Running head: THE LAG EFFECT IN SECONDARY SCHOOL CLASSROOMS

Published in *Instructional Science*:
http://link.springer.com/article/10.1007/s11251-013-9285-2

The Lag Effect in Secondary School Classrooms: Enhancing Students' Memory for Vocabulary

Carolina E. Küpper-Tetzel, Edgar Erdfelder, and Oliver Dickhäuser

Department of Psychology
School of Social Sciences
University of Mannheim
68131 Mannheim, Germany
Email: kuepper-tetzel@psychologie.uni-mannheim.de
Email: erdfelder@psychologie.uni-mannheim.de
Email: oliver.dickhaeuser@uni-mannheim.de

Corresponding author's address:

Carolina E. Küpper-Tetzel
Department of Psychology
School of Social Sciences
University of Mannheim
Schloss, Ehrenhof Ost
68131 Mannheim, Germany
Phone: +49-621-181 2145
Fax: +49-621-181 3997
Email: kuepper-tetzel@psychologie.uni-mannheim.de

Abstract

Educators often face serious time constraints that impede multiple repetition lessons on the same material. Thus, it would be useful to know when to schedule a single repetition unit to maximize memory performance. Laboratory studies revealed that the length of the retention interval (i.e., the time between the last learning session and the final memory test) dictates the optimal lag between two learning sessions. The present study tests the generalizability of this finding to vocabulary learning in secondary school. Sixth-graders were retaught English-German vocabulary after lags of 0, 1, or 10 days and tested 7 or 35 days later. In line with our predictions, we found that the optimal lag depends on the retention interval: Given a 7-day retention interval, students performed best when relearning occurred after 1 day. When vocabulary was tested after 35 days, however, students benefited from lags of both 1 and 10 days. Model-based analyses show that enhanced encoding processes and stronger resistance to forgetting – but not better retrieval processes – underlie the benefits of optimal lag. Our findings have practical implications for classroom instruction and suggest that review units should be planned carefully by taking the time of the final test into consideration.

Keywords: Lag effect; Long-term memory; Secondary school students; Classroom-based learning; Vocabulary learning

The Lag Effect in Secondary School Classrooms: Enhancing Students' Memory for Vocabulary

A large part of the knowledge that students acquire in school is quickly forgotten and cannot be accessed when it is needed later on (Bahrick & Hall, 1991). How can teachers address this problem? Researchers in cognitive psychology have revealed efficient learning methods that improve retention of previously learned information (Pashler, Rohrer, Cepeda, & Carpenter, 2007). For example, laboratory studies have demonstrated that long-term retention of a wide range of to-be-learned materials can be enhanced when multiple restudying units are not massed together, but rather distributed over time (e.g., mathematics learning: Rohrer & Taylor, 2007; text passages: Rawson & Kintsch, 2005; vocabulary pairs: Kornell, 2009). This phenomenon is called the *spacing effect* (i.e., massed[1] versus spaced practice). It has also been established that long-term memory benefits more from multiple relearning units that are separated by long lags instead of short lags (e.g., vocabulary pairs: Bahrick, Bahrick, Bahrick, & Bahrick, 1993; Bahrick & Hall, 2005). This is referred to as the *lag effect* (i.e., differences in effectiveness of nonzero lags, e.g., 1-day lag compared to a 10-day lag). Although the lag effect is related to the spacing effect, it has been argued that it is important to distinguish between them (see Cepeda, Pashler, Vul, Wixted, and Rohrer, 2006; Delaney, Verkoeijen, & Spirgel, 2010).

Optimal distribution of practice is not only easily implemented, but also produces remarkable effects on learning outcomes. In his comprehensive synthesis of meta-analyses, Hattie (2009) reported that spaced rather than massed learning clearly enhanced students' learning (Cohen's $d = 0.71$). Moreover, a recent experimental study by Küpper-Tetzel and Erdfelder (2012) revealed large effect sizes for the difference between massed and optimally distributed learning sessions in cued recall (Cohen's $d \geq 1.13$) and also for the difference

[1] Massed practice means that the entire study time is crammed into one single learning session and the same material is repeatedly studied over and over (i.e., studying the same material for 4 hours on Tuesday). Spaced practice allocates the same study time to different learning sessions which, for example, take place on different days (i.e., studying 2 hours on Monday and 2 hours on Tuesday).

between non-optimal and optimal distributions of learning sessions in cued recall (Cohen's *d* ≥ 0.66). Thus, the systematic distribution of learning and relearning sessions bears the potential to provide an extremely helpful and effective instruction method in the school context. Küpper-Tetzel and Erdfelder (2012) demonstrated that participants' long-term memory performance on delayed cued recall tests (e.g., after one week or one month) is increased by up to 89% when learning and relearning sessions are optimally distributed across time instead of condensed into a single learning episode and by up to 29% when learning sessions are separated by optimal lags compared to lags that are non-optimal.

How can we explain these effects? Why does memory performance improve after optimal lags compared to non-optimal or zero lags between learning sessions? Three types of explanations have been suggested that differ in regard to the underlying memory processes. First, there are explanations that attribute the lag effect to enhanced *encoding processes* during relearning (e.g., the study-phase retrieval theory, cf. Thios & D'Agostino, 1976). Then, there are explanations that propose improved *maintenance processes to the time of testing* to be responsible for the lag effect. In other words, repeating the to-be-learned material after an optimal lag is assumed to establish memory traces that are more resistant against forgetting (e.g., the Multiscale Context Model, cf. Mozer, Pashler, Cepeda, Lindsey, & Vul, 2009). And, lastly, there are explanations assuming that a repetition of the to-be-learned material after adequate lags leads to better *retrieval processes at test* (e.g., the contextual variability theory, cf. Glenberg, 1979). Recently, Küpper-Tetzel and Erdfelder (2012) used Multinomial Processing Tree (MPT) modeling (Batchelder & Riefer, 1999; Erdfelder et al., 2009) to disentangle encoding, maintenance, and retrieval contributions to the lag effect. Their findings point to the conclusion that the lag effect is largely driven by encoding and maintenance processes: Whereas encoding benefits from relative short (but nonzero) lags, maintenance in memory benefits from long lags, the more so the longer the

retention interval (i.e., the time between the last learning session and the final test). In contrast, retrieval processes seem to play only a minor role for understanding the lag effect.

The generalizability of the spacing effect to authentic learning settings has been tested in a few applied studies. Bloom and Shuell (1981), for example, had high-school students learn French vocabulary in a massed (30-minute unit on a single day) or a spaced fashion (10-minute units on three consecutive days) during their regular French class. In line with laboratory findings, students with spaced learning outperformed students with massed learning on a test administered four days later. Other field studies demonstrated beneficial spacing effects in preschoolers for enhancing reading ability (Seabrook, Brown, & Solity, 2005) and for promoting the acquisition of complex sentence construction (Ambridge, Theakston, Lieven, & Tomasello, 2006).

Thus, laboratory and field studies alike suggest to use multiple repetition units and to distribute them over time to boost long-term retention. However, teachers, who must accomplish comprehensive curricula, often face serious time constraints that impedes multiple repetition sessions of previously taught material. Thus, if school curricula allow only for a small number of repetition units, for example, for one repetition session only, the optimal timing of this unit is of major interest. Thus, the main interest is not in comparing massed versus spaced learning, but rather to compare lags of different lengths (i.e., the lag effect). The question is: How much time should elapse between initial teaching of the to-be-learned material and the repetition of this material in order to enhance memory performance in the long run?

Recent laboratory and web-based studies suggest that the answer to this question is complex (Cepeda et al., 2009; Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008). In the study of Cepeda et al. (2009) undergraduate university students studied Swahili-English vocabulary during an initial learning session and restudied the vocabulary after lags of 0, 1, 2, 4, 7, or 14 days. All participants were tested 10 days after the restudy session. They found that memory

performance on the final test was best when the lag between initial study and restudy session was 1 day. For lags shorter or longer than 1 day, correct vocabulary retention was decreased 10 days after practice. Thus, the appropriate timing of a repetition unit matters.

Moreover, Cepeda et al. (2008) and Cepeda et al. (2009) examined whether the optimal lag between two learning episodes changes as a function of the retention interval. Again, several lags were used and memory performance was assessed 168 days (Cepeda et al., 2009) or up to 350 days (Cepeda et al., 2008) following the end of the practice phase. To avoid floor effects due to massive forgetting during these long intervals, they used more meaningful study material than vocabulary in these experiments (i.e., largely unknown but true trivia facts). They found that the optimal time for relearning depends on the length of the retention interval. More precisely, for any given retention interval, memory performance follows an inverted-U-shaped function by first increasing with lag until reaching an *optimal lag* and then decreasing again. The optimal lag is dictated by the length of the retention interval and increases with longer retention intervals: For retention after 7 days the optimal lag was 1 day, for retention after 35 days the optimal time for relearning was 11 days, and for retention after a long retention interval of 350 days the optimal lag was 21 days. Furthermore, Cepeda et al. (2008) showed that the ratio of optimal lag to retention interval length decreases with longer retention intervals. The results of Küpper-Tetzel and Erdfelder (2012) provide converging evidence for these lag effect trends.

These findings may have important implications for classroom instruction because they emphasize the appropriate scheduling of a repetition unit and reveal that a lag which is either too short or too long may have detrimental effects on retention of the to-be-learned material across a pre-defined retention interval. In addition to classroom instruction, the results might also be important for students' self-regulated learning as during phases of self-regulation, students can schedule the point of time for repetition units on their own.

A recent study by Bird (2010) investigated the interaction between specific lags (3- vs. 14-day lag) and retention intervals (7- vs. 60-day retention interval) in a classroom setting for second language syntax learning in university students. He found that memory performance 60 days after practice benefited more from a 14-day lag than from a 3-day lag between learning sessions. After a 7-day test interval no difference was detected between the two lag conditions that he examined. The latter finding is not surprising. Previous studies have repeatedly shown that people will perform best on a final memory test administered one week later if they relearn the material *one day* after initial learning – not earlier or later (see Ausubel, 1966; Glenberg & Lehmann, 1980; Cepeda et al., 2008; Küpper-Tetzel & Erdfelder, 2012). The interesting aspect of Bird's study is, however, that he evaluated the lag effect in an ecologically more valid learning environment by having participants study meaningful content in a classroom setting rather than in a laboratory. But, as all studies did so far, Bird investigated the lag effect in the standard population that is usually used in laboratory studies, namely, university students.

For at least two reasons, it is important to examine the lag effect dynamics also in younger student populations, especially in young secondary school students. First and most importantly, it is unknown yet whether the results previously obtained with university students and adults generalize to younger students in the school context. Second, if teachers can rely on long-term maintenance of previously taught material in students, they can avoid unplanned and costly review sessions, when instead new and advanced material is scheduled. This promotes the effective use of classroom instruction time.

Thus, secondary school instruction and learning may benefit from research-based optimization techniques of learning across time if and only if it can be demonstrated that previous laboratory findings also hold for secondary school classroom learning. There is one study that has investigated the spacing effect with *two learning sessions* in secondary school classrooms. In Sobel, Cepeda, and Kapler (2011), students learned GRE vocabulary during

two learning sessions that were either massed in time or separated by a lag of 7 days. Five weeks later, students performed better on vocabulary that had been practiced in a spaced fashion than on vocabulary that had been practiced in a massed fashion. However, to date, no study has examined the effect of lags of different lengths between two learning sessions in a secondary school classroom setting when authentic school material is used. Therefore, the goal of our study was to examine the lag effect in secondary school vocabulary learning and, particularly, to test whether the interaction between lag and retention interval as revealed in previous experimental studies (Cepeda et al., 2008; Cepeda et al., 2009) generalizes to real-world educational settings[2] and materials. In accordance with Ulrich Neisser's advice we aimed at investigating "cognition as it occurs in the *ordinary environment* and in the context of natural *purposeful* activity" (Neisser, 1976, p. 7, own emphasis). Thus, we implemented the lag effect intervention into the classroom during the regular lessons and, most importantly, used material that was meaningful for the students. In most laboratory studies, the material that participants learn has no further implications for their future academic performance. This might foster contextual influences such as lag effects on memory performance. Therefore, it is possible that the lag effects are attenuated when meaningful material is learned in an authentic setting in which regular assessments of students' performance impacts their future. Students may adopt strategies that lead to deeper and better encoding of the material which, in turn, diminishes the effect of distributed learning. Hence, it is not certain at all that the lag effect trends – as found in the laboratory – generalize to such authentic educational environments when material is learned that has immediate and future relevance for the population under investigation.

[2] Note that the research on the lag effect should be distinguished from a line of work that focuses on the benefits of blocked versus nonblocked teaching. In the latter line of research, different pieces of information are presented either within a single large session or allocated to multiple, but shorter sessions (Randler, Kranich, & Eisele, 2008; Lawrence & McPherson, 2000). In the current paper, in contrast, we investigate after which lag newly learned information should be repeated given that the goal is to retrieve this information after a pre-defined retention interval without further study.

To test the benefits and limitations of the lag effect, we conducted a field experiment in an authentic secondary school classroom setting and had German sixth graders practice and re-practice new German-English vocabulary from advanced chapters of their textbook in two learning sessions separated by a 0-day (massed), 1-day, or 10-day[3] lag. Students were tested either 7 or 35 days later on their memory performance for the vocabulary pairs. Based on previous empirical findings (Cepeda et al., 2008; Küpper-Tetzel & Erdfelder, 2012), we expected that students who were tested 7 days after the last learning session would show better vocabulary recall when their two learning sessions were separated by a 1-day lag than when the two learning sessions were separated by a 0- or 10-day lag. Hence, we assumed that memory performance would follow an inverted-U-shaped trend with increasing lag in the 7-day retention interval group. In contrast, after a 35-day retention interval, we predicted better memory for vocabulary when the second learning session occurs after a lag longer than 1 day, resulting in a trend that increases beyond a lag of 1 day and perhaps up to a lag of 10 days. Students' final memory performance for vocabulary was assessed with a cued recall test.

In addition to memory performance data, we also analyzed the memory processes underlying these data using Küpper-Tetzel and Erdfelder's (2012) MPT model to test for converging evidence in regard to the importance of encoding and maintenance processes for the lag effect. In order to run these model-based analyses we assessed students' memory performance with a free recall test that was administered right before the cued recall test.

Method

Participants

A total of 76 sixth-graders from a secondary school participated in the study. Data from eight students had to be excluded from all analyses because they did not attend all

[3] To revisit, Cepeda et al.'s (2008) findings suggest that the optimal lag for a test administered 35 days after practice is 11 days. However, due to the predetermined school schedule, it was not possible to realize a relearning session 11 days after the initial learning session. Therefore, the longest lag was 10 days instead.

learning sessions or failed to appear on the test session. Data from one student who was diagnosed with dyslexia was dropped because the test scores were based on correctly written words only. Due to experimenter error, cued recall data were not collected from one student on the final test session. Finally, data from one participant were not included in the analyses because she failed to follow the testing instructions. These exclusions led to 65 students[4]. They were on average 11.45 years old (range, 11-13 years). Of all students, 50.8% were male. Students came from three classrooms.

Materials

Since the study was conducted during the regular English lessons and the to-be-learned material should be relevant to the students, 26 German-English vocabulary pairs were selected from advanced units of the English exercise book. All words were concrete nouns (see Appendix).

Design

We realized two learning sessions separated by a 0-, 1-, or 10-day lag and one test session occurring after a retention interval of 7 or 35 days. This resulted in a 3 x 2 between-subjects design. As it is often the case in applied studies, individual students could not be randomly assigned to the different lag conditions (e.g., Seabrook et al., 2005; Randler et al., 2008). We had to respect the classroom structure because the study was realized during their regular English lessons. Thus, a whole classroom was assigned to a lag condition by taking into consideration their school schedule. This resulted in 27 students in the 0-day lag group, 22 students in the 1-day lag group, and 16 students in the 10-day lag group. However, the

[4] Three of the excluded participants were in the 10_7 condition (i.e., 10 days lag and 7 days retention interval), three were in the 0_35 condition, three were in the 10_35 condition, and two were in the 1_35 condition. We ran analyses on 7 out of the 11 excluded students for which we collected valid cued recall performance at the end of the first learning session. We compared their mean in cued recall at the end of the first learning session (M = 18.14) to the mean of the students that were used in the final analyses (M = 19.05). There was no systematic difference in regard to their initial memory performance, $t(70) = -0.43$, $p = .672$.

retention interval was experimentally manipulated within each classroom by randomly assigning one half of the students to the 7-day condition (n = 35 across the three lag conditions) and the other half to the 35-day condition (n = 30 across the three lag conditions).

Procedure

The study consisted of two learning sessions and one final test session. All sessions were run as group sessions.

Learning sessions

The first learning session encompassed two study-test trials and lasted 45-60 minutes. The second learning session took place after the respective lag and consisted of one study-test trial which lasted 25-30 minutes. A study-test trial involved the presentation of the German-English vocabulary, a recognition test, a cued recall test, and a picture quiz.

During vocabulary presentation, 26 German-English vocabulary were presented on the front wall of the classroom with a portable LCD projector. Students were instructed to pay attention to each word pair and to watch out for the orthography of the English words in particular. They were not allowed to take notes or rehearse vocabulary aloud. The presentation started after ensuring that the students understood the instructions. A German word appeared for two seconds alone on the left side of the projection. The experimenter read out the German word. Then, the English translation appeared on the right side and the experimenter read out the English word. Both words of a vocabulary pair were displayed for eight seconds. Word pairs were presented in a different random order for each vocabulary presentation.

After vocabulary presentation, students worked for five minutes on a paper-pencil three alternative forced choice recognition test. The test consisted of 26 rows and each row contained a target word (English vocabulary) from the presentation and two distractors. The distractors were English words that featured a high orthographical similarity to the English

target (see Appendix). Students were instructed to circle the target word. The order of the rows and the words within each row were printed in a random order on each recognition test. To prevent cheating, four parallel versions of the recognition test were used that differed with regard to the random order of rows and words. Upon completion students were asked to turn the recognition test sheet over and the paper-pencil cued recall test was handed out to them.

On the cued recall test, all German words were printed in random order one below the other. The students were allotted five minutes to recall and write the English translation next to each German word. Again, four parallel versions of the cued recall test were used that differed regarding the random order of German words. After all students had turned the cued recall sheet over, the picture quiz started.

The picture quiz was used as a feedback tool. Since it was not possible to give individual feedback on the recognition and cued recall tests because of the group setting, the picture quiz represented a good way to provide feedback. In addition, its interactive format motivated the students, which enhanced their compliance to the study. For the picture quiz, actual pictures of each target word were projected on the wall along with three English words. One of the words was the target word that correctly identified the depicted picture. The other two words were distractor words that were orthographically similar to the target word. All distractor words were new and had not been shown before. On each trial, students saw a picture and three words that were labeled with a red, a blue, and a yellow dot, respectively. Each student had a red, a blue, and a yellow card. They were instructed to indicate the target word that correctly described the picture by holding up the card with the respective color. Afterwards, the correct target word was revealed to them. The assignment from color to word and the order of the pictures were randomized for each picture quiz. The picture quiz took 5-8 minutes.

Test session

The test session occurred either 7 or 35 days after the last learning session. Students were instructed that the German-English vocabulary would not be presented to them and that they had to retrieve the vocabulary from memory instead. The test session involved a free recall test of German-English word pairs immediately followed by a cued recall test. For the free recall test, students were instructed to recall as many vocabulary pairs as they could. They were told that this was a hard task and were encouraged to write down all words they could remember from the learning phase, even if they could only remember single words, that is, only the German or English word of a vocabulary pair. They were allotted 5 minutes for the free recall test. The subsequent final cued recall test was identical to the one students received during their learning sessions except that the German words were printed in a different random order. After completion of the test session, the students were thanked for their help and informed that they would receive feedback on their test performance once all tests were checked. All students received a study booklet that contained not only their test scores, but also concrete suggestions on how to distribute their learning in order to improve long-term retention.

Results

In analyzing the memory performance data, we focus on the cued recall performance in the final vocabulary test as this is the practically relevant dependent variable in applied contexts and in educational settings in particular. In addition, free recall performances will become important in our additional data analyses in the framework of the MPT model that allows us to disentangle the contributions of encoding, maintenance, and retrieval processes to overall memory performance (Küpper-Tetzel & Erdfelder, 2012).

As mentioned before, the school setting did not allow the random assignment of individual students to lag conditions. In fact, a whole classroom was assigned to a specific

lag. In order to rule out classroom-dependent factors, we controlled statistically for these factors (see also Randler et al., 2008). We argue that students from the three classrooms should not differ in their memory performance at the end of the first learning session. Any difference in memory at this point of the study must be due to classroom-dependent factors rather than the lag manipulation because lag was initiated only after the first learning session. Possible factors that could have varied between classrooms and influenced memory performance at the end of the first learning session are, for example, learning ability or learning motivation. To control for these possible classroom-dependent effects, we used the cued recall performance assessed at the end of the first learning session as an additional predictor (i.e., covariate) in all analyses. An α-level of .05 was assumed for all analyses. All *p* values reported below refer to two-tailed tests, even in case of directed predictions.

Final cued recall performance

Averaged across lag conditions, students recalled more vocabulary after a 7-day retention interval (M = 71%, SD = 21%) than after a 35-day retention interval (M = 51%, SD = 18%), $t(62) = -6.12$, $p < .001$, $\eta^2 = 0.38$. Of greatest interest, however, were the different memory functions resulting from increasing lags in the 7-day and 35-day retention interval condition, respectively. To revisit, we expected that, in the 7-day retention interval group, memory performance would follow an inverted-U-shaped trend with lags increasing from 0 to 10 days, that is, producing a peak at a 1-day lag. In contrast, given a long retention interval of 35 days we predicted an increasing trend with lag instead. The percentage of correctly recalled vocabulary on the final cued recall test is presented in Figure 1. In the 7-day retention interval condition a significant negative quadratic trend emerged, $t(58) = 2.32$, $p = .024$, $\eta^2 = 0.08$. The linear trend was not significant, $t(58) = 0.39$, $p = .702$, $\eta^2 < 0.01$. In the 35-day retention interval condition, the reverse finding occurred. Here, a significant positive linear trend was detected, $t(58) = 2.00$, $p = .05$, $\eta^2 = 0.06$, but the negative quadratic

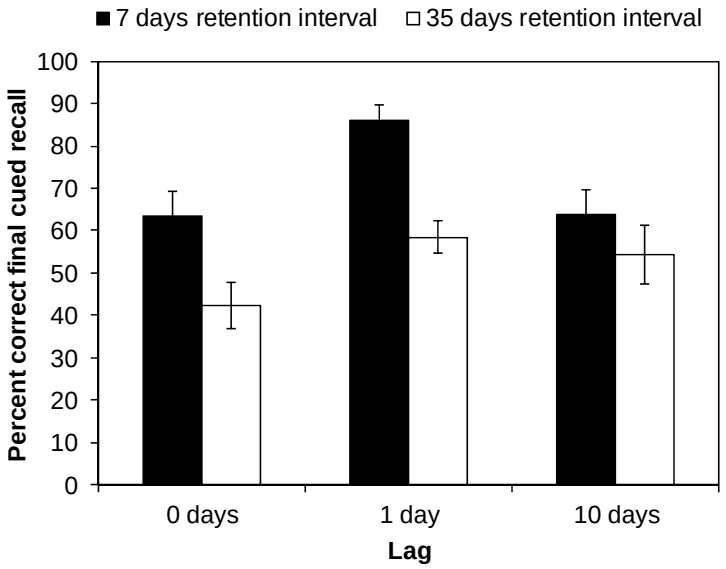

Figure 1. Mean and standard errors of correctly recalled vocabulary on the final cued recall test as a function of lag and retention interval.

was not significant, $t(58) = 0.04$, $p = .970$, $\eta^2 < 0.01$. Thus, as expected, we find that memory for foreign vocabulary tested 7 days after practice is severely impaired if the lag is shorter (massed) or longer (i.e., 10 days) than 1 day. The significant negative quadratic trend clearly shows that a 1-day lag is optimal given a 7-day retention interval. In contrast, we find a significant linear trend with increasing lag in the 35-day retention interval condition and no significant negative quadratic trend. This means that given a 35-day retention interval memory performance benefits from lags of 1 day and longer.

Multinomial Processing Tree analyses

Previous studies have used the combination of a test that depends heavily on retrieval processes (e.g., free recall) and a test that depends less on retrieval processes (e.g., cued recall) to separate the contributions of storage and retrieval processes to a memory phenomenon (see, e.g., Hogan & Kintsch, 1971; Thomson & Tulving, 1970). Following this approach, we also applied a free recall test in addition to the final cued recall test to disentangle contributions of encoding, maintenance, and retrieval processes to memory

performance. To measure these three types of processes, we used the Encoding-Maintenance-Retrieval multinomial model for free-then-cued-recall recently proposed by Küpper-Tetzel and Erdfelder (2012). This model uses performance data at different points in time (i.e., during practice and during the final test session) and from different tests (i.e., free and cued recall) to estimate seven parameters representing underlying memory processes: one probability of associative encoding (e), two probabilities of associative maintenance in memory until the final test (m_s and m_u for maintenance following successful vs. unsuccessful cued recall during practice, respectively), two probabilities of successful retrieval in free and cued recall (r_f and r_c, respectively), and two probabilities of single word retrieval in free recall in case of successful vs. unsuccessful associative encoding or maintenance (s and u, respectively). For a detailed model description we would like to refer to Küpper-Tetzel and Erdfelder (2012) since a full exposition goes beyond the scope of this work.

The multiTree software (Moshagen, 2010) was used for all MPT model analyses. The Type I error level was set to $\alpha = .05$ for all model-based analyses. A sensitivity analysis was performed using G*Power 3.1 (Faul, Erdfelder, Buchner, & Lang, 2009). This analysis showed that with $N = 1{,}419$ data points, a significance level of $\alpha = .05$, and a desired power of $1\text{-}\beta = .95$, the detectable effect size for G^2 goodness-of-fit tests based on $df \leq 35$ is $\omega \leq 0.16$ (i.e., a small effect; cf. Cohen, 1988). Thus, all G^2 model tests reported below allowed detecting already small deviations from the model.

Following Küpper-Tetzel and Erdfelder (2012), we restricted the free-then-cued-recall MPT model to obtain a parsimonious specification with only one maintenance parameter m. This was achieved by setting the maintenance probabilities m_s (maintenance after successful cued recall at the end of practice) and m_u (maintenance after unsuccessful cued recall at the end of practice) equal in each condition. This model version fit the data ($G^2(30) = 41.16$, $p = .084$). Furthermore, we tested the additional restriction that the probability of associative retrieval in cued recall, r_c, is equal across experimental conditions. Indeed, the G^2 difference

test was not significant, $\Delta G^2(5) = 4.11$, $p = .534$, with r_c being estimated to .96. Thus, our model-based findings are based on this restricted model version. The overall goodness-of-fit test indicates a good fit to the data ($G^2(35) = 45.27$, $p = .115$).

Of greatest interest for the evaluation of the theories are the probability estimates for associative encoding *e*, associative maintenance *m*, and associative retrieval r_f. Maximum likelihood estimates and standard errors for these three parameters are summarized in Figure 2. As shown in Figure 2A, the associative encoding parameter *e* followed an inverted-U-shaped trend with increasing lag in the 7-day retention interval condition. More precisely, associative encoding increased significantly between the 0-day and the 1-day lag condition, $\Delta G^2(1) = 21.94$, $p < .001$, and decreased between the 1-day lag and the 10-day lag condition, $\Delta G^2(1) = 24.18$, $p < .001$. In the 35-day retention interval condition, we found descriptively the same inverted-U-shaped trend with increasing lag. However, the only significant effect was the decrease in associative encoding between the 1-day and the 10-day lag condition, $\Delta G^2(1) = 4.67$, $p = .031$. The increase between the 0-day and the 1-day lag condition did not reach significance, $\Delta G^2(1) = 1.27$, $p = .260$.

As illustrated in Figure 2B, the parameter for associative maintenance *m* was affected differently by the length of the retention interval. In the 7-day retention interval condition, associative maintenance increased between the 0-day lag and the 1-day lag, $\Delta G^2(1) = 26.66$, $p < .001$, and decreased again between the 1-day and 10-day lag condition, $\Delta G^2(1) = 12.50$, $p < .001$. There was no difference in associative maintenance between the 0-day and the 10-day lag, $\Delta G^2(1) = 0.73$, $p = .392$. In the 35-day retention interval condition, associative maintenance increased significantly between both the 0-day lag and the 1-day lag, $\Delta G^2(1) = 13.21$, $p < .001$, and between the 0-day and the 10-day lag, $\Delta G^2(1) = 13.75$, $p < .001$. We detected no difference between the two spaced conditions for associative maintenance, $\Delta G^2(1) = 0.15$, $p = .700$.

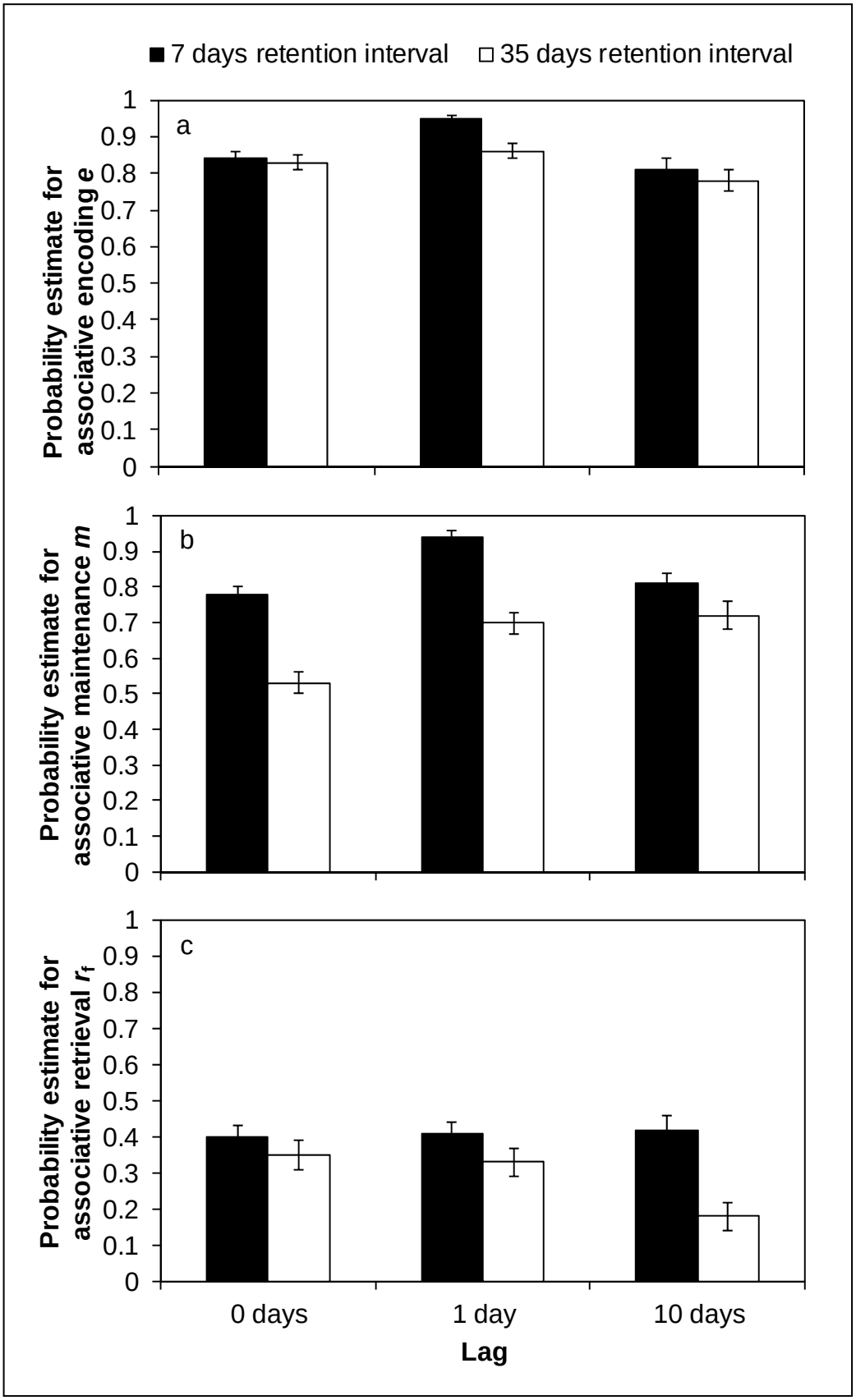

Figure 2. Parameter estimates and standard errors for the probability of associative encoding *e* (2a), for the probability of associative maintenance *m* (2b), and for the probability of associative retrieval r_f (2c) as a function of lag and retention interval.

Results for associative retrieval r_f during free recall are displayed in Figure 2C. Retrieval was equal across all lag conditions in the 7-day retention interval condition, $\Delta G^2(2) = 0.11$, $p = .949$. In the 35-day retention interval condition, we detected a significant decrease in associative retrieval between the 0-day and 10-day lag condition, $\Delta G^2(1) = 9.23$, $p = .002$, as well as the 1-day and 10-day lag condition, $\Delta G^2(1) = 7.57$, $p = .006$.

Discussion

The current field experiment examined the effect of different lags between two learning sessions on memory performance for foreign language vocabulary in sixth graders after 7 and 35 days. The findings are in line with our predictions.

In essence, students' memory for German-English vocabulary that was assessed one week after practice benefited most from a 1-day lag between initial study session and restudy session. In line with the predictions, lags of shorter (massed practice) or longer (10-day lag) length led to lower students' performance. However, when vocabulary memory was measured about one month after practice, students were best off in the two distributed practice conditions (i.e., 1-day and 10-day lag). Thus, we conclude that the optimal lag for reviewing vocabulary that is tested after 35 days is located beyond a 1-day lag, with a 10-day lag leading to comparable benefits for memory performance. At first this seems to be at odds with the findings of Cepeda et al. (2008) who revealed a significant increase between a 0-day up to an 11-day lag in the 35-day retention interval group. However, it is important to keep the sample in mind. Whereas Cepeda et al.'s (2008) sample consisted of adults only; we investigated young students explicitly in our field experiment. It makes sense to assume that the optimal time for relearning depends not only on the length of the retention interval, but in part also on learner characteristics (e.g., working memory skills (Gathercole, Pickering, Ambridge, & Wearing, 2004; Gatherhole, Lamont, & Alloway, 2006) or forgetting rates (Brainerd, Reyna, Howe, Kingma, & Guttentag, 1990)). Our findings hint at this possibility.

Additional lag effect studies (laboratory and field experiments), however, are needed to obtain a better understanding of learning in young students. Those studies should use a broader variation of lag and retention interval to shed light on the systematic dependency of optimal lag and retention interval for secondary school vocabulary learning.

In order to assess the practical significance of the obtained effects, we calculated Cohen's *d* effect size measures between the massed (lag = 0 days) and the best lag condition for each retention interval separately. In the 7-day retention interval condition, there was a 35% increase in correct vocabulary recall between the massed and the optimal 1-day lag. Stated differently, students in the 1-day lag condition remembered on average nine words more than students in the massed practice condition. This translates to a very large effect size (Cohen's $d = 1.69$). However, increasing the lag to 10 days led to a decline in performance of 34%. This means that students recalled on average nine vocabulary words less in the 10-day lag condition than in the optimal 1-day lag condition. This results in a large effect size of Cohen's $d = 2.07$. In the 35-day retention interval condition, we found a 28% and 38% increase in memory performance between the massed and the 10-day and the massed and the 1-day condition, respectively. Students in the two distributed lag conditions recalled on average 7 to 10 vocabulary words more than students in the massed condition. Again, this results in large effect sizes (Cohen's $d = 1.41$ for the comparison with the 1-day lag group and Cohen's $d = 0.87$ for the comparison with the 10-day lag group).

These are remarkable effects that allow us to make promising suggestions to educators and learners. Also, as proposed by Dempster (1988), we obtained these effects in a real-world educational environment by using relevant material and by keeping the classroom setting as naturalistic as possible by using group learning sessions and the integration of the field experiment in ongoing lessons. Both points should encourage teachers to implement the lag effect as instruction method in the classroom. Other applied studies (e.g., Reynolds & Glaser, 1964; Seabrook et al., 2005; Sobel et al., 2011) have already demonstrated beneficial

spacing effects in the classroom. Our field study extends this line of research by investigating the effect of different lags. We reveal an important boundary condition for classroom instruction. More precisely, teachers who face time constraints should take the retention interval into account when planning a repetition unit. Choosing a too long or a too short interval between study sessions can lead to detrimental effects on memory performance depending on the length of the retention interval. For example, if a surprise test of new vocabulary is due one week after the end of the practice phase (i.e., without interim learning), teachers can boost students' performance by introducing the vocabulary eight days before the test and program a repeating unit one day after initial learning. Given the restriction of only two learning sessions, they should refrain from introducing the new vocabulary at an earlier point in time, say two and a half weeks before the final assessment, and initiating a repeating lesson one week before the test. The inverted-U-shaped trend with increasing lag in the 7-day retention interval condition clearly shows that a further extension of the lag beyond one day has substantial negative effects on students' vocabulary memory.

The second aim of the present paper was to examine the underlying memory processes of these lag effect trends and to test different explanations of lag and spacing effects. Therefore, we applied the Encoding-Maintenance-Retrieval (EMR) multinomial processing tree model for lag effect data that has recently been proposed and validated by Küpper-Tetzel and Erdfelder (2012). We found that the inverted-U-shaped trend in the 7-day retention interval condition is produced by an increase in encoding and maintenance processes between the 0-day and the 1-day lag condition and a decrease of these processes for a lag of 10 days. In contrast, retrieval processes are not affected by different lags in the 7-day retention interval condition. Furthermore, the linear increasing trend in memory performance in the 35-day retention interval condition is produced by enhanced maintenance processes and not by better encoding or retrieval processes. In other words, maintenance processes are primarily responsible for the differences in memory performance trends with increasing lag

between the 7-day and the 35-day retention interval conditions. Better maintenance of the material to the time of testing explains why performance remains stable in the 35-day retention interval group and drops in the 7-day retention interval group for a lag beyond 1 day. Thus, the increase of the optimal lag with increasing retention interval is largely due to *stronger resistance to forgetting* induced by relearning after long as compared to short lags.

In summary, theories that focus on encoding and maintenance processes in explaining lag effect trends (i.e., study-phase retrieval theory and Multiscale Context Model) are corroborated by our findings. Theories emphasizing the role of retrieval processes for the lag effect (i.e. contextual variability theory), in contrast, are not in line with the EMR model findings. If anything, there was a decrease rather than the predicted increase in retrieval probabilities across the different lag conditions. The current findings are similar to those obtained in the laboratory study by Küpper-Tetzel and Erdfelder (2012) with respect to the underlying processes.

To conclude, our field study clearly shows that vocabulary learning in secondary school can benefit from adequate distribution of review units. In line with previous experiments (Cepeda et al., 2008; Cepeda et al., 2009), we reveal that the optimal lag increases as a function of retention interval. Given the circumstances under which the field study was conducted (i.e., heterogeneous student population and group learning sessions), these robust findings are encouraging and allow us to provide teachers with valid and useful suggestions. Based on the current findings, we recommend that when only one repeating lesson is feasible (e.g., due to time constraints), then the timing of the first learning lesson should be chosen appropriately by taking the desired length of the retention interval into consideration. This means that shorter lags between first and second learning should be chosen if the pre-defined retention interval is short and longer lags are appropriate when it is long.

Of course, not only educators can benefit from our findings, but also young students can be instructed to distribute their learning properly and boost their memory performance. They have acquired the necessary cognitive resources to understand such learning strategies and to apply them (e.g., Brehmer, Li, Müller, von Oertzen, & Lindenberger, 2007; Pressley & Hilden, 2006). In the present study, we created teacher and student booklets to inform teachers and students about the study findings and their implications. These booklets contained detailed information on the study and the results, as well as hands-on suggestions for classroom instruction and self-regulated learning. In addition, our empirical findings and analyses of the underlying cognitive processes have important implications for the development of computer-based learning tools. Currently, Mozer and colleagues are developing a web-based tutor for learning facts or vocabulary. This tool is based on assumptions of the Multiscale Context Model (Mozer et al., 2009) – which are in agreement with our findings. The tool prompts students individually as to when to review specific vocabulary in order to enhance memory performance on a final test at a predetermined time in the future. Thus, this learning tool incorporates an appropriate theory of human memory (i.e., Multiscale Context Model) which considers the complex interaction between optimal lag and retention interval. The overall benefit of this learning tool is currently being evaluated[5].

Our study contributes to applied human learning research in educational contexts. Similar to Seabrook et al. (2005) or Randler et al. (2008), we were not allowed to assign students from different classrooms randomly to their lag condition. We are aware of this limitation which, in the present case, could not be avoided due to the restricted freedom of scheduling and due to the strict classroom structure that had to be obeyed. To cope with this problem, we controlled for possible classroom effects statistically. Using this approach, we revealed robust lag effects in foreign vocabulary learning similar to those found in previous

[5] For detailed information see http://www.cs.colorado.edu/~mozer/index.php?dir=/Research/Projects/Optimization%20of%20learning/

completely randomized experiments. This enables us to provide teachers with better recommendations for their classroom instruction. Future studies should follow this line of classroom-based research and examine whether the lag effect as found for verbal learning transfers to other educational domains as, for instance, learning in mathematics and physics. Rohrer and Taylor (2006, 2007) revealed reliable spacing effects for geometry and permutation problems in the laboratory and Grote (1995) demonstrated beneficial spacing effects for physics learning in an authentic classroom setting. However, the generalizability of lag effects and potential interactions with the retention interval has not yet been examined for mathematics and physics learning. This should be the focus of future studies. In general, more applied studies in authentic classroom settings are needed since they broaden the evidence and validity of well-known memory effects for naturalistic learning environments. Although these studies are extensive and challenging, they promise to have the greatest impact on everyday educational routines.

Acknowledgements

The authors express their gratitude to the school principal, Mr. Michael Hohenadel, to the teachers, and the students of the Elisabeth secondary school in Mannheim for making this study possible. We thank the graduate students of the first author's service learning seminar, Dagmar Klein, Martin Knab, Sharmila Pushpakanthan, Sonja Sobott, and Sarah Zelt, for data collection and four anonymous reviewers for helpful comments on an earlier version of the manuscript.

References

Ambridge, B., Theakston, A. L., Lieven, E. V., & Tomasello, M. (2006). The distributed learning effect for children's acquisition of an abstract syntactic construction. *Cognitive Development*, *21*, 174–193. doi: 10.1016/j.cogdev.2005.09.003

Ausubel, D. P. (1966). Early versus delayed review in meaningful learning. *Psychology in the Schools*, *3*, 195-198. doi: 10.1002/1520-6807(196607)3:3<195::AID-PITS2310030302>3.0.CO;2-X

Bahrick, H. P., & Hall, L. K. (1991). Lifetime maintenance of high school mathematics content. *Journal of Experimental Psychology: General*, *120*, 20-33. doi: 10.1037/0096-3445.120.1.20

Bahrick, H. P., & Hall, L. K. (2005). The importance of retrieval failures to long-term retention: A metacognitive explanation of the spacing effect. *Journal of Memory and Language*, *52*, 566–577. doi: 10.1016/j.jml.2005.01.012

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. doi: 10.1111/j.1467-9280.1993.tb00571.x

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review*, *6*, 57–86. doi: 10.3758/BF03210812

Bird, S. (2010). Effects of distributed practice on the acquisition of second language English syntax. *Applied Psycholinguistics, 31,* 635-650. doi: 10.1017/S0142716410000172

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.

Brainerd, C. J., Reyna, V. F., Howe, M. L., Kingma, J., & Guttentag, R. E. (1990). The Development of Forgetting and Reminiscence. *Monographs of the Society for Research in Child Development, 55*, 1–109. doi: 10.2307/1166106

Brehmer, Y., Li, S.-C., Müller, V., von Oertzen, T., & Lindenberger, U. (2007). Memory plasticity across the life span: Uncovering children's latent potential. *Developmental Psychology, 43*, 465-478. doi: 10.1037/0012-1649.43.2.465

Cepeda, N. J., Coburn, N., Rohrer, D., Wixted, J. T., Mozer, M. C., & Pashler, H. (2009). Optimizing distributed practice: Theoretical analysis and practical implications. *Experimental Psychology*, *56*, 236–246. doi: 10.1027/1618-3169.56.4.236

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.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2.th ed.). Hillsdale, NJ: Erlbaum.

Delaney, P. F., Verkoeijen, P. P. J. L., & Spirgel, A. (2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. In B. H. Ross (Ed.), *Psychology of Learning and Motivation: The psychology of learning and motivation: Advances in research and theory* (pp. 63–147). Academic Press.

Dempster, F. N. (1988). The spacing effect: A case study in the failure to apply the results of psychological research. *American Psychologist*, *43*, 627–634. doi: 10.1037/0003-066X.43.8.627

Erdfelder, E., Auer, T.-S., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models: A review of the literature. *Zeitschrift für Psychologie / Journal of Psychology*, *217*, 108–124. doi: 10.1027/0044-3409.217.3.108.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149–1160. doi: 10.3758/BRM.41.4.1149

Gathercole, S. E., Lamont, E., & Alloway, T. P. (2006). Working Memory in the Classroom. In Susan J. Pickering (Ed.), *Working Memory and Education* (pp. 219–240). Burlington: Academic Press.

Gathercole, S. E., Pickering, S. J., Ambridge, B., & Wearing, H. (2004). The Structure of Working Memory From 4 to 15 Years of Age. *Developmental Psychology, 40*, 177–190. doi: 10.1037/0012-1649.40.2.177

Glenberg, A. M. (1979). Component-levels theory of the effects of spacing of repetitions on recall and recognition. *Memory & Cognition*, *7*, 95–112. doi: 10.3758/BF03197590

Glenberg, A. M., & Lehmann, T. S. (1980). Spacing repetitions over 1 week. *Memory & Cognition, 8*, 528-538. doi: 10.3758/BF03213772

Grote, M. G. (1995). Distributed versus massed practice in high school physics. *School Science and Mathematics, 95,* 97-101. doi: 10.1111/j.1949-8594.1995.tb15736.x

Hattie, J. (2009). *Visible Learning: A synthesis of over 800 meta-analyses relating to achievement.* New York, NY: Routledge.

Hogan, R. M., & Kintsch, W. (1971). Differential effects of study and test trials on long-term recognition and recall. *Journal of Verbal Learning and Verbal Behavior*, *10*, 562–567. doi: 10.1016/S0022-5371(71)80029-4.

Kornell, N. (2009). Optimising learning using flashcards: Spacing is more effective than cramming. *Applied Cognitive Psychology, 23*, 1297-1317. doi: 10.1002/acp.1537

Küpper-Tetzel, C. E., & Erdfelder, E. (2012). Encoding, maintenance, and retrieval processes in the lag effect: A multinomial processing tree analysis. *Memory*, *20*, 37-47. doi: 10.1080/09658211.2011.631550

Lawrence, W. W., & McPherson, D. D. (2000). A comparative study of block scheduling and traditional scheduling on academic achievement. *Journal of Instructional Psychology, 27,* 178-182.

Moshagen, M. (2010). multiTree: A computer program for the analysis of multinomial processing tree models. *Behavior Research Methods*, *42*, 42–54. doi: 10.3758/BRM.42.1.42.

Mozer, M. C., Pashler, H., Cepeda, N. J., Lindsey, R., & Vul, E. (2009). Predicting the optimal spacing of study: A multiscale context model of memory. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems* (pp. 1321–1329). La Jolla, CA: NIPS Foundation.

Neisser, U. (1976). *Cognition and reality: Principles and implications of cognitive psychology.* New York, NY US: Freeman.

Pashler, H., Rohrer, D., Cepeda, N. J., & Carpenter, S. K. (2007). Enhancing learning and retarding forgetting: Choices and consequences. *Psychonomic Bulletin & Review, 14*, 187-193. doi: 10.3758/BF03194050

Pressley, M., & Hilden, K. (2006). Cognitive Strategies. In D. Kuhn, R. S., Siegler, W. Damon, & R. M. Lerner (Eds.), *Handbook of child psychology: Vol 2, Cognition, perception, and language* (pp. 511-556). Hoboken, NJ, US: John Wiley & Sons Inc.

Randler, C., Kranich, K., & Eisele, M. (2008). Block scheduled versus traditional biology teaching-an educational experiment using the water lily. *Instructional Science, 36,* 17-25. doi: 10.1007/s11251-007-9020-y

Rawson, K. A., & Kintsch, W. (2005). Rereading effects depend on time of test. *Journal of Educational Psychology, 97,* 70-80. doi: 10.1037/0022-0663.97.1.70

Reynolds, J. H., & Glaser, R. (1964). Effects of repetition and spaced review upon retention of a complex learning task. *Journal of Educational Psychology*, *55*, 297–308. doi: 10.1037/h0040734.

Rohrer, D., & Taylor, K. (2006). The Effects of overlearning and distributed practise on the retention of mathematics knowledge. *Applied Cognitive Psychology, 20*, 1209–1224. doi: 10.1002/acp.1266.

Rohrer, D., & Taylor, K. (2007). The shuffling of mathematics problems improves learning. *Instructional Science, 35*, 481-498. doi: 10.1007/s11251-007-9015-8

Seabrook, R., Brown, G. D., & Solity, J. E. (2005). Distributed and massed practice: From laboratory to classroom. *Applied Cognitive Psychology*, *19*, 107–122. doi: 10.1002/acp.1066.

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

Thios, S. J., & D'Agostino, P. R. (1976). Effects of repetition as a function of study-phase retrieval. *Journal of Verbal Learning & Verbal Behavior*, *15*, 529–536. doi: 10.1016/0022-5371(76)90047-5

Thomson, D. M., & Tulving, E. (1970). Associative encoding and retrieval: Weak and strong cues. *Journal of Experimental Psychology*, *86*, 255–262. doi: 10.1037/h0029997.

Appendix

List of vocabulary word pairs and distractor words

	Cue word	Target word	Distractor words in recognition test	
1	Burg	castle	cartel	cattle
2	Dieb	thief	belief	chief
3	Eisenbahn	railway	doorway	motorway
4	Engel	angel	bangle	tangle
5	Feuer	fire	dire	wire
6	Fluss	river	diver	liver
7	Frosch	frog	fog	food
8	Fuchs	fox	box	lox
9	Handtuch	towel	tower	town
10	Hof	yard	dart	lard
11	Holz	wood	rood	good
12	Hügel	hill	bill	pill
13	Kehle	throat	road	goat
14	Küste	coast	coach	coal
15	Kuh	cow	low	row
16	Landkarte	map	gap	nap
17	Mauer	wall	call	mall
18	Mond	moon	mood	noon
19	Müll	rubbish	rubber	rumbler
20	Schaf	sheep	deep	sleep
21	Spiegel	mirror	marrow	narrow
22	Stein	stone	alone	clone
23	Stern	star	staff	starch
24	Suppe	soup	group	loup
25	Träne	tear	deer	gear
26	Zucker	sugar	sucker	suffer