# Spacing, Feedback, and Testing Boost Vocabulary Learning in a Web Application

Angelo Belardi[a], Salome Pedrett[a], Nicolas Rothen[a], Thomas P. Reber[a,b]

[a] *Faculty of Psychology, Swiss Distance University Institute, Brig, Switzerland*
[b] *Department of Epileptology, University of Bonn, Bonn, Germany*

**Abstract**

As information and communication technology (ICT) becomes more prevalent in education its efficacy in general and that of specific learning applications in particular has not been fully established yet. One way to further improve learning applications could be to use insights from fundamental memory research. We here assess whether four established learning principles (spacing, feedback, testing, and multimodality) can be translated into an applied ICT context to facilitate vocabulary learning in a self-developed web application. Effects on the amount of newly learned vocabulary were assessed in a mixed factorial design ($3\times2\times2\times2$) with the independent variables Spacing (between-subjects; one, two, or four sessions), Feedback (within-subjects; with or without), Testing (within-subjects, 70% or 30% retrieval trials), and Multimodality (within-subjects; unimodal or multimodal). Data from 79 participants was analyzed and revealed significant main effects for Spacing ($F[2, 76] = 8.51$, $p = 0.0005$, $\eta_p^2 = 0.18$) and Feedback ($F[1, 76] = 21.38$, $p < 0.001$, $\eta_p^2 = 0.22$), and a significant interaction between Feedback and Testing ($F[1, 76] = 14.12$, $p = 0.0003$, $\eta_p^2 = 0.16$). Optimal Spacing and the presence of corrective Feedback in combination with Testing together boost learning by 29% as compared to non-optimal realizations (massed learning, testing with lack of corrective feedback). Our findings indicate that established learning principles derived from basic memory research can successfully be implemented in web applications to optimize the acquisition of new vocabulary.

*Keywords:* distance education, online learning, web application, memory, language learning

## 1. Introduction

Information and communication technology (ICT) changes how we access information and the way we learn. Smartphones, tablets, and desktop-computers become ubiquitous in living rooms and classrooms, transforming how learners of all ages perceive and interact with learning material. Identifying how ICT may improve learning is vital to ensure successful adaptation of educational practices for the digital age (Sung et al., 2016). In the current work, we investigate this general question by addressing the following specific research gaps: 1) Can some of the best researched learning principles originating from basic memory research be applied to optimize computer-assisted learning environments? 2) How do these learning principles interact? We do that in the setting of vocabulary learning because it is a central task in classes of foreign languages in schools. Vocabulary learning lends itself well to assess these questions because the transfer between basic memory research and its application seems rather close: Vocabulary learning essentially entails long-term storage of memories for paired-associates (i.e., a word and its associated translation in the foreign language), a well-researched phenomenon in the domain of basic memory research (Steinel et al., 2007).

A meta-meta-analysis in 2011 summarized the findings of 25 meta-analyses and found a small to moderate effect favoring the use of computer technology in the classroom to support teaching and learning, but also great variability among the results (Tamim et al., 2011). The efficacy of interactive learning applications on computers and mobile devices to improve learning in classrooms is similarly promising and unclear (Sung et al., 2016). Some meta-analyses report beneficial effects initially that fade after 6 to 12 month of continuous ICT use (Sung et al., 2016; Cheung and Slavin,

2013). This fading of effects might be due an initial boost of motivation and novelty when ICT is first introduced (Sung et al., 2016; Cheung and Slavin, 2013).

As plans to invest more resources for ICT in classrooms emerge (Roediger and Pyc, 2012; European Commission, 2019; Futuresource Consulting Press, 2016), more research is needed that investigates not just whether but *how* ICT can be successfully applied in education. There is high potential for studies that invest into a deeper understanding of the sources for the variation of outcomes reported earlier, and that try to identify individual features of apps and learning situations contributing to successful implementations of ICT in education.

One approach to scrutinize underlying mechanisms of ICT success is to take the perspective of a researcher interested in fundamental memory processes. This perspective has surprisingly seldom been taken to the extent that it influenced common educational practices in regular classrooms or digital learning applications (Roediger and Pyc, 2012; Reber and Rothen, 2018). Therefore, we recently proposed to focus on the research of four established learning principles known to facilitate learning in laboratory situations which are also straightforward to implement in digital learning applications (Reber and Rothen, 2018).

Probably the most research-backed of the four principles is derived form the *spacing effect* (also distributed practice effect or distributed learning). Spacing refers to splitting up the learning time into several short sessions and distributing them over time (Carpenter et al., 2012; Kornell et al., 2010). Learning is improved when we space out the learning time into separate distributed sessions, in contrast to cramming it into one session, also called massing (for reviews, see Cepeda et al., 2006; Benjamin and Tullis, 2010; Delaney et al., 2010).

A second principle concerns giving *corrective feedback* about mistakes in comparison to no feedback or simple right/wrong feedback (Metcalfe, 2017). To be effective, feedback must include the correct answer, merely stating whether the answer was correct or not is not sufficient (Pashler et al., 2005). From a cognitive perspective, corrective feedback leads to a "prediction-error" signal in the brain (Wilkinson et al., 2014), which in turn catalyzes learning by switching brain regions relevant for long-term memory into a more receptive encoding rather than retrieval mode (Lisman and Grace, 2005; Greve et al., 2017).

The third principle builds on the *testing effect* (also test-enhanced learning or retrieval practice). When people have to reproduce or answer questions about the studied material, they remember more than when they study it repeatedly (for a review, see Rowland, 2014). Better performance due to testing has been explained on one hand by the *transfer-appropriate processing framework*, which posits that memory is better when learning and test situations are similar rather than different (Morris et al., 1977). That is, being able to recall information is more likely when recalling information was practiced in comparison to restudied. On the other hand, testing situations afford more effort which may lead to deeper encoding of material according to the *desired difficulties framework* (Bjork, 1994; Bjork and Kroll, 2015).

Finally, presenting the learning material *multimodally*, i.e. to multiple senses simultaneously, benefits learning as well (Shams and Seitz, 2008; Kast et al., 2007, 2011). Multimodal presentation is inarguably closer to how we perceive the world and learn everyday, without deliberate effort (incidental learning), than to present learning material for only one sensory channel. Furthermore, e.g. audiovisual presentations of learning materials recruit larger regions of the brain - namely the ones processing auditory *and* the ones processing visual information - as compared to unimodal presentations (auditory or visual stimuli alone). These "many routes" (Bjork, 1975) by which a stimulus is processed for encoding are then thought to also facilitate retrieval by making use of redundant information stored in distributed brain regions (Murray and Sperdin, 2010).

While extensive data on these four principles exist, few studies assessed how these principles interact (Weinstein et al., 2018). A notable exception is a study by Cull (2000) which looked at the interaction between spacing and testing. In a word pair learning task, testing improved learning success and this effect was even higher when the learning time was spaced beyond mere addition of these two main effects (Cull, 2000, experiment 1).

This is also interesting considering that popular language learning and general learning tools available online already implement some learning principles we investigated: Duolingo (`www.duolingo.com`), Rosetta Stone (`www.rosettastone.com`), Memrise (`www.memrise.com`), Anki (`https://apps.ankiweb.net`), and Quizlet (`www.quizlet.com`) for example all implement testing, feedback,

and multimodality in some way or another. The flashcard-style learning applications Anki and Quizlet both further implement spacing based on the so-called Leitner system, which is an algorithm to space and prioritize flash-cards (Godwin-Jones, 2010). Duolingo applies a self-developed procedure for spaced repetition using Half-Life Regression (Settles and Meeder, 2016). Please note that what we mean in the context of this manuscript by "spacing" is slightly different from "spacing" in the Leitner system. We refer to spacing of individual learning sessions, whereas the mentioned spacing algorithms refer to the scheduling of individual learning items within and across individual learning sessions.

The above learning principles were mostly researched using traditional learning methods (no use of ICT) in laboratory or classroom settings. The purpose of the current study is therefore to investigate whether these principles also improve learning efficiency in the context of a web application in a home environment. A further aim is to explore whether and how these principles interact with each other. Such interactions have seemingly not been investigated in language learning applications before.

We implemented a web application that allows for independent variation of presence or absence and/or parametrization of all four learning principles. German-speaking participants used the app to learn foreign (Finnish) language vocabulary and we tested their recall of the learned word pairs two days after their last learning session.

Our main research question was: "Can established learning principles be used to optimize learning of vocabulary with a web application?" (RQ1). Consequently, our hypotheses were: Learning success, as measured in a cued recall test, is improved when...

1. the time spent learning is spaced vs. massed (H1).
2. corrective feedback is given vs. no feedback is given (H2).
3. more testing/retrieval trials are presented for a specific word pair (H3).
4. stimuli were presented multimodally vs. unimodally (H4).

Our second research question (RQ2) was : "Are there any pairwise interactions between these principles?".

## 2. Material and Methods

### *2.1. Design*

The study was set up as a $3\times2\times2\times2$ mixed factorial design with the independent variables *Spacing* (between-subjects, one, two, and four learning sessions), *Feedback* (within-subjects, with and without feedback), *Testing* (within-subjects, proportions of retrieval and learning trials were 70%/30% or 30%/70%), and *Multimodality* (within-subjects, unimodal [visual only] and multimodal [audiovisual]). The dependent variable was the proportion of correct translations recalled in the testing session. Additionally, we varied the independent variable *learning direction* (within-subjects) in the learning phase, and the independent variable *testing direction* (within-subjects) in the test phase.

### *2.2. Participants*

Participants were recruited among friends and acquaintances of the students in a class on experimental research in the bachelor's program in psychology, conducted at the Swiss Distance University Institute, in the autumn semester of 2018. Bachelor-students in psychology acted as experimenters ($N_e = 22$) and recruited a total of $N_p = 87$ participants. Participants received no compensation for taking part in the experiment, but a small thank-you gift was made by some of the experimenters. Participants gave written informed consent.

The final analyses were conducted with data from 79 participants (43 female, 30 male, 6 did not declare their gender; age ranged between 16 and 77 years [$M = 39.7, SD = 15.5$]). We excluded 8 participants according to the following criteria: 3 had not completed the learning phase, 3 had not adhered to the scheduled gaps between sessions required for proper operationalization of the spacing factor, 1 had a long gap (over 20 min) during the testing session, and for 1 participant age data was not available.

All participants were either native German speakers or had mastered the language to at least an advanced degree (73 natives, 4 near-native, 1 proficient, 1 advanced). None of the participants reported any previous knowledge of the Finnish language, though 8 did not respond to this question. Neither of the participants indicated knowledge in any language closely related to Finnish, such as Hungarian or Estonian.

As is common practice in the literature of language learning and teaching, we use the terms *L1* and *L2* to refer to the native language (German) and foreign language (Finnish), respectively. Although for some of the participants, German is not the native but a language in which they are advanced, we apply the term L1 also to them for simplicity.

Fourteen participants held a bachelor's degree, 21 a master's degree, 25 had finished an apprenticeship, 8 held a higher education entrance qualification, 5 had finished compulsory education, and 6 reported another form of education (or were still in school).

### *2.3. Materials*

#### *2.3.1. Stimuli*

We used 48 Finnish-German word pairs as stimuli, gathered from various lists of frequently used words in Finnish and English, lyrics of Finnish pop songs, and words from a Finnish online dictionary. We filtered an initial list of 250 words and removed Finnish words that seemed too similar to a German word, ambiguous terms, compound words, interrogatives, personal pronouns (which were difficult to translate), and terms that were subjectively too complicated or too simple. This resulted in 214 stimuli, out of which the final set of 48 word pairs was selected randomly. A list of all stimuli is available in the Supplement to this article (see Table S7). Audio files for the stimuli were created with the text-to-speech software Balabolka (v. 2.14.0.676, Ilya Morozov).

#### *2.3.2. Learning phase*

For the learning phase, we developed a web application written in the R programming language (R Core Team, 2018), and we used the "shiny" (Chang et al., 2018) and "ShinyPsych" (Steiner et al., 2018) packages. We hosted the applications with the open source version of "Shiny Server" (v. 1.5.9.923, RStudio, Boston, MA, U.S.) on a self-administered virtual server running Linux Debian (v. 4.9.110).

For the implementation of the between-subjects factor Spacing, we kept the overall learning time equal for all participants, but split it into either one learning session of 80 minutes, two sessions of 40 minutes, or four sessions of 20 minutes. The sessions were further split into 20-minute learning blocks. Thus, every participant conducted four 20-minute learning blocks altogether. The levels of the variable Spacing (1, 2, or 4 sessions) were equally distributed among the experimenters, but the participant allocation to the levels was not done entirely at random: While most participants were allocated randomly, the rigid scheduling of several sessions would have made it impossible for some to participate. Thus, before participants knew about the exact content and procedure of the experiment, their preferences were considered in regard to having two, three, or five sessions with the experimenter, as the experiment entailed either one, two, or four learning sessions and always one additional test session. The participants did not know beforehand that they would be learning the same vocabulary regardless of the number of learning sessions. Furthermore, they did not know any specifics about the experiment or study procedure at the point on which the sessions were scheduled and when they generally expressed interest in participation to the experiment. The overall procedure was only explained to them after scheduling the sessions, during the first learning session. After participants knew about that, they could still decide not to participate in the study, but they could then not change the Spacing condition to which they were allocated.

The 48 word pairs were assigned to the 8 factor-combinations of the three within factors (Feedback, Testing, Multimodality). This assignment was randomized for each participant. The condition to which a word pair belonged did not change during the learning phase.

Testing varied in the proportion of *learning* vs. *retrieval* trials. *Learning trials* entailed the presentation of a word pair in both languages and *retrieval trials* entailed presentation of one word as cue (German or Finnish) and an input field in which participants were prompted to input the translation of the word. The two types of trials are illustrated with screenshots, available in the Supplement to this publication (Figure S8). Each word pair was presented in several trials during the whole learning phase. Each trial was either a learning or retrieval trial. Among all trials of one word pair, the proportion of retrieval trials and learning trials was set to either "70% of retrieval trials and 30% of learning trials" or '30% of retrieval trials and 70% of learning trials'.

Feedback was varied in that corrective feedback was provided for some translations but not for others. Feedback entailed showing the correct solution along with the cue word and the answer given by the

participant, after a participant entered an incorrect answer to a retrieval trial. If the answer was correct, the feedback was "correct" and the participant could proceed to the next trial.

Multimodality entailed trials with multimodal vs. unimodal presentation. For multimodal (audio-visual) stimulus presentation, a word was displayed in either German or Finnish while an audio file of the word spoken by a computer voice was played simultaneously in the same language. In learning trials, the audio recording was played only for the word displayed on top of the screen, not for the translation in the other language shown below. In unimodal trials, no audio recording was played.

We controlled for potential effects of learning direction. Here, a word pair could either be learned in the direction from L1 (German) to L2 (Finnish) or the other way around (*L2-to-L1*). In learning trials, the first word was at the top of the screen, and the second word was below. In retrieval trials, the first word was at the top, and the input field into which the participants could enter the translation was below.

#### *2.3.3. Test phase*

Like the learning application, the test application was programmed in R. This application was used during the testing session and displayed only retrieval trials. We varied the independent variable *testing direction*: Each of the 48 word pairs was tested once in either direction (L1-to-L2 and L2-to-L1), resulting in 96 trials. Participants received no feedback on individual trials.

#### *2.3.4. Questionnaires*

We further created a questionnaire with LimeSurvey (v. 3.14.3+180809, LimeSurvey GmbH, Hamburg, Germany) to assess sociodemographic information and motivation. To measure motivation, we applied the Questionnaire on Current Motivation (QCM) in its German version "Fragebogen zur Erfassung aktueller Motivation in Lern- und Leistungssituationen" (FAM; Rheinberg et al., 2001).

### *2.4. Procedure*

For the individual learning and testing sessions, one experimenter met individually with one participant at a time. Experimenters followed a written guideline (available in the supporting materials online repository at `https://osf.io/djxmr`). The experiment was either conducted at the experimenter's or the participant's home and participants could use their own computer or one provided by the experimenter. There was an exception for four experimenters (and thus 16 participants), who were allowed to test their participants without being physically present. Instead, they kept contact with the participants via Skype during the experiment on an additional device. Whether participants were tested remotely using Skype or not had no effect on the conclusions of the experiment, as analyses excluding these participants were virtually identical to the main analyses presented in the Results section (Supplement Table S6).

The experiment started with the online questionnaire. Next, the participants began learning with the web application in their first study session. The first screen contained information on how to interact with the application and a query to check the audio settings.

Depending on the level of Spacing, the participants performed multiple 20-minute blocks in the same learning session and could take short breaks in-between (about 5-10 minutes). Within each 20-minute block, there were three phases of equal duration, during which a set of 16 word pairs was learned, one word for each of the 8 within-subject factor-combinations, in both learning directions. The three sets were presented in the same order in all 20-minute blocks.

At the beginning of each trial, a word pair was randomly chosen from the active set of 16 word pairs. If the word pair was chosen for the first time, it was presented as learning trial; otherwise, it was presented as learning or retrieval trial with a probability according to the Testing condition. Learning trials proceeded by button press or mouse click; retrieval trials by submitting a response via keyboard, followed by feedback depending on the condition.

Overall, all participants learned the same 48 word pairs and had a total learning time of 80 minutes. Due to randomized presentation of word pairs, the number of trials per word pair and participant varied (mean = 22.3 [SD = 0.687]), but an ANOVA showed that while there were slight differences in the number of trials between the factor conditions, those are unlikely to account for our results (see Supplement Tables S9 and S10).

Between learning sessions and between the last learning session and the testing session, a gap of two days was planned and the actual mean gap time ranged from 42.8 h to 77.2 h between subjects

(Med = 69.7, IQR = 22.8). We will refer to the gap between individual learning sessions as *inter-study interval* (ISI) and to that between the last learning session and the testing session as *retention interval* (Cepeda et al., 2006).

In the testing session, participants conducted a cued recall test of all learned translations using the testing application. The 96 trials were presented in randomized order (differently for each participant). The testing session ended as soon as the participant answered all 96 trials.

## 3. Results

### *3.1. Learning principles*

To assess the effects of the four learning principles on recall performance during the testing session, we conducted a four-way $3\times2\times2\times2$ mixed design analysis of variance (ANOVA) with factors Spacing (1, 2, or 4 learning sessions), Feedback (with or without feedback), Testing (70% retrieval trials or 30%), and Multimodality (unimodal or multimodal; Figures 1, 2). Dependent variable was the proportion of correctly recalled words during the testing session.

We found a main effect for the factor Spacing ($F[2, 76] = 8.51, p = .0005, \eta_p^2 = 0.18$). In support of H1, participants in which learning was distributed the most (4 sessions of 20 min each) had the highest recall performance ($M = 77.2, SD = 29$). Performance was intermediate in participants who learned during two sessions of 40 minutes each ($M = 59.1, SD = 35.3$). Lowest performance was recorded in participants in the massed learning condition (1 learning session of 80 minutes; M = 52.5, SD = 37; see Figure 1A). We therefore conclude that spacing of learning episodes is also beneficial when using web-applications to learn vocabulary. The ANOVA also revealed a main effect of Feedback ($F[1, 76] = 21.38, p < .0001, \eta_p^2 = 0.22$; Figure 1B). As hypothesized (H2), the recall performance was higher ($M = 65.6, SD = 35$) on translations to which corrective feedback was provided in the learning phases than on translations without feedback during learning ($M = 60.4, SD = 35.9$). Corrective feedback is hence a further beneficial ingredient for the design of vocabulary learning apps.

To our surprise, the main effects for the factors Testing (H3) and Multimodality (H4) were both found to be insignificant ($F_{Testing}[1, 76] = 0.31, p_{Testing} = .58, \eta^2_{pTesting} = 0.004; F_{Multimodality}[1, 76] = 0.26, p_{Multimodality} = .61, \eta^2_{pMultimodality} = 0.003$; Figure 1 C&D). One potential explanation for the absence of an effect of testing may arise from considering the two-way interactions of the ANOVA (RQ2). Here, a significant interaction between the factors Testing and Feedback was found ($F[1, 76] = 14.12, p = .0003, \eta_p^2 = 0.16$). Recall performance is higher in the feedback vs. no-feedback condition only when combined with a high rate of retrieval trials (0.7) administered during learning. No such difference is found for a low rate of retrieval trials (Figure 2E). Thus, as there is no main effect of Testing, it appears that Testing nevertheless improves learning performance, but only in situations when Testing is combined with Corrective Feedback. No other two-way interaction reached significance. For an overview of all effects in the ANOVA, see Table 1.

Rather than merely looking at the statistical significance, we think that specifically in an applied context it is crucial to consider the effect sizes. Spacing led to 24.7% higher recall when participants learned in four spaced sessions instead of in one massed session. Feedback increased recall by 5.2%. Due to the combination of feedback and testing, recall gained another 5.8%. The optimal combination of factors levels was four learning sessions, feedback, and 70% retrieval trials. The observed means of our sample show that this combination and the one with 30% retrieval trials were at the top, with almost identical values of 78.2% and 78.7% correctly recalled words. The least efficient combination for learning consisted of one learning session, no feedback, and 70% retrieval trials and led to 49.5% recall. The difference between the observed best and worst combination was thus a boost of 29%.

### *3.2. Exploratory analyses: Direction of learning and direction of testing*

Each word pair was learned in one direction, either L1-to-L2 or L2-to-L1. Furthermore, since during our testing sessions participants performed the recall task in both directions, we could also assess effects of testing direction and the interaction between learning direction and testing direction. This resulted in an additional ANOVA with six factors, adding learning direction and testing directions to the model.

Learning direction had a substantial effect on recall performance ($F[1, 76] = 28.61, p < .0001, \eta_p^2 =$

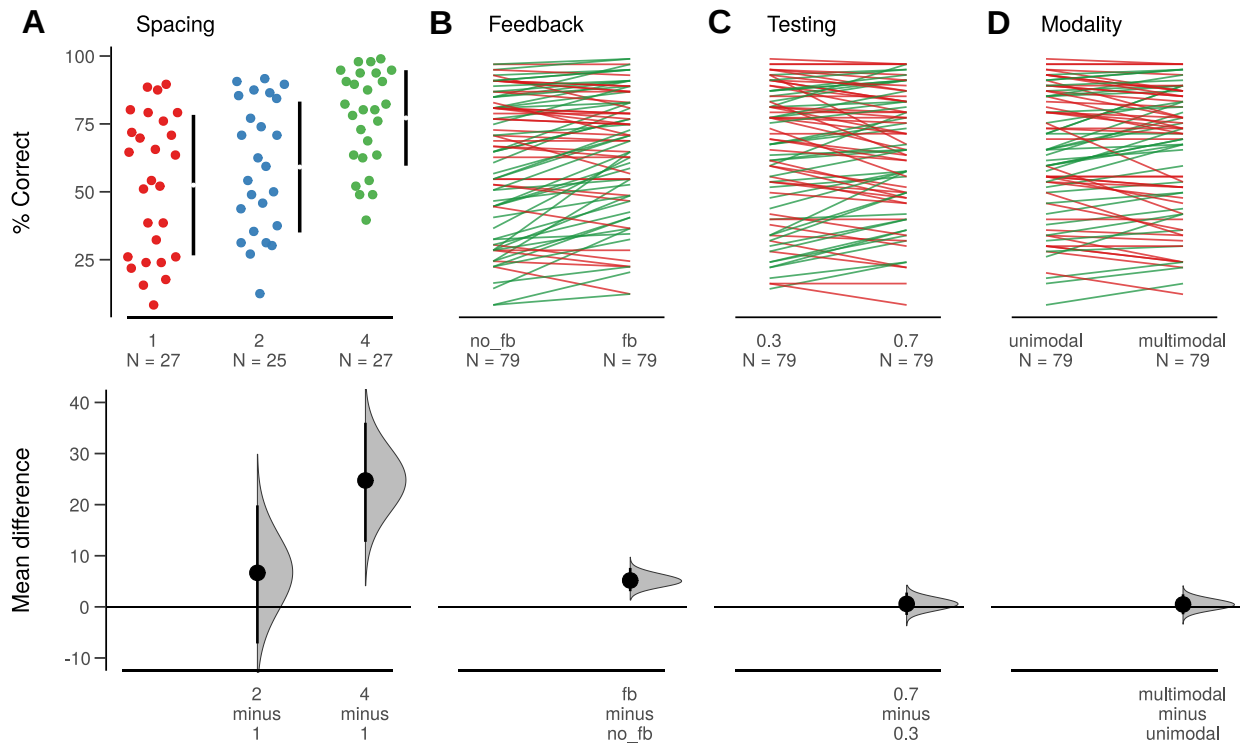


Figure 1: Estimation plots for the learning principles via Cumming plots. Upper row shows individual participant data in a swarmplot for unpaired data and a slopegraph for paired data. For unpaired data the mean ± standard deviation are shown as gapped lines. Lower row shows unpaired or paired mean differences as a bootstrap sampling distribution, with the dot indicating the mean difference and the ends of the error bars the 95% confidence interval.

0.27, see Figure 3A), where the words which participants learned in the direction L1-to-L2 ($M = 66.2, SD = 34.4$) were recalled better than those in the direction L2-to-L1 ($M = 59.9, SD = 36.3$). Adding learning direction to the design features described above we observe a difference of 38% between best and worst combinations of features of the learning app (see Table 2). Regarding testing direction, recall performance was generally higher for the direction L2-to-L1 ($M = 72.0, SD = 31.9$) as compared to L2-to-L1 ($M = 54.0, SD = 36.7, F[1, 76] = 233.38, p < .0001, \eta_p^2 = 0.75$).

We further found an interaction between learning direction and testing direction ($F[1, 76] = 105.74, p < .0001, \eta_p^2 = 0.58$, see Figure 3C), given that words which had been learned in the direction L2-to-L1 were recalled much better when the testing direction matched. For words learned in the direction L1-to-L2, the recall difference was much smaller and recall was actually higher when the testing direction did not match. The complete results table for this exploratory analysis is available in the Supplement (Table S1).

### *3.3. Covariates age, number of trials, and motivation*

We checked the influence of the potential covariates age, number of trials, and motivational factors, which could have effected the main results of the learning principles. As learning and memory is linked to aging, performance in cued recall tasks usually dwindles with higher age (e.g. Park et al., 1996). To control for potential age effects, given that we had a broad variance of age in our sample (range from 16 to 77 years), we ran an ANCOVA adding age as a covariate to our main model of learning principles. There was a significant age effect ($F[1, 73] = 21.68, p < .0001, \eta_p^2 = 0.23$), but the main results of the learning principles remained virtually identical when controlling for age. The complete results table of this model is available in the Supplement (Table S2).

The number of trials the participants saw during their learning sessions depended on how quickly

Table 1: ANOVA learning principles

| *Effect* | *df* | *MSE* | *F* | $\eta_p^2$ | *p-value* |
|---|---|---|---|---|---|
| spacing | 2, 76 | 4143.08 | 8.51 | .18 | .0005 *** |
| modality | 1, 76 | 156.01 | 0.26 | .003 | .61 |
| spacing:modality | 2, 76 | 156.01 | 0.50 | .01 | .61 |
| testing | 1, 76 | 185.77 | 0.31 | .004 | .58 |
| spacing:testing | 2, 76 | 185.77 | 1.85 | .05 | .16 |
| feedback | 1, 76 | 204.98 | 21.38 | .22 | <.0001 *** |
| spacing:feedback | 2, 76 | 204.98 | 3.11 | .08 | .05 |
| modality:testing | 1, 76 | 87.46 | 0.37 | .005 | .54 |
| spacing:modality:testing | 2, 76 | 87.46 | 1.74 | .04 | .18 |
| modality:feedback | 1, 76 | 125.67 | 2.83 | .04 | .10 |
| spacing:modality:feedback | 2, 76 | 125.67 | 1.14 | .03 | .32 |
| testing:feedback | 1, 76 | 178.79 | 14.12 | .16 | .0003 *** |
| spacing:testing:feedback | 2, 76 | 178.79 | 1.21 | .03 | .30 |
| modality:testing:feedback | 1, 76 | 134.12 | 2.36 | .03 | .13 |
| spacing:modality:testing:feedback | 2, 76 | 134.12 | 1.28 | .03 | .28 |

*Note.* df = degrees of freedom, MSE = mean square error, $\eta_p^2$ = partial eta-squared effec size. Asterisks indicate statistical significance (* = p <0.05; ** = p <0.01; *** = p <0.001

they pressed the button to continue to the next trial (in learning trials) or entered their answers (in retrieval trials). Therefore, the number of trials varied substantially between subjects ($M = 1071, SD = 386$, range: 338-2213). To check whether the number of trials had an effect on recall itself and whether it influenced the findings of our main model, we ran another ANCOVA, adding number of trials (summed up over all learning sessions) as a covariate. The results were virtually identical with the age covariate: While we found a significant effect of number of trials ($F[1, 73] = 7.56, p = .008, \eta_p^2 = 0.09$), the other results remained similar to the main model. The results table of this model is available in the Supplement (Table S3).

Before the first learning session, we assessed motivation related to the learning task. Two of the motivational factors in the questionnaire we used, namely fear of failure and success seeking, are related to tasks described as *question-led fact learning*, a definition into which our vocabulary learning task seems to fit (Rheinberg et al., 2001). We consequently ran two additional models including each of these factors in turn as a covariate, but there were no significant effects of fear of failure ($F[1, 69] = 0, p = .95, \eta_p^2 < 0.0001$) or success seeking ($F[1, 62] = 0.13, p = 0.72, \eta_p^2 = 0.002$) and the general results were similar to those of the main model. You can find the complete results of the model in the Supplement (Tables S4 and S5).

## 4. Discussion

We investigated whether four learning principles (Spacing, Feedback, Testing, Multimodality) derived from fundamental memory research can be used to optimize a web-application in a real-world digital context for vocabulary learning. Varying the presence/absence or parameters of each of these principles independently, we find that Spacing and the presence of corrective Feedback & Testing together significantly boost learning by 29%. Our results hence demonstrate that informing the development of ICT applications with knowledge from basic memory research can significantly ameliorate their efficiency.

### *4.1. Learning principles*

*Spacing.* With an increased recall of approximately 25% due to Spacing (four learning sessions compared to one session), the effect we found seems remarkably large compared to the majority of previous spacing literature. It is not that extraordinary, though, when set besides specific studies with similar experiments. In their extensive review, Delaney et al. (2010) considered increases of merely 15% already as large spacing effects. However, this review contains results of verbal list learning tasks, not paired-associate learning tasks, which are

Table 2: Proportions of correctly recalled word pairs in combinations of Spacing, Feedback, Testing, and learning direction

| | | | Spacing | | |
|---|---|---|---|---|---|
| **Feedback** | **Testing** | **Learning direction** | *1* | *2* | *4* |
| *No feedback* | *30%* | *L2-to-L1* | 50.31 | 54.67 | 75.62 |
| *No feedback* | *70%* | *L2-to-L1* | **44.14** | 48.67 | 71.91 |
| *No feedback* | *30%* | *L1-to-L2* | 52.47 | 58.67 | 80.25 |
| *No feedback* | *70%* | *L1-to-L2* | 54.94 | 56 | 75.93 |
| *Feedback* | *30%* | *L2-to-L1* | 46.6 | 57.67 | 75.62 |
| *Feedback* | *70%* | *L2-to-L1* | 54.94 | 63 | 74.38 |
| *Feedback* | *30%* | *L1-to-L2* | 55.25 | 62.67 | 81.79 |
| *Feedback* | *70%* | *L1-to-L2* | 61.11 | 71.67 | **82.1** |

*Note.* Mean proportions of correctly recalled word pairs in specific factor combinations. Minimal and maximal values are set in bold font.

more similar to foreign language vocabulary learning (Steinel et al., 2007). When looking at the paired-associate literature or studies with vocabulary learning paradigms, we see comparable or even higher effect sizes than those we found:

One study which showed an even higher spacing effect was conducted by Bloom and Shuell (1981). In their experiment on French vocabulary learning by English native speakers, either in a massed learning session over 30 minutes or in three spaced sessions over 10 minutes on three successive days, they found 35% higher recall in the spaced condition in a test taken four days later. Sobel et al. (2011) also compared one 30 minutes learning session to three sessions lasting 10 minutes, but reported an increase of only about 13%, though with a much larger retention interval of five weeks and in a sample of fifth-graders (aged around 10 years). Cepeda et al. (2009, experiment 1) ran an experiment in computer-based Swahili vocabulary learning (40 words) with either massed learning or two learning sessions with inter-study intervals (ISIs) between 1 and 14 days and a retention interval of 10 days. They reported an almost 19% increase in recall with an ISI of one day compared to the massed condition, though this advantage decreased to 14% in the condition with ISI of two days. While these increases are smaller in comparison to our effect, this difference might come about because they only used two learning sessions. In fact, our gain from two spaced sessions compared to one session was 6.6%.

One experiment reported in the publication by Cull (2000, experiment 3) used similar conditions to our study (3 day retention interval; fixed ISI of 2 days; 4 learning sessions; first learning session lasted about 30 minutes; 40 word pairs; computer-based flash-card app) and found a difference of almost 50% in recall between the uniformly distributed and massed learning conditions. In contrast to our study, they used uncommon-common word pairs of their native language and not foreign language vocabulary and the learning time was not fixed.

Overall, the size of our spacing effect ranges within those of earlier studies working with paired-associate learning tasks and foreign language vocabulary (Bloom and Shuell, 1981; Cepeda et al., 2009), but were smaller than reported effects in studies without foreign vocabulary (Cull, 2000).

*Feedback.* Our Feedback factor led to 5.2% higher recall during the testing session. Whether and to which degree feedback supports learning, depends on what aspect of learning we focus on. In oral language production, for example, it is important that learners notice their mistakes. Techniques helping them to do so are thus vital (Mackey, 2006). Other important factors are the delay from making a mistake to getting feedback and the nature of the feedback (Metcalfe, 2017). Overall, making and using errors with appropriate feedback helps learners of all ages, according to Metcalfe (2017).

Only few studies investigated feedback in vocabulary learning and we found only one which used foreign language vocabulary (Pashler et al., 2005). This study assessed five different feedback conditions in English speakers who learned vocabulary of the Luganda language, but were unable to find a significant effect (Pashler et al., 2005). In Metcalfe et al. (2009), a vocabulary learning task is reported which used new or difficult vocabulary of a language the participants already spoke. This study focused on the difference between delayed and immediate

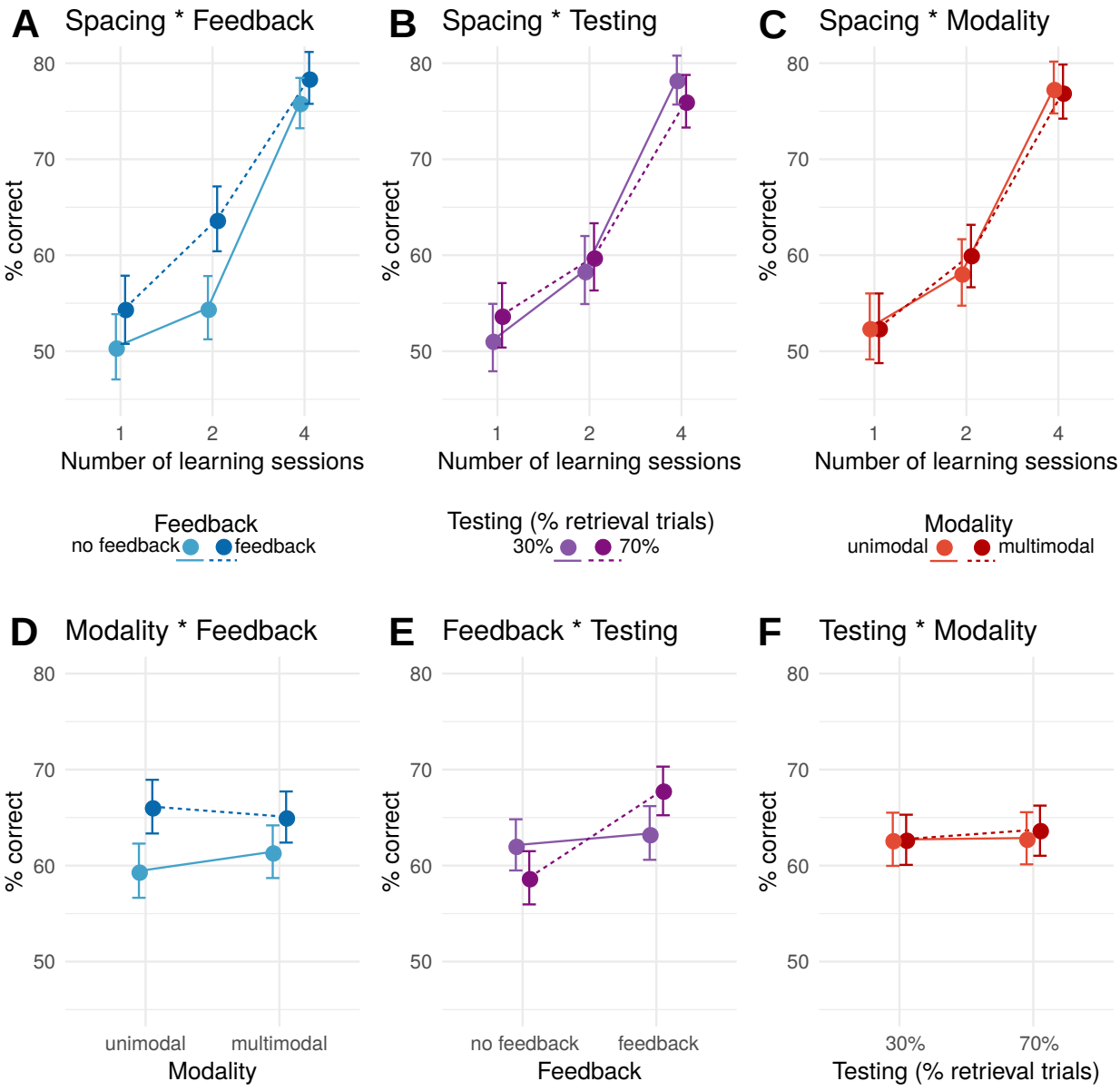


Figure 2: Pairwise interaction plots between learning principles. Significant interactions are highlighted. Values are offset horizontally to avoid overplotting (of error bars). Error bars indicate nonparametically bootstrapped 95% confidence intervals.

feedback, but also reported the increases in recall performance by immediate feedback, which is more comparable to our experiment, as 11% (experiment 1, with sixth-graders) and 18% (experiment 2, with college students).

*Testing.* In our results, Testing influenced performance in an interaction together with Feedback. Participants could thus probably only profit from retrieval trials when they received feedback. To explore the interaction between Feedback and Testing further, one could also incorporate more levels for each of these factors, for example an option with simple right/wrong feedback (non-corrective) or with a rewrite variant, where subjects have to write out the correct answer directly after they got the corrective feedback. This might lead to deeper

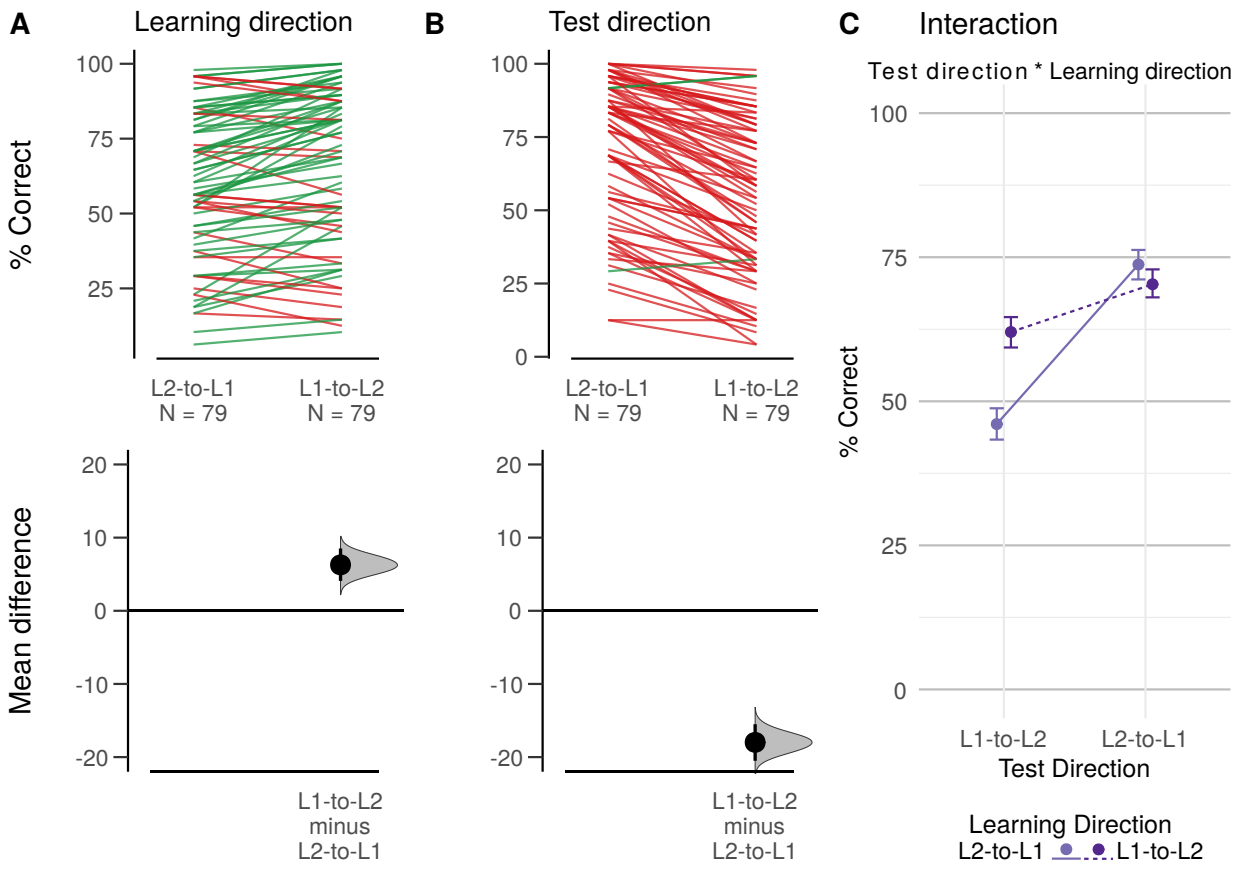


Figure 3: Estimation plots for the learning and testing direction (A&B) and interaction plot (C). Upper row shows individual participant data in a swarmplot for unpaired data and a slopegraph for paired data. Lower row shows paired mean differences as a bootstrap sampling distribution, with the dot indicating the mean difference and the ends of the error bars the 95% confidence interval. Values are offset horizontally to avoid over-plotting (of error bars) in the interaction plot. Error bars there indicate bootstrapped 95% confidence intervals.

processing of the feedback.

*Multimodality.* In our experiment, multimodality did not improve recall. It is important to note the difference between presentation mode and sensory modality: Presentation mode describes the code used to represent the information (e.g. verbal, non-verbal) and sensory modality denotes the sense through which participants perceive the information (Moreno and Mayer, 2007).

Moreno and Mayer (2007) stated that the most supportive learning environments combine both verbal and non-verbal representations of the learning materials, based on the modality principle of instructional design (e.g. Moreno and Mayer, 2002; Moreno, 2006; Mayer, 2014; Mayer and Fiorella, 2014; Low and Sweller, 2014). They also described a cognitive theory for multimedia learning, drawing on multiple cognitive theories (Mayer and Moreno, 2002). Adding a non-verbal representation of the vocabulary via images, should thus support learning even better than having two different modalities of verbal representation as in our experiment. Moreno and Mayer (2007) also cautioned against delivering both verbal and non-verbal stimuli through the same modality (e.g. written word and images), since this could overload the learners' cognitive capacity (Mayer and Fiorella, 2014; Low and Sweller, 2014). In accordance with this, we avoided presenting two forms of the stimuli through the same modality by targeting two different modalities (vision, audio). We did not, however, combine verbal and non-verbal presentation modes. That we had two different verbal representations of our word pairs might help to understand why we did

not observe the expected multimodality effect.

In the field of computer-assisted language learning, multimodality has been investigated since the 1990's, especially looking at different modalities of glosses and annotations to support vocabulary learning (e.g. Chun and Plass, 1996; Kim and Gilman, 2008; Yanguas, 2009). While glosses and annotations are not exactly the same as our simple word items, insights from these studies can add to the discussion. Chun and Plass (1996) for example found that combined text and image annotations outperformed those with text only, but adding videos did not. They also already emphasized the need to isolate the types of annotations in further studies and suggested the audio modality for further investigation. Kim and Gilman (2008) found further support for the use of images together with written definitions in vocabulary learning but their participants performed poorer when spoken text was added to written text instructions. They studied Korean learners of English and the authors theorized that the problem might be that their participants were used to learning new vocabulary without knowing the pronunciation and thus the additional information might have distracted rather than helped. This might not be transferable to native German speakers in Switzerland who are used to focusing on the pronunciation of new vocabulary in language classes. Yanguas (2009) reported no difference between text-only, image-only and combined text and image glosses. Overall, these mixed findings regarding multimodality are consistent with our results indicating no significant difference in the multimodality condition.

Widespread commercial language learning applications (e.g. Duolingo, Rosetta Stone) often apply multiple different modalities, most commonly audio clips and images, accompanying the written word. For beginning courses it is even common to use auditory and image representations by themselves – a combination which satisfies the above-mentioned suggestions for optimal learning environments described by Moreno and Mayer (2007). While this might help in vocabulary learning, completely replacing written information with audio recordings might not be useful for all learning materials, as Tabbers et al. (2004) report in a study in which learners who read the information outperformed those who had gotten the same information as audio clips.Further studies could also present vocabulary in three modalities at the same time. Dubois and Vial (2000) reported an advantage in language learning when adding a third medium, but also highlighted that an additional medium only helps if it encourages more in-depth processing.

### *4.2. Learning and testing direction*

Our exploratory analyses showed an effect of learning direction. Others suggested that the increased difficulty when learning in the direction L1-to-L2 leads to poorer learning performance initially (during the learning phase and when tested immediately after learning) but improves long-term retention several days later (Hummel, 2010; Steinel et al., 2007). The L1-to-L2 direction is more difficult because one has to completely reproduce a newly learned word from memory and to write it, compared to merely recognising the word and writing the translation in one's native language. Thus, learning in this direction requires deeper mental processing of new vocabulary, and is expected to benefit learning in the long run.

Schneider et al. (2002) reported an experiment with English native speakers who learned French vocabulary. The authors hypothesized that the more difficult the circumstances to learn, the better the retention. In support of their hypothesis, they found that participants who had learned in the direction L1-to-L2 recalled less in an immediate test, but then performed marginally better in a test one week later in comparison to those in the L2-to-L1 condition (Schneider et al., 2002, experiment 2). These results are in line with conclusions of another study that found the L1-to-L2 direction to be overall preferable when one learns for both, comprehension and production of the new vocabulary (Griffin and Harley, 1996). Our findings add new evidence for the advantage of the L1-to-L2 learning direction in a delayed recall test.

In line with our results, Steinel et al. (2007) found an interaction of learning and testing direction: the apparently more difficult learning direction (L1-to-L2) helped in later recall only if the testing direction matched. Together these results may suggest that when a word pair is studied the easier way (L2-to-L1), participants have a hard time recalling and producing the word correctly in the difficult direction (L1-to-L2).

### *4.3. Limitations*

Participants had some control over their allocation to levels of the spacing variable when the planned schedule of sessions did not fit with the

schedules of participants. As noted by a reviewer, we can thus not completely rule out the possibility that some subjects that knew of the beneficial effect of spacing on memory may have influenced scheduling of sessions such that they ended up in the more distributed conditions to maximize their learning outcome. However, three reasons speak against such a bias affecting our results. First, we think such highly informed subjects are rather the exception if they existed at all. Second, random allocation of subjects to conditions was only altered individually when problems arose in scheduling the participants to the learning sessions. These issues are more likely to happen in the more distributed learning conditions as there are more sessions to find a date for than in the massed condition. Thirdly, information given to participants about the experiment when scheduling the sessions were very limited such that the participants inference on the study design and research questions could have only been very limited as well. We thus think that the fact that subjects' allocation to the Spacing condition was not completely random is highly unlikely to affect the findings and conclusions of the study.

### *4.4. Future research*

Future studies could further investigate the role of the spacing time between learning sessions (lag effect) and between learning and testing sessions, in addition to the number of learning sessions (spacing effect). Effects of different ISI have already been dissected in a review by Cepeda et al. (2006), which focused on studies with verbal memory tasks and recall as performance measure. This review reported that longer ISI generally led to better recall and that for a given retention interval there is an optimal ISI, which increases with larger retention intervals. Thus, the relationship between ISI, retention interval, number of sessions, and other learning principles remains a large field open for scrutiny.

The purpose of all vocabulary learning is the retention of learned material for the long-term. In the current study, we assessed a rather brief retention period of two days which is also the case for much of the previous literature we are aware of. One fruitful direction for further research suggested by a reviewer is to investigate whether the learning principles and interactions between them differentially affect longer retention intervals of weeks or even months.

In this study, we focused on four learning principles which are in our perception of the literature very prominent and established and can easily be implemented in a web application (Metcalfe, 2017; Reber and Rothen, 2018; Roediger and Pyc, 2012). There are other factors which could potentially improve a specific learning application and it is one of our goals for future research to test the usability of more findings from basic memory research for this purpose, especially also in combination with one another.

### *4.5. Practical implications*

What are these findings telling us about how to develop learning apps? For one, we see that it is possible to use established learning principles in the context of online learning applications. It also tells us that this might not work as simply in all cases. While two of four implemented learning principles supported learning as planned and one did so in an interaction, the fourth did not.

One reason behind this might be that the learning principles can often be implemented in various ways. For example, feedback can be given immediately or delayed (Metcalfe et al., 2009); spacing can vary depending on the length of the ISI and retention interval or the number of learning sessions (Cepeda et al., 2006); and multimodality can be achieved by combining different modalities and also by changing the mode of representation. For any real-world application the optimal implementation or combination of these learning principles might consequently vary.

Thus, we are well-advised to implement, test, and then update the applications accordingly, possibly through several iterations, to develop better learning tools. That iterative process is what software developers have been doing for decades (e.g. Larman and Basili, 2003). However, one might argue that the features a software company wants to optimize might not be those teachers would optimize in a software they hand their pupils.

Apps, like any component of ICT, remain tools – they are not goals themselves. A successful application of such tools in the classroom or at home should be measured by outcomes like the degree to which it eases the workload of teachers and supports learners. Schools or governments can certainly profit from checking for scientifically tested tools, or if such data is not available, to have them tested, before acquiring and distributing them. Our study was motivated by this thought and shows

that turning to basic memory research to inform app-design can boost learning quite significantly.

### *4.6. Conclusion*

Three established learning principles, Spacing, corrective Feedback, and Testing in combination with Feedback improved vocabulary learning performance in the context of a web application with which German speakers learned Finnish vocabulary – a language with which they had no prior experience. Recall improved by 29% when participants could use the learning principles. These findings support our notion that knowledge from fundamental memory research can inform the development of learning applications to improve them.

## Declaration of interests

The authors declare no conflicts of interest.


## Acknowledgments

We thank all students of the class "M08: Experimentelle Übungen" at the Swiss Distance University Institute of the autumn semester 2018, who recruited the participants and carefully conducted the experiments.


## References


Benjamin, A. S. and Tullis, J. (2010). What makes distributed practice effective? *Cognitive Psychology*, 61(3):228–247, ISSN: 0010-0285, DOI: 10.1016/J.COGPSYCH.2010.05.004, https://www.sciencedirect.com/science/article/pii/S0010028510000332.

Bjork, R. A. (1975). Retrieval as a Memory Modifier: An Interpretation of Negative Recency and Related Phenomena. In Solso, R. L., editor, *Information processing and cognition: The Loyola symposium*, pages 123–144. Halsted Press, ISBN: 978-0470812303.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In Metcalfe, J. and Shimamura, A. P., editors, *Metacognition: Knowing about knowing*, pages 185–205. MIT Press, Cambridge, MA, ISBN: 978-0262631693.

Bjork, R. A. and Kroll, J. F. (2015). Desirable Difficulties in Vocabulary Learning. *The American journal of psychology*, 128(2):241–52, ISSN: 0002-9556, http://www.ncbi.nlm.nih.gov/pubmed/26255443http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4888598.

Bloom, K. C. and Shuell, T. J. (1981). Effects of massed and distributed practice on the learning and retention of second-language vocabulary. *Journal of Educational Research*, 74(4):245–248, ISBN: 00220671, ISSN: 19400675, DOI: 10.1080/00220671.1981.10885317, http://www.tandfonline.com/doi/abs/10.1080/00220671.1981.10885317.

Carpenter, S. K., Cepeda, N. J., Rohrer, D., Kang, S. H. K., and Pashler, H. (2012). Using Spacing to Enhance Diverse Forms of Learning: Review of Recent Research and Implications for Instruction. *Educational Psychology Review*, 24(3):369–378, ISSN: 1040-726X, DOI: 10.1007/s10648-012-9205-z, http://link.springer.com/10.1007/s10648-012-9205-z.

Cepeda, N. J., Coburn, N., Rohrer, D., Wixted, J. T., Mozer, M. C., and Pashler, H. (2009). Optimizing Distributed Practice. *Experimental Psychology*, 56(4):236–246, ISSN: 1618-3169, DOI: 10.1027/1618-3169.56.4.236, https://econtent.hogrefe.com/doi/10.1027/1618-3169.56.4.236.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., and Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, 132(3):354–380, ISBN: 0033-2909 (Print), ISSN: 00332909, DOI: 10.1037/0033-2909.132.3.354.

Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. (2018). *shiny: Web Application Framework for R*, https://CRAN.R-project.org/package=shiny. R package version 1.2.0.

Cheung, A. C. and Slavin, R. E. (2013). The effectiveness of educational technology applications for enhancing mathematics achievement in K-12 classrooms: A meta-analysis. *Educational Research Review*, 9:88–113, ISSN: 1747-938X, DOI: 10.1016/J.EDUREV.2013.01.001, https://www.sciencedirect.com/science/article/pii/S1747938X13000031.

Chun, D. M. and Plass, J. L. (1996). Effects of multimedia annotations on vocabulary acquisition. *Modern Language Journal*, 80(2):183–198, ISSN: 15404781, DOI: 10.1111/j.1540-4781.1996.tb01159.x.

Cull, W. L. (2000). Untangling the benefits of multiple study

opportunities and repeated testing for cued recall. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 14(3):215–235, ISSN: 0888-4080.

Delaney, P. F., Verkoeijen, P. P., and Spirgel, A. (2010). Spacing and Testing Effects: A Deeply Critical, Lengthy, and At Times Discursive Review of the Literature. In *Psychology of Learning and Motivation*, volume 53, pages 63–147. Academic Press, ISBN: 9780123809063, ISSN: 0079-7421, DOI: 10.1016/S0079-7421(10)53003-2, https://www.sciencedirect.com/science/article/pii/S0079742110530032.

Dubois, M. and Vial, I. (2000). Multimedia design: the effects of relating multimodal information. Technical report, https://onlinelibrary.wiley.com/doi/pdf/10.1046/j.1365-2729.2000.00127.x.

European Commission (2019). *2nd Survey of schools: ICT in education - Objective 1: Benchmarking progress in ICT in schools*. Publications Office of the European Union, 2019, Luxembourg, ISBN: 978-92-79-99675-7, DOI: 10.2759/23401, https://ec.europa.eu/digital-single-market/en/news/2nd-survey-schools-ict-education.

Futuresource Consulting Press (2016). Education Technology Hardware Spend in K-12 Increases. https://futuresource-consulting.com/press-release/education-technology-press/education-technology-hardware-spend-in-k-12-increases-by-7-in-2015/.

Godwin-Jones, R. (2010). Emerging technologies from memory palaces to spacing algorithms: Approaches to second-language vocabulary learning. *Language Learning and Technology*, 14(2):4–11, ISSN: 10943501.

Greve, A., Cooper, E., Kaula, A., Anderson, M. C., and Henson, R. (2017). Does prediction error drive one-shot declarative learning? *Journal of Memory and Language*, 94:149–165, ISSN: 0749-596X, DOI: 10.1016/j.jml.2016.11.001, http://www.sciencedirect.com/science/article/pii/S0749596X16301899.

Griffin, G. and Harley, T. A. (1996). List learning of second language vocabulary. *Applied Psycholinguistics*, 17(04):443, ISSN: 0142-7164, DOI: 10.1017/S0142716400008195, http://www.journals.cambridge.org/abstract{_}S0142716400008195.

Hummel, K. M. (2010). Translation and short-term L2 vocabulary retention: Hindrance or help? *Language Teaching Research*, 14(1):61–74, ISSN: 1362-1688, DOI: 10.1177/1362168809346497, http://journals.sagepub.com/doi/10.1177/1362168809346497.

Kast, M., Baschera, G. M., Gross, M., Jäncke, L., and Meyer, M. (2011). Computer-based learning of spelling skills in children with and without dyslexia. *Annals of Dyslexia*, 61(2):177–200, ISBN: 07369387, ISSN: 07369387, DOI: 10.1007/s11881-011-0052-2, http://link.springer.com/10.1007/s11881-011-0052-2.

Kast, M., Meyer, M., Vögeli, C., Gross, M., and Jäncke, L. (2007). Computer-based multisensory learning in children with developmental dyslexia. *Restorative neurology and neuroscience*, 25(3-4):355–69, ISSN: 0922-6028, http://www.ncbi.nlm.nih.gov/pubmed/17943011.

Kim, D. and Gilman, D. A. (2008). Effects of text, audio, and graphic aids in multimedia instruction for vocabulary learning. *Educational Technology and Society*, 11(3):114–126, ISSN: 11763647.

Kornell, N., Castel, A. D., Eich, T. S., and Bjork, R. A. (2010). Spacing as the friend of both memory and induction in young and older adults. *Psychology and Aging*, 25(2):498–503, ISBN: 1939-1498; 0882-7974, ISSN: 19391498, DOI: 10.1037/a0017807, http://web.williams.edu/Psychology/Faculty/Kornell/Publications/Kornell.Castel.Eich.Bjork.2010.pdf.

Larman, C. and Basili, V. (2003). Iterative and incremental developments. a brief history. *Computer*, 36(6):47–56, ISSN: 0018-9162, DOI: 10.1109/MC.2003.1204375, http://ieeexplore.ieee.org/document/1204375/.

Lisman, J. E. and Grace, A. A. (2005). The Hippocampal-VTA Loop: Controlling the Entry of Information into Long-Term Memory. *Neuron*, 46(5):703–713, ISSN: 08966273, DOI: 10.1016/j.neuron.2005.05.002, https://linkinghub.elsevier.com/retrieve/pii/S0896627305003971.

Low, R. and Sweller, J. (2014). The Modality Principle in Multimedia Learning. In Mayer, R. E., editor, *The Cambridge Handbook of Multimedia Learning*, chapter 9. Cambridge University Press, New York, 2nd edition, ISBN: 978-1-107-03520-1.

Mackey, A. (2006). Feedback, Noticing and Instructed Second Language Learning. *Applied Linguistics*, 27(3):405–430, ISSN: 1477-450X, DOI: 10.1093/applin/ami051, http://academic.oup.com/applij/article/27/3/405/182957/Feedback-Noticing-and-Instructed-Second-Language.

Mayer, R. E. (2014). Introduction to Multimedia Learning. In Mayer, R. E., editor, *The Cambridge Handbook of Multimedia Learning*, chapter 1. Cambridge University Press, New York, 2nd edition, ISBN: 978-1-107-03520-1.

Mayer, R. E. and Fiorella, L. (2014). Principles for managing essential processing multimedia learning: Segmenting, pretraining, and modality principles. In Mayer, R. E., editor, *The Cambridge Handbook of Multimedia Learning*, chapter 13. Cambridge University Press, New York, 2nd edition, ISBN: 978-1-107-03520-1.

Mayer, R. E. and Moreno, R. (2002). Aids to computer-based multimedia learning. *Learning and Instruction*, 12(1):107–119, ISSN: 0959-4752, DOI: 10.1016/S0959-4752(01)00018-4, https://www.sciencedirect.com/science/article/pii/S0959475201000184.

Metcalfe, J. (2017). Learning from Errors. *Annual Review of Psychology*, 68(1):465–489, ISSN: 0066-4308, DOI: 10.1146/annurev-psych-010416-044022, http://www.annualreviews.org/doi/10.1146/annurev-psych-010416-044022.

Metcalfe, J., Kornell, N., and Finn, B. (2009). Delayed versus immediate feedback in children's and adults' vocabulary learning. *Memory & Cognition*, 37(8):1077–1087, ISSN: 0090-502X, DOI: 10.3758/MC.37.8.1077, http://www.springerlink.com/index/10.3758/MC.37.8.1077.

Moreno, R. (2006). Does the modality principle hold for different media? A test of the method-affects-learning hypothesis. ISSN: 02664909, DOI: 10.1111/j.1365-2729.2006.00170.x, http://doi.wiley.com/10.1111/j.1365-2729.2006.00170.x.

Moreno, R. and Mayer, R. (2007). Interactive Multimodal Learning Environments. *Educational Psychology Review*, 19(3):309–326, ISSN: 1040-726X, DOI: 10.1007/s10648-007-9047-2, http://link.springer.com/10.1007/s10648-007-9047-2.

Moreno, R. and Mayer, R. E. (2002). Verbal redundancy

in multimedia learning: When reading helps listening. *Journal of Educational Psychology*, 94(1):156–163, ISSN: 00220663, DOI: 10.1037/0022-0663.94.1.156, https://pdfs.semanticscholar.org/5870/17a482a0625514da944f5783c4a150cf1140.pdf.

Morris, C. D., Bransford, J. D., and Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, 16(5):519–533, ISSN: 0022-5371, DOI: 10.1016/S0022-5371(77)80016-9, http://www.sciencedirect.com/science/article/pii/S0022537177800169.

Murray, M. M. and Sperdin, H. F. (2010). Single-Trial Multisensory Learning and Memory Retrieval. In Kaiser, J. and Naumer, M. J., editors, *Multisensory Object Perception in the Primate Brain*, pages 191–208. Springer New York, New York, NY, DOI: 10.1007/978-1-4419-5615-6.

Park, D. C., Smith, A. D., Lautenschlager, G., Earles, J. L., Frieske, D., Zwahr, M., and Gaines, C. L. (1996). Mediators of long-term memory performance across the life span. *Psychology and aging*, 11(4):621–37, ISSN: 0882-7974, DOI: 10.1037//0882-7974.11.4.621, http://www.ncbi.nlm.nih.gov/pubmed/9000294.

Pashler, H., Cepeda, N. J. N., Wixted, J. T., Rohrer, D., . . . , J. W. . . . . P. L., and 2005, U. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(1):3–8, ISBN: 4156687503, ISSN: 1939-1285, DOI: 10.1037/0278-7393.31.1.3, https://psycnet.apa.org/fulltext/2004-22496-001.htmlhttp://www.ncbi.nlm.nih.gov/pubmed/15641900http://doi.apa.org/getdoi.cfm?doi=10.1037/0278-7393.31.1.3.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/.

Reber, T. P. and Rothen, N. (2018). Educational App-Development needs to be informed by the Cognitive Neurosciences of Learning & Memory. *npj Science of Learning*, 3(1):22, ISSN: 2056-7936, DOI: 10.1038/s41539-018-0039-4, http://www.nature.com/articles/s41539-018-0039-4.

Rheinberg, F., Vollmeyer, R., and Burns, B. D. (2001). FAM: Ein Fragebogen zur Erfassung aktueller Motivation in Lern- und Leistungssituationen12 (Langversion, 2001). page 17.

Roediger, H. L. and Pyc, M. A. (2012). Inexpensive techniques to improve education: Applying cognitive psychology to enhance educational practice. *J. Appl. Res. Mem. Cogn.*, 1(4):242–248, ISSN: 2211-3681, DOI: 10.1016/j.jarmac.2012.09.002.

Rowland, C. A. (2014). The effect of testing versus restudy on retention: a meta-analytic review of the testing effect. *Psychol. Bull.*, 140(6):1432–1463, ISSN: 0033-2909, DOI: 10.1037/a0037559.

Schneider, V. I., Healy, A. F., and Bourne, L. E. (2002). What Is Learned under Difficult Conditions Is Hard to Forget: Contextual Interference Effects in Foreign Vocabulary Acquisition, Retention, and Transfer. *Journal of Memory and Language*, 46(2):419–440, ISSN: 0749-596X, DOI: 10.1006/JMLA.2001.2813, https://www.sciencedirect.com/science/article/pii/S0749596X0192813X.

Settles, B. and Meeder, B. (2016). A trainable spaced repetition model for language learning. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 1848–1858. ACL, DOI: 10.18653/v1/P16-1174, http://www.aclweb.org/anthology/P16-1174.

Shams, L. and Seitz, A. R. (2008). Benefits of multisensory learning. *Trends in Cognitive Sciences*, 12(11):411–417, ISSN: 1364-6613, DOI: 10.1016/J.TICS.2008.07.006, https://www.sciencedirect.com/science/article/pii/S1364661308002180.

Sobel, H. S., Cepeda, N. J., and Kapler, I. V. (2011). Spacing effects in real-world classroom vocabulary learning. *Applied Cognitive Psychology*, 25(5):763–767, ISSN: 08884080, DOI: 10.1002/acp.1747, http://doi.wiley.com/10.1002/acp.1747.

Steinel, M. P., Hulstijn, J. H., and Steinel, W. (2007). SECOND LANGUAGE IDIOM LEARNING IN A PAIRED-ASSOCIATE PARADIGM: Effects of Direction of Learning, Direction of Testing, Idiom Imageability, and Idiom Transparency. *Studies in Second Language Acquisition*, 29(03):449–484, ISSN: 0272-2631, DOI: 10.1017/S0272263107070271, http://www.journals.cambridge.org/abstract{_}S0272263107070271.

Steiner, M., Phillips, N., and Trutmann, K. (2018). *ShinyPsych: An easy way to program psychology experiments using Shiny*, https://github.com/mdsteiner/ShinyPsych. R package version 0.2.3.

Sung, Y.-T., Chang, K.-E., and Liu, T.-C. (2016). The effects of integrating mobile devices with teaching and learning on students' learning performance: A meta-analysis and research synthesis. *Computers & Education*, 94:252–275, ISSN: 0360-1315, DOI: 10.1016/J.COMPEDU.2015.11.008, https://www.sciencedirect.com/science/article/pii/S0360131515300804.

Tabbers, H. K., Martens, R. L., and Merriënboer, J. J. G. V. (2004). Multimedia instructions and Cognitive Load Theory: split-attention and modality effects. *British Journal of Educational Psychology*, 74:71–81.

Tamim, R. M., Bernard, R. M., Borokhovski, E., Abrami, P. C., and Schmid, R. F. (2011). What Forty Years of Research Says About the Impact of Technology on Learning. *Review of Educational Research*, 81(1):4–28, ISSN: 0034-6543, DOI: 10.3102/0034654310393361.

Weinstein, Y., Madan, C. R., and Sumeracki, M. A. (2018). Teaching the science of learning. *Cognitive Research: Principles and Implications*, 3(1), ISSN: 2365-7464, DOI: 10.1186/s41235-017-0087-y.

Wilkinson, L., Tai, Y. F., Lin, C. S., Lagnado, D. A., Brooks, D. J., Piccini, P., and Jahanshahi, M. (2014). Probabilistic classification learning with corrective feedback is associated with in vivo striatal dopamine release in the ventral striatum, while learning without feedback is not. *Human Brain Mapping*, 35(10):5106–5115, ISSN: 1097-0193, DOI: 10.1002/hbm.22536, https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.22536.

Yanguas, I. (2009). Multimedia glosses and their effect on L2 text comprehension and vocabulary learning. *Language Learning and Technology*, 13(2):48–67, ISSN: 10943501.