Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

Speaking to remember: Model-based adaptive vocabulary learning using automatic speech recognition

Thomas Wilschut [a,*], Florian Sense [b], Hedderik van Rijn [a]

[a] *Department of Experimental Psychology, University of Groningen, Grote Kruisstraat 2/1, 9712 TS, Groningen, Netherlands*
[b] *Infinite Tactics, LLC, 1430 Oak Ct Ste 203, 45430, Beavercreek, OH, United States of America*

ARTICLE INFO

Dataset link: https://osf.io/cpfsq/

Keywords:
Adaptive learning
ACT-R
Automatic speech recognition (ASR)
Memory
Response Times (RT)
Speech
Typing
Vocabulary

ABSTRACT

Memorizing vocabulary is a crucial aspect of learning a new language. While personalized learning- or intelligent tutoring systems can assist learners in memorizing vocabulary, the majority of such systems are limited to typing-based learning and do not allow for speech practice. Here, we aim to compare the efficiency of typing- and speech based vocabulary learning. Furthermore, we explore the possibilities of improving such speech-based learning using an adaptive algorithm based on a cognitive model of memory retrieval. We combined a response time-based algorithm for adaptive item scheduling that was originally developed for typing-based learning with automatic speech recognition technology and tested the system with 50 participants. We show that typing- and speech-based learning result in similar learning outcomes and that using a model-based, adaptive scheduling algorithm improves recall performance relative to traditional learning in both modalities, both immediately after learning and on follow-up tests. These results can inform the development of vocabulary learning applications that–unlike traditional systems–allow for speech-based input.

1. Introduction

Over the last decades, advancements in cognitive psychology and educational data mining have led to the development of personalized or adaptive learning systems, which aim to improve the process of fact and vocabulary learning by tailoring learning processes to the needs of individual learners (Lindsey et al., 2014; Papousek et al., 2014; Van Rijn et al., 2009; Settles and Meeder, 2016; Wozniak and Gorzelanczyk, 1994). Although the exact mechanisms employed by these systems differ, they all aim to track the actions of individual learners using behavioral indices such as accuracy scores, and in the case of the system presented in this paper, reaction times (RTs). These indices are used to estimate and continually update a set of parameters in mathematical equations that estimate various aspects of the learning process (e.g., the activation of the item in the learner's memory, or the rate at which the learner forgets the item). These parameters are assumed to capture differences in memory processes between learners and are used to predict performance later in the learning session. Subsequently, this information can be used to provide specific feedback, select appropriate practice problems, or optimize item repetition schedules. Employing the above-described adaptive repetition protocols generally results in better retention of the studied materials compared to learning with traditional, non- or less-adaptive systems (Lindsey et al., 2014; Mettler et al., 2016; Papousek et al., 2014; Van Rijn et al., 2009; Sense et al., 2016; Van der Velde et al., 2021a; Wozniak and Gorzelanczyk, 1994).

Most current adaptive learning systems are based on typed responses or the selection of response options using keyboard, touch, or mouse, and do not support speech input. Although some language learning systems that are currently on the market use speech

[*] Corresponding author.
E-mail address: t.j.wilschut@rug.nl (T. Wilschut).

https://doi.org/10.1016/j.csl.2023.101578
Received 12 September 2022; Received in revised form 19 October 2023; Accepted 20 October 2023
Available online 31 October 2023

recognition software[1] to provide feedback to the learner (usually in the form of 'correct/incorrect' feedback), to our knowledge, no learning system uses automatic speech assessment or speech-related behavioral measures such as RTs for refined item-level adaptation of learning processes.

Despite the lack of focus on speech-related behavioral measures in current adaptive learning systems, relying on spoken rather than typed input has several potential benefits. First, speech-based systems allow users to efficiently practice speech and pronunciation while learning vocabulary, which is an important part of language acquisition that is largely omitted in traditional textbook- or typing-based approaches. Second, speech-based applications allow for studying vocabulary in cases where typing or reading is not possible. Such systems could be used by people who physically lack the ability or the opportunity to type (e.g., physically challenged people, or people who are driving a car), making them applicable in a wide range of settings. Finally, although individual differences in the speed and accuracy of speech production may influence RTs in spoken vocabulary learning, speech-based RTs are independent of individual differences in typing skills. Because of the various advantages of speech-based, we will here explore the possibilities of using personalized learning algorithms that were initially designed for typing-based learning, to improve speech-based learning.

2. Theory

Before examining potential benefits of speech-based learning, it is important to consider the functional differences between the cognitive mechanisms involved in speech- and typing-based learning that could potentially complicate the usage of spoken input in adaptive learning systems that were developed for typing-based learning. For example, speech-based responses involve the complex coordination of jaw, tongue and lip movements that together result in a specific sound pattern, where typing-based learning involves a very different motor coordination to result in a sequence of keystrokes on the keyboard. As mentioned above, most adaptive learning systems aim to estimate memory parameters for individual learners using behavioral indices like accuracy and RTs. Furthermore, typing-based learning involves the storage and retrieval of the orthographic representation of words (i.e., their spelling), whereas speech-based learning involves storing and retrieving their phonological representation (i.e., their sound).

Although a range of perceptual, cognitive and motor processes are involved in processing a vocabulary item, retrieving its correct foreign-language translation, and producing a typed response, only the *memory components* of this process are modeled in adaptive learning systems. Some of the major differences between typing- and speech-based learning lie in the perceptual and motor components involved in the production of a response, and some research suggests that the memory components involved in retrieving a spoken versus typed response are functionally similar. This notion is supported by the prototypical models of the mental lexicon in psycholinguistics, which assume that the phonological (pronunciation) and orthographic (spelling) representations of a word are functionally similar components of formal representations we store for each word (Aitchison (2012), Jiang (2000) and Levelt (1999), but see Plag et al. (2017)). Wilschut et al. (2021) empirically substantiated these ideas by showing that RT and accuracy score distributions were similar in typing- and speech-based learning, and that both behavioral measures for speech- and typing-based learning could be used to reliably estimate memory model parameters such as the rate at which a learner forgets an item. Therefore, we here assume that, from the perspective of the current application, there are sufficient functional similarities between the memory processes involved in vocabulary learning using spoken and typed input.

Besides theoretical considerations, using speech input in adaptive learning systems adds a technological challenge. For one, the automatic speech recognition (ASR, see Yu and Deng, 2016) system has to be based on real-time decoding algorithms as immediate feedback is required. Given this requirement, employing ASR systems was practically impossible until recently due to low reliability and high speech-to-noise ratios, which explains the lack of scientific research into speech-based adaptive learning (see Litman et al., 2018, for an overview of historical developments). However, contemporary ASR technology allows for accurately and reliably transcribing speech to text in real-time (Litman et al., 2018; Nassif et al., 2019; Shadiev et al., 2020; Shadiev and Liu, 2022). With real time decoding, all behavioral indices used to determine item scheduling in typing- or keypress-based learning have a speech-based equivalent: Response times defined by the first keypress in typing-based systems can be translated to response latencies in speech-based systems and accuracy scores can be computed by treating ASR-transcribed text strings as if they were a typed response (for more details, see Materials). Because of the similarities between behavioral indices in typing- and speech-based learning, using ASR output in existing typing-based adaptive learning systems is technically feasible.

In short, applying existing typing-based adaptive scheduling algorithms to speech-based learning is a propitious possibility to exploit, as typing-based adaptive systems have been proven to be successful, and adaptive memory models are likely to generalize to speech-based learning. The current study will be the first to test the effectiveness of speech-based adaptive learning by (1) directly comparing typing-based learning to speech-based learning using state-of-the-art ASR technology and (2) comparing adaptive learning benefits relative to using a less-adaptive, Leitner-inspired[2] scheduling algorithm for both input modalities. Importantly, we will not only consider memory performance immediately after the learning session, but we will also test long-term retention after an interval of 4–8 days. Because of the assumed functional similarities between the memory processes involved in typing- and speech-based learning, we hypothesize that spoken and typed vocabulary learning will result in similar learning performance. Furthermore, we assume that adaptive learning benefits will generalize from typing-based learning to speech-based learning.

[1] For example, see Duolingo, www.duolingo.com, Graphogame, www.graphogame.com, Rosetta Stone, www.rosettastone.com, ProTutor (Epp and McCalla, 2011), or Alex (Munteanu et al., 2010, 2014).

[2] An item repetition protocol that repeats items that have been answered incorrectly more often than then items that have been answered correctly, but does not take response times into account, see Methods.

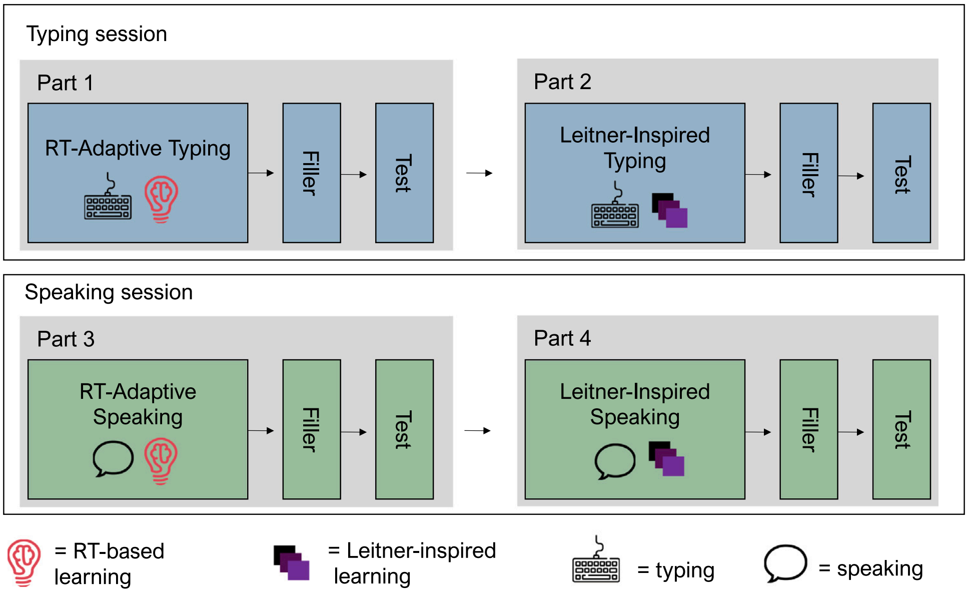

Fig. 1. Study design. Participants completed all four parts. Within each session, an RT-adaptive, MemoryLab studying condition was followed by a Leitner-inspired, accuracy-adaptive studying condition. Between the first test and the second learning session, there was a self-paced break.

3. Methods

3.1. Participants

In total, 50 first-year psychology students (age 17–29) took part in this experiment, of whom 46 completed both experimental sessions (see Design and Procedure). Participants were native Dutch speakers and indicated that they were fluent in English. Participants gave informed consent and the study was approved by the ethical committee of the department of Psychology at the University of Groningen (study approval code: PSY-2021-S-0025).

3.2. Design and procedure

The study consisted of a total of four conditions, which participants completed in two separate sessions scheduled 4–8 days apart, see Fig. 1. There was one typing session and one speaking session. Half of the participants (n = 26) started with the typing session, and completed the speaking session 4–8 days later. For the other half of the participants (n = 24), this order was reversed.

Both sessions contained two experimental conditions. Each condition had the same general structure: An 11-min[3] learning block, in which participants studied the English translation of Swahili vocabulary items (see Materials) was followed by a 3-min filler block, in which participants were asked to complete simple integer sequences. Each condition ended with a test of all items that were presented during the preceding learning session. The four experimental conditions differed (1) in user response modality (either typing or speaking) and (2) item repetition protocol (either using an RT-based adaptive algorithm (discussed below) or using a less adaptive, Leitner-inspired flashcard algorithm (also discussed below)). The response modality on the test matched the modality during learning. Before the start of the second session, participants were tested on the materials they studied during the first session.

We used the *MemoryLab* (Van Rijn et al., 2009; Sense et al., 2016, see http://memorylab.nl/en) adaptive fact-learning algorithm, which was originally developed for typing-based learning, to determine item repetition schedules in the RT-based, adaptive learning conditions (RT-adaptive). MemoryLab is based on the ACT-R architecture's model of human declarative memory (Anderson et al., 1998; Pavlik and Anderson, 2008) and it functions by measuring RTs and accuracy scores to determine optimal item repetition schedules for individual learners. The system is based on the assumption that RTs are a good proxy for the strength of fact representations in memory: The quicker the learner produces a correct response, the stronger the memory representation is assumed to be Anderson and Schooler (1991), Jescheniak and Levelt (1994), Levelt (1999), Van Rijn et al. (2009) and Sense and Van Rijn (2022). An abundance of experimental data supports this link: Faster responses are generally associated with more accurate responses and a stronger association between cue and response compared to slower responses (Byrne and Anderson, 1998; Settles et al., 2018). The MemoryLab adaptive learning system is based on two well-known findings in cognitive psychology. First, it enhances learning

[3] This study duration was chosen because prior studies suggested that it was long enough to find differences in learning performance between learners and conditions, but short enough to have learners complete two learning blocks in a single experimental session.

by exploiting the *testing effect*: the idea that testing oneself is one of the most effective ways of learning. Testing has been found to be more effective than reading and note-taking (Rummer et al., 2017). Several studies have shown that actively and successfully recalling an item from memory results in stronger memory representations for that item (Moreira et al., 2019; Roediger and Karpicke, 2006). Second, the system uses the *spacing effect*, which describes that one learns better when learning is spread over several learning moments (Cepeda et al., 2008; Karpicke and Bauernschmidt, 2011; Kornell, 2009; Nakata, 2017). By repeating a piece of information just before it is estimated to be forgotten, the MemoryLab system balances the beneficial effects of spacing and retrieval practice: it ensures a maximal time period in between successful item repetitions. For a more detailed description of the model, see Van Rijn et al. (2009) and Sense et al. (2016). The beneficial effects of using RTs and accuracy scores to create item repetition schedules tailored to individual learners are demonstrated in both controlled lab studies (Mettler et al., 2011; Sense et al., 2016; Van der Velde et al., 2021a; Zhou et al., 2021) and in real-world classroom situations (Van Rijn et al., 2009; Van der Velde et al., 2021b; Sense et al., 2021): Learning facts with the MemoryLab system results in up to 10 percent better recall of studied materials compared to learning with less adaptive, accuracy-based algorithms in which RTs are not taken into account.

The *MemoryLab typing condition* used the MemoryLab algorithm to determine the scheduling of the items, see above. At its first presentation, a Swahili word was shown in text on a computer screen, together with its written English translation. In subsequent presentations, only the Swahili word was shown, participants were asked to type the correct translation, and received corrective feedback ('correct!' if the typed response was correct; 'Incorrect, the correct answer was [correct answer]' if the typed response was incorrect, and 'Too slow! The correct answer was [correct answer]' if the participants took more than 15 s to respond). RTs were defined as the time elapsed between the start of the presentation of the cue and the first keypress. If the user deleted the first keypress to correct the answer, the response was considered invalid and not used to determine further item scheduling. On each trial, the MemoryLab algorithm determined whether a previously-encountered item would be repeated or a whether a new item would be introduced, based on earlier responses (RTs and correctness scores) by the learner (for more details, see Sense et al., 2016). As a result, the number of items presented by the MemoryLab algorithm in the MemoryLab learning conditions differed between participants depending on performance.

In the *Leitner-inspired typing condition*, the presentation of the items, correctness scoring, feedback, and RT operationalization were the same as in the above-described MemoryLab typing condition. For each participant, the number of words that had to be studied was set to the number of words that were presented in the MemoryLab typing condition. The item repetition schedule was determined by a Leitner-inspired flashcard system (Mubarak and Smith, 2008), which groups words into three virtual boxes: All words start in Box 1 and move to the next box if answered correctly. If a word is answered incorrectly, it moves back to the previous box. Words in Box 1 are presented first, followed by words in Box 2, followed by words in Box 3. If all items are in Box 3 (and if they are all answered correctly) the items are repeated in the order of first presentation until the learning time is over. This flashcard system allows for difficult items to be rehearsed more often than easy items and is a frequently used and effective study strategy (Bryson, 2012).

In the *MemoryLab speaking condition*, for the first presentation of an item, participants saw a Swahili word on the computer screen in text, together with the written translation of this word. Additionally, the spoken the English translation of the word was presented through headphones. In all subsequent presentations, only the Swahili word was shown, and participants were instructed to speak the correct English translation, after which they received written and auditory feedback (only after incorrect responses). Voice utterances were transcribed to text automatically and in real time using the Google Web Speech API (see Materials). As in the MemoryLab typing condition, the correctness scores and RTs were used by the MemoryLab algorithm to determine the scheduling of the responses.

In the *Leitner-inspired speaking condition*, the item presentation, voice recording, correctness scoring, feedback and RT measurement were the same as in the MemoryLab speaking condition. The item scheduling by the Leitner-inspired algorithm was the same as in the Leitner-adaptive typing condition.

3.3. Materials

The full list of word pairs, exemplar pronunciations, and filler materials used in this study can be found in the supplementary materials at https://osf.io/cpfsq/.

Hardware and software. The experiment was conduced in a dimly-lit and quiet lab room. The experiment was built with JavaScript and HTML5 using the jsPsych online experiment library (De Leeuw, 2015). In the speaking session, participants wore USB headsets which were used to record and play audio. Participants' voice utterances were recorded and transcribed to speech automatically and in real time using the Google Web Speech API (see https://wicg.github.io/speech-api/; Këpuska and Bohouta, 2017, for a comparison). The Google Web Speech API is a JavaScript programming interface that can be used for speech recognition and synthesis, and can transcribe both brief (one-shot) speech input and continuous speech input. It has been used in various learning applications (e.g., see Daniels, 2015) and is one of the most accurate speech-to-text API's currently on the market (Fendji et al., 2022; Filippidou and Moussiades, 2020; Kimura et al., 2019). The speaker RT was defined as the time elapsed between the presentation of the cue, and the moment at which the first spoken response was given, and was determined by the ASR system. Speech-to-text transcription procedures are generally considered error-prone processes that are highly dependent on the circumstances in which transcriptions are made: It works best with high-quality audio for utterances that contain sentence-level context. In the current study, we transcribed context-free, single-word utterances to text. Especially for homophonic or similar-sounding words, erroneous transcriptions can be an issue. To assess whether transcription errors influence the results of this study, we conducted a pilot experiment to validate the speech-to-text API (see below).

Items and answer scoring. The study materials were taken from a word list containing 100 paired-associate Swahili–English word pairs (Nelson and Dunlosky, 1994) (in the current study, only 88 word pairs were used, see Validating Automatic Speech Assessment). This list was selected because (1) items were relatively short (8 letters or less, never more than 3 syllables), making them easy to pronounce; (2) because participants were unlikely to be familiar with any of the Swahili words (due to the general low familiarity of the Swahili language in the participant population, and because the word list contains no English or Dutch loan words Nelson and Dunlosky, 1994); and (3) because normative difficulty estimations were available for each word on the list. The word list was divided into four subsets of equal size and normative difficulty scores (as specified in the Nelson and Dunlosky (1994) word list). Subsequently, for each participant, one word subset was assigned to one experimental condition. The order in which word subsets were distributed over conditions was fully counterbalanced. Within each condition, to-be-studied materials were randomly drawn from the assigned word subsets.

To prevent that minor typing during learning errors would result in scoring the response as incorrect, responses were considered correct if the Levenshtein's edit distance from response to answer (see Yujian and Bo, 2007) was equal to or less than 2. The specific edit distance threshold was chosen because it allows for a substantial correction in typing, without the possibility of scoring incorrect translations as correct: We ensured that all English translations of the items within each word subset were at least 3 edit distances apart. The choice of edit distance threshold is dependent on the specific item set and goals of the learner. For example, if a learning aims to study the correct spelling of vocabulary items, implementing an edit distance may not be a suitable solution. Similar to the typing conditions, transcribed voice signals were compared to the correct written response, and were considered correct if they were 2 or less than 2 Levenshtein's edit units apart, also see Validating Automatic Speech Assessment. It is good to note that an edit distances on typed responses on one hand, and edit distances on ASR-transcribed text on the other hand, operate on somewhat different levels, and where introduced for different reasons. In the typing-based system, the edit distance was use to filter typing- and spelling mistakes, in the speech-based system, the edit distance was used to correct for minor transcription errors (see Validating Automatic Speech Assessment).

Exemplar pronunciations. The correct exemplar pronunciations that were provided to the participants at the first presentation of an item, and as corrective feedback, were generated by Google's WaveNet text-to-speech algorithm (http://cloud.google.com/text-to-speech) in British English.

Filler materials. In the two-minute filler task, participants completed integer sequences in an open-question format (e.g., '3-6-12-24-?' requires response "48").

3.4. Validating automatic speech assessment

A pilot study was conducted to examine the accuracy of the Google Web Speech API when native Dutch-speaking participants pronounced the study materials used in this study. Thirteen participants completed the pilot experiment. Written English translations of all 100 words in the Swahili-item list where presented to the participants. Participants cycled through the full word list three times. At the first presentation of each word, participants heard the correct pronunciation of the English word through headphones. Participants were instructed to read out loud/pronounce the words presented on the screen after the exemplar pronunciation. For the second and third presentation of each item, the exemplar pronunciation was not presented. No feedback was provided after the response. This provided voice recordings of the kind we expected participants to produce when they knew the correct English translation of the Swahili cue during the study.

We used the Google Web Speech API to transcribe the voice utterances to text. In 75,3% of all utterances, there was an exact match between the text string that the participants were asked to pronounce and the transcription. Non-matching trials could be divided into two categories: (1) incorrect responses that were accurately transcribed to text, or (2) correct responses that were inaccurately transcribed to text. The majority of incorrect transcriptions concerned utterances that contained errors in number (e.g., 'dogs' instead of 'dog') or tense (e.g., 'played' instead of 'play'). Average transcription accuracy varied markedly between different items on the list. In order to ensure maximal transcription accuracy, we removed 12 words from the word list that had exceptionally low transcription accuracy (lower than 40%). Furthermore, we applied the same criteria used to allow for typing mistakes (see Design and Procedure) in the typing-based conditions to the transcribed text in the speech-based conditions. When considering transcriptions with an edit distance of 2 or lower a match, 98.4% of the remaining transcriptions matched the written word that the participants were asked to pronounce.

3.5. Analyses

Data preprocessing and statistical analyses were conducted in R 3.4.1 (R Core Team, 2020), with the mixed-effects modeling package lme4 1.1-28 (Bates et al., 2015). Incorrect responses in the speech-based learning block were manually checked after the experiment. Of all responses, 1.7% were scored as incorrect by the adaptive learning system, as a consequence of a mismatch between the uttered response and the ASR transcription (i.e., the participant gave the correct answer, but a transcription error resulted in a mismatch between the transcribed answer text and the correct answer). These responses were included in the analyses. In the speech conditions, 12.56% of trials were scored correct because of the implementation of the edit distance (i.e., they would be scored incorrect without the implementation of the edit distance). Although this is a high number, the majority of these corrections were transcriptions that were simply incorrect in number (e.g., transcription: 'ornaments'; correct answer: 'ornament'). Other common errors were single-letter swaps (e.g., transcription: 'long'; correct answer: 'lung'; or transcription: 'blue'; correct answer: 'glue'). In the typing conditions, 5.21% of trials were scored correct because of the implementation of the edit distance. These corrections typically

Table 1
Logistic mixed-effects model explaining accuracy during learning from response modality and repetition protocol.

Accuracy during learning	β	SE	z	p
Intercept	1.897	0.080	23.695	<0.001***
Repetition protocol	0.822	0.044	18.704	<0.001***
Response modality	0.046	0.045	1.024	0.306
Repetition protocol × Response modality	0.535	0.087	6.135	<0.001***

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

involved typing or spelling errors (e.g., typed response: 'mattresas'; correct answer: 'mattress'). The mixed effects models reported in this study include item repetition protocol (deviance coded, Leitner-inspired = −0.5; MemoryLab = 0.5) and response modality (also deviance coded, Speaking = −0.5; Typing = 0.5) as fixed-effects factors, and accuracy (logistic mixed effects models) or RTs (linear mixed effects models) as dependent variables. Participant- and item id were added as random intercepts to all models (Baayen et al., 2008). Mixed effects models including study session (indicating the difference between the first and the second session) and study time (minutes elapsed since the start of the study session until the current trial) as fixed-effects variables were considered, but not reported because these models had lower log-likelihood values compared to the models without these factors. The data was visualized using ggplot2 (Wickham, 2016). The data generated in this study and analysis scripts can be found in the online supplementary materials: https://osf.io/cpfsq/.

4. Results

This study aims to address two research questions. First, we will compare typing-based learning to speech-based learning. Second, we will consider adaptive learning benefits and compare MemoryLab learning to Leitner-inspired learning for both input modalities. Before discussing the results of these two research questions, which all focus on the *outcomes* of learning as measured by performance on the tests, we will discuss how the different item repetition protocols and learning modalities influenced performance *during* learning.

4.1. Performance during learning

Fig. 2 shows correctness during learning. Dots represent performance by individual participants, color represents modality. Most dots are above the diagonal line that represents equal performance in the MemoryLab and Leitner-inspired conditions: Correctness during learning was higher in the MemoryLab condition for both modalities. In the typing conditions, participants encountered on average 17.0 (SD = 4.01) items, and in the speaking conditions, participants encountered an average of 15.9 (SD = 4.23) items. The number of items in the MemoryLab and Leitner-inspired conditions was equal by design. We conducted a logistic mixed effects regression analysis to test the differences in accuracy during learning as a function of modality (speaking versus typing), item repetition protocol (Leitner-inspired versus MemoryLab) and the interaction between these two factors. The results are shown in Table 1[4] and corroborate what is apparent in Fig. 2: Accuracy was higher in the MemoryLab conditions than in the Leitner-inspired conditions ($z = 23.69$, $p < 0.001$). Modality had no significant effect on accuracy ($z = 1.02$, $p < 0.001$). The interaction effect of repetition protocol and response modality was significant, indicating that the difference between accuracy during MemoryLab learning and during Leitner-inspired learning was larger for typing-based learning compared to speech-based learning ($z = 6.14$, $p < 0.001$).

The proportion of correct responses as a function of item repetition, separated for the four learning conditions, is shown in Fig. 3. Since the correct answer was provided at the first presentation of an item, the proportion of correct responses during this first presentation (in the figure: repetition 0) is close to 1. For the MemoryLab conditions, average correctness drops to around 75%–95% in the trials that follow the first item presentation, and then remains relatively stable.

In the Leitner-inspired conditions, the proportion of correct responses fluctuates over repetitions in an alternating (saw-tooth) pattern. The pattern of alternating accuracy can be explained by the nature of the repetition schedules used in the Leitner-inspired learning condition: Items are repeated sooner when they are answered incorrectly than when they are answered correctly.[5] After an incorrect response on the first repetition, corrective feedback is provided and the item is repeated quickly. The short time interval between the corrective feedback and the item repetition results in a high number of correct responses on the second repetition. After a correct second repetition, it takes much longer for the item to be repeated again. This long time interval between a correct response for an item and its next repetition results in a lower accuracy on the next repetition, etc.

In short, the analyses reported in this section show that the MemoryLab system results in repetition schedules in which lag times and correctness scores are kept relatively constant across item repetitions. In contrast, accuracy-based Leitner-inspired learning results in accuracy scores that fluctuate over repetitions. As a consequence, average accuracy during learning was higher in the MemoryLab conditions than in the Leitner-inspired conditions. In the next sections, we will examine the performance on the test that followed the learning sessions discussed above.

[4] The logistic regression coefficients in Tables 1 and 2 can be converted to probabilities using an inverse logit transform. For example, in Table 1: typing-based MemoryLab learning $= exp(1.897 + (0.5 * 0.822) + (0.5 * 0.046))/(1 + exp(1.897 + (0.5 * 0.822) + (0,5 * 0.046))) = 0.911$, compared to speech-based, Leitner-inspired learning $= exp(1.897 + (-0.5 * 0.822) + (0.046 * -0.5))/(1 + exp(1.897 + (-0.5 * 0.822) + (-0.5 * 0.822))) = 0.812$.

[5] The difference in time (lag) between repetitions is shown in Supplementary Figure 1, see https://osf.io/y2fsc/.

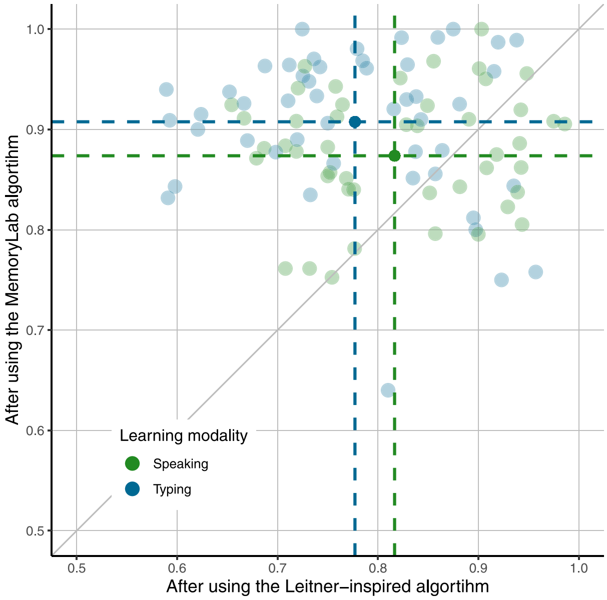

Fig. 2. Accuracy during learning. Dots show the average correctness for individual participants. The diagonal line represents equal performance in the MemoryLab and Leitner-inspired conditions. Dotted lines show averages for the typing-based MemoryLab (horizontal blue line); the speech-based MemoryLab (horizontal green line); the typing-based Leitner-inspired (vertical blue line) and the Leitner-inspired speech-based conditions (vertical green line).

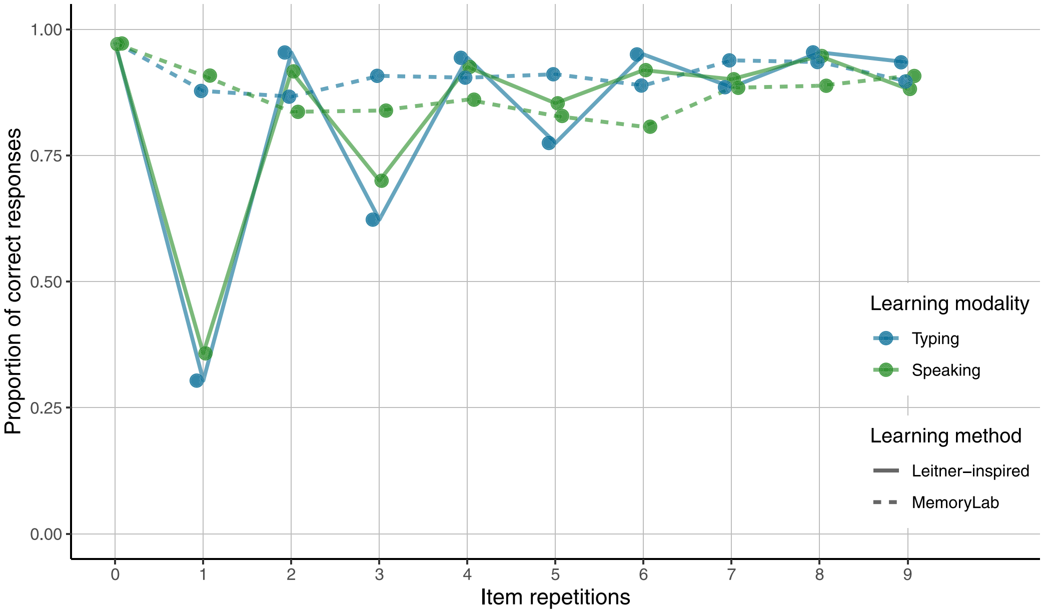

Fig. 3. Correctness as a function of item repetitions during the learning phase, separated for the four learning conditions.

4.2. Performance on test

Fig. 4A and B summarize test performance by showing the number of items that were recalled correctly in each condition, at immediate test and after 4–8 days, respectively.[6] Again, dots represent performance by individual participants, and dotted lines represent average performance in each condition. We conducted two item-level logistic mixed effects models to examine the effects

[6] Note that these figures show test performance in absolute numbers. For the figures showing performance as proportions, please see the supplementary Figure 2 (https://osf.io/92vjs).

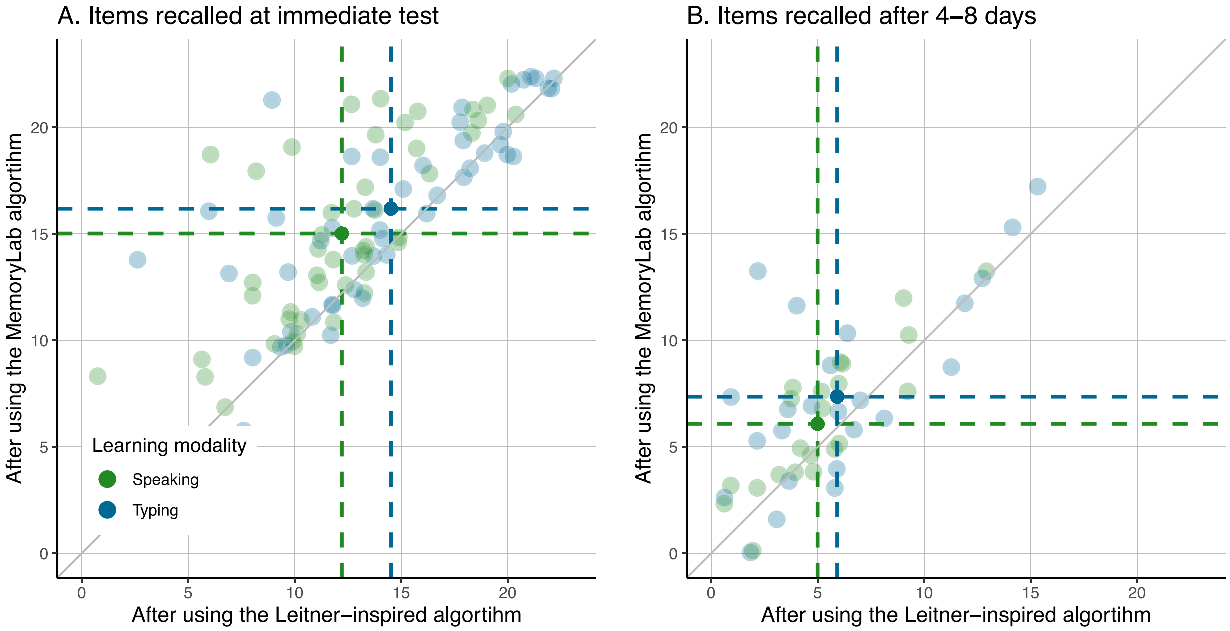

Fig. 4. Number of items correctly recalled on test, separated for response modality (speaking in green and typing in blue) and item repetition protocol (MemoryLab learning on the y-axis, Leitner-inspired learning on the x-axis). Dots show the number of items recalled by individual participants. Dotted colored lines represent average performance by modality, diagonal gray lines represent equal performance in the MemoryLab- and Leitner-inspired conditions. **A** shows the total number of items that were successfully recalled at the test that immediately followed the learning session. **B** shows the number of items that were recalled after 4–8 days.

Table 2
Logistic mixed-effects models explaining accuracy from item repetition protocol and response modality.

A. Accuracy on immediate test	β	SE	z	p
Intercept	2.767	0.176	15.761	<0.001***
Repetition protocol	1.480	0.145	10.186	<0.001***
Response modality	0.134	0.147	0.912	0.362
Repetition protocol × Response modality	−0.289	0.288	−1.001	0.317
B. Accuracy after 4–8 days	β	SE	z	p
Intercept	−0.565	0.164	−3.445	<0.001***
Repetition protocol	0.519	0.123	4.220	<0.001***
Response modality	0.066	0.303	0.216	0.829
Repetition protocol × Response modality	−0.043	0.242	−0.179	0.858

of response modality and item repetition protocol on correctness during test: one for the immediate test and one for the test that followed after 4–8 days, see Table 2.[7] The results of the analyses can be summarized in three main points.

First, Fig. 4 also shows that for most participants, the number of items recalled after MemoryLab scheduling was higher than after Leitner-inspired scheduling: Most dots lie above the diagonal line that represents equal performance for both scheduling systems (participants recalled on average 3.0 more items after MemoryLab scheduling than after Leitner-inspired scheduling in the speaking condition at immediate test, 1.7 more items in the typing condition at immediate test, 1.1 more items in the speaking condition at the followup test, and 1.4 more items in the typing condition at the followup test). Corroborating the pattern apparent in Fig. 4, the main effect of MemoryLab scheduling was significant in both mixed effects models, indicating that MemoryLab learning resulted in a significantly higher number of correctly recalled items, both on the immediate test and the long-term retention test ($z = 10.19$, $p < 0.001$; $z = 4.22$, $p < 0.001$, respectively).

Second, Fig. 4A and B show that for both item repetition protocols, typing-based learning resulted in a higher average number of items recalled compared to speech-based learning: Dotted blue lines lie higher and further to the right than dotted green lines. Despite these numerical differences, the main effects of Typing were not significant, indicating that the proportion of correct responses did not differ significantly between both input modalities, both at the immediate test and at the followup test ($z = 0.91$, $p = 0.362$; $z = 0.22$, $p = 0.829$, respectively).

[7] A logistic mixed effects model that considers both immediate and long-term testing in the same model, can be found in the online supplementary materials, see https://osf.io/y7eph For ease of interpretation, the qualitatively similar separate models are reported here.

Table 3
Linear mixed-effects models explaining RTs for correct answers on test from learning condition and modality.

RT (ms)	β	SE	df	t	p
Intercept	2142.69	67.53	74.95	31.731	<0.001***
Repetition protocol	−487.39	59.86	2561.31	−8.143	<0.001***
Response modality	−141.69	60.97	2598.27	−2.324	0.020*
Protocol × Modality	603.52	119.90	2570.94	5.034	<0.001***

Third, the interaction effects of item repetition protocol and response modality were not significant, in both the immediate test model and the long-term retention model, which indicates that the adaptive learning benefits did not significantly differ between input modalities ($z = -1.00$, $p < 0.317$; $z = -0.179$, $p = 0.857$, respectively).

Since RTs for correct responses may be used as an additional measure of performance, we conducted a linear mixed-effects model to explore the effects of repetition protocol and modality on RTs for correct items.[8] Table 3 shows that RTs for correct responses were, on average, 974 ms faster after MemoryLab learning than after Leitner-inspired learning ($t(2561) = -8.143$, $p < 0.001$). The effect of modality was also significant, indicating that typed responses were on average 283 ms faster than spoken responses ($t(2598) = -2.324$, $p = 0.020$). Finally, the interaction effect of repetition protocol and modality was significant, indicating that the difference in RTs between MemoryLab and Leitner-inspired was larger after speech-based learning than after typing-based learning ($t(2551) = 5.03$, $p < 0.001$). These results show that the benefits of RT-adaptive learning relative to Leitner-inspired learning in terms of retrieval speeds for correct answers were larger after speech-based learning than after typing-based learning.

5. Discussion

The two goals of this study were (1) to compare typing-based learning to speech-based learning, and (2) to explore adaptive learning benefits for each of these input modalities. The results of this study can be summarized in four main points—the first addressing (1) and the other three speaking to (2). First, we show relatively similar overall performance for typing- and speech-based learning: We were unable to detect significant differences in correctness between typing- and speech-based learning. Second, we replicate earlier studies by demonstrating that adaptive learning results in higher learning efficiency for typing-based learning, as conveyed by lower RTs and higher accuracy on (long-term retention) test. Third, we show that these adaptive learning benefits generalize to speech-based learning. Finally, adaptive learning benefits did not differ significantly in size between typing- and speech-based learning. Overall, these results demonstrate that speech-based vocabulary learning using real-time ASR technology is a practically feasible alternative to typing-based vocabulary learning, and that it is possible to personalize and improve such speech-based learning systems using an RT-based adaptive scheduling algorithm. Below, we first discuss these results in more detail, and then consider their theoretical and practical implications.

There are a number of clear differences between the processes involved in the production of spoken versus typed retrieval attempts. For example, typing involves the retrieval of the orthographic representation of a word, whereas speaking involves the retrieval of its phonological representation. Notwithstanding these obvious differences, based on studies in psycholinguistics (Aitchison, 2012; Jiang, 2000) and cognitive psychology (Anderson and Schooler, 1991; Jescheniak and Levelt, 1994; Levelt, 1999; Wilschut et al., 2021), we expected some functional similarity between the memory processes involved in storing and retrieving spoken and typed representations of words. Indeed, we found relatively similar learning outcomes for both typed and spoken word learning: Correctness on both tests was not significantly different between typing- and speech-based learning. However, we did find longer average RTs for spoken compared to typed responses. We speculate that these longer RTs for spoken recalls do not necessarily indicate poorer memory performance, but could also reflect higher effort and demands on attention (Hopman and MacDonald, 2018), as well as additional preparatory processing[9] associated with producing spoken responses. Future studies should focus specifically on examining differences in mechanisms underlying the production of responses to memory cues for typing- and speech-based. If spoken response times are indeed reliably slower in comparison to typed response times because producing spoken responses is more effortful than producing a typed response, the implementation of modality-specific RT corrections (such that slightly slower RTs for speech result in the same memory activation as slightly higher RTs for typing) might be beneficial. In conclusion, as hypothesized, we failed to detect convincing evidence in favor of one of the modalities in learning performance.

The second goal of this study was to compare adaptive learning benefits for typing- and speech-based learning. Given the assumed similarities between memory processes, we hypothesized that the MemoryLab system can successfully improve item scheduling by using response times to estimate when a learner would forget an item, for both typing- and speech-based inputs. The results confirmed our hypothesis: We found consistently better performance after learning with MemoryLab relative to learning with the Leitner-inspired system, both on measured on a test immediately after the learning session and on a test after 4–8 days. To our knowledge, we are the first to demonstrate that these adaptive learning benefits effects hold both for typing and for ASR-driven learning, and both for performance on the immediate test and for long-term retention. The adaptive learning benefits found for typing-based learning were not significantly different from the adaptive learning benefits for speech-based learning in size, further

[8] See Supplementary Figure 3 at https://osf.io/28zaj/ for a graphical representation of RT distributions for correct responses on test.

[9] For example, Torreira et al. (2016) and Indefrey and Levelt (2004) show that spontaneous voice responses (like the responses required in the current speech-based learning setup) often require preparatory ‘planning’.

stressing the universal applicability of the RT-based scheduling algorithm. The learning session analyses reported above explain why MemoryLab adaptive learning was successful in all conditions: In contrast to the Leitner-inspired flashcard algorithm, the RT-based MemoryLab algorithm was able to estimate the strength of item representations in memory, and use this information to exploit the beneficial effects of maximal retrieval practice and spacing repetitions over time.

In this study, we are the first to present a fully adaptive, standalone speech-based learning system and demonstrate its effectiveness on immediate and long-term retention of vocabulary items. A number of factors should be taken into account when interpreting the results of this study. First, it is important to note that the current study was conducted in a controlled, quiet environment, and that the performance of the ASR system in real-world (classroom) situations should be examined in future studies. With the implementation of the Levenshtein's edit distance, we could here correct minor ASR-transcription errors and achieve relatively high speech recognition accuracy. It should be noted, however, that this solution only works with a limited, predefined item set, in which all items are sufficiently distinct from one another. For more scalable applications with larger item sets, other solutions have to be considered. Finally, the idea that using RTs contributes to predicting memory retrieval success has proved to be useful for vocabulary items. Future research should examine its usefulness for more complex or non-declarative information.

Even when carefully interpreted, the results of this study are important in two ways. First, the fact that the RT-adaptive MemoryLab algorithm was able to successfully estimate the strength of item representations in memory for both typing- and speech-based learning provides further insights in the mechanisms underlying both input modalities. This finding suggests that, while typing- and speech-based learning differ in various aspects and rely on separate preparatory and production processes, they share functionally similar mechanisms that can be simulated using the same cognitive model of declarative memory. Next to their theoretical implications, the results of this study can assist the development of educationally highly relevant language learning applications that focus on practicing speech. Importantly, the current system is the adaptation of an existing system that was originally developed for typing-based learning. Here, we only changed the response modality; the used model remained unmodified. We argue that it is reasonable to assume that the methods used in the current study (i.e., directly replacing typed input by ASR) should yield successful results in other existing typing-based learning systems that rely on the same general assumptions. Therefore, our findings are valuable for a wide range of intelligent tutoring systems, cognitive tutors or other systems that use behavioral learning indices to improve learning processes that are now limited to typed or keypress input.

5.1. Conclusion

In summary, in this study we present a speech-based vocabulary learning system that adapts to the needs of individual users. We show that typing- and speech-based inputs result in relatively similar learning outcomes, and that using RTs and accuracy scores to personalize learning sessions improves both typing- and speech-based learning to an equal extent. These results open the way for the development of educationally relevant speech-based language learning applications that facilitate practicing speech during vocabulary learning.

CRediT authorship contribution statement

Thomas Wilschut: Designed the study, Data curation, Formal analysis, Writing – original draft. **Florian Sense:** Designed the study, Writing – review & editing. **Hedderik van Rijn:** Designed the study, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

All data/code is available at: https://osf.io/cpfsq/.

References

Aitchison, J., 2012. Words in the Mind: An Introduction to the Mental Lexicon, fourth ed. John Wiley & Sons.
Anderson, J.R., Bothell, D., Lebiere, C., Matessa, M., 1998. An integrated theory of list memory. J. Mem. Lang. 38 (4), 341–380.
Anderson, J.R., Schooler, L.J., 1991. Reflections of the environment in memory. Psychol. Sci. 2 (6), 396–408.
Baayen, R.H., Davidson, D.J., Bates, D.M., 2008. Mixed-effects modeling with crossed random effects for subjects and items. J. Mem. Lang. 59 (4), 390–412.
Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. J. Stat. Softw. 67 (1), 1–48.
Bryson, D., 2012. Using flashcards to support your learning. J. Vis. Commun. Med. 35 (1), 25–29.
Byrne, M.D., Anderson, J.R., 1998. Perception and action. In: The Atomic Components of Thought, Vol. 16. pp. 23–28.
Cepeda, N.J., Vul, E., Rohrer, D., Wixted, J.T., Pashler, H., 2008. Spacing effects in learning: A temporal ridgeline of optimal retention. Psychol. Sci. 19 (11), 1095–1102.
Daniels, P., 2015. Using web speech technology with language learning applications. Jalt Call J. 11 (2), 177–187.
De Leeuw, J.R., 2015. jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. Behav. Res. Methods 47 (1), 1–12.
Epp, C.D., McCalla, G., 2011. ProTutor: Historic open learner models for pronunciation tutoring. In: International Conference on Artificial Intelligence in Education. Springer, pp. 441–443.

Fendji, J.L.K.E., Tala, D.C., Yenke, B.O., Atemkeng, M., 2022. Automatic speech recognition using limited vocabulary: A survey. Appl. Artif. Intell. 36 (1), 2095039.

Filippidou, F., Moussiades, L., 2020. A benchmarking of IBM, Google and wit automatic speech recognition systems. In: Artificial Intelligence Applications and Innovations: 16th IFIP WG 12.5 International Conference, AIAI 2020, Neos Marmaras, Greece, June 5–7, 2020, Proceedings, Part I 16. Springer, pp. 73–82.

Hopman, E.W., MacDonald, M.C., 2018. Production practice during language learning improves comprehension. Psychol. Sci. 29 (6), 961–971.

Indefrey, P., Levelt, W.J., 2004. The spatial and temporal signatures of word production components. Cognition 92 (1–2), 101–144.

Jescheniak, J.D., Levelt, W.J., 1994. Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. J. Exp. Psychol: Learn. Mem. Cogn. 20 (4), 824.

Jiang, N., 2000. Lexical development and representation in a second language. Appl. Linguist. 21 (1), 47–77.

Karpicke, J.D., Bauernschmidt, A., 2011. Spaced retrieval: absolute spacing enhances learning regardless of relative spacing. J. Exp. Psychol: Learn. Mem. Cogn. 37 (5), 1250.

Këpuska, V., Bohouta, G., 2017. Comparing speech recognition systems (Microsoft API, Google API and CMU Sphinx). Int. J. Eng. Res. Appl. 7 (03), 20–24.

Kimura, T., Nose, T., Hirooka, S., Chiba, Y., Ito, A., 2019. Comparison of speech recognition performance between Kaldi and Google cloud speech API. In: Recent Advances in Intelligent Information Hiding and Multimedia Signal Processing: Proceeding of the Fourteenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, November, 26-28, 2018, Sendai, Japan, Volume 2 14. Springer, pp. 109–115.

Kornell, N., 2009. Optimising learning using flashcards: Spacing is more effective than cramming. Appl. Cogn. Psychol.: Off. J. Soc. Appl. Res. Mem. Cogn. 23 (9), 1297–1317.

Levelt, W.J., 1999. Models of word production. Trends Cogn. Sci. 3 (6), 223–232.

Lindsey, R.V., Shroyer, J.D., Pashler, H., Mozer, M.C., 2014. Improving students' long-term knowledge retention through personalized review. Psychol. Sci. 25 (3), 639–647.

Litman, D., Strik, H., Lim, G.S., 2018. Speech technologies and the assessment of second language speaking: Approaches, challenges, and opportunities. Lang. Assess. Q. 15 (3), 294–309.

Mettler, E., Massey, C.M., Kellman, P.J., 2011. Improving adaptive learning technology through the use of response times. In: Proceedings of the 33rd Annual Meeting of the Cognitive Science Society, Vol. 1. pp. 2532–2537.

Mettler, E., Massey, C.M., Kellman, P.J., 2016. A comparison of adaptive and fixed schedules of practice. J. Exp. Psychol. [Gen.] 145 (7), 897.

Moreira, B.F.T., Pinto, T.S.S., Starling, D.S.V., Jaeger, A., 2019. Retrieval practice in classroom settings: a review of applied research. Front. Educ. 4, 5.

Mubarak, R., Smith, D.C., 2008. Spacing effect and mnemonic strategies: A theory-based approach to E-learning. In: E-Learning. pp. 269–272.

Munteanu, C., Lumsden, J., Fournier, H., Leung, R., D'Amours, D., McDonald, D., Maitland, J., 2010. ALEX: mobile language assistant for low-literacy adults. In: Proceedings of the 12th International Conference on Human Computer Interaction with Mobile Devices and Services. pp. 427–430.

Munteanu, C., Molyneaux, H., Maitland, J., McDonald, D., Leung, R., Fournier, H., Lumsden, J., 2014. Hidden in plain sight: low-literacy adults in a developed country overcoming social and educational challenges through mobile learning support tools. Pers. Ubiquitous Comput. 18 (6), 1455–1469.

Nakata, T., 2017. Does repeated practice make perfect? The effects of within-session repeated retrieval on second language vocabulary learning. Stud. Second Lang. Acquis. 39 (4), 653–679.

Nassif, A.B., Shahin, I., Attili, I., Azzeh, M., Shaalan, K., 2019. Speech recognition using deep neural networks: A systematic review. IEEE Access 7, 19143–19165.

Nelson, T.O., Dunlosky, J., 1994. Norms of paired-associate recall during multitrial learning of Swahili-English translation equivalents. Memory 2 (3), 325–335.

Papousek, J., Pelánek, R., Stanislav, V., 2014. Adaptive practice of facts in domains with varied prior knowledge. In: Educational Data Mining 2014. pp. 6–13.

Pavlik, P.I., Anderson, J.R., 2008. Using a model to compute the optimal schedule of practice. J. Exp. Psychol.: Appl. 14 (2), 101.

Plag, I., Homann, J., Kunter, G., 2017. Homophony and morphology: The acoustics of word-final S in English1. J. Linguist. 53 (1), 181–216.

R Core Team, 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL: https://www.R-project.org/.

Roediger, H.L., Karpicke, J.D., 2006. The power of testing memory: Basic research and implications for educational practice. Perspect. Psychol. Sci. 1 (3), 181–210.

Rummer, R., Schweppe, J., Gerst, K., Wagner, S., 2017. Is testing a more effective learning strategy than note-taking? J. Exp. Psychol.: Appl. 23 (3), 293.

Sense, F., Behrens, F., Meijer, R.R., Van Rijn, H., 2016. An individual's rate of forgetting is stable over time but differs across materials. Top. Cogn. Sci. 8 (1), 305–321.

Sense, F., Van Rijn, H., 2022. Optimizing fact-learning with a response-latency-based adaptive system. PsyArXiv. January, 27.

Sense, F., Van der Velde, M., Van Rijn, H., 2021. Predicting university students' exam performance using a model-based adaptive fact-learning system. J. Learn. Anal. 1–15.

Settles, B., Brust, C., Gustafson, E., Hagiwara, M., Madnani, N., 2018. Second language acquisition modeling. In: Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications. pp. 56–65.

Settles, B., Meeder, B., 2016. A trainable spaced repetition model for language learning. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1848–1858.

Shadiev, R., Chien, Y.-C., Huang, Y.-M., 2020. Enhancing comprehension of lecture content in a foreign language as the medium of instruction: comparing speech-to-text recognition with speech-enabled language translation. SAGE Open 10 (3), 2158244020953177.

Shadiev, R., Liu, J., 2022. Review of research on applications of speech recognition technology to assist language learning. ReCALL 1–15.

Torreira, F., Bögels, S., Levinson, S.C., 2016. Breathing for answering. The time course of response planning in conversation. Front. Psychol. 6.

Van Rijn, H., Van Maanen, L., Van Woudenberg, M., 2009. Passing the test: Improving learning gains by balancing spacing and testing effects. In: Proceedings of the 9th International Conference of Cognitive Modeling, Vol. 2. pp. 7–6.

Van der Velde, M., Sense, F., Borst, J., Van Rijn, H., 2021a. Alleviating the cold start problem in adaptive learning using data-driven difficulty estimates. Comput. Brain Behav. 4 (2), 231–249.

Van der Velde, M., Sense, F., Spijkers, R., Meeter, M., Van Rijn, H., 2021b. Lockdown learning: Changes in online foreign-language study activity and performance of dutch secondary school students during the COVID-19 pandemic. In: Frontiers in Education. p. 294.

Wickham, H., 2016. Ggplot2: Elegant Graphics for Data Analysis. springer.

Wilschut, T., Sense, F., Van der Velde, M., Fountas, Z., Maass, S., Van Rijn, H., 2021. Benefits of adaptive learning transfer from typing-based learning to speech-based learning. In: Frontiers in AI and Big Data, Vol. 4. Frontiers Media.

Wozniak, P.A., Gorzelanczyk, E.J., 1994. Optimization of repetition spacing in the practice of learning. Acta Neurobiol. Exp. 54, 59.

Yu, D., Deng, L., 2016. Automatic Speech Recognition. Springer.

Yujian, L., Bo, L., 2007. A normalized Levenshtein distance metric. IEEE Trans. Pattern Anal. Mach. Intell. 29 (6), 1091–1095.

Zhou, P., Sense, F., Van Rijn, H., Stocco, A., 2021. Reflections of idiographic long-term memory characteristics in resting-state neuroimaging data. Cognition 212, 104660.