace theory itself. It is likely that the central notion that testing activates semantically related information does not fully correspond to reality. From a broader perspective, this would also mean that this type of elaborative retrieval accounts (e.g. Carpenter, 2009, 2011; Pyc & Rawson, 2010) is not corroborated, which is in line with a number of other studies (e.g. Karpicke et al., 2014; Lehman & Karpicke, 2016).

A different theory that might explain our findings is the bifurcation model (Halamish & Bjork, 2011; Kornell, Bjork, & Garcia, 2011). This theory predicts that the more difficult the final test, the larger the benefits of testing. According to this framework, items that are successfully recalled during testing are strengthened more in memory than items that are restudied. This implies that when the final test is sufficiently difficult, tested items will more often meet the criterion for retrieval in the final test than restudied items. Applied to the present study, the reduction in surface features overlap did not require participants to rely more on gist than on verbatim, but simply made the final test more difficult. However, the findings in Experiment 2 do not square well with the bifurcation framework. In the condition with the lowest average performance (the synonym condition), no benefit of testing emerged. Also, according to the bifurcation framework, one would have expected the benefit of testing to increase with a decreasing level of performance in the different final test conditions of Experiment 1. However, the data do not show such a pattern. In the two most difficult final test conditions, the difference between testing and restudying is absent (the scrambled condition) or statistically nonsignificant (the background scrambled condition). Furthermore, performance in the image condition was lower in Experiment 4b than in Experiment 4a, while the advantage of testing compared to restudy was somewhat larger in Experiment 4a than in Experiment 4b. Taken together, the bifurcation model cannot account for the findings in the present study.

All things considered, the outcomes of this study do not provide strong support for the fuzzy trace theory of the testing effect. The theory predicts that a short-term testing effect will arise when the overlap in surface cues between the learning phase and the final test is limited. This idea was substantiated in Experiments 4a and 4b, and partly in Experiment 3. Because the effect size estimates in these experiments were small and somewhat variable, it would be interesting to conduct a large-scale replication study to obtain more precise estimates, and shed more light on the question whether the fuzzy trace theory reveals one of the mechanisms underlying the testing effect.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research was funded by the Nederlandse Organisatie voor Wetenschappelijk Onderzoek [grant number 411-10-912].

References

Anderson, J. R. (1974). Verbatim and propositional representations of sentences in immediate and long-term memory. *Journal of Verbal Learning and Verbal Behavior*, *13*, 149–162. doi:10.1016/S0022-5371(74)80039-3

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. doi:10.1016/j.jml.2011.02.005

Brainerd, C. J., & Reyna, V. F. (2004). Fuzzy-trace theory and memory development. *Developmental Review*, *24*, 396–439. doi:10.1016/j.dr.2004.08.005

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk. A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*, 3–5. doi:10.1177/1745691610393980

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. doi:10.1037/a0017021

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. doi:10.1037/a0024140

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior*, *29*, 2156–2160. doi:10.1016/j.chb.2013.05.009

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. doi:10.1080/20445911.2011.507188

Coursey, D., Hovis, J., & Schulze, W. (1987). The disparity between willingness to accept and willingness to pay measures of value. *The Quarterly Journal of Economics*, *102*, 679–690.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7–29. doi:10.1177/0956797613504966

De Groot, A. M. B. (2002). Lexical representation and lexical processing in the second language user. In V. Cook (Ed.), *Portraits of the L2 user* (pp. 29–63). Clevedon: Multilingual Matters.

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology*, *58*, 17–22. doi:10.1037/h0046671

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: Academic Press. doi:10.1016./S0079-7421(10)53003-2

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*, 4–58. doi:10.1177/1529100612453266

Gerard, L. D., & Scarborough, D. L. (1989). Language-specific access of homographs by bilinguals. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 305–315. doi:10.1037/0278-7393.15.2.305

Gámez, E., Diaz, J. M., & Marrero, H. (2011). The uncertain universality of the Macbeth effect with a Spanish sample. *The Spanish Journal of Psychology*, *14*, 156–162.

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. doi:10.1037/a0023219

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. doi:10.1016/S0022-5371(71)80029-4

Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental Economics*, *14*, 399–425. doi:10.1007/s10683-011-9273-9

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards Logit mixed models. *Journal of Memory and Language*, *59*, 434–446. doi:10.1016/j.jml.2007.11.007

Johnson, C. J., Paivio, A., & Clark, J. M. (1996). Cognitive components of picture naming. *Psychological Bulletin*, *120*, 113–139. doi:10.1037/0033-2909.120.1.113

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: An episodic context account. In B. H. Ross (Ed.), *The psychology of learning and motivation* (Vol. 61, pp. 237–284). doi:10.1016/B978-0-12-800283-4.00007-1

Kintsch, W., Welsch, D., Schmalhofer, F., & Zimny, S. (1990). Sentence memory: A theoretical analysis. *Journal of Memory and Language*, *29*, 133–159. doi:10.1016/0749-596X(90)90069-C

Kirsner, K., Brown, H. L., Abrol, S., Chadna, N. K., & Sharma, N. K. (1980). Bilingualism and lexical representation. *Quarterly Journal of Experimental Psychology*, *32*, 585–594. doi:10.1080/14640748008401847

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., … Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, *45*, 142–152. doi:10.5334/jopd.ad

Kline, R. B. (2004). *Beyond significance testing. Reforming data analysis methods in behavioral research*. Washington, DC: APA Books.

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*, 85–97. doi:10.1016/j.jml.2011.04.002

Lakens, D., & Etz, A. (2017). Too true to be bad: When sets of studies with significant and non-significant findings are probably true. *Social Psychological and Personality Science*. doi:10.1177/1948550617693058

Lehman, M., & Karpicke, J. D. (2016). Elaborative retrieval: Do semantic mediators improve memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*, 1573–1591. doi:10.1037/xlm0000267

Morey, R. D., & Lakens, D. (2016). Why most of psychology is statistically unfalsifiable. (Submitted).

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, *5*, 411–419.

Perea, M., & Rosa, E. (2002). The effect of associative and semantic priming in the lexical decision task. *Psychological Research*, *66*, 180–194. doi:10.1007/s00426-002-0086-5

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science*, *330*, 335–335. doi:10.1126/science.1191465

Rand, D. G. (2012). The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology*, *299*, 172–179. doi:10.1016/j.jtbi.2011.03.004

Rawson, K. A., Vaughn, K. E., & Carpenter, S. K. (2015). Does the benefit of testing depend on lag, and if so, why?

Evaluating the elaborative retrieval hypothesis. *Memory & Cognition*, *43*, 619–633. doi:10.3758/s13421-014-0477-z

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*, 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 retention. *Psychological Science*, *17*, 249–255. doi:10.1111/j.1467-9280.2006.01693.x

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 803–814. doi:10.1037/0278-7393.21.4.803

Ross, J., Irani, L., Six Silberman, M., Zaldivar, A., & Tomlinson, B. (2010). Who are the crowdworkers? Shifting demographics in Mechanical Turk. In *Proceedings of CHI 2010* (pp. 2863–2872). Atlanta, GA: ACM.

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*, 1432–1463. doi:10.1037/a0037559

Sachs, J. S. (1967). Recognition memory for syntactic and semantic aspects of connected discourse. *Perception & Psychophysics*, *2*, 437–442. doi:10.3758/BF03208784

Scarborough, D. L., Gerard, L., & Cortese, C. (1984). Independence of lexical access in bilingual word recognition. *Journal of Verbal Learning and Verbal Behavior*, *23*, 84–99. doi:10.1016/S0022-5371(84)90519-X

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, *26*, 559–569. doi:10.1177/0956797614567341

Verkoeijen, P. P. J. L., Bouwmeester, S., & Camp, G. (2012). A short term testing effect in cross-language recognition. *Psychological Science*, *23*, 567–571. doi:10.1177/0956797611435132

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

Zeelenberg, R., & Pecher, D. (2003). Evidence for long-term cross-language repetition priming in conceptual implicit memory tasks. *Journal of Memory and Language*, *49*, 80–94. doi:10.1016/S0749-596X(03)00020-2

Appendix A

Appendix B