LANGUAGE LEARNING
A Journal of Research in Language Studies
Language Learning ISSN 0023-8333

EMPIRICAL STUDY

Contextual Richness and Word Learning: Context Enhances Comprehension but Retrieval Enhances Retention

Gesa S. E. van den Broek,[a,d] Atsuko Takashima,[a,b,c] Eliane Segers,[a] and Ludo Verhoeven[a]

[a]Behavioural Science Institute, Radboud University, [b]Max Planck Institute for Psycholinguistics, [c]Donders Institute for Brain, Cognition, and Behaviour, Radboud University, and [d]Department of Education, Utrecht University

Learning new vocabulary from context typically requires multiple encounters during which word meaning can be retrieved from memory or inferred from context. We compared the effect of memory retrieval and context inferences on short- and long-term retention in three experiments. Participants studied novel words and then practiced the words either in an uninformative context that required the retrieval of word meaning from memory ("I need the *funguo*") or in an informative context from which word meaning could be inferred ("I want to unlock the door: I need the *funguo*"). The

This research was supported by the National Initiative Brain & Cognition, Netherlands Organization for Scientific Research (NWO Grant number 056-33-014) and by a *Language Learning* Dissertation Grant. We thank Paul K. Gerke, Hubert Voogd, and Wendy van Hintum for technical support and contributions.

This article has been awarded Open Materials and Open Data badges. Stimuli and datasets are publicly accessible via the Open Science Framework at https://osf.io/eujyn. Learn more about the Open Practices badges from the Center for Open Science: https://osf.io/tvyxz/wiki.

Correspondence concerning this article should be addressed to Gesa van den Broek, Utrecht University, Department of Pedagogical and Educational Sciences, Heidelberglaan 1,3584 CS, Utrecht, Netherlands. E-mail: g.s.e.vandenbroek@uu.nl

DOI: 10.1111/lang.12285

informative context facilitated word comprehension during practice. However, later recall of word form and meaning and word recognition in a new context were better after successful retrieval practice and retrieval practice with feedback than after context-inference practice. These findings suggest benefits of retrieval during contextualized vocabulary learning whereby the uninformative context enhanced word retention by triggering memory retrieval.

Keywords retrieval; contextual inference; testing effect; vocabulary learning; paired-associate learning; second language

Introduction

Learning vocabulary in a second language (L2) is a gradual process that often requires repetition (e.g., Webb, 2007b). The way in which a word is processed during these repetitions predicts how well the word is remembered over time (e.g., Hulstijn & Laufer, 2001; Nation, 2001). To support acquisition, words are often presented in context, which allows learners to infer word meaning from contextual clues. In addition to using contextual clues, learners can also understand the meaning of words by retrieving knowledge gained during previous encounters with the word from memory (Nation, 2015; Schmitt, 2008). Both of these processes—context inferences and memory retrieval—are potentially beneficial for the long-term retention of words (e.g., Folse, 2006; Hulstijn, 1992; Nation, 2001). However, it is unclear whether the long-term retention of words is influenced by the degree to which a text stimulates readers to engage in context inferences or in memory retrieval. The present study was therefore conducted to examine the effect of these two processes more closely. Through three experiments, we investigated whether word retention is better when learners can infer word meaning from rich contextual clues or when learners must engage in the retrieval of word meaning from memory because the context is uninformative.

Background Literature

Word Learning Through Inferences From Context

Successful context inferences allow readers to establish the meaning of hitherto unknown words, which is necessary to create a form–meaning association (e.g., Li, 1988; Webb, 2008). Beyond this effect on understanding, inferences may also influence word retention. First, the processing of a word, together with relevant contextual information, could create semantic associations and enhance retention, compared to processing of words without context (e.g., Schouten-van Parreren, 1989). Second, the inference process could enhance word learning because of deeper (i.e., more effortful, elaborate) processing

of words during inferences, compared to other ways to gain access to word meaning, such as consulting a glossary (e.g., Grace, 1998; Hulstijn, 1992). This is especially likely when the inferences are difficult (Haastrup, 1991; Hu & Nassaji, 2012).

A substantial number of studies have focused on learners' understanding of novel words in context, for example, describing the contextual clues and comprehension strategies that facilitate inferences (Beck, McKeown, & McCaslin, 1983; Fukkink & de Glopper, 1998; Kuhn & Stahl, 1998). In comparison, less is known about the effects of context inferences on word retention. Some studies have reported better word retention after words are studied in context than without context (e.g., Baleghizadeh & Shahry, 2011), but others have reported no effect of contextual information or even an advantage of learning words without context (e.g., Choi, Kim, & Ryu, 2014; Prince, 1996; Webb, 2007a). Similarly, contradicting results were found in studies that compared the retention of inferred and given word meaning. Some experiments showed better word retention in an inference condition, compared to a condition in which words were presented in the same context but with the word meaning given (Carpenter, Sachs, Martin, Schmidt, & Looft, 2012; Hulstijn, 1992, Experiment 5). Others found no benefits of inferences (Hulstijn, 1992, Experiments 1 and 2), even when participants spent more time processing each word by inferring the meaning than when the meaning was provided (Mondria, 2003). Taken together, previous research has shown that readers can use inference processes to understand unknown words in a text, but there is limited evidence that exposure to contextual information and the cognitive processes involved in inferring word meaning from context also have benefits for retention.

A possible explanation for limited benefits of context inferences for retention is that learners might insufficiently process the word form while making inferences (Lawson & Hogben, 1996; Pressley, Levin, & McDaniel, 1987). Although learning L2 vocabulary involves the acquisition of many different aspects of word knowledge, the encoding of the novel word form, its spelling and pronunciation, and the association of this word form with meaning are crucial (Deconinck, Boers, & Eyckmans, 2015). Inferences may not always strengthen form–meaning associations. Consider the following sentence for illustration: "I want to unlock the door. I need the ___." Here, the missing word "key" can be guessed even when no word form is present. Contextual information can thus enable readers to infer the meaning of a word without paying attention to its orthographic or phonological characteristics (Hu & Nassaji, 2012; Hulstijn, Hollander, & Greidanus, 1996). Such a focus on semantics can lead to reduced encoding of the word form and,

consequently, weaken form–meaning associations (Barcroft, 2002). Thus, although inferences may involve effortful processing of word meaning, the processing of the word form while making inferences may be insufficient to create and retain strong form–meaning association (Pressley et al., 1987).

Word Learning Through Retrieval

Word learning often requires repetition, and after a while readers can access word meaning not only through inferences from context but also increasingly through the retrieval of word meaning from memory (Nation, 2015). Inferences and memory retrieval are, to some extent, competing processes because information that readers infer from the context is not searched for and retrieved from memory, and vice versa. The degree to which learners engage in retrieval is relevant for word learning because repeated successful memory retrieval leads to better retention over time (Roediger & Karpicke, 2006). For example, Karpicke and Roediger (2008) showed that when learners remembered the meaning of a new L2 word, practicing the retrieval of the word meaning from memory significantly enhanced performance on a translation test 1 week later, compared to a restudy condition in which words were repeatedly studied with translation but not retrieved from memory. Such positive effects of memory retrieval, compared to other practice conditions, are referred to as testing effects. The cognitive mechanism thought to underlie these testing effects is that information—in this case, the word meaning—is remembered better if it is retrieved from memory through an intentional mental search that involves the recall of knowledge encoded earlier than if it is presented to the learner (see also Karpicke & Zaromb, 2010).

Testing effects have been documented in numerous studies in cognitive psychology (for reviews, see Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Roediger & Butler, 2011; Roediger & Karpicke, 2006; Rowland, 2014), including multiple studies that used L2 vocabulary or rare words from the first language (L1) as stimuli (for an overview, see Goossens, Camp, Verkoeijen, & Tabbers, 2014). Vocabulary researchers have also acknowledged that retrieval is beneficial for word learning (e.g., Folse, 2006; Nation, 2001). For example, retrieval practice can be incorporated in vocabulary learning by using pictures (Barcroft, 2007) or a cloze task during reading (Barcroft, 2015) to trigger the retrieval of words from memory. In both studies, retrieval enhanced word learning, compared to a practice condition in which the translations were presented to the learners.

Prominent applications of memory retrieval to vocabulary learning are flashcards that allow learners to test themselves (e.g., Pimsleur, 1967) and

computer-assisted language learning programs with repeated, spaced translation exercises (e.g., Lindsey, Shroyer, Pashler, & Mozer, 2014; Sense, Behrens, Meijer, & van Rijn, 2016). These exercises involve explicit recall activities similar to tasks employed in psychological research (e.g., Karpicke & Roediger, 2008). However, testing effects might also be evoked more incidentally when a learner is in a situation that requires the activation of word knowledge from memory (Barcroft, 2015). For example, a reader who encounters a newly learned word for which the meaning cannot be derived from its context might try to retrieve word knowledge from memory and thereby improve the word retention over time. In other words, the number of contextual clues about a word's meaning might influence to what extent readers engage in memory retrieval or context inferences. Given the positive effects of retrieval on long-term retention, this raises the question of whether and how contextual richness influences word retention.

Effect of Contextual Richness on Word Retention

So far, only a limited number of studies have experimentally tested the effect of contextual richness on L2 word retention (Mondria & Wit-de Boer, 1991; Webb, 2008).[1] Mondria and Wit-de Boer conducted a study with Dutch high school students who guessed the meaning of French words from sentences and later reviewed the words in context with translations provided. When the initial practice sentences contained information about the function of the target words, students guessed the word meaning correctly more often than when the sentences did not contain this information. However, students were less likely to recall the words' meaning on a test after learning. Extra contextual information that made it easier to guess the word meaning during practice thus reduced later recall. The authors suggested that learners may have processed the target words less thoroughly in the richer context condition.

Unlike Mondria and Wit-de Boer (1991), Webb (2008) found positive effects of contextual information on word learning. He compared word learning between two groups of Japanese advanced learners of English who were presented with target words first in an informative sentence and then in two more sentences that were either "more informative" or "less informative" regarding word meaning (p. 236). On an extensive test immediately after learning, the recall and recognition of word meanings were better for words practiced in the more informative condition, but the recall and recognition of word forms were similar in both conditions. From this, the researcher concluded that the informative context may have increased the acquisition of word meaning but not of word form.

An important characteristic of both these studies is their focus on the initial presentation of words. Contextual information enables readers to understand the meaning of unknown words, which is important during readers' first encounter with a word. However, contextual information may have a different effect during later repetitions of words when readers have already acquired (partial) word knowledge that can be retrieved from memory instead of inferred from context. At this stage of learning, an uninformative context could become a beneficial trigger for retrieval. Results from Webb (2008) support the idea that readers' encounters with words in an uninformative context indeed become beneficial during later repetitions of words: Whereas presentations of novel words in uninformative sentences had no measurable effect on retention after their first presentation, significant benefits emerged after seven presentations. It is possible that the uninformative context triggered the retrieval of (aspects of) word meaning from memory. After a single exposure, this retrieval likely failed; however, after seven exposures to a word, learners might have gained some word knowledge, such that uninformative sentences could trigger successful retrieval and thus produce a testing effect (Kornell, Bjork, & Garcia, 2011; van den Broek, Segers, Takashima, & Verhoeven, 2014).

The Present Study

The central research question of the present study was whether repetitions of words in context enhance retention more when the context stimulates learners to retrieve word meaning from memory than when it allows learners to infer word meaning from context. To the best of our knowledge, no previous study has elicited testing effects in vocabulary learning through a manipulation of the context in which words appear. To address this question, adult participants learned the meaning of selected words from a previously unknown language and then further practiced these words either in an uninformative L1 context that required memory retrieval (the retrieval condition, as in "Look at the *anga*!") or in an informative L1 context that facilitated meaning inference (the context-inference condition, as in "There is not a single cloud today. Look at the *anga*!"). There is substantial evidence for beneficial effects of memory retrieval on the retention of information over time (Roediger & Butler, 2011; Roediger & Karpicke, 2006; Rowland, 2014) but only limited evidence for benefits of context inferences. Therefore, we predicted that retrieval would enhance word retention over time in comparison to context inferences. Although context inferences might involve beneficial semantic processing (Hulstijn, 1992; Schouten-van Parreren, 1989), we assumed that context inferences would direct readers' attention to word meaning rather than the form–meaning association

(Pressley et al., 1987) and would therefore lead to weaker retention, compared to retrieval practice.

Experiment 1

The overarching hypothesis for the three experiments reported here was that practicing words in uninformative sentences that triggered memory retrieval would lead to better word retention than practicing words in informative sentences from which word meaning could be inferred. In Experiment 1, this hypothesis was tested by manipulating contextual richness after pretraining, during which learners gained (partial) word knowledge. Such prior exposures are necessary to obtain testing effects in the absence of feedback because retrieval practice is only beneficial if learners can indeed retrieve information from memory or receive feedback after failed retrieval attempts (e.g., Kornell et al., 2011; Rowland, 2014). Otherwise, learners are not reexposed to the information to be learned and cannot benefit from retrieval practice.

The effect of retrieval and context-inference practice was tested in several ways to establish whether the predicted testing effects generalized across different measures of word learning. First, tests were administered both immediately and 7 days after practice because testing effects sometimes only become visible over time (Toppino & Cohen, 2009). Therefore, we predicted that benefits of the retrieval condition might be more pronounced on the delayed test than on the immediate test. Second, participants translated words both into their L1 and into the L2. This allowed us to test whether context-inference and retrieval practice affect both receptive and productive knowledge of L2 words (measured as recall of the word meaning and L2 word form, respectively). Productive word knowledge is typically more difficult to acquire than receptive knowledge—likely because it involves the formation of new lexical representations—whereas receptive knowledge "requires only discriminable, but not necessarily complete, representations of the new L2 words" (Schneider, Healy, & Bourne, 2002, p. 420). One reason to include both types of recall was that recall of word meaning and form might benefit from different retrieval tasks (Nakata, 2016). Moreover, it was unclear if retrieval of the word meaning during practice would also benefit later recall of the word form (see also Carpenter, Pashler, & Vul, 2006). Previously, Webb (2008) suggested that contextual information may be particularly beneficial for the retention of word meaning but not of word form. Mondria and Wit-de Boer (1991), however, found that contextual information reduced the recall of word meaning. Therefore, we did not formulate specific hypotheses about changes in productive and receptive word knowledge but expected that

retrieval practice would lead to better word retention than context-inference practice for all outcome measures. This testing effect would be driven by those words that participants translated successfully during practice, given the importance of retrieval success for testing effects (Kornell et al., 2011; Rowland, 2014).

Method

Participants

Forty-five undergraduate students (64.4% female) from a Dutch university took part in the experiment. All participants ($M_{age} = 23.8$ years, $SD = 8.8$) spoke Dutch fluently (88.9% native speakers), and none had prior knowledge of Swahili. In all three experiments, participants received partial course credits or monetary compensation (€10/hour).

Design

The study involved a 2 × 2 within-subjects design, with practice condition (retrieval, context inference) and testing time (immediate, delayed) as within-subjects factors and the proportion of words that were translated correctly on the tests of receptive and productive word knowledge as dependent variables. The assignment of words to the two conditions (52 words per condition) and to the immediate or delayed test (25 and 27 words from each condition, respectively) was random, as was the order of retrieval and the sequence of context-inference trials during practice.

Materials

The participants studied 104 Swahili nouns with Dutch translations that were pronounceable for Dutch speakers, such as *anga* ("sky"), *bustani* ("garden"), *kichwa* ("head"), and *samaki* ("fish"). Most words were taken from a norming study (Nelson & Dunlosky, 1994); additional words were found in an online dictionary. During practice, target words were presented in sentences. In the retrieval condition, these sentences contained only limited information and required memory retrieval to translate the target word (e.g., "We do not have any *mkate* left"). In the context-inference condition, an additional sentence made it possible to infer the word meaning (e.g., "I'll go to the bakery. We do not have any *mkate* left," where the word meaning is "bread").[2] The practice sentences were piloted to ensure that someone without prior knowledge of the target words could still derive word meaning from the context-inference sentences but not from the retrieval sentences (for further information, see Appendix S1 in the Supporting Information online).

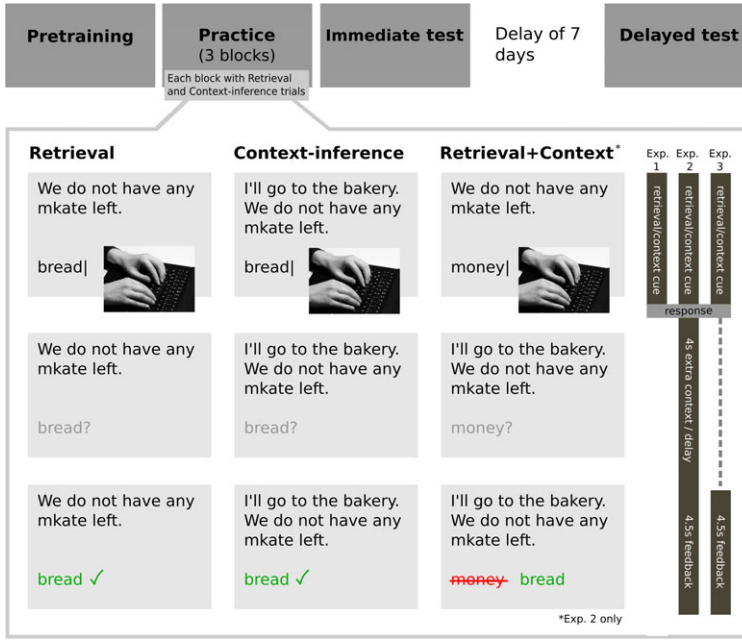

Figure 1 Overview of experimental procedure. In all three experiments, participants first completed pretraining in which Swahili words were studied together with their Dutch translations. The within-subjects experimental manipulation took place during the practice phase. Words were pseudorandomly assigned to the retrieval condition, context-inference condition, or retrieval-plus-context condition (Experiment 2). In the retrieval condition, participants practiced with sentences that provided only limited information about the Swahili word. In the context-inference condition, more information was provided to allow learners to infer the meaning of the target word from context. In the retrieval-plus-context condition (Experiment 2), participants first responded to a retrieval sentence and were then presented with contextual information. The diagram on the right indicates differences between practice trials in the three experiments: After participants responded, either the next trial began (Experiment 1) or the response remained visible on the screen in gray font for 4 seconds followed by feedback (Experiment 2) or feedback was shown directly (Experiment 3). The figure illustrates feedback for two correct responses (tick mark after correct word) and one incorrect response (strikethrough response, display of correct word). [Color figure can be viewed at wileyonlinelibrary.com]

Procedure

The experiment consisted of two sessions (see Figure 1). Session 1 included an initial encoding phase (pretraining) followed by retrieval and context-inference practice and the immediate test. Session 2, 7 days later, included the delayed test. Session 1 took about 2.5 hours while Session 2 took about 1 hour to complete.

Pretraining. The purpose of pretraining was to ensure that participants learned the meaning of the majority of the Swahili words before the practice phase. Participants intentionally studied the Swahili words together with translations in four different tasks. This was done using a repeated study procedure employed previously to ensure high initial encoding without providing retrieval opportunities (van den Broek et al., 2014; van den Broek, Takashima, Segers, Fernández, & Verhoeven, 2013). During the third task, each word was included in spaced repetitions until participants indicated twice that they already knew the word. On average, participants saw the words from both conditions 5.4 times over the course of the complete pretraining, both in the retrieval condition ($M = 5.4$, $SD = 2.1$) and in the context-inference condition ($M = 5.4$, $SD = 2.1$).

Practice With Retrieval and Context-Inference Sentences. The practice phase immediately followed pretraining. It consisted of three blocks. In each block, 52 words were presented in the retrieval condition and 52 words were presented in the context-inference condition. Sentences were presented one by one, and participants typed in the translation of the included Swahili word (see Figure 1). The sentences remained visible until the participants submitted a response. No feedback was provided, and a fixation cross was shown for 1.5 seconds before the next sentence was presented. The same 104 sentences were presented in all three practice rounds; presentation order was randomized.

Immediate and Delayed Memory Tests. A translation test was administered for 25 words from each condition directly after practice in Session 1, and for the other 27 words 7 days after practice in Session 2. Swahili words were presented one by one, and participants were asked to type the Dutch translation (a test of receptive word knowledge). After a short distractor task (an iconic memory task that took about 1 minute), participants were then asked to translate the same words from Dutch to Swahili (a test of productive word knowledge). In Session 2, the translation test was preceded by a picture-naming test. Participants were shown three complex pictures and were instructed to type in any Swahili word that described an element of the picture. Performance on this test was at floor level; therefore, the data are not reported in this article. After the picture test, participants engaged in distractor tasks for 3 minutes before completing the translation tests. The order in which items were tested was always randomized.

Data Analysis

Responses on the translation tests were categorized as either correct or incorrect, with spelling errors counted as correct in the test of receptive knowledge

(e.g., *fahter* instead of *father* was counted as correct). Responses in Swahili during the test of productive knowledge were counted as correct when they had an edit distance of two or lower from the correct answer, which means that no more than two letters had to be added or removed to get to the perfect answer (e.g., *keja* or *keah* instead of *keha* were counted as correct). Repeated-measures analyses of variance (ANOVAs) were run in SPSS (version 22.0.0.1) to examine the effect of practice condition (retrieval, context inference) and testing time (immediate, delayed) on the proportion of words that were translated correctly, as aggregated per participant. For all analyses, data met assumptions of normality, heteroskedasticity, and sphericity. Confidence intervals of the difference scores for pairwise contrasts are included in Figure 2; partial eta squared (η_p^2) is reported as a measure of effect size for omnibus tests, and Cohen's *d* is provided for pairwise comparisons, using Formula 3 in Dunlap, Cortina, Vaslow, and Burke (1996, p. 171). Because it is increasingly recommended to use mixed-effects modeling in psycholinguistics (Baayen, Davidson, & Bates, 2008), all analyses reported here were replicated using mixed logit models with crossed random effects for items and participants, using the glmer function in the lme4 package (Bates, Mächler, Bolker, & Walker, 2015) in R (version 3.1.2; R Core Team, 2014). All effects revealed in factorial analyses using the aggregate data were also significant in the mixed models. In addition, Bayesian analyses were used to quantify the evidence for and against the null hypothesis. These analyses indicated strong or very strong evidence for all significant effects; these findings are reported in Appendix S2 in the Supporting Information online. Given the sample size of 40–45 participants in the three experiments, a statistical power of 80% was reached for the pairwise comparisons of simple contrasts if the effect size d_z was between 0.37 and 0.40, which corresponded to the critical *t* value of 1.68, determined through a sensitivity analysis with Gpower (Version 3.1.9.2; Faul, Erdfelder, Lang, & Buchner, 2007). An effect size of this magnitude is considered small in L2 research (Plonsky & Oswald, 2014).

Results

Effects of Retrieval and Context Inference on Recall

Descriptive statistics for the proportion of word forms and meanings that were translated correctly on the final memory tests (immediately and seven days after learning) are reported in Table 1 and illustrated in Figure 2. For receptive word knowledge, there was a significant main effect of testing time, which reflected a decline in performance over time, $F(1, 44) = 181.00$, $p < .001$, $d = 0.82$, but no significant main effect of practice condition, $F(1, 44) = 0.004$,

Table 1 Mean proportion of correct responses (standard deviation) during practice and in the final tests in Experiment 1 ($N = 45$)

	Practice			Immediate recall		Delayed recall	
Practice condition	Block 1	Block 2	Block 3	Receptive	Productive	Receptive	Productive
Retrieval	.76 (.20)	.77 (.21)	.79 (.20)	.78 (.22)	.65 (.21)	.43 (.22)	.42 (.21)
Context inference	.96 (.07)	.96 (.07)	.96 (.06)	.77 (.21)	.64 (.22)	.41 (.22)	.40 (.19)
Successful retrieval	.92 (.07)	.93 (.10)	.95 (.07)	.90 (.13)	.73 (.16)	.50 (.20)	.48 (.19)
Successful context inference	.98 (.03)	.98 (.04)	.99 (.03)	.79 (.20)	.65 (.21)	.42 (.22)	.40 (.20)

Note. For the analysis of successful retrieval and successful context-inference practice, all included items were correctly translated in at least one of the three practice blocks.

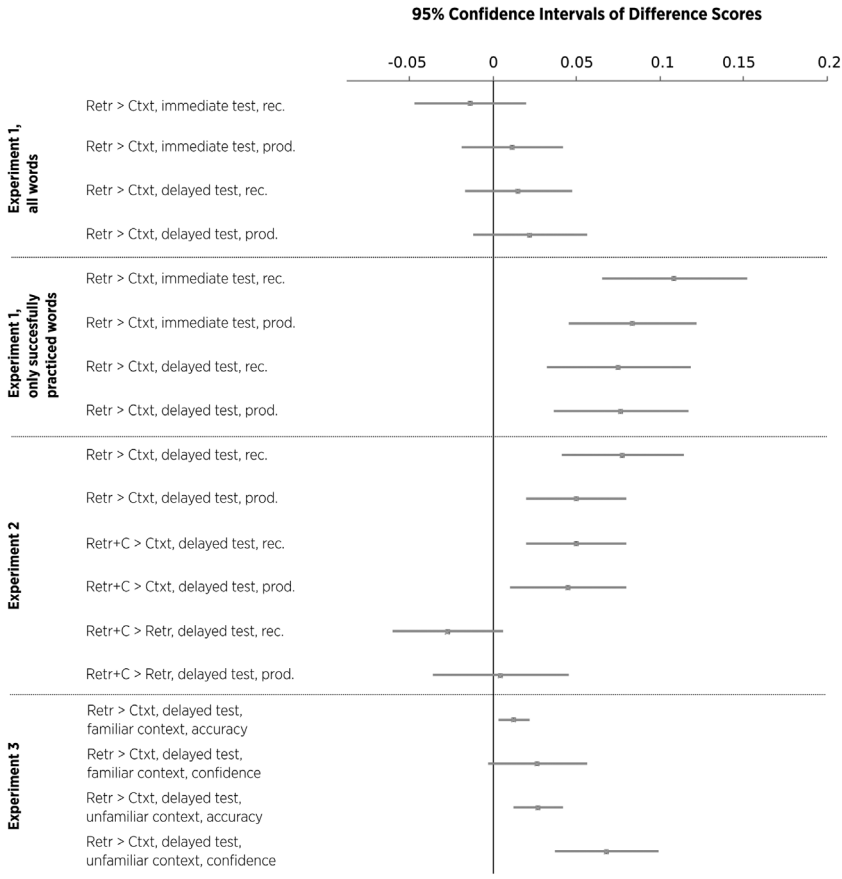

Figure 2 Ninety-five percent confidence intervals (CI) for the difference scores between final test performance after practice in the retrieval (Retr), context inference (Ctxt), and retrieval-plus-context (Retr + C) conditions (direction of comparisons is indicated by ">"). CIs are shown for accuracy of recall on tests of receptive (rec.) and productive (prod.) word knowledge, and for response accuracy and confidence in the sentence judgment test in Experiment 3. CIs that do not overlap with 0 indicate significant differences at $p < .05$.

$p = .95$, $d = 0.008$, nor an interaction between practice condition and testing time, $F(1, 44) = 1.62$, $p = .21$, $\eta_p{}^2 = 0.035$. For productive word knowledge, the pattern of results was the same, with a significant effect of testing time due to a decline in performance over time, $F(1, 44) = 109.14$, $p < .001$, $d = 0.67$, but no significant main effect of practice condition, $F(1, 44) = 1.94$, $p = .17$,

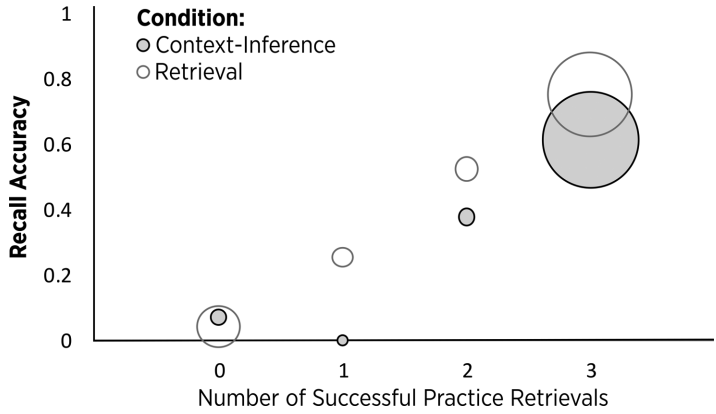

Figure 3 Recall accuracy for word meaning (receptive word knowledge) in Experiment 1 as a function of the number of successful practice responses (0, 1, 2, or 3) and practice condition (context inference or retrieval). Data from the immediate and the delayed test are combined. The surface of the circles represents the number of items in each category; the center of the circles represents mean recall accuracy averaged across item-level observations. See Appendix S3 in the Supporting Information online for the exact values illustrated in this figure.

$d = 0.09$, nor an interaction between practice condition and testing time, $F(1, 44) = 0.27$, $p = .62$, $\eta_p^2 = .006$.

Word Knowledge After Successful Retrieval and Context Inference

We measured participants' performance during practice because retrieval success is a requirement for obtaining testing effects. Although participants successfully translated the majority of words during practice, they incorrectly typed the translation of 18.2% of the words in the retrieval condition and of 2.4% of the words in the context-inference condition. Figure 3 summarizes the number of correct practice responses in the two conditions and illustrates the relationship between the number of correct practice responses with receptive word knowledge after practice. There are two important patterns to note. First, as indicated by differences in the surface area of the gray and the white circles, more words were successfully translated during practice in the context-inference condition than in the retrieval condition. A chi-square test of independence showed that this association between the practice condition and the number of correct responses during practice was significant, $\chi^2(3) = 834.62$, $p < .001$. Second, for words that had been translated successfully during at least one of the three practice rounds, receptive

word knowledge was higher for the retrieval condition than for the context-inference condition, as indicated by the height of the white circles over the gray circles.

Separate ANOVAs were carried out with participants' performance aggregated for only the words that had been translated correctly at least once during practice. These analyses revealed that the practice condition had a large significant effect on later recall, $F(1, 44) = 30.8$, $p < .001$, $d = 0.51$, with higher accuracy on the test of receptive word knowledge after successful retrieval practice ($M_{est} = 0.70$) than after successful context-inference practice ($M_{est} = 0.61$), with confidence intervals presented in Figure 2. Accuracy on the test of productive knowledge was also better after successful retrieval practice ($M_{est} = 0.61$) than after successful context-inference practice ($M_{est} = 0.51$), $F(1, 44) = 27.27$, $p < .001$, $d = 0.46$.

Discussion

In Experiment 1, participants repeatedly translated Swahili words presented either in an informative context from which word meaning could be inferred or in an uninformative context that required memory retrieval. Tests given immediately and 7 days after practice showed no difference between the two conditions in recall accuracy. However, whereas participants almost always filled in the correct translation during practice in the context-inference condition, they failed to provide the correct translation in the retrieval condition for about 18% of the words. When this difference in correct practice responses was controlled for, a benefit of retrieval practice became visible: Recall was significantly higher for words that had been successfully practiced in the retrieval condition than for words that had been successfully practiced in the context-inference condition. This result was found for all subscores—on the immediate and the delayed test and for productive and receptive word knowledge. Benefits of retrieval practice were not more pronounced on the delayed test than on the immediate test.

One interpretation of the results of Experiment 1 is that a testing effect exists during reading but only if the retrieval is successful (e.g., Halamish & Bjork, 2011; Kornell et al., 2011). It is plausible that participants only benefited from retrieval if they translated the target words successfully because, otherwise, they did not have access to the correct word meaning. However, this restriction of the analysis to successfully retrieved items may have introduced an unwanted bias (see Karpicke, Lehman, & Aue, 2014). Specifically, the 81.8% of words that were translated correctly during retrieval practice may have been inherently easier for participants than the 97.6% of words that were translated

correctly during context-inference practice. After all, during retrieval practice, participants could only translate those words from memory that they remembered from pretraining, whereas during context-inference practice, they could translate almost all of the words by inferring their meaning. Experiment 2 was therefore conducted to replicate the comparison between retrieval and context-inference practice while controlling for item selection effects by ensuring that learners always had access to the correct word meaning in both conditions via the use of feedback.

An alternative interpretation of the results of Experiment 1 is that extensive pretraining may have reduced the effect of practice. The size of the practice effect ($d = 0.46$ for receptive recall and $d = 0.51$ for productive recall) was relatively small (Plonsky & Oswald, 2014). There was no baseline measurement, but performance in the first round of retrieval practice can be considered a rough estimate of the proportion of translations that participants could recall after pretraining but before practice (76%; see Table 1). Comparing this estimate to performance on the immediate test after practice (77% in the context-inference condition, 78% in the retrieval condition) showed no significant improvement of performance after practice in either condition (both $d < 0.10$). This result was surprising for the context-inference condition, in which participants correctly translated almost all words repeatedly during practice and therefore could have learned additional words. The similar results before and after practice suggest that repeated successful context inferences had no or only minimal benefits for later recall. To rule out the possibility that these limited benefits were due to participants' performance reaching a plateau through extensive pretraining and to make the effect of practice more measurable, pretraining was shortened for Experiment 2.

Experiment 2

Experiment 2 was conducted to rule out the possibility that findings in Experiment 1 were driven by an item selection bias and to test, again, whether the retrieval condition enhanced word retention compared to the context-inference condition. For this purpose, we added feedback to the practice phase in Experiment 2. Feedback is beneficial for contextual word learning, especially when contextual support is weak (Frishkoff, Collins-Thompson, Hodges, & Crossley, 2016). In particular, feedback allows learners to encode information that they cannot retrieve from memory and therefore reduces the impact of retrieval failures (e.g., Rowland & DeLosh, 2015). Adding feedback in Experiment 2 made it unnecessary to include retrieval success in the analyses, which thus removed potential bias through item selection.

Two other major changes were made to the testing paradigm in comparison to Experiment 1. First, to make the effect of practice more measurable, pretraining was shortened in Experiment 2. Second, a third practice condition was added. In this retrieval-plus-context condition, participants first responded to an uninformative (retrieval) sentence and then were exposed to the contextual information from the context-inference condition so that they could evaluate their response. The inclusion of this combined condition allowed us to study possible additive benefits of the two conditions because both memory retrieval and context inferences are supposedly beneficial for word learning. We expected the combined condition to further enhance performance, in comparison to practice with only retrieval or only context inferences.

Method

This experiment had a similar structure to Experiment 1 (see Figure 1), but pretraining was shortened, a retrieval-plus-context condition was added, and feedback was included in the practice trials.

Participants

Forty-four undergraduate students ($M_{age} = 18.6$ years, $SD = 0.8$) took part in the experiment. All participants (84% female) spoke Dutch fluently (93% native speakers), and none of them had participated in Experiment 1 or reported prior knowledge of Swahili.

Design

This experiment involved a within-subjects design, with practice condition (retrieval, context inference, retrieval plus context) as the within-subjects factor and accuracy on delayed tests of receptive and productive word knowledge as dependent variables. The assignment of words to the three conditions was pseudorandom (based on pretraining), and the order of retrieval, context-inference, and retrieval-plus-context trials during practice was random.

Materials

For this experiment, we used 102 of the 104 Swahili nouns from Experiment 1, which were distributed across the retrieval condition, the context-inference condition, and the retrieval-plus-context condition per participant. The retrieval sentences and the context-inference sentences were identical to Experiment 1; in the newly added retrieval-plus-context condition, participants first saw the retrieval sentence (e.g., "Where is the *funguo*?"), made a response, and were then presented with the full contextual information from the context-inference

condition (e.g., "Where is the *funguo*? I would like to unlock the door"), as shown in Figure 1.

Procedure

Pretraining. Pretraining was similar to Experiment 1, but the third task was shortened. During the third task, participants now saw each word only once and rated how well they already knew the word on a continuous scale from 1 ("not at all") to 5 ("perfectly"). These ratings were used to assign words of similar difficulty to the three practice conditions for every participant, first by ranking words based on rating, then randomly distributing groups of three words across the three conditions. As a result, the average ratings in the three conditions were highly similar in the retrieval condition ($M = 2.23$, $SD = 1.04$), in the context-inference condition ($M = 2.24$, $SD = 1.04$), and in the retrieval-plus-context condition ($M = 2.24$, $SD = 1.03$).

Practice With Retrieval, Context-Inference, and Retrieval-Plus-Context Sentences. As in Experiment 1, participants typed in the translation of each Swahili word upon seeing the word in a sentence. The sentences remained visible until participants submitted a response. Next, the informative context sentence was added to the display in the retrieval-plus-context condition and remained visible for a fixed duration of 4 seconds. To ensure that trials in the three conditions were of the same length, the sentence(s) and the submitted response also remained visible on the screen for 4 seconds in the other two conditions, before feedback (i.e., the correct translation) was shown for 4.5 seconds in all three conditions (see Figure 1).

Immediate and Delayed Memory Tests. Due to the addition of a third condition, it was not possible to test sufficient items in both immediate and delayed tests. We therefore focused on recall performance on the delayed test in this experiment. To give participants some experience with the test situation, four words from each condition were presented in the immediate test directly after practice in Session 1; the other 30 words per condition were presented in the delayed test, 7 days after practice. The order of the translation tasks and distracter activities was the same as in Experiment 1 but with no picture-description task.

Data Analysis

Only the data from the delayed test were used in two repeated-measures ANOVAs, with practice condition (retrieval, context, retrieval plus context) as a within-subjects factor and accuracy on tests of receptive and productive

word knowledge as dependent variables. Data from the immediate test were excluded due to the low number of test trials, but exploratory analyses with data from both testing moments showed the same main effect of practice condition on the immediate test as on the delayed test and no interaction of practice condition and testing time. All analyses were also carried out through mixed-effects modeling, as described in Experiment 1.

Results

Receptive Word Knowledge

Descriptive statistics for the data from this experiment are summarized in Table 2. For receptive word knowledge, a repeated-measures ANOVA revealed a significant main effect of practice condition, $F(2, 86) = 11.15$, $p = < .001$, $\eta_p^2 = .21$. Pairwise comparisons for the three practice conditions showed that performance was lower in the context-inference condition ($M_{est} = 0.35$, $SE = 0.03$) than in the retrieval condition ($M_{est} = 0.42$, $SE = 0.03$, $p < .001$, $d = 0.38$) and in the retrieval-plus-context condition ($M_{est} = 0.40$, $SE = 0.03$, $p = .002$, $d = 0.25$), as illustrated in Figure 2. Numerically, performance was higher in the retrieval condition than in the retrieval-plus-context condition, but this difference did not reach significance ($p = .10$, $d = 0.13$). However, in a mixed logit model, the difference between the retrieval and the retrieval-plus-context condition was significant ($p < .05$), indicating higher performance in the retrieval condition than in the retrieval-plus-context condition.[3]

Productive Word Knowledge

For productive word knowledge, there was a significant main effect of practice condition, $F(2, 86) = 5.28$, $p = .007$, $\eta_p^2 = .11$. Pairwise comparisons revealed that, as with the receptive test, performance was lower in the context-inference condition ($M_{est} = 0.37$, $SE = 0.03$) than in the retrieval condition ($M_{est} = 0.42$, $SE = 0.03$, $p = .001$, $d = 0.24$) and in the retrieval-plus-context condition ($M_{est} = 0.42$, $SE = 0.03$, $p = .008$, $d = 0.22$). However, performance did not differ significantly between the two retrieval conditions ($p = .824$, $d = 0.02$).

Discussion

In Experiment 2, both retrieval conditions led to significantly higher productive and receptive word knowledge 7 days after practice, compared to the context-inference condition. With feedback, uninformative sentences that required memory retrieval led to better retention than informative sentences from which word meaning could be inferred. These results provide further evidence

Table 2 Mean proportion of correct responses (standard deviation) during practice and in the final tests in Experiment 2 ($N = 44$)

	Practice			Immediate recall		Delayed recall	
Practice condition	Block 1	Block 2	Block 3	Receptive	Productive	Receptive	Productive
Retrieval	.68 (.18)	.89 (.11)	.96 (.06)	.87 (.19)	.74 (.25)	.43 (.20)	.42 (.21)
Context inference	.96 (.08)	.99 (.02)	.99 (.01)	.74 (.25)	.70 (.28)	.35 (.20)	.37 (.21)
Retrieval-plus-context inference	.68 (.18)	.88 (.13)	.95 (.08)	.84 (.22)	.70 (.30)	.40 (.20)	.42 (.21)

Note. The data from the immediate test in Experiment 2 were not included in statistical analyses because the number of observations was too low (four items per condition) but are included here for descriptive purposes.

that reducing the amount of contextual information during practice can enhance L2 word retention by triggering retrieval. In Experiment 1, this testing effect was only found for those items that were successfully translated. In Experiment 2, however, after the addition of feedback, a testing effect was found for all items. This result strengthens the tentative conclusion from Experiment 1 that retrieval practice leads to better word retention than context-inference practice if learners have access to the meaning of words. By adding corrective feedback in Experiment 2, access to word meaning was guaranteed, and retrieval practice led to better performance than context inferences. This effect was found on tests of both productive and receptive word knowledge, as in Experiment 1.

An unexpected finding in Experiment 2 was the trend toward a negative effect of providing contextual information after retrieval. The difference between the retrieval and the retrieval-plus-context condition was small and reached statistical significance in mixed-effects modeling but not in the analysis of the aggregate data (where $p = .10$). Nevertheless, this result is noteworthy, because it contradicted our prediction that participants in the combined condition would benefit from each of the two conditions if they first tried to translate words encountered in the uninformative retrieval sentences from memory and then processed additional contextual information to infer word meaning and evaluate their answer. In the retrieval condition, participants could not infer word meaning from context and, instead, waited for 4 seconds after responding. Still, the retrieval condition led to better learning outcomes than the retrieval-plus-context condition, at least in terms of receptive knowledge. This finding provides further evidence that additional contextual information does not always enhance word learning and can sometimes even have negative effects.

Possibly, participants paid more attention to the target words in the period after submitting their response in the retrieval condition than in the retrieval-plus-context condition, where they may have instead focused on the sentence context. The contextual information could also have changed how participants approached the retrieval task in the second and third practice rounds if participants translated the words by recognizing or recalling the context rather than by activating the form–meaning association. In both cases, the focus on word meaning might have reduced encoding of the word form (Barcroft, 2002; Deconinck et al., 2015; Hu & Nassaji, 2012). Irrespective of the specific mechanism underlying the effect, additional contextual information had a negative effect on word retention in this experiment, as it resulted both in lower performance in the context-inference condition compared to the retrieval condition

and in lower performance in the retrieval-plus-context condition compared to the retrieval condition.

The unexpected results from the combined retrieval-plus-context condition raised the question whether inferring a word's meaning from context has benefits at all for word retention over time once a learner can understand the word. Again comparing translation accuracy in the first round of retrieval practice to recall on the immediate test, we found significantly better performance after practice than before practice in all three conditions. According to Plonsky and Oswald's (2014) guidelines, this effect was of a medium to large size both in the retrieval condition ($d = 1.03$) and in the retrieval-plus-context condition ($d = 0.79$), but this effect was small in the context-inference condition, $t(43) = 1.9, p = .03, d = 0.26$ (one tailed). These results suggest that, once learners understand a word, repetition in an uninformative context that triggers retrieval is more beneficial for retention than repetition in a rich context from which word meaning can easily be guessed. Repetition in a rich context had only a weak effect on word retention in Experiment 2.

Experiment 3

Experiments 1 and 2 showed better recall of word form and meaning after learners had practiced words repeatedly in an uninformative context that required memory retrieval than after learners had practiced words in an informative context from which word meaning could be inferred. We attribute this result to the beneficial effects of memory retrieval on retention (e.g., Roediger & Butler, 2011). However, an alternative explanation is that retrieval practice and the final translation tests both required that learners activate form–meaning associations from memory, whereas the context inference condition may have focused learners' attention more on word and context meaning. This difference in overlap between processing during practice and test may have biased results in favor of the retrieval condition due to transfer-appropriate processing, that is, the phenomenon that practice tends to have larger benefits when it involves similar cognitive processes as the final performance test (Morris, Bransford, & Franks, 1977; Veltre, Cho, & Neely, 2015; Winstanley, 1996). Indeed, L1 studies suggest that the benefits of practicing words in context can become visible when tests are sensitive to semantic associations or require the use of words in context, even when recall tests show no such benefits (Frishkoff, Perfetti, & Collins-Thompson, 2011).

Experiment 3 was therefore conducted to see if the benefits of retrieval practice, compared to context inferences, could be replicated with a final test that was more sensitive to semantic associations and was more similar to

context-inference practice. We constructed a test in which participants had to judge whether the practiced words were appropriately used in different sentences. Some test items presented the target words in a sentence that included words and semantic concepts from the context-inference condition; other test items presented the target words in a new, unrelated context. Based on the idea that the overlap between practice and final test enhances performance, we expected to find smaller or no benefits of retrieval practice over context-inference practice in this test, compared to the previous experiments. Furthermore, differences in performance on familiar and unfamiliar test items were investigated to determine to what extent benefits of context-inference practice were restricted to the specific context from practice or transferred to a new context. The answer scale measured both accuracy and confidence of responses to measure word learning both objectively and subjectively.

Method

Experiment 3 included only the retrieval condition and the context-inference condition. Pretraining was identical to that in Experiment 2. Practice trials were also similar to those used in Experiment 2 but were shortened, and the test format was changed (see Figure 1 for an overview of the differences between experiments).

Participants

Forty-one university students (M_{age} = 20.0 years, SD = 2.3) took part in Experiment 3. Again, all participants spoke Dutch fluently (40 female, 92.7 % native speakers), and none had prior knowledge of Swahili or had participated in Experiments 1 or 2.

Materials and Procedure

The target words included 100 of the 102 words from Experiment 2.

Retrieval and Context-Inference Practice. Immediately after participants submitted a response, the same feedback (i.e., correct translation) from Experiment 2 was displayed for 4.5 seconds.

Sentence Judgment Test. For each Swahili word, four test sentences were constructed (see Table 3 for examples). These were two sentences in which the Swahili words fit into the context (fit) and two sentences in which the Swahili words did not fit (no-fit). For each word, one of the two fit test sentences and one of the two no-fit test sentences were semantically related to the practice sentences from the context-inference condition (i.e., familiar). The other fit and no-fit test sentences were different from practice (i.e., unfamiliar).

Table 3 Example items from the sentence judgment test in Experiment 3

	Test sentence type			
	Familiar		Unfamiliar	
Sample item	Fit	No-fit	Fit	No-fit
mkate ("bread")	I'll quickly go to the bakery to get some *mkate*.	During an asthma attack, *mkate* cannot enter the lungs freely.	Fresh *mkate* tastes best.	He walks with crutches because he hurt his *mkate*.
hewa ("air")	During an asthma attack, *hewa* cannot enter the lungs freely.	She knits a scarf of fine *hewa*.	Many factories in this area pollute the *hewa*.	The *hewa* was sharing my friend's bike.
Answer scale	1. I am sure that the word does not fit in this context. 2. I think the word does not fit in this context. 3. I don't know but my guess is that the word does not fit in this context. 4. I don't know but my guess is that the word fits in this context. 5. I think the word fits in this context. 6. I am sure that the word fits in this context.			
Correct response	The word fits in this context.	The word does not fit in this context.	The word fits in this context.	The word does not fit in this context.

Notes. During the test, all words were presented in each of the four types of test sentences (familiar fit, unfamiliar fit, familiar no-fit, unfamiliar no-fit). The familiar fit sentences were created based on the practice sentences from the context-inference condition (see Figure 1). For the familiar no-fit items, a Swahili word was inserted into a sentence that belonged to a different Swahili word, such that the word did not fit into the context.

The familiar fit sentences were constructed using words or concepts from the context-inference practice sentences. For the familiar no-fit sentences, Swahili words were inserted into the familiar fit sentence of a different Swahili word, thereby creating a test sentence in which the word did not fit. The unfamiliar sentences were constructed using words and topics that did not occur in the practice context. The presentation order of the test items was random, but it ensured that for half of the words from each condition, the familiar fit sentences were presented first, and for the other half of the words, the unfamiliar fit sentences were presented first for each participant.

Accuracy and Confidence Measures. Participants rated each test item on a 6-point scale that indicated whether they thought that the word fit the context (left half of the scale) or not (right half of the scale) and how confident they were in their answer (three levels on each half of the scale, from 1 = *guess* to 3 = *I am sure*), as illustrated in Table 3. Response accuracy (i.e., answering on the left or right half of the scale given the fit of the word into the sentence), confidence, and confidence for accurate responses only were then aggregated per participant.

Data Analysis

Repeated-measures ANOVAs were carried out with practice condition (retrieval, context inference) and familiarity of test context (familiar, unfamiliar) as within-subjects factors and participant means for accuracy and confidence ratings as dependent variables. Separate mixed-effects models were also conducted using accuracy and confidence measures; these analyses replicated the reported significance effects.

Results

Accuracy

Descriptive statistics for response accuracy and confidence are summarized in Table 4. The accuracy of judgments of words in context was higher in the retrieval condition than in the context-inference condition, $F(1, 40) = 19.32$, $p < .001$, $d = 0.24$. Participants also showed higher accuracy for the familiar test items than for the unfamiliar test items, $F(1, 40) = 89.93$, $p < .001$, $d = 0.50$. There was no interaction between practice condition and familiarity, $F(1, 40) = 2.56$, $p = .12$, $\eta_p^2 = .06$.

Confidence

Similar to the results for accuracy, confidence was higher in the retrieval condition than in the context inference condition, $F(1, 40) = 12.94$, $p < .001$,

Table 4 Mean proportion of correct responses (standard deviation) during practice and response accuracy and confidence ratings (0–3 scale) for word recognition in familiar and unfamiliar contexts in Experiment 3 ($N = 41$)

	Practice			Recognition: familiar		Recognition: unfamiliar	
Practice condition	Block 1	Block 2	Block 3	Accuracy	Confidence	Accuracy	Confidence
Retrieval	.68 (.18)	.88 (.12)	.96 (.08)	.89 (.08)	2.65 (0.24)	.85 (.09)	2.55 (0.32)
Context inference	.96 (.05)	.99 (.01)	.99 (.01)	.88 (.07)	2.63 (0.26)	.83 (.08)	2.51 (0.31)

$d = 0.14$, and higher for the familiar than for the unfamiliar test items, $F(1, 40) = 206.14$, $p < .001$, $d = 0.43$. Additionally, there was a significant interaction between practice condition and familiarity of test context, $F(1, 40) = 8.31$, $p = .006$, $\eta_p^2 = .17$, reflecting significantly greater confidence when participants rated words from the retrieval condition than when they rated words from the context-inference condition for the unfamiliar test items ($p < .001$, $d = 0.19$) but not for the familiar test items ($p = .075$, $d = 0.09$).[4] To ensure that these results were not driven by differences in accuracy, confidence ratings were also aggregated for accurate responses only. This analysis led to the same pattern of results as the analysis of all confidence data.

Discussion

Learners more accurately and more confidently recognized words in context 7 days after retrieval practice than after context-inference practice. Experiment 3 thus replicated the benefits of retrieval practice found in Experiments 1 and 2, now with a test that involved the presentation of words in a sentential context. Judgments on this test were more accurate after prior retrieval practice than after prior context-inference practice both when test items were familiar because they resembled the sentences used during context-inference practice and when test items were unfamiliar. In addition, participants were more confident when they judged words from retrieval practice than when they judged words from context-inference practice. This effect was more pronounced when words were presented in an unfamiliar context than when words were presented in a familiar context ($p = .075$ in the ANOVA but $p < .05$ in the corresponding mixed-effects model).

As we had predicted, familiar sentences were rated more accurately and more confidently than unfamiliar sentences, possibly due to transfer-appropriate processing (Veltre et al., 2015; Winstanley, 1996). However, the familiar test items were constructed based on the sentences from context-inference practice and were therefore only familiar to participants for the words practiced in that condition. This greater overlap between context-inference practice and familiar test context than between retrieval practice and familiar test context might explain why the data showed stronger evidence for benefits of the retrieval condition for the unfamiliar than for the familiar test items. For the latter, benefits of retrieval may have been counteracted by the greater overlap between test items and context-inference practice. Independent of familiarity, accuracy was better for the words practiced in the retrieval condition than in the context-inference condition for both familiar and unfamiliar test items. Overall, although transfer-appropriate processing might have influenced how

well participants recognized words in context, benefits of retrieval practice over context-inference practice were robust and existed even on a test that presented words in context.

General Discussion

Summary of Findings

Given that language learners often practice words in context, it is important to understand the effect of textual characteristics on word retention. This study focused on contextual richness as a source of context inferences and memory retrieval during intentional vocabulary practice. In three experiments, words were remembered better after practice with an uninformative context that required memory retrieval to access word meaning, compared to practice with an informative context from which word meaning could be inferred. In Experiment 1, this testing effect was found only for words that participants had translated successfully during practice: Performance was higher after successful retrieval practice than after successful context-inference practice, both immediately after learning and after 7 days, on tests of productive and receptive recall. In Experiment 2, feedback was added to the practice phase and a testing effect was found for all items. This confirmed that the testing effect in Experiment 1 was not an artifact of item selection but was, indeed, related to benefits of successful retrieval for retention. Moreover, Experiment 2 showed that a combined retrieval-plus-context condition did not enhance performance, compared to a pure retrieval condition, suggesting that benefits of contextual inferences are limited once learners can retrieve the meaning of words from memory. Finally, in Experiment 3, the testing effect was obtained with a final test that presented words in a sentence context. Both response accuracy and confidence were higher after retrieval practice than after context-inference practice, showing that the testing effect was not restricted to a specific recall test.

Overall, the three experiments showed that memory retrieval enhances long-term retention of novel L2 vocabulary to a greater extent than context inferencing, as had been predicted based on the extensive literature targeting testing effects (e.g., Roediger & Butler, 2011; Rowland, 2014) and the comparably limited empirical support for benefits of context inferences on word retention (Mondria, 2003; Mondria & Wit-de Boer, 1991). The fact that reducing the amount of contextual information to trigger memory retrieval had a consistent positive influence on word retention confirms that testing effects can be evoked indirectly by creating a need to retrieve information from memory when that information is not accessible from context.

Context Influences on Word Comprehension and Retention

The present results appear at odds with the widely held view that an informative context is conducive to word learning because contextual clues facilitate the inference of word meaning (e.g., Seibert, 1945) and that understanding a word's meaning is necessary to establish a form–meaning connection (Li, 1988). However, the comprehension of words in context (e.g., during reading) and the retention of words over time are distinct processes (Lawson & Hogben, 1996; Verspoor & Lowie, 2003). As a case in point, the present study showed that contextual information affected comprehension and retention in different ways: Contextual information increased the chance that learners found the correct word meaning during practice, but it reduced the retention of these words over time. This somewhat counterintuitive finding parallels other learning conditions that facilitate practice but lead to worse long-term outcomes, such as massed repetition, as opposed to spaced repetition, and continuous practice with the same task, as opposed to practice with varying tasks (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Nakata, 2017; Nakata & Webb, 2016). Such conditions lead to high performance during practice but allow learners to bypass the effort and engagement necessary for durable learning, resulting in worse long-term outcomes (Bjork, 1994; Yan, Clark, & Bjork, 2016). In contrast, conditions like retrieval constitute so-called desirable difficulties—a term coined by Bjork (1994)—that require more effortful, often slower and more error-prone processing of the information to be learned but also lead to better long-term outcomes. In alignment with this framework, reducing contextual information in our experiments likely created a desirable difficulty because learners had to engage in effortful retrieval, whereas rich contextual information gave learners easy access to word meaning and involved only superficial processing of the form–meaning association.

Control analyses showed that the retrieval condition did not lead to longer processing times, compared to the context inference condition. On the contrary, response times were longer in the context inference condition than in the retrieval condition in the first practice block in all experiments (see Appendix S4 in the Supporting Information online). Thus, benefits of retrieval compared to context inferencing seem to be driven by the type of processing rather than the duration of processing. These results demonstrate that it is crucial to consider the effects of context manipulation not only on comprehension of words but also on the way in which learners process and subsequently remember these words.

Although the present study showed that reducing contextual information during practice can enhance word learning, it should not be seen as an argument

for words to always be presented in an uninformative rather than in an informationally rich context. Whether difficulties during learning are desirable depends on learner capabilities and their prior knowledge (McNamara, Kintsch, Songer, & Kintsch, 1996). Here, we focused on the effect of contextual richness during later repetition of words, when learners most likely have acquired some word knowledge that must be consolidated through further repetition. This is supposed to be "one of the most important phases in vocabulary learning which has not been researched sufficiently" (Peters, Hulstijn, Sercu, & Lutjeharms, 2009, p. 118). The present study showed that in this phase of learning, a reduction of contextual information can be beneficial if learners succeed at retrieving word meaning from memory. In contrast, during initial repetition, learners would be less likely to successfully retrieve word meaning from memory. In this case, the trade-off between facilitating comprehension with more context, on the one hand, and enhancing retention over time with less context, on the other, poses a greater challenge. The previous finding that exposure to an uninformative context may become more beneficial after repeated prior exposures to words (Webb, 2008) supports this idea. One solution to ensure that learners can benefit from retrieval opportunities earlier in practice could be to provide feedback as in Experiments 2 and 3. Another solution may be to reduce contextual richness gradually over the course of several repetitions (see also Finley, Benjamin, Hays, Bjork, & Kornell, 2011).

Given the importance of feedback in the present experiments, it is noteworthy that the literature on testing effects is not limited to retrieval practice for consolidation of previously learned materials. It also describes so-called test-potentiated encoding. This refers to the finding that the information provided after learners have (unsuccessfully) attempted to guess or retrieve this information from memory is remembered better than the information that is directly presented to learners (e.g., Kornell, Hays, & Bjork, 2009; Potts & Shanks, 2014; Richland, Kornell, & Kao, 2009). Retrieval attempts might enhance learner involvement. In fact, similar concepts have been discussed in the language literature in terms of the need to process a word (Laufer & Hulstijn, 2001). Alternatively, retrieval attempts might lead to a more thorough inspection of available cues, such as word form. In any case, studies on test-potentiated learning suggest that the retrieval condition could be beneficial even in the absence of extensive prior training because learners may benefit from retrieval attempts even when the retrieval fails, as long as corrective feedback is available (Rowland & DeLosh, 2015; van den Broek et al., 2014). In the present study, we did not distinguish the indirect effects of retrieval on feedback processing from the effects of the retrieval itself,[5] but the mechanisms of

feedback processing could be interesting for follow-up research. For example, one practical question is whether feedback after a retrieval attempt has to be explicit or whether presenting a context sentence from which word meaning can be derived is similarly effective. The magnitude of significant differences reported in this study were comparably small in terms of effect size values (Plonsky & Oswald, 2014), and more research is needed to establish under which conditions a reduction of contextual clues is beneficial.

Limitations and Future Research

A number of directions for future research can be derived from the design, materials, and procedures of this study. First, learners practiced with single sentences, typed the translation of target words, and saw the word translations as feedback to their responses. These are characteristics of intentional vocabulary practice. An interesting avenue for future studies would be to test whether reducing contextual information can also be used to trigger retrieval and enhance word retention in more incidental learning situations, such as during the study of text passages or free reading. It is unclear whether retrieval can be triggered in the same way in these situations. For instance, learners pay more attention to novel words in sentences than in passages (Wochna & Juhasz, 2013). Moreover, learners regularly ignore novel words during free reading (e.g., Hulstijn et al., 1996). On the other hand, an overt response may not be necessary to obtain benefits of retrieval. Covert retrieval—thinking of an answer but not providing an overt response—produces similar benefits for retention as overt retrieval (Smith, Roediger, & Karpicke, 2013). Therefore, it would be interesting to see if reading materials for language learners, such as short texts in handbooks or guided readers, could also be adapted to elicit the retrieval of target word meaning. Feedback could be realized through glossaries or by providing contextual information a few sentences after the retrieval cue, similar to the combined retrieval-plus-context condition in Experiment 2.

Second, in this study, learners were exposed to L2 words in a L1 context, which allowed us to manipulate contextual richness while ensuring that all target words were unknown and all remaining words were known to the learners. This manipulation, however, may have had a benefit especially for the context-inference condition because it made it more likely that learners understood the contextual clues. Although it is unlikely that retrieval benefits were due to the choice of language in the present experiments, a text in the target language may be useful for L2 learners to also strengthen their knowledge of the words that constitute the context, in addition to the specific experimental target words.

It is therefore a relevant question to ask if the effect of contextual richness found here would be comparable in a situation when learners read texts in the target L2. Moreover, it remains to be tested whether contextual richness has the same effect when learners try to acquire conceptually complex words. Studies on L1 word learning suggest that, in this case, more presentations in an informative context might be beneficial—at least until comprehension has been achieved (Frishkoff, Perfetti, & Collins-Thompson, 2010; Frishkoff et al., 2011).

Finally, we focused on words presented in either informative or neutral, uninformative contexts to isolate the effect of memory retrieval from the effect of context inferencing. In reality, the context surrounding a word falls on a continuum from defining, to uninformative, to misleading (see also Webb, 2008). This raises additional questions, for example, as to whether retrieval is also beneficial if it is elicited in a distracting or irrelevant context and whether retrieval from an uninformative context is beneficial, compared to decontextualized word practice or compared to more effortful context inferences. Some researchers have argued that the effort involved in inferencing may increase deeper processing and lead to greater retention (e.g., Haastrup, 1989, as cited in Nation, 2001; Hu & Nassaji, 2012). A related point is that deeper or more beneficial processing may occur if word meaning is inferred from different context sentences instead of the same context repeatedly. Encoding variability is thought to enhance memory by creating additional associations that enrich memory representations and make reactivation easier (Benjamin & Tullis, 2010), and context variability has been specifically shown to be beneficial for learning word meaning (e.g., Bolger, Balass, Landen, & Perfetti, 2008). On the other hand, varying inference contexts may further draw participants' attention to comprehension instead of novel word forms and therefore may lead to weaker form–meaning associations. These issues need to be addressed in future research.

Conclusion

The present study focused on the influence of contextual richness on word learning. Three experiments showed that practice with newly learned words in an uninformative context that required memory retrieval improved word retention, compared to context-inference practice in an informative context. These testing effects were obtained using different outcome measures, such as recall of word forms and meanings as well as recognition of words in context, both immediately and 7 days after learning. Reducing contextual information—particularly after initial encoding of novel word forms—creates desirable difficulties during

vocabulary practice and functions as a trigger for memory retrieval, leading to enhanced long-term retention of novel L2 words.

Final revised version accepted 9 November 2017

Notes

1 There are also a limited number of correlational studies that describe the relation between the informativeness of context and word retention from reading specific texts. Zahar, Cobb, and Spada (2001), for example, analyzed which target words most readers of a text did or did not learn from reading and found no difference in the informativeness of the context that surrounded acquired and nonacquired words. These results must be interpreted cautiously, however, because contextual informativeness was not experimentally manipulated.

2 The degree to which learners engage in memory retrieval or contextual inferences may vary on a continuum rather than categorically. The names of the conditions indicate the way in which learners most likely accessed word meaning in the two conditions: The retrieval condition required word retrieval from memory; the context-inference condition facilitated inferences by providing rich contextual information that made memory retrieval unnecessary. However, if learners managed to detect the Swahili word while ignoring the context, they may have also engaged in memory retrieval, to some extent, in the context-inference condition.

3 Mixed logit models were fitted using the glmer function in the lme4 package (Bates et al., 2015) in R (version 3.1.2; R Core Team, 2014), with accuracy on the delayed receptive recall test as a binary outcome variable (correct or incorrect). The model with the best fit as determined by the maximum likelihood criterion was a model with random intercepts for participants and words and a fixed effect of the practice condition. The dummy-coded regression coefficients showed that the odds that receptive recall was successful were significantly higher in the retrieval condition than in the context-inference condition ($OR = 1.58$) and were higher in the retrieval-plus-context condition than in the context-inference condition ($OR = 1.26$). These contrasts replicate the results from the repeated-measures ANOVAs. However, the mixed model also showed a significant difference between the two retrieval conditions: The odds for correct recall were significantly higher in the retrieval condition than in the retrieval-plus-context condition ($OR = 1.25$). Confidence intervals obtained with bootstrapping and p values obtained with Satterthwaite's approximation in lmerTest indicated that this effect was significant at $.01 < p < .05$. This was the only contrast in the mixed model that led to a conclusion different from those based on the ANOVA run with the aggregate data.

4 In contrast, mixed-effects modeling indicated that this effect was significant at $.01 < p < .05$. The odds that confidence was high were significantly higher in the retrieval condition than in the context-inference condition for familiar ($OR = 1.32$) as well as unfamiliar test items ($OR = 1.43$).

5 For further information about the cognitive mechanisms that might underlie benefits of retrieval practice for retention, we refer readers to discussions in the recent literature (e.g., Carpenter & Yeung, 2017; Whiffen & Karpicke, 2017; for overview publications, see Roediger & Butler, 2011, and Rowland, 2014; for information on retrieval-induced suppression in mixed-list designs, see Rowland, Littrell-Baez, Sensenig, & DeLosh, 2014).

References

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Baleghizadeh, S., & Shahry, M. N. N. (2011). The effect of three consecutive context sentences on EFL vocabulary-learning. *TESL Canada Journal*, *28*, 74–89. https://doi.org/10.18806/tesl.v28i2.1073

Barcroft, J. (2002). Semantic and structural elaboration in L2 lexical acquisition. *Language Learning*, *52*, 323–363. https://doi.org/10.1111/0023-8333.00186

Barcroft, J. (2007). Effects of opportunities for word retrieval during second language vocabulary learning. *Language Learning*, *57*, 35–56. https://doi.org/10.1111/j.1467-9922.2007.00398.x

Barcroft, J. (2015). Can retrieval opportunities increase vocabulary learning during reading? *Foreign Language Annals*, *48*, 236–249. https://doi.org/10.1111/flan.12139

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Beck, I. L., McKeown, M. G., & McCaslin, E. S. (1983). Vocabulary development: All contexts are not created equal. *The Elementary School Journal*, *83*, 177–181. https://doi.org/10.1086/461307

Benjamin, A. S., & Tullis, J. (2010). What makes distributed practice effective? *Cognitive Psychology*, *61*, 228–247. https://doi.org/10.1016/j.cogpsych.2010.05.004

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

Bolger, D. J., Balass, M., Landen, E., & Perfetti, C. A. (2008). Context variation and definitions in learning the meanings of words: An instance-based learning approach. *Discourse Processes*, *45*, 122–159. https://doi.org/10.1080/01638530701792826

Carpenter, S. K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin & Review*, *13*, 826–830. https://doi.org/10.3758/BF03194004

Carpenter, S. K., Sachs, R. E., Martin, B., Schmidt, K., & Looft, R. (2012). Learning new vocabulary in German: The effects of inferring word meanings, type of feedback, and time of test. *Psychonomic Bulletin & Review*, *19*, 81–86. https://doi.org/10.3758/s13423-011-0185-7

Carpenter, S. K., & Yeung, K. L. (2017). The role of mediator strength in learning from retrieval. *Journal of Memory and Language*, *92*, 128–141. https://doi.org/10.1016/j.jml.2016.06.008

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, *132*, 354–380. https://doi.org/10.1037/0033-2909.132.3.354

Choi, S., Kim, J., & Ryu, K. (2014). Effects of context on implicit and explicit lexical knowledge: An event-related potential study. *Neuropsychologia*, *63*, 226–234. https://doi.org/10.1016/j.neuropsychologia.2014.09.003

Deconinck, J., Boers, F., & Eyckmans, J. (2015). "Does the form of this word fit its meaning?" The effect of learner-generated mapping elaborations on L2 word recall. *Language Teaching Research*, *21*, 31–53. https://doi.org/10.1177/1362168815614048

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods*, *1*, 170–177. https://doi.org/10.1037/1082-989X.1.2.170

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. https://doi.org/10.1177/1529100612453266

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191.

Finley, J. R., Benjamin, A. S., Hays, M. J., Bjork, R. A., & Kornell, N. (2011). Benefits of accumulating versus diminishing cues in recall. *Journal of Memory and Language*, *64*, 289–298. https://doi.org/10.1016/j.jml.2011.01.006

Folse, K. S. (2006). The effect of type of written exercise on L2 vocabulary retention. *TESOL Quarterly*, *40*, 273–293. https://doi.org/10.2307/40264523

Frishkoff, G. A., Collins-Thompson, K., Hodges, L., & Crossley, S. (2016). Accuracy feedback improves word learning from context: Evidence from a meaning-generation task. *Reading and Writing*, *29*, 609–632. https://doi.org/10.1007/s11145-015-9615-7

Frishkoff, G. A., Perfetti, C. A., & Collins-Thompson, K. (2010). Lexical quality in the brain: ERP evidence for robust word learning from context. *Developmental Neuropsychology*, *35*, 376–403. https://doi.org/10.1080/87565641.2010.480915

Frishkoff, G. A., Perfetti, C. A., & Collins-Thompson, K. (2011). Predicting robust vocabulary growth from measures of incremental learning. *Scientific Studies of Reading*, *15*, 71–91. https://doi.org/10.1080/10888438.2011.539076

Fukkink, R. G., & de Glopper, K. (1998). Effects of instruction in deriving word meaning from context: A meta-analysis. *Review of Educational Research*, *68*, 450–469. https://doi.org/10.2307/1170735

Goossens, N. A. M. C., Camp, G., Verkoeijen, P. P. J. L., & Tabbers, H. K. (2014). The effect of retrieval practice in primary school vocabulary learning. *Applied Cognitive Psychology*, *28*, 135–142. https://doi.org/10.1002/acp.2956

Grace, C. A. (1998). Retention of word meanings inferred from context and sentence-level translations: Implications for the design of beginning-level call software. *The Modern Language Journal*, *82*, 533–544. https://doi.org/10.2307/330223

Haastrup, K. (1991). *Lexical inferencing procedures, or, talking about words: Receptive procedures in foreign language learning with special reference to English*. Tuebingen, Germany: Narr.

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. https://doi.org/10.1037/a0023219

Hu, H. M., & Nassaji, H. (2012). Ease of inferencing, learner inferential strategies, and their relationship with the retention of word meanings inferred from context. *Canadian Modern Language Review*, *68*, 54–77. https://doi.org/10.1353/cml.2011.0036

Hulstijn, J. H. (1992). Retention of inferred and given word meanings: Experiments in incidental vocabulary learning. In P. J. L. Arnaud & H. Béjoint (Eds.), *Vocabulary and applied linguistics* (pp. 113–125). London: Macmillan.

Hulstijn, J. H., Hollander, M., & Greidanus, T. (1996). Incidental vocabulary learning by advanced foreign language students: The influence of marginal glosses, dictionary use, and reoccurrence of unknown words. *The Modern Language Journal*, *80*, 327–339. https://doi.org/10.1111/j.1540-4781.1996.tb01614.x

Hulstijn, J. H., & Laufer, B. (2001). Some empirical evidence for the involvement load hypothesis in vocabulary acquisition. *Language Learning*, *51*, 539–558. https://doi.org/10.1111/0023-8333.00164

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning. *Psychology of Learning and Motivation*, *61*, 237–284. https://doi.org/10.1016/B978-0-12-800283-4.00007-1

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science*, *319*, 966–968. https://doi.org/10.1126/science.1152408

Karpicke, J. D., & Zaromb, F. M. (2010). Retrieval mode distinguishes the testing effect from the generation effect. *Journal of Memory and Language*, *62*, 227–239. https://doi.org/10.1016/j.jml.2009.11.010

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. https://doi.org/10.1016/j.jml.2011.04.002

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 989–998. https://doi.org/10.1037/a0015729

Kuhn, M., & Stahl, S. (1998). Teaching children to learn word meanings from context: A synthesis and some questions. *Journal of Literacy Research*, *30*, 119–138. https://doi.org/10.1080/10862969809547983

Laufer, B., & Hulstijn, J. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics*, *22*, 1–26. https://doi.org/10.1093/applin/22.1.1

Lawson, M. J., & Hogben, D. (1996). The vocabulary-learning strategies of foreign-language students. *Language Learning*, *46*, 101–135. https://doi.org/10.1111/j.1467-1770.1996.tb00642.x

Li, X. (1988). Effects of contextual cues on inferring and remembering meanings of new words. *Applied Linguistics*, *9*, 402–413. https://doi.org/10.1093/applin/9.4.402

Lindsey, R. V., Shroyer, J. D., Pashler, H., & Mozer, M. C. (2014). Improving students' long-term knowledge retention through personalized review. *Psychological Science*, *25*, 639–647. https://doi.org/10.1177/0956797613504302

McNamara, D., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction*, *14*, 1–43. https://doi.org/10.1207/s1532690xci1401_1

Mondria, J.-A. (2003). The effects of inferring, verifying, and memorizing on the retention of L2 word meanings. *Studies in Second Language Acquisition*, *25*, 473–499. https://doi.org/https://doi.org/10.1017/S0272263103000202

Mondria, J.-A., & Wit-de Boer, M. (1991). The effects of contextual richness on the guessability and the retention of words in a foreign language. *Applied Linguistics*, *12*, 249–267. https://doi.org/10.1093/applin/12.3.249

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*, 519–533. https://doi.org/10.1016/S0022-5371(77)80016-9

Nakata, T. (2016). Effects of retrieval formats on second language vocabulary learning. *International Review of Applied Linguistics in Language Teaching*, *54*, 257–289. https://doi.org/10.1515/iral-2015-0022

Nakata, T. (2017). Does repeated practice make perfect? The effects of within-session repeated retrieval on second language vocabulary learning. *Studies in Second Language Acquisition*, *39*, 653–679. https://doi.org/10.1017/S0272263116000280

Nakata, T., & Webb, S. (2016). Does studying vocabulary in smaller sets increase learning? The effects of part and whole learning on second language vocabulary acquisition. *Studies in Second Language Acquisition*, *38*, 523–552. https://doi.org/10.1017/S0272263115000236

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge, UK: Cambridge University Press.

Nation, I. S. P. (2015). Principles guiding vocabulary learning through extensive reading. *Reading in a Foreign Language*, *27*, 136–145.

Nelson, T. O., & Dunlosky, J. (1994). Norms of paired-associate recall during multitrial learning of Swahili-English translation equivalents. *Memory*, *2*, 325–335. https://doi.org/10.1080/09658219408258951

Peters, E., Hulstijn, J. H., Sercu, L., & Lutjeharms, M. (2009). Learning L2 German vocabulary through reading: The effect of three enhancement techniques compared. *Language Learning*, *59*, 113–151. https://doi.org/10.1111/j.1467-9922.2009.00502.x

Pimsleur, P. (1967). A memory schedule. *The Modern Language Journal*, *51*, 73–75. https://doi.org/10.2307/321812

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*, 878–912. https://doi.org/10.1111/lang.12079

Potts, R., & Shanks, D. R. (2014). The benefit of generating errors during learning. *Journal of Experimental Psychology: General*, *143*, 644–667. https://doi.org/10.1037/a0033194

Pressley, M., Levin, J. R., & McDaniel, M. A. (1987). Remembering versus inferring what a word means: Mnemonic and contextual approaches. In M. G. McKeown & M. E. Curtis (Eds.), *The nature of vocabulary acquisition* (pp. 107–127). Hillsdale, NJ: Erlbaum.

Prince, P. (1996). Second language vocabulary learning: The role of context versus translations as a function of proficiency. *The Modern Language Journal*, *80*, 478–493. https://doi.org/10.1111/j.1540-4781.1996.tb05468.x

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing Vienna, Austria. http://www.R-project.org/

Richland, L. E., Kornell, N., & Kao, L. S. (2009). The pretesting effect: Do unsuccessful retrieval attempts enhance learning? *Journal of Experimental Psychology: Applied*, *15*, 243–257. https://doi.org/10.1037/a0016496

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

Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*, 181–210. https://doi.org/10.1111/j.1745-6916.2006.00012.x

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. https://doi.org/10.1037/a0037559

Rowland, C. A., & DeLosh, E. L. (2015). Mnemonic benefits of retrieval practice at short retention intervals. *Memory*, *23*, 403–419. https://doi.org/10.1080/09658211.2014.889710

Rowland, C. A., Littrell-Baez, M. K., Sensenig, A. E., & DeLosh, E. L. (2014). Testing effects in mixed- versus pure-list designs. *Memory & Cognition*, *42*, 912–921. https://doi.org/10.3758/s13421-014-0404-3

Schmitt, N. (2008). Review article: Instructed second language vocabulary learning. *Language Teaching Research*, *12*, 329–363. https://doi.org/10.1177/1362168808089921

Schneider, V. I., Healy, A. F., & Bourne, L. E. (2002). What is learned under difficult conditions is hard to forget: Contextual interference effects in foreign vocabulary acquisition, retention, and transfer. *Journal of Memory and Language*, *46*, 419–440. https://doi.org/10.1006/jmla.2001.2813

Schouten-van Parreren, C. A. (1989). Vocabulary learning through reading: Which conditions should be met when presenting words in texts. *AILA Review*, *6*, 75–85.

Seibert, L. C. (1945). A study on the practice of guessing word meanings from a context. *The Modern Language Journal*, *29*, 296–322. https://doi.org/10.2307/318219

Sense, F., Behrens, F., Meijer, R. R., & van Rijn, H. (2016). An individual's rate of forgetting is stable over time but differs across materials. *Topics in Cognitive Science*, *8*, 305–321. https://doi.org/10.1111/tops.12183

Smith, M. A., Roediger, H. L., & Karpicke, J. D. (2013). Covert retrieval practice benefits retention as much as overt retrieval practice. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 1712–1725. https://doi.org/10.1037/a0033569

Toppino, T. C., & Cohen, M. S. (2009). The testing effect and the retention interval: Questions and answers. *Experimental Psychology*, *56*, 252–257. https://doi.org/10.1027/1618-3169.56.4.252

van den Broek, G. S. E., Segers, E., Takashima, A., & Verhoeven, L. (2014). Do testing effects change over time? Insights from immediate and delayed retrieval speed. *Memory*, *22*, 803–812. https://doi.org/10.1080/09658211.2013.831455

van den Broek, G. S. E., Takashima, A., Segers, E., Fernández, G., & Verhoeven, L. (2013). Neural correlates of testing effects in vocabulary learning. *NeuroImage*, *78*, 94–102. https://doi.org/10.1016/j.neuroimage.2013.03.071

Veltre, M. T., Cho, K. W., & Neely, J. H. (2015). Transfer-appropriate processing in the testing effect. *Memory*, *23*, 1229–1237. https://doi.org/10.1080/09658211.2014.970196

Verspoor, M., & Lowie, W. (2003). Making sense of polysemous words. *Language Learning*, *53*, 547–586. https://doi.org/10.1111/1467-9922.00234

Webb, S. (2007a). Learning word pairs and glossed sentences: The effects of a single context on vocabulary knowledge. *Language Teaching Research*, *11*, 63–81. https://doi.org/10.1177/1362168806072463

Webb, S. (2007b). The effects of repetition on vocabulary knowledge. *Applied Linguistics*, *28*, 46–65. https://doi.org/10.1093/applin/aml048

Webb, S. (2008). The effects of context on incidental vocabulary learning. *Reading in a Foreign Language*, *20*, 232–245.

Whiffen, J. W., & Karpicke, J. D. (2017). The role of episodic context in retrieval practice effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*, 1036–1046. https://doi.org/10.1037/xlm0000379

Winstanley, P. A. D. (1996). Generation effects and the lack thereof: The role of transfer-appropriate processing. *Memory*, *4*, 31–48. https://doi.org/10.1080/741940667

Wochna, K. L., & Juhasz, B. J. (2013). Context length and reading novel words: An eye-movement investigation. *British Journal of Psychology*, *104*, 347–363. https://doi.org/10.1111/j.2044-8295.2012.02127.x

Yan, V. X., Clark, C. M., & Bjork, R. A. (2016). Memory and metamemory considerations in the instruction of human beings revisited: Implications for optimizing online learning. In J. C. Horvath, J. M. Lodge, & J. Hattie (Eds.), *From the laboratory to the classroom: Translating science of learning for teachers* (pp. 61–78). London: Routledge.

Zahar, R., Cobb, T., & Spada, N. (2001). Acquiring vocabulary through reading: Effects of frequency and contextual richness. *Canadian Modern Language Review*, *57*, 541–572. https://doi.org/10.3138/cmlr.57.4.541

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Appendix S1. Pilot Procedure to Construct Practice Sentences.
Appendix S2. Additional Statistical Analyses (Bayesian Models).
Appendix S3. Further Information on Data Reported in Figure 3.
Appendix S4. Analyses of Response Times During Practice.