Applied Cognitive Psychology, Appl. Cognit. Psychol. **30**: 700–712 (2016)
Published online 6 June 2016 in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/acp.3245

Distributed Practice and Retrieval Practice in Primary School Vocabulary Learning: A Multi-classroom Study

NICOLE A. M. C. GOOSSENS[1,2]*, GINO CAMP[2,1], PETER P. J. L. VERKOEIJEN[1,3], HUIB K. TABBERS[1], SAMANTHA BOUWMEESTER[1] and ROLF A. ZWAAN[1]

[1]*Erasmus University Rotterdam, Rotterdam, the Netherlands*
[2]*Open University of the Netherlands, Heerlen, the Netherlands*
[3]*Avans University of Applied Sciences Breda, the Netherlands*

Summary: Distributed practice and retrieval practice are promising learning strategies to use in education. We examined the effects of these strategies in primary school vocabulary lessons. Grades 2, 3, 4, and 6 children performed exercises that were part of the regular curriculum. For the distributed practice manipulation, the children performed six exercises distributed within 1 week (short-lag repetition) or across 2 weeks (long-lag repetition). For the repetition type manipulation, children copied a part of the description of a word (restudy) or recalled the description (retrieval practice). At the end of each week, the children received a cued-recall vocabulary test. After 1 to 11 weeks they received a multiple-choice vocabulary test. Both on the cued-recall test and on the multiple-choice test no benefits of long-lag repetition and retrieval practice were found. These results put into question the practical value of long-lag repetition and retrieval practice in real-life primary school vocabulary lessons. Copyright © 2016 John Wiley & Sons, Ltd.

Vocabulary development is considered to be important for enhancing reading comprehension (e.g., Anderson & Freebody, 1981) as well as for increasing world knowledge (Stahl & Nagy, 2006). The larger one's vocabulary knowledge, the more precisely and complexly one can speak and think about the world (Stahl & Nagy, 2006). For instance, a person who knows the terms *searing*, *stewing*, and *poaching* will think about cooking in a different way (and will maybe even *cook* in a different way) than a person who is limited to *baking*, *boiling*, and *roasting* (after examples of Stahl & Nagy, 2006, p. 5). Although teachers are aware of the importance of vocabulary knowledge in primary schools, some studies have shown that only a small amount of the school time is devoted to vocabulary learning (e.g., Scott, Jamieson-Noel, & Asselin, 2003). Also, the time spent on vocabulary learning is not always used effectively. For example, teachers may instruct their pupils to copy word definitions from dictionaries, even though this has been shown to be ineffective (Scott et al., 2003).

Vocabulary teaching and learning should be guided by well-supported instructional principles, such as pre-teaching the words, using dictionaries and glossaries during reading texts, and using contexts to guess the words in a text (Nation, 2004). Furthermore, a number of memory strategies have consistently been shown to have beneficial effects on the learning of word meanings in the in the psychological laboratory (for an overview see Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013). However, there is a paucity of research that investigates these strategies and their effects on vocabulary learning in the classroom. The aim of the current study was therefore to examine in a classroom setting the effects of two memory strategies that produce robust effects in the psychological laboratory: distributed practice and retrieval practice.

*Correspondence to: Nicole Goossens, Welten Institute, Open University of the Netherlands, P.O. Box 2690, 6401 DL Heerlen, The Netherlands.
E-mail: Nicole.Goossens@ou.nl

Many studies have demonstrated that distributed practice (i.e., spacing learning sessions over time) leads to better retention than massed practice (i.e. cramming learning sessions in immediate succession). This so-called spacing effect has been demonstrated in more than 300 published experiments (for reviews, see e.g., Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Delaney, Verkoeijen, & Spirgel, 2010). Most of these studies were conducted in a laboratory setting with adults learning word-pairs. However, the beneficial effect of distributed practice has also been found in primary school children learning pictures (e.g., Toppino & DiGeorge, 1984; Toppino, Kasserman, & Mracek, 1991), words (e.g., Toppino & DeMesquita, 1984), or a combination of pictures and words (e.g., Cahill & Toppino, 1993; Rea & Modigliani, 1987; Toppino, 1993). Moreover, some studies have demonstrated that distributed practice enhances foreign vocabulary learning in undergraduate students (e.g., Bahrick, Bahrick, Bahrick, & Bahrick, 1993; Bloom & Shuell, 1981).

Only a few studies have demonstrated a benefit of distributed practice in first language vocabulary learning (e.g., Goossens, Camp, Verkoeijen, Tabbers, & Zwaan, 2012; Kornell, 2009; Sobel, Cepeda, & Kapler, 2011). For example, Kornell (2009) had undergraduate students study 40 flashcards, each with an unknown word and its synonym (e.g., *effulgent–brilliant*). Twenty cards were studied via distributed practice, namely, twice in four consecutive sessions, and 20 flashcards were studied via massed practice, namely, eight times in one single session. On the final test a beneficial effect of distributed practice was found. Goossens et al. (2012) had children in Grade 3 learn 30 words by performing three different exercises from regular learning materials of Grade 4. The exercises were presented on the same day (massed practice) or on three consecutive days (distributed practice). On the final cued-recall tests after 1 week and after 5 weeks, the children performed better on the words learned via distributed practice than on the words learned via massed practice. Thus, there is some evidence that distributed practice is effective for children who are learning new vocabulary.

Furthermore, only a few studies have investigated distributed practice with lags of 1 week or more, which is more consistent with how vocabulary exercises are distributed over time in primary school education (e.g., Gerbier, Toppino, & Koenig, 2014; Kang, Lindsey, Mozer, & Pashler, 2014; Küpper-Tetzel, Erdfelder, & Dickhäuser, 2014; Küpper-Tetzel, Kapler, & Wiseheart, 2014). However, the studies using longer time intervals were conducted in the laboratory or in secondary school education; thus, the question remains whether these results can be translated to primary school education. Also, in some studies it was shown that long-lag practice does not always lead to better retention than short-lag practice (e.g., Toppino & Gracen, 1985). Therefore, it is important to investigate distributed practice in primary school vocabulary learning over longer time lags.

The second memory strategy we investigated is retrieval practice, also commonly referred to as *the testing effect*. The testing effect is the phenomenon that retrieval practice leads to better long-term retention than additional study (for reviews see e.g., Rawson & Dunlosky, 2011; Roediger & Butler, 2011; Roediger, Putnam, & Smith, 2011; Roediger & Karpicke, 2006; Rowland, 2014). As in the studies conducted on distributed practice, most experiments on retrieval practice have been conducted using word lists or word pairs (e.g., Bouwmeester & Verkoeijen, 2011; Carpenter, Pashler, & Vul, 2006; Carpenter, Pashler, Wixted, & Vul, 2008; Toppino & Cohen, 2009; Tulving, 1967; Wheeler, Ewers, & Buonanno, 2003) and foreign vocabulary pairs (e.g., Carpenter et al., 2008; Carrier & Pashler, 1992; Fritz, Morris, Acton, Voelkel, & Etkind, 2007; Karpicke, 2009; Karpicke & Roediger, 2008; Pashler, Cepeda, Wixted, & Rohrer, 2005; Pyc & Rawson, 2007, 2009, 2011; Toppino & Cohen, 2009). Some studies have found a beneficial effect of retrieval practice for first language vocabulary learning, with adults (e.g., Cull, 2000; Karpicke & Smith, 2012), as well as with primary school children (e.g., Goossens, Camp, Verkoeijen, & Tabbers, 2014; Goossens, Camp, Verkoeijen, Tabbers, & Zwaan, 2014; Metcalfe, Kornell, & Son, 2007). However, none of these studies has been performed in the classroom using materials from the actual vocabulary curriculum.

To summarize, it has been shown that both distributed practice and retrieval practice can facilitate word learning in adults and in primary school children. In the current study, the research question was whether the use of distributed practice and retrieval practice is beneficial when it is integrated in primary school vocabulary lessons. The goal was to provide an ecologically valid test of the possible benefits of distributed practice and retrieval practice to determine the relevance for application of these learning strategies in the classroom.

To ensure a high level of ecological validity, we took several measures. First, the current study was conducted within real-life classroom vocabulary lessons using words and exercises that were part of the curriculum. This means that the study was entirely integrated within the normal vocabulary lessons.

Second, in many experiments on distributed practice, a distributed condition is compared with a massed condition, in which words are learned in a single learning session (see for example Goossens et al., 2012; Sobel et al., 2011). However, in (Dutch) vocabulary learning curricula, vocabulary exercises are usually not massed in a single learning session, but rather distributed over several learning sessions during the week (e.g., Janssen & Van Ooijen, 2012; Van de Gein, Van de Guchte, & Kouwenberg, 2008). Thus, it is more relevant for educational practice to investigate the benefit of distributed practice by comparing *lags* of different length, rather than to compare distributed practice with massed practice. In the present study, we compared a short-lag condition in which practice was distributed across 1 week (similar to current classroom practice) to a long-lag condition in which practice was distributed across 2 weeks.

Third, in the current study we attempted to make repeated practice more similar to regular educational practices. In previous experiments, participants simply repeatedly studied the same word pairs (see for example Goossens, Camp, Verkoeijen, & Tabbers, 2014). However, in (Dutch) vocabulary learning in primary education, words are not only restudied by simple repetition, but practiced with different types of exercises (e.g., Janssen & Van Ooijen, 2012; Van de Gein et al., 2008). Therefore, we investigated the benefit of distributed practice and retrieval practice in a situation, in which the words were practiced with different exercises from the regular learning material.

Fourth, in contrast to studies performed with only one age group or grade (e.g., Goossens, Camp, Verkoeijen, & Tabbers, 2012; Goossens, Camp, Verkoeijen, Tabbers, & Zwaan, 2014; Sobel et al., 2011), we conducted experiments in Grades 2, 3, 4, and 6 to be able to generalize our results to different age groups.

Fifth, we used multiple-choice tests to measure long-term effects, because in (Dutch) educational practice, multiple-choice tests are most often used to establish vocabulary size (e.g., Van Berkel & Alberts, 2009; Van Berkel & Hilte, 2009; Verhoeven & Vermeer, 1993).

Sixth and finally, the current study combined distributed practice and retrieval practice within a single study. To our knowledge, only a few studies have investigated these strategies simultaneously (e.g., Cull, 2000). Importantly, distributed practice and retrieval practice have not yet been investigated simultaneously in primary school vocabulary learning. It is important to look at the combination of these two strategies, because it mirrors teaching practice: teachers often use distributed practice and retrieval practice together in one lesson.

So, our main interest was to investigate whether the benefits of distributed practice and retrieval practice found in previous studies (e.g., Goossens, Camp, Verkoeijen, & Tabbers, 2014; Sobel et al., 2011) would generalize to actual classroom practice and whether the benefits of these two strategies can be combined.

In short, in the current experiment we investigated the effects of distributed practice and retrieval practice in primary school children in Grades 2, 3, 4, and 6. The experiment was integrated in the regular vocabulary-learning curriculum. According to the literature, distributed practice and retrieval practice are strong and robust memory strategies that should generalize to an ecologically valid learning situation (e.g., Dunlosky et al., 2013). However, lag effects are in general smaller than distributed practice effects, and sometimes lag effects are not found at all (e.g., Toppino & Gracen, 1985). Therefore, we expected to find a relatively small main effect

of distributed practice when a long-lag condition is compared with a short-lag condition. Thus, we expected that long lags would benefit primary school vocabulary learning more than short lags and that the addition of retrieval practice exercises would benefit primary school vocabulary learning more than additional restudy exercises.

METHOD

Participants

Children of the Dutch Grades 4, 5, 6, and 8, equivalent to Grades 2, 3, 4, and 6 in the USA, were all recruited from the same school in the Netherlands. Because of some extracurricular activities in Grade 5 (according to US grades), the children of that grade were not able to participate in the experiment.

This study is the first to investigate the effect of study type (retrieval practice versus restudy) and repetition (short lag versus long lag) and their interaction on vocabulary learning in an ecologically valid learning situations with ecologically valid learning tasks using sample from different primary school grades. Therefore, we could not formulate strong a priori estimates of the effect sizes for the individual factors and their interaction within the different grades. This, in turn, prevented us from performing meaningful power calculations. To remedy this problem, we aimed at testing as many children as possible. Moreover, to provide researchers, who might want to follow up on our results, with anchors for a power calculation, we will report 95% confidence intervals of relevant effects, which encompass not only the magnitude of the effect but also the precision of the estimate (Cumming, 2014).

The participating classrooms (two or three for each grade) contained a total number of 237 children: 60 children in Grade 2, 55 children in Grade 3, 53 children in Grade 4, and 69 children in Grade 6. For 36 children, parents gave no consent to use the children's data. Another 69 children were not able to participate during all sessions, and the data of three other children could not be used because of procedural errors. This resulted in a final sample of 129 participants (64 boys, 65 girls), with a mean age of 9.90 years ($SD = 1.67$). The group consisted of 33 participants in Grade 2 (15 boys) with a mean age of 7.93 years (range 7.42 to 8.50, $SD = 0.29$), 32 participants in Grade 3 (15 boys) with a mean age of 9.26 years (range 8.25 to 11.17, $SD = 0.61$), 31 participants in Grade 4 (15 boys) with a mean age of 10.14 years (range 9.42 to 11.42, $SD = 0.51$), and 33 participants in Grade 6 (19 boys) with a mean age of 12.27 years (range 11.67 to 13.42, $SD = 0.48$). In total 125 children took the multiple-choice test, namely, 31 children of Grade 2, 32 children of Grade 3, 30 children of Grade 4, and 32 children of Grade 6. The children knew they participated in an experiment, and their parents had given informed consent for using the children's data. In the sample there were no children with established learning problems.

Design

We used for four grades a 2×2 design, in which both distribution of learning (short lag versus long–lag) and repetition type (restudy versus retrieval practice) was manipulated within-subjects. Distribution of learning was manipulated by distributing the exercises of the same words over two sessions in 1 week (short-lag condition) or over four sessions in 2 weeks (long-lag condition). That is, in the long-lag condition half of the exercises were performed in the first week, and the other half of the exercises were performed in the second week, while in the short-lag condition all exercises were performed in a single week. Repetition type was manipulated by providing two additional exercises for each word that required either copying a part of the word description (restudy condition) or recalling the word description (retrieval practice condition). This resulted in four learning conditions: (i) short-lag restudy, (ii) short-lag retrieval practice, (iii) long-lag restudy, and (iv) long-lag retrieval practice.

For each grade, the 40 stimulus words were split into four lists of 10 words each, in which five easy and five difficult words were included (the qualification of the words in 'easy' and 'difficult' was provided by the authors of the textbook). These four lists were assigned to the four learning conditions (short-lag restudy, short-lag retrieval practice, long-lag restudy, and long-lag retrieval practice) by counterbalancing. For practical purposes, we used only four different counterbalancing versions. We used two counterbalancing versions for the distributed practice manipulation and two counterbalancing versions for the repetition type manipulation. At the final tests we used another list order than the order, in which the words were presented in the textbook. At the multiple-choice test, including all words, the lists were mixed, but we used six versions with different list orders to prevent copying.

There were two dependent variables. The first dependent variable was the performance on a cued-recall test given 1, 2, or 3 days after the final learning session, in which the children were given the vocabulary words and for each word they had to write down the correct description. The second dependent variable was the performance on a multiple-choice test, in which the children had to choose for each vocabulary word the correct description out of four options. This test was intended to measure long-term retention and was given 1 week (for Grade 6), 2 weeks (for Grade 3), or 11 weeks (for Grades 2 and 4) after the last learning session. For the children of Grade 3 and Grade 6, the multiple-choice test session came relatively shortly after the learning phase because of the fact that the learning phase for these grades was at the end of the school year, whereas for the children of Grades 2 and 4 the learning phase occurred earlier in the school year.

Materials

For each grade, 40 words and their exercises were selected from the regular vocabulary learning materials that were currently used in that grade within the school. These 40 words were presented in thematic lists of five words and were introduced in blocks of 20 words. Each word was introduced in the vocabulary-learning book by means of an illustration, a definition and a context sentence. For each word we selected four different exercises from the textbook and the workbook. For a small number of words, we found fewer than four

exercises in the vocabulary learning materials, so in these cases we constructed similar exercises ourselves.

Tables 1, 2, 3, and 4 show the Dutch words, their English translations, and their characteristics (their word type, number of letters, and word frequency) used in Grades 2, 3, 4, and 6, respectively. Table 5 shows the characteristics of the four vocabulary lists of 40 words. The median word frequency is based on the Dutch Measure of Lexical Richness for primary school materials (Schrooten & Vermeer, 1994).

The words were first described in one or two sentences (e.g., *A pedestrian crossing—A place to cross the street safely. There are white stripes on the road.*). There were different types of textbook exercises: questions in which the word had to be associated with the correct picture, fill-in-the-blank questions (e.g., *We cross the street on ...*), or questions in which the words had to be related to (a part of) their description (e.g., *White stripes on the road*).

We also constructed additional exercises in which repetition of learning was manipulated. In the restudy exercises children had to copy a part of the description (e.g., *A pedestrian crossing is a place to cross the street safely. There are white stripes on the road.—A pedestrian crossing is a place to cross the street safely. There are white ...*). In the retrieval practice exercises the children had to write down the description themselves (e.g., *A pedestrian crossing—...*).

The cued-recall test consisted of 40 questions, in which the children had to give the description of the word (e.g., *A pedestrian crossing—...*). The children could receive a maximum of two points per correctly answered question. They received 0 points (e.g., *a pedestrian crossing is white*), either 1 point (e.g., *a pedestrian crossing is on the street*) or 2 points (e.g., *a pedestrian crossing is a place to cross the street safely*) for an answer. The multiple-choice test consisted of 40 questions, in which the children had to choose the correct description of the word. Some of the questions asked what was important about the word, some questions asked which item was not related to the word, and other questions asked which one or three items were most strongly related to a word (e.g., *Which words are most related to the*

Table 1. All Dutch words used in Grade 2 with English translations and their characteristics

Dutch word	English translation	Word type	Letters	Frequency
oversteken	to cross over	verb	10	89
fietspad	cycle track	noun	8	18
stoep	sidewalk	noun	5	159
zebrapad	pedestrian crossing	noun	8	6
papierbak	waste-paper basket	noun	9	2
in de buurt	nearby	adverb	9	196
veilig	safe	adjective	6	136
omweg	detour	noun	5	8
plantsoen	public garden	noun	9	13
station	station	noun	7	66
de weg vragen	to ask for directions	verb	11	-
verdwalen	to get lost	verb	9	65
brug	bridge	noun	4	89
kruispunt	intersection	noun	9	12
bushalte	bus stop	noun	8	6
in de buurt blijven	to stay close	verb	16	196
openbaar vervoer	public transport	noun	15	-
verkeersdrempel	speed bump	noun	15	-
wegwijzer	sign post	noun	9	19
oprit	driveway	noun	5	2
grond	ground	noun	5	21
ijskoud	ice-cold	adjective	7	23
luchtje	odor	noun	7	23
vleugel	wing	noun	7	122
meer	lake	noun	4	152
fris	fresh	adjective	4	14
kenmerk	characteristic	noun	7	2
natuur	nature	noun	6	172
omgeving	environment	noun	8	48
een onderzoek doen	to investigate	verb	16	47
rommel	mess	noun	6	89
spons	sponge	noun	5	23
vuilniszak	garbage bag	noun	10	40
vuilnisman	garbage collector	noun	10	9
vuilniswagen	garbage truck	noun	12	19
afval	garbage	noun	5	78
een kijkje nemen	to take a look	verb	14	16
proefje	experiment	noun	7	38
temperatuur	temperature	noun	11	54
zo goed mogelijk	as good as possible	adverb	14	-

Note: The English translations can deviate from the original Dutch meaning.

Table 2. All Dutch words used in Grade 3 with English translations and their characteristics

Dutch word	English translation	Word type	Letters	Frequency
fototoestel	camera	noun	11	18
fotograaf	photographer	noun	9	28
muzikant	musician	noun	8	33
kunstschilder	painter	noun	13	11
gitaar	guitar	noun	6	34
acrobaat	acrobat	noun	8	16
publiek	audience	noun	7	86
dwarsfluit	German flute	noun	10	17
applaus	applause	noun	7	20
model	model	noun	5	1
tentoonstelling	exhibition	noun	15	28
viool	violin	noun	5	78
piano	piano	noun	5	56
zanger	singer	noun	6	17
danseres	female dancer	noun	8	13
bibliotheek	library	noun	11	31
reiziger	traveler	noun	8	21
taxi	taxi	noun	4	18
optreden	performance	noun	8	1
taalprobleem	language problem	noun	12	-
kwaken	to croak	verb	6	44
in de hoek zetten	to be put in the corner of the classroom for punishment	verb	14	-
paling	eel	noun	6	12
met mes en vork eten	to eat with a knife and fork	verb	16	-
verschrikkelijk	terrible	adjective	15	85
grazen	to graze	verb	6	27
dobberen	to float	verb	8	17
rakker	rascal	noun	6	3
ruif	rack	noun	4	2
de trom slaan	to beat the drum	verb	11	-
japanner	Japanese	noun	8	3
chinees	Chinese	noun	7	30
zweed	Swede	noun	5	-
marokkaan	Moroccan	noun	9	15
surinamer	Surinamer	noun	9	11
cello	cello	noun	5	1
contrabas	double-bass	noun	9	-
saxofoon	saxophone	noun	8	3
keyboard	keyboard	noun	8	9
harp	harp	noun	4	-

Note: The English translations can deviate from the original Dutch meaning.

meaning of ***a pedestrian crossing****? A. car—bike—pedestrian; B. road—to walk—zebra; C. white stripes—to walk—pedestrian; D. zebra—white stripes—to cycle*[1]). These questions and the distractors of these multiple-choice items were based on items from a vocabulary test used in primary schools in the Netherlands (e.g., Van Berkel et al., 2010).

Procedure

The experiment took place in the classroom and was the same for every grade. The learning phase lasted 4 weeks and was divided into two blocks of 2 weeks that were procedurally identical. In each week there was a learning session on Monday, another learning session on Tuesday, Wednesday, or Thursday, and a test session on Friday.[2] A number of weeks after the learning phase, a final test phase took place, in which the multiple choice test was administered. For an overview of the complete procedure, we refer to Table 6.

Procedure in week 1

For the procedure in the first week see the first row in Table 6. In the first learning session of the first week, usually on Mondays, the experimenter (the first author) showed an illustration from the textbook in which 20 words were presented, of which 10 words were assigned to the short-lag conditions and 10 words to the long-lag conditions. Next to the illustration, the 20 words and their definitions were given. At first, the children viewed the illustration and talked with the experimenter about what was shown in the pictures. Then, the experimenter explained each word that was illustrated, by giving a definition, a context sentence for the word, and some additional information about the meaning of the word. After all words were explained, the children had to perform two rounds of exercises. In the first round, the children had to

[1] The word 'pedestrian crossing' in Dutch is 'zebrapad', resulting in the connotation of 'zebra' with ' pedestrian crossing' in Dutch.
[2] For Grade 2 and Grade 4 and for some children of Grade 6 in 1 week the first learning session took place on Tuesday instead of on Monday, and for Grade 2 and 4 in 1 week the test session took place on Thursday instead of on Friday.

Table 3. All Dutch words used in Grade 4 with English translations and their characteristics

Dutch word	English translation	Word type	Letters	Frequency
slopen	to demolish	verb	6	7
tractor	tractor	noun	7	17
kiepauto	dump truck	noun	8	1
hekwerk	fencing	noun	7	-
bouwmaterialen	construction materials	noun	14	-
riolering	sewerage	noun	9	1
moker	sledgehammer	noun	5	-
het puin afvoeren	to transport rubble	verb	15	5
sloper	demolisher	noun	6	4
wals	roller	noun	4	4
dertien hoog	on the 14th floor	adverb	11	-
twee onder één kap	semi-detached house	noun	15	-
woonboot	houseboat	noun	8	6
beton	concrete	noun	5	20
balkon	balcony	noun	6	41
landhuis	country cottage	noun	8	2
villa	villa	noun	5	23
oprijlaan	entranceway	noun	9	6
kraakpand	squat	noun	9	-
bungalow	bungalow	noun	8	9
herstellen	to mend	verb	10	6
helaas	unfortunately	adverb	6	46
verstoren	to disturb	verb	9	4
verspillen	to waste	verb	10	6
afhankelijk	dependent	adjective	11	11
elektriciteitscentrale	power station	noun	22	6
grondstof	raw material	noun	9	29
aantasten	to affect	verb	9	10
recycling	recycling	noun	9	-
aardgas	natural gas	noun	7	14
stiltegebied	quiet area	noun	12	10
natuurreservaat	nature reserve	noun	15	14
wild	wildlife	noun	4	21
hei	heather	noun	3	14
moeras	swamp	noun	6	68
natuurbehoud	nature conservation	noun	12	1
fazant	pheasant	noun	6	7
natuurlandschap	nature landscape	noun	15	17
zandverstuiving	sand drifts	noun	15	5
poel	pool	noun	4	4

Note: The English translations can deviate from the original Dutch meaning.

do Exercise 1a for five short-lag words and for five long-lag words. After having completed the first round, the children checked their performance with an answer sheet and corrected wrong answers. Then, the children continued with the second round of exercises, doing Exercise 2a for the same five short-lag words, and doing Exercise 1b for the other five long-lag words. Again, after the second round, the children checked their performance with an answer sheet.

In the second learning session of the week, usually on Wednesdays, the children started with two rounds of repetition exercises (A and B) on the words already practiced in the first learning session. In round A, five words were practiced by restudy (two or three short-lag words and two or three long-lag words), and five words were practiced by retrieval practice (two or three short-lag words and two or three long-lag words). Round B was identical to round A, except that the same five short-lag words, and the other five long-lag words were practiced. After the two rounds of repetition exercises, the children again performed two rounds of regular textbook exercises. As in the first learning session, the children practiced the same list of five short-lag words with two different exercises (Exercise 3a and Exercise 4a), and the two different lists of five long-lag words with the same exercise (Exercise 2a and Exercise 2b). Again, the children used an answer sheet to check their performance. Thus, at the end of the second learning session, the children had performed two repetition exercises and four regular textbook exercises on 5 of the 10 short-lag words and one repetition exercise and two regular textbook exercises on all 10 long-lag words.

In the test session at the end of the first week, usually on Friday, the children were tested on the five short-lag words that had been practiced during the week. After the children finished the test, they could continue with their own schoolwork.

Procedure after week 1

For the procedure in the second, third, and fourth weeks, see respectively the second, the third, and the fourth rows in Table 6. The procedure of the second week was the same

Table 4. All Dutch words used in Grade 6 with English translations and their characteristics

Dutch word	English translation	Word type	Letters	Frequency
redactie	editorial staff	noun	8	5
permanent	permanent	adjective	9	-
herinnering	memory	noun	11	24
opstellen	to draw up	verb	9	16
anoniem	anonymous	adjective	7	-
coördineren	to coordinate	verb	11	-
inspiratie	inspiration	noun	10	-
kopij	copy	noun	5	4
rode draad	theme	noun	9	-
aandenken	keepsake	noun	9	1
tenzij	unless	adverb	6	7
toekomen aan iets	to get round to something	verb	15	1
principe	principle	noun	8	1
vermenigvuldigen	to multiply	verb	16	1
tijdelijk	temporary	adjective	9	6
in je achterhoofd houden	to keep in mind	verb	21	-
weemoed	melancholy	noun	7	2
deadline	deadline	noun	8	-
overzichtelijk	well-ordered	adjective	14	4
uitgave	publication	noun	7	3
faalangst	fear of failure	noun	9	-
planning	schedule	noun	8	-
voortzetten	to continue	verb	11	8
administratie	administration	noun	13	1
kaften	to cover a book	verb	6	1
spieken	to peek	verb	7	4
pressen	to press	verb	7	-
tussenuur	odd hour	noun	9	-
voortijdig	premature	adjective	10	-
raadplegen	to consult	verb	10	2
stranden	to get bogged down	verb	8	4
havo	higher general secondary education	noun	4	20
vmbo	preparatory secondary vocational education	noun	4	-
vwo	pre-university education	noun	3	5
mentor	tutor	noun	6	-
afhaken	to drop out	verb	7	-
falen	to fail	verb	5	1
registreren	to register	verb	11	-
vakkenpakket	curriculum	noun	12	-
benjamin	Benjamin	noun	8	-

Note: The English translations can deviate from the original Dutch meaning.

Table 5. Characteristics of the vocabulary lists used in Grades 2, 3, 4, and 6

Grade	Noun	Verb	Adj.	Adv.	Letters	MLR words	Frequency[a]
2	29	6	3	2	8.55 (3.43)	36	35.5 (range 2 to 196)
3	33	6	1	0	8.25 (3.17)	33	17 (range 1 to 86)
4	31	6	1	2	8.98 (4.02)	33	7 (range 1 to 68)
6	20	14	5	1	8.93 (3.48)	22	4 (range 1 to 24)

The number of nouns, verbs, adjectives, and adverbs; the average number of letters in the words (SD in parentheses); the number of words that were in the Measure of Lexical Richness list (out of 40); and the median frequency of the words.
[a]Frequency based on the Dutch Measure of Lexical Richness for primary school materials (Schrooten & Vermeer, 1994).

as the first week, with a few exceptions. First, there was no introduction during the first learning session; the children immediately started with the exercises. Second, the children practiced with five new short-lag words and with the 10 long-lag words already practiced during the first week. Third, in the test session the children were tested on both the five new short-lag words, as well as on the 10 long-lag words. At the end of the second week, the children had performed in total four regular textbook exercises and two repetition exercises for both the 10 long-lag words as well as for the 5 new short-lag words.

The procedure of the third and fourth weeks was identical to the procedure in the first and second week. In the third and fourth weeks 20 new vocabulary words were learned.

A final test phase took place 1 week (Grade 6), 2 weeks (Grade 3), or 11 weeks (Grades 2 and 4) after the learning phase. In the final test phase, the children were tested again on all 40 words with the multiple-choice test.

Table 6. Procedure of the experiment for one of the four counterbalancing versions

Week no.	Session 1 Short-lag items	Session 1 Long-lag items	Session 2 Short-lag items	Session 2 Long-lag items	Session 3
1	*Introduction* 1–10	*Introduction* 11–20	*Repetition Exercises* Round A: 1–2–3–4–5 Round B: 1–2–3–4–5	*Repetition Exercises* Round A: 11–12–13–14–15 Round B: 16–17–18–19–20	*Tests* 1–5
	Textbook exercises		*Textbook exercises*		
	1a: 1–5	1a: 11–15	3a: 1–5	2a: 11–15	
	2a: 1–5	1b: 16–20	4a: 1–5	2b: 16–20	
2			*Repetition exercises* Round A: 6–7–8–9–10 Round B: 6–7–8–9–10	*Repetition exercises* Round A: 11–12–13–14–15 Round B: 16–17-18–19–20	*Test* 6–20
	Textbook exercises		*Textbook exercises*		
	1b: 6–10	3a: 11–15	3b: 6–10	4a: 11–15	
	2b: 6–10	3b: 16–20	4b: 6–10	4b: 16–20	
3	*Introduction* 21–30	*Introduction* 31–40	*Repetition exercises* Round A: 21–22–23–24–25 Round B: 21–22–23–24–25	*Repetition exercises* Round A: 31–32–33–34–35 Round B: 36–37–38–39–40	*Test* 21–25
	Textbook exercises		*Textbook exercises*		
	1c: 21–25	1c: 31–35	3c: 21–25	2c: 31–35	
	2c: 21–25	1d: 36–40	4c: 21–25	2d: 36–40	
4			*Repetition exercises* Round A: 26–27–28–29–30 Round B: 26–27–28–29–30	*Repetition exercises* Round A: 31–32–33–34–35 Round B: 36–37–38–39–40	*Tests* 26–40
	Textbook exercises		*Textbook exercises*		
	1d: 26–30	3c: 31–35	3d: 26–30	4c: 31–35	
	2d: 26–30	3d: 36–40	4d: 26–30	4d: 36–40	
5 – 15					*MC Test* 1–40

Every row represents a week. Every column represents a learning session in which short-lag and long-lag items were learned. In Session 2, half of the Repetition Exercises A and B were restudy exercises and half of them were retrieval practice exercise.

Data analysis

For the final cued-recall questions the children could receive a maximum of two points per correct answer. For the final multiple-choice questions, the children received one point per correct answer. Hence, for the cued-recall test, the scoring range was 0–20 points and for the multiple-choice test, it was 0–10 points.

One rater scored all cued-recall tests, and another independent rater scored 20% of the cued-recall tests, by using the same scoring form, to check the interrater reliability. Weighted Kappa between the two raters was .615, indicating a relatively low interrater reliability.[3]

RESULTS

Retrieval practice during the learning phase

Before we report the main analyses, we will first explore the data from the retrieval practice exercises during the learning phase (see Table 6 for the procedure). In exploring these data we had to consider that there were differences for the short-lag and the long-lag condition in the number of textbook exercises the children had performed before they did the two retrieval practice exercises. For the short-lag words, the children completed two textbook exercises before they did the first and second retrieval practice exercise. For the long-lag words, the children completed one textbook exercise before they did the first retrieval practice exercise, and they completed three textbook exercises before they did the second retrieval practice exercise. The mean scores on these retrieval practice exercises are presented in Table 7.

The overall proportion correct on the retrieval practice exercises was .57 ($SE = 0.01$). In the first retrieval practice exercise, average performance was .56 ($SE = 0.01$), whereas in the second retrieval practice exercise, average performance was .58 ($SE = 0.01$). The mean proportion correct on retrieval practice exercises for the short-lag words was .57 ($SE = 0.02$), and on retrieval exercises for the long-lag words it was .57 as well ($SE = 0.01$).

Performance on the cued-recall and multiple-choice tests

We calculated the mean proportion correct scores and the standard deviations for the four conditions (i.e., short-lag restudy, short-lag retrieval practice, long-lag restudy, and long-lag retrieval practice) on the cued-recall test (Table 8) and on the multiple-choice test (Table 9).

Before we will analyze performance for each grade separately, we will first examine overall performance on the cued-recall and multiple-choice tests. Table 8 shows that the overall mean proportion score on the cued-recall test was .59 correct ($SE = 0.01$). The overall mean proportion correct score for the short-lag condition was .61 ($SE = 0.01$), and for the long-lag condition it was .57 ($SE = 0.01$). The overall mean proportion correct score for restudy was .61 ($SE = 0.01$), and for retrieval practice it was .57 ($SE = 0.01$). Overall, the average proportion correct score on the cued-recall test was quite similar across the different grade levels.

[3] We did not incldue an estimate of the reliability of the multiple-choice test. The reason is that this estimate would be very imprecise, because it would be based on only 33 or fewer children for each grade, resulting in very wide confidence intervals. Therefore, this estimate would not be very informative about the reliability of the multiple-choice test.

Table 7. Proportion correct on the words in the first and second retrieval practice exercise for the four grades and the children of the four grades together (*SE* in parentheses)

Grade	Short-lag first exercise	Short-lag second exercise	Long-lag first exercise	Long-lag second exercise
2	.52 (.03)	.48 (.03)	.48 (.03)	.52 (.04)
3	.65 (.03)	.64 (.03)	.63 (.03)	.65 (.02)
4	.62 (.03)	.63 (.03)	.59 (.03)	.61 (.03)
6	.52 (.03)	.54 (.03)	.52 (.02)	.57 (.02)
All	.57 (.02)	.57 (.02)	.55 (.01)	.58 (.02)

Table 8. Proportion correct on the cued-recall test in the four learning conditions for the four grades and the children of the four grades together (*SE* in parentheses)

Grade	Short-lag restudy	Short-lag retrieval practice	Long-lag restudy	Long-lag retrieval practice
2	.63 (.04)	.57 (.04)	.57 (.04)	.52 (.04)
3	.67 (.03)	.62 (.02)	.66 (.03)	.61 (.02)
4	.65 (.03)	.65 (.03)	.60 (.03)	.52 (.03)
6	.57 (.03)	.53 (.03)	.56 (.03)	.53 (.03)
All	.63 (.02)	.59 (.02)	.60 (.02)	.55 (.02)

Table 9. Proportion correct on the multiple-choice test in the four learning conditions for the four grades and the children of the four grades together (*SE* in parentheses)

Grade	Short-lag restudy	Short-lag retrieval practice	Long-lag restudy	Long-lag retrieval practice
2	.71 (.04)	.72 (.04)	.73 (.03)	.73 (.03)
3	.80 (.02)	.76 (.02)	.80 (02)	.75 (.03)
4	.76 (.04)	.76 (.04)	.74 (.04)	.72 (.04)
6	.70 (.03)	.66 (.03)	.69 (.03)	.73 (.03)
All	.74 (.02)	.72 (.02)	.74 (.02)	.73 (.02)

Table 9 shows that the overall mean proportion correct score on the multiple-choice test was .73 ($SE = 0.01$). The overall mean proportion correct score for the short-lag condition was .73 ($SE = 0.01$), and for the long-lag condition it was .74 ($SE = .01$). Also, the overall mean proportion correct score for the restudy condition was .74 ($SE = 0.01$), and for the retrieval practice condition it was .73 ($SE = 0.01$). Overall, the average proportion correct score on the multiple-choice test was quite similar across the different grades.

We analyzed the scores on the final tests for each grade separately, for several reasons. First, the retention interval between the last learning session and the multiple-choice test was different for the four grades. Second, in Grades 3 and 6 the experiment was conducted at the end of the school year, in contrast to in Grades 2 and 4. Finally, the words learned in the different grades were qualitatively different.

Analysis of the scores on the cued-recall test for each grade

We analyzed the scores on the cued-recall tests using a 2×2 repeated measures ANOVA with distribution of learning (lag) and repetition of learning (repetition) as within-subject factors. We did this for each grade separately. The assumptions of these 2×2 repeated measures ANOVAs were not violated.[4]

[4] For some analyses the Kolmogorov–Smirnov test and the Shapiro–Wilk test showed significant results, but in these analyses the central limit theorem applies because the sample sizes were above 30. 'This theorem states that the sampling distribution will take the shape of a normal distribution regardless of the shape of the population from which the sample was drawn' (Field, 2009, p. 782).

For Grade 2, mean cued recall performance was significantly better in the short-lag condition than in the long-lag condition [$M_D = 0.06$, $SE = 0.02$, 95% CI (0.024, 0.088)], $F(1, 32) = 13.01$, $p = .001$, $\eta^2 = .07$. The restudy condition outperformed the retrieval practice condition [$M_D = 0.05$, $SE = 0.03$, 95% CI (0.000, 0.103)], but the effect was only marginally significant, $F(1, 32) = 40.08$, $p = .052$, $\eta^2 = .06$. There was no significant interaction between lag and repetition ($F < 1$), $\eta^2 = .00$.

For Grade 3, there was no significant difference between the short-lag condition and the long-lag condition on cued recall [$M_D = 0.01$, $SE = 0.01$, 95% CI (-0.017, 0.036)], $F < 1$, $\eta^2 = .00$. There was a significant benefit of restudy over retrieval practice [$M_D = 0.05$, $SE = 0.02$, 95% CI (0.014, 0.076)], $F(1, 31) = 8.93$, $p = .005$, $\eta^2 = .09$. Furthermore, there was no significant interaction between lag and repetition ($F < 1$), $\eta^2 = .00$.

For Grade 4, mean cued recall performance was significantly better in the short-lag condition than in the long-lag condition [$M_D = 0.09$, $SE = 0.02$, 95% CI (0.058, 0.125)], $F(1, 30) = 30.81$, $p < .001$, $\eta^2 = .23$. There was no significant difference between the restudy condition and the retrieval practice condition [$M_D = 0.04$, $SE = 0.02$, 95% CI (-0.009, 0.081)], $F(1, 30) = 2.72$, $p = .109$, $\eta^2 = .04$. Furthermore, there was a significant interaction between lag and repetition, indicating that the benefit of the short-lag condition over the long-lag condition was larger in the retrieval practice condition than in the restudy condition [$M_D = 0.08$, $SE = 0.03$, 95% CI (0.020, .132)], $F(1, 30) = 7.70$, $p = .009$, $\eta^2 = .05$.

For Grade 6, there was no significant difference between the short-lag condition and the long-lag condition [$M_D = 0.00$, $SE = 0.02$, 95% CI (−0.033, 0.036)], $F < 1$, $\eta^2 = .00$. Also, there was no significant difference between the restudy condition and the retrieval practice condition [$M_D = 0.03$, $SE = 0.02$, 95% CI (−0.008, 0.075)], $F(1, 32) = 2.65$, $p = .114$, $\eta^2 = .03$. There was no significant interaction between lag and repetition ($F < 1$) $\eta^2 = .00$.

To sum up, there were almost no significant effects of lag and repetition, and the effects found were not in the expected direction. In Grade 2 and in Grade 4 there were benefits of the short-lag condition. Also, in Grade 3 there was a benefit of the restudy condition.

Analysis of the scores on the multiple-choice test for each grade separately

We analyzed the scores on the multiple-choice test using a 2 × 2 repeated measures ANOVA with distribution of learning (lag) and repetition of learning (repetition) as within-subject factors. We did this for each grade separately. The assumptions of these 2 × 2 repeated measures ANOVAs were not violated.[5]

For Grade 2, there was no significant difference between the short-lag condition and the long-lag condition [$M_D = 0.02$, 95% CI (−0.07, 0.03)], $F < 1$, $\eta^2 = .01$. Also, there was no significant difference between the restudy condition and the retrieval practice condition [$M_D = 0.01$, 95% CI (−0.04, 0.03)], $F < 1$, $\eta^2 = .00$. There was no significant interaction between lag and repetition ($F < 1$, $\eta^2 = .00$).

For Grade 3, there was no significant difference between the short-lag condition and the long-lag condition [$M_D = 0.01$, 95% CI (−0.04, 0.06)], $F < 1$, $\eta^2 = .00$. There was a significant benefit of restudy over retrieval practice [$M_D = .05$, 95% CI (0.00, 0.09)], $F(1, 31) = 4.50$, $p = .042$, $\eta^2 = .05$. Further, there was no significant interaction between lag and repetition ($F < 1$, $\eta^2 = .00$).

For Grade 4, there was no significant difference between the short-lag condition and the long-lag condition [$M_D = 0.03$, 95% CI (−0.05, 0.11)], $F < 1$, $\eta^2 = 0.01$. There was no significant difference between the restudy condition and the retrieval practice condition [$M_D = 0.01$, 95% CI (−0.03, 0.06)], $F < 1$, $\eta^2 = .00$. There was no significant interaction between lag and repetition ($F < 1$, $\eta^2 = .00$).

For Grade 6, there was no significant difference between the short-lag condition and the long-lag condition [$M_D = 0.03$, 95% CI (−0.09, 0.03)], $F < 1$, $\eta^2 = .01$. Also, there was no significant difference between the restudy condition and the retrieval practice condition [$M_D = 0.01$, 95% CI (−0.04, 0.05)], $F < 1$, $\eta^2 = .00$. Furthermore, there was no significant interaction between lag and repetition, $F(1, 31) = 1.83$, $p = .186$, $\eta^2 = .02$.

To sum up, there were no significant effects of lag. In Grades 2, 4, and 6 there were no significant effects of repetition type, but there was a significant benefit of restudy in Grade 3. This means that the only effect found was not in the expected direction.

[5] See footnote 2.

DISCUSSION

The aim of our experiment was to investigate whether distributed practice and retrieval practice can enhance primary school vocabulary learning. In contrast to earlier conducted studies (e.g., Goossens et al., 2012; Goossens, Camp, Verkoeijen, & Tabbers, 2014; Goossens, Camp, Verkoeijen, Tabbers, & Zwaan, 2014; Sobel et al., 2011), the experiment was fully integrated in the vocabulary learning curriculum. The children had to perform their own vocabulary learning exercises that were scheduled differently (for the distributed practice manipulation) and that were supplemented with additional exercises (for the repetition type manipulation). Furthermore, we investigated these memory strategies in four different grades (Grades 2, 3, 4, and 6). The effect of distributed practice was investigated by comparing a short-lag condition in which words were learned during two sessions in 1 week to a long-lag condition in which words were learned during four sessions in 2 weeks. The effect of repetition type was investigated by comparing a restudy condition in which children had to copy parts of the descriptions of the words to a retrieval practice condition in which children had to recall the descriptions of the words. To establish the effects of distributed practice and retrieval practice, we administered two final tests: a cued-recall test on the short term in which children had to give descriptions of the words and a multiple-choice test on the long term in which children had to choose the right descriptions of the words.

In contrast to our expectations, these experiments showed significant benefits of short-lag practice in two grades and of restudy practice in one grade on the cued-recall test. Furthermore, no benefits were found on the multiple-choice test, except for an effect of restudy practice in one grade. Thus, contrary to earlier investigations of distributed practice and retrieval practice in primary school vocabulary learning (e.g., Goossens et al., 2012; Goossens, Camp, Verkoeijen, & Tabbers, 2014; Goossens, Camp, Verkoeijen, Tabbers, & Zwaan, 2014; Sobel et al., 2011), we did not find any benefits of these memory strategies.

How can we explain this set of findings? We will first describe two possible procedural explanations for these results, and then we will describe the practical implications of the current study.

One procedural difference between the distributed practice manipulation of the current study and of other studies (e.g., Goossens et al., 2012; Kornell, 2009; Sobel et al., 2011) is that even in our short-lag condition, practice was distributed over several days, instead of being concentrated within the same day. As a result, our study was more about the optimal study *lag* rather than about the effect of spaced versus massed practice. Perhaps, the retention interval of the cued-recall test (1 to 3 days) suited the short-lag condition (also 1 to 3 days) better than the long-lag condition (1 to 6 days). It has been shown that the optimal lag between learning sessions is dependent on the retention interval (e.g., Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008; Cepeda et al., 2006; Küpper-Tetzel et al., 2014). When the retention interval increases, the optimal lag between learning sessions increases as well. However, on the multiple choice tests, which were

administered after long retention intervals up to 11 weeks, no benefit of a long lag was found either.

Another procedural difference between the retrieval practice manipulation of the current study and of other studies (e.g., Goossens, Camp, Verkoeijen, & Tabbers, 2014; Karpicke & Smith, 2012) is that in addition to the retrieval and restudy exercises, our participants also performed textbook exercises. Thus, there was no 'pure' comparison between restudy practice and retrieval practice. Although in one study additional vocabulary exercises were performed, these exercises were only performed in the first learning session, in which the actual retrieval practice manipulation did not take place (Goossens, Camp, Verkoeijen, Tabbers, & Zwaan, 2014). In the current study, retrieval practice was implemented in only a small part of the exercises (33.3%), which might have diluted the positive effect of retrieval practice. The children may have put more effort into the regular exercises than into the retrieval practice exercises and/or restudy exercises, which might explain the absence of a benefit of retrieval practice. However, the fact that in all grades performance in the restudy conditions was even better than in the retrieval conditions (with some effect sizes as big as .09) cannot be explained by merely a reduced benefit of retrieval practice.

So why do we not only fail to find any benefits of retrieval practice and distributed practice, benefits that have been repeatedly found in the laboratory, but in some grades even find a trend in the opposite direction? The answer may reside in the real-life setting of our study. Applying both strategies in actual primary school vocabulary lessons has high ecological validity. However, this realistic setting may have negatively impacted the chances of finding an effect. First, there was less experimental control over factors such as time on task and environmental noise than in the laboratory. Duration of the lessons differed among the different grade levels and classes and among different lessons within a classes as well. Furthermore, it was quite difficult for the children to stay quiet during the lessons, and the noisiness of some of the children may have interfered with other children's concentration. Also, the fact that the children had to correct their own answers with an answering sheet made the classroom atmosphere quite lively, and therefore, not all children may have inspected their answers with an equal amount of attention.

While the lack of experimental control enhances measurement error variances, it cannot explain the absence of systematic differences between conditions. However, concentration problems may have been more harmful in the long-lag condition and in the retrieval practice condition than in the short-lag condition and in the restudy condition, respectively, because the latter conditions require less effortful attention. The longer the lag, the more effortful study-phase retrieval will be (e.g., Cepeda et al., 2006). Also, retrieving a description will be more effortful than rewriting parts of a description, and therefore, it might be possible that the children focused more attention on the restudy exercises than on the retrieval practice exercises. Some children did not fill in any of the retrieval practice exercises, because they found it quite difficult to retrieve the answers themselves. In sum, our study shows in real-life primary school vocabulary lessons, other factors might come into play that interact with the potential benefits of strategies such as retrieval practice.

Finally, an important factor to note is the low interrater reliability in scoring the cued-recall test. The low interrater reliability might have influenced the power of the study and might have led to random measurement error, decreasing the systematic variance of the data. However, the effects we found were not in the expected direction. Thus, even if the interrater reliability would have been higher, we probably would have only found even larger effects in the opposite direction. Furthermore, although asking for word descriptions seems an ecologically valid manner to measure vocabulary knowledge, and is preferable to other measures like recognition tasks or fill-in-the-blank tasks, the low interrater reliability shows that judging the quality of word descriptions is quite subjective. Even with relatively easy words, and with a relatively extensive answer form, experimenters disagree on many of the descriptions of the words given by the primary school children. In our experiment, we used a short scoring protocol in which a few examples were given to specify the range of acceptable responses. However, if researchers want to use a similar kind of cued-recall test to ours to assess vocabulary knowledge, it might be better to include a broader range of acceptable responses in the protocol as for example is performed in the WISC-IV. To sum up, the low degree of interrater reliability that emerged from this study shows two things. First, it is quite a challenge to measure vocabulary knowledge in an ecologically valid manner. Second, it is difficult for children to give a description of a word.

CONCLUSION

We investigated distributed practice (i.e. short-lag practice versus long-lag practice) and repetition type (i.e. restudy practice versus retrieval practice) in real-life primary school vocabulary learning settings in four grades. Contrary to our expectations, and contrary to earlier experimental findings concerning primary school vocabulary learning (e.g., Goossens et al., 2012; Goossens, Camp, Verkoeijen, & Tabbers, 2014; Goossens, Camp, Verkoeijen, Tabbers, & Zwaan, 2014; Sobel et al., 2011), we found benefits of the short-lag condition on the cued-recall test in Grade 2 and in Grade 4, and of the restudy condition on both the cued-recall test and the multiple-choice test in Grade 3. However, the differences between conditions were generally small to medium. These results show that we cannot simply translate memory strategies from the laboratory to a regular classroom situation and that we cannot simply advise teachers to use long-lag practice and retrieval practice in primary school vocabulary lessons. Further applied research into the conditions under which these two memory strategies are beneficial is needed. For example, we need to know more about the relation between lag and retention interval (e.g., Cepeda et al., 2006; Cepeda et al., 2008; Küpper-Tetzel et al., 2014) and about the number of retrieval practice exercises needed. Also, it is important to investigate whether the effects are dependent on the kind of words being learned in the classroom, and the children's prior knowledge of these words. Furthermore, it is important to investigate the magnitude of the differences between the conditions. In the laboratory the differences between spaced practice and massed practice and the differences between retrieval practice and restudy practice are in general much larger than in the classroom. On the other hand,

the results of the current study are more ecologically valid and thus more generalizable to a classroom situation. Nonetheless, there is room for optimism, as we think that strategies like distributed practice and retrieval practice can be implemented in the classroom in a more effective way than in the current study. However, we do need further empirical evidence that shows us under which conditions and to what extent long-lag practice and retrieval practice can be beneficial in real-life primary school vocabulary lessons.

ACKNOWLEDGEMENTS

The Board of Public Education Rotterdam (Stichting BOOR) provided financial support for this research. The authors would like to thank the director and teachers of the school for their cooperation. Furthermore, we thank the children for participating in this research. Also, we thank Pim Renkema for scoring the final cued-recall tests.

REFERENCES

Anderson, R. C., & Freebody, P. (1981). Vocabulary knowledge. In J. T. Guthrie (Ed.), *Comprehension and teaching: Research reviews* (pp. 71-117). Newark, Del.: International Reading Association.

Bahrick, H. P., Bahrick, L. E., Bahrick, A. S., & Bahrick, P. E. (1993). Maintenance of foreign language vocabulary and the spacing effect. *Psychological Science*, *4*, 316–321.

Bloom, K. C., & Shuell, T. J. (1981). Effects of massed and distributed practice on the learning and retention of second-language vocabulary. *Journal of Educational Research*, *74*, 245–248.

Bouwmeester, S., & Verkoeijen, P. P. J. L. (2011). Why do some children benefit more from testing than others? Gist trace processing to explain the testing effect. *Journal of Memory and Language*, *65*, 32–41. DOI:10.1016/j.jml.2011.02.005.

Cahill, A., & Toppino, T. C. (1993). Young children's recognition as a function of the spacing of repetitions and the type of study and test stimuli. *Bulletin of the Psychonomic Society*, *31*, 481–484.

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. DOI:10.3758/BF03194004.

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

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition*, *20*, 633–642. DOI:10.3758/BF03202713.

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. DOI:10.1037/0033-2909.132.3.354.

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning. A temporal ridgeline of optimal retention. *Psychological Science*, *19*, 1095–1102. DOI:10.1111/j.1467-9280.2008.02209.x.

Cull, W. L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology*, *14*, 215–235. DOI:10.1002/(SICI)1099-0720(200005/06)14:3<215::AID-ACP640>3.0.CO;2-1.

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

Delaney, P. F., Verkoeijen, P. P. J. L., & Spirgel, A. (2010). Spacing and the testing effects: A deeply critical, lengthy, and at times discursive review of the literature. *Psychology of Learning and Motivation*, *53*, 63–147. DOI:10.1016/S0079-7421(10)53003-2.

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

Field, A. P. (2009). *Discovering statistics using SPSS: and sex and drugs and rock 'n' roll* (third edn). London: Sage publications.

Fritz, C. O., Morris, P. E., Acton, M., Voelkel, A. R., & Etkind, R. (2007). Comparing and combining retrieval practice and the keyword mnemonic for foreign vocabulary learning. *Applied Cognitive Psychology*, *21*, 499–526. DOI:10.1002/acp.1287.

Gerbier, E., Toppino, T. C., & Koenig, O. (2014). Optimizing retention through multiple study oppotunities over days: The benefit of an expanding schedule of repetitions. *Memory*. Advance online publication. DOI:10.1080/09658211.2014.944916.

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. DOI:10.1002/acp.2956.

Goossens, N. A. M. C., Camp, G., Verkoeijen, P. P. J. L., Tabbers, H. K., & Zwaan, R. A. (2012). Spreading the words: A spacing effect in vocabulary learning. *Journal of Cognitive Psychology*, *24*, 965–971. DOI:10.1080/20445911.2012.722617.

Goossens, N. A. M. C., Camp, G., Verkoeijen, P. P. J. L., Tabbers, H. K., & Zwaan, R. A. (2014). The benefit of retrieval practice over elaborative restudy in primary school vocabulary learning. *Journal of Applied Research in Memory and Cognition*, *3*, 177–182. DOI:10.1016/j.jarmac.2014.05.003.

Janssen, M., & Van Ooijen, M. (Eds.) (2012). *Taal Actief, 4, 5, 6, 7, 8*. Den Bosch: Malmberg.

Kang, S. H. K., Lindsey, R. V., Mozer, M. C., & Pashler, H. (2014). Retrieval practice over the long term: Should spacing be expanding or equal interval? *Psychonomic Bulletin & Review*, *21*, 1544–1550. DOI:10.3758/s13423-014-0636-z.

Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General*, *138*, 469–486. DOI:10.1037/a0017341.

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

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

Kornell, N. (2009). Optimizing learning using flash-cards: Spacing is more effective than cramming. *Applied Cognitive Psychology*, *23*, 1297–1317. DOI:10.1002/acp.1537.

Küpper-Tetzel, C. E., Erdfelder, E., & Dickhäuser, O. (2014). The lag effect in secondary school classrooms: Enhancing students' memory for vocabulary. *Instructional Science*, *42*, 373–388. DOI:10.1007/s11251-013-9285-2.

Küpper-Tetzel, C. E., Kapler, I. V., & Wiseheart, M. (2014). Contracting, equal, and expanding learning schedules: The optimal distribution of learning sessions depends on retention interval. *Memory & Cognition*, *42*, 729–741. DOI:10.3758/s13421-014-0394-1.

Metcalfe, J., Kornell, N., & Son, L. K. (2007). A cognitive-science based programme to enhance study efficacy in a high and low risk setting. *European Journal of Cognitive Psychology*, *19*, 743–768. DOI:10.1080/09541440701326063.

Nation, I. S. P. (2004). Vocabulary learning and intensive reading. *English Australia Journal*, *21*(2), 20–29.

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 3–8. DOI:10.1037/0278-7393.31.1.3.

Pyc, M. A., & Rawson, K. A. (2007). Examining the efficiency of schedules of distributed retrieval practice. *Memory & Cognition*, *35*, 1917–1927. DOI:10.3758/BF03192925.

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, *60*, 437–447. DOI:10.1016/j.jml.2009.01.004.

Pyc, M. A., & Rawson, K. A. (2011). Costs and benefits of dropout schedules of test-restudy practice: Implications for student learning. *Applied Cognitive Psychology*, *25*, 87–95. DOI:10.1002/acp.1646.

Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? *Journal of Experimental Psychology: General*, *140*, 283–302. DOI:10.1037/a0023956.

Rea, C. P., & Modigliani, V. (1987). The spacing effect in 4- to 9-year-old children. *Memory and Cognition*, *15*, 436–443.

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

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

Roediger, H. L., Putnam, A. L., & Smith, M. A. (2011). Ten benefits of testing and their applications to educational practice. In J. Mestre & B. Ross (Eds.), *Psychology of Learning and Motivation: Cognition in education* (pp. 1–36). Oxford: Elsevier. DOI:10.1016/B978-0-12-387691-1.00001-6.

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

Schrooten, W., & Vermeer, A. (1994). *Woorden in het basisonderwijs. 15.000 woorden aangeboden aan leerlingen*. Tilburg: Tilburg University Press.

Scott, J., Jamieson-Noel, D., & Asselin, M. (2003). Vocabulary instruction throughout the school day in 23 Canadian upper-elementary classrooms. *The Elementary School Journal*, *103*, 269–286.

Sobel, H. S., Cepeda, N. J., & Kapler, I. V. (2011). Spacing effects in real-world classroom vocabulary learning. *Applied Cognitive Psychology*, *25*, 763–767. DOI:10.1002/acp.1747.

Stahl, S. A., & Nagy, W. E. (2006). *Teaching word meanings*. Mahwah, NJ: Lawrence Erlbaum.

Toppino, T. C. (1993). The spacing effect in preschool children's free recall of pictures and words. *Bulletin of the Psychonomic Society*, *31*, 27–30.

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

Toppino, T. C., & DeMesquita, M. (1984). Effects of spacing repetitions on children's memory. *Journal of Experimental Child Psychology*, *37*, 637–648.

Toppino, T. C., & DiGeorge, W. (1984). The spacing effect in free recall emerges with development. *Memory and Cognition*, *12*, 118–122.

Toppino, T. C., & Gracen, T. F. (1985). The lag effect and differential organization theory: Nine failures to replicate. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*, 185–191. DOI:10.1037/0278-7393.11.1.185.

Toppino, T. C., Kasserman, J. E., & Mracek, W. A. (1991). The effect of spacing repetitions on the recognition memory of young children and adults. *Journal of Experimental Child Psychology*, *51*, 123–138.

Tulving, E. (1967). The effects of presentation and recall of material in free-recall learning. *Journal of Verbal Learning and Verbal Behavior*, *6*, 175–185. DOI:10.1016/S0022-5371(67)80092-6.

Van Berkel, S., & Alberts, N. (2009). *LOVS Woordenschat groep 4*. Arnhem: Cito.

Van Berkel, S., & Hilte, M. (2009). *LOVS Woordenschat groep 5*. Arnhem: Cito.

Van Berkel, S., Hilte, M., Engelen, R., Kamphuis, F., Kleintjes, F., & Krom, R. (2010). *Woordenschat groep 3 t/m 5. Wetenschappelijke verantwoording*. Arnhem: Cito.

Van de Gein, J., Van de Guchte, C., & Kouwenberg, B. (2008). *Zin in Taal Nieuw A, B, C, D, E*. Tilburg: Zwijsen.

Verhoeven, L., & Vermeer, A. (1993). *Taaltoets Allochtone Kinderen Bovenbouw*. Tilburg: Zwijsen.

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