Optimize learning with reaction time based spacing

By modifying the order of items in a learning session

Wendy van Thiel
s1466917
March 2010

Supervisors:
Dr. Hedderik van Rijn (Artificial Intelligence)
Dr. Niels Taatgen (Artificial Intelligence)

Abstract

Can we optimize learning efficiency by modifying the order of items in a learning session?

By taking into account well-known memory phenomena, we can improve learning. However, in practice, learning methods that take into account memory effects such as primacy, recency and spacing are not often used. Especially the characteristics of the spacing effect, which refers to enhanced learning when trials are spaced over time instead of massed in a short time, are only rarely applied. This study proposes an adaptive cognitive model that takes these effects into account and is easily applicable in practice. This model lets people learn facts effectively and is a refined version of models from previous research by Van Woudenberg (2008) and Pavlik and Anderson (2008). Just as in these studies, the new model keeps a representation of the strengths of items in memory. As in the study of Van Woudenberg (2008), the memory strengths are based on response times of the user, but now based on a more direct and continuous measure. On the basis of the strength representations and a forecast of the development of these strengths, the model optimizes the word order. This model is compared to a standard teaching schedule in an experiment done with three Havo/VWO classes. In a learning session of fifteen minutes, students studied twenty Dutch-French word pairs. The next day, performance of these word pairs was tested. The training data of the experiment showed that the model's prediction of response times, e.g. its representation of strength of items in memory, improved as repetition increased. Analyses of the test data showed that participants in the adaptive condition scored on average 1.1 point higher on a scale of 1 to 10 than the control condition. Although more refinements are still possible, this work confirmed that spacing through adaptation based on reaction times yields an effective learning method.

Acknowledgement

First of all I would like to thank my supervisor, Hedderik van Rijn, for his support. He was always willing to answer questions when possible, took the time for every discussion until both parties understood and agreed with the decisions made and always respected personal circumstances. Secondly, I would like to thank Wim Woudman and Fred Stevens (dr. Aletta Jacobs College) who enabled me to conduct my research experiments. Without their help I would not have been able to apply this research in a real-life setting.

Contents

1. Introduction

In 1885, Hermann Ebbinghaus taught himself lists of nonwords and discovered that he remembered them better when learning them over a period of time, known as spaced presentation, instead of memorizing them intensively over a short period, known as massed presentation (1964, 1885). This phenomenon is called the spacing effect, and refers to enhanced learning when (study) trials are spaced.
Spaced learning is often compared with *massed learning*, which occurs when trials of one item are presented without interruption of either a time interval or other trials (e.g. *aaabbbccc*). In s*paced learning*, an item always has an interval that consists of a pause, other trial sessions – or both – before it is rehearsed again (e.g. *abc abc*). When spacing multiple sessions over time, distinction is made between interstudy interval (ISI) and retention interval (RI). The ISI is the interval between study sessions of the same material. The RI is the time separating the last study session and the test moment.
Thanks to extensive studies, it is now known that the spacing effect occurs in different task types, such as in the learning of vocabulary, facts and motor tasks (Cepeda et al., 2006; Donovan & Radosevich, 1999). Prior to the current decade, the effect had not yet been demonstrated in more complex tasks or tasks with perceptual properties (Donovan & Radosevich, 1999; Rickard et al., 2008). However, tasks such as the learning of vocabulary or facts are very common in daily practice, especially for students. Many students could therefore benefit from spaced learning, although it is rarely applied.
In his case study, Dempster (1988) listed several reasons why application of the spacing effect failed in practice. He states for instance that '*the phenomenon has not been demonstrated satisfactorily in the classroom'* and that '*the phenomenon cannot be linked to issues of current concern to educators'*. The first argument no longer holds true since recent successful demonstrations of learning with spaced methods in classrooms (Van Woudenberg, 2008; Bloom, K.C. & Shuell, T.J., 1981). Neither is the second argument completely valid because the rise of computerization in education has made it easier to use computer programs which take the spacing effect into account. But the arguments of Dempster still hold true in the sense that all results combined do not provide teachers with a method that can easily be implemented in classroom settings.
Finding an application that exploits the spacing effect *and* is easy to use in real-life situations is complex. Firstly, the effect does not follow a consistent pattern. The optimal spacing depends on the number of training sessions, the study time per session, and on the moment that the knowledge is needed for recall. For example, when the retention interval lengthens, there is more benefit if the ISIs are also longer (Pavlik & Anderson, 2008). On the other hand, the effect decreases when intervals are too long (Capeda et al., 2006). Furthermore, when the test is conducted directly after a learning session, spaced presentations offer no advantage at all (Dempster, 1988).
Secondly, data results can differ because of heterogeneity in tasks, number of learning sessions, amount of knowledge, study intervals, session lengths and retention intervals. This makes it complex to draw a one-size-fits-all conclusion. Although some researchers are attempting to create a model of long-term memory that fits the results of more studies (Mozer, Pashler, Lindsey & Vul, submitted), they do not account for all specifications above and are therefore only applicable to a limited extent.
Thirdly, if a formula that takes these specifications into account were to exist, it would still be a generalization that does not match individual characteristics *per se*.
Lastly, even if a model could account for tasks and individual specifications, this does not mean that the model would, in practice, produce useful schedules. After all, not all individuals are able to adapt their real-life situations to a learning schedule, however optimal.

In conclusion, finding an optimally spaced schedule seems possible only if specifications for the task, the individual and the practice situation are given. However, the length and number of training sessions are decided by the individuals themselves and are possibly not known in advance. Should individuals not incorporate a practice scheme into their daily routine, the quality of their learning will depend on the quality of the learning sessions and their number. The only benefit achievable would then depend on the order of study trials in a learning session itself. This can be done with a model that is a good representation of the strength and speed of decay of items in memory. Predictions about the strength of the items in memory made according to such a representation allow one to anticipate which item(s) need to be learned to prevent forgetting. The representation should additionally adapt to individual characteristics. These adaptable models were only created during this decade (De Boer, 2003; Van Woudenberg, 2008; Pavlik & Anderson, 2005; Pavlik & Anderson, 2008), and showed promising results.

The models of Van Woudenberg (2008) and Pavlik and Anderson (2008) create user-dependent representations of the strength of items in memory and attempt to optimize the word order in study sessions. They are similar and robust, but can still be further improved on. This thesis proposes a more flexible and refined model for the learning of word pairs based on earlier work by Van Woudenberg (2008) and Pavlik (2005, 2008). It tries to answer the question of whether these refinements lead to representations that can make accurate predictions about the strength of items in memory and whether modifying the word order can further optimize retention.

2. Introduction of a latency-based ACT-R model

This chapter describes the means used to optimize the word order in a session. We aim to develop a cognitive model that captures memory effects, adapts to users and thus is able to modify the word order with better retention. The cognitive model is based on the ACT-R architecture. This architecture can capture memory effects, which will be explained in section 2.1. Existing cognitive models of Van Woudenberg (2008) and Pavlik and Anderson (2008) had promising results and will therefore be compared to define advantages and disadvantages of these models for further optimization. Section 2.3 introduces a revision of the dynamic spacing/reaction time condition of Van Woudenberg (2008).

2.1 Spacing in ACT-R

ACT-R (Adaptive Character of Thought - Rationale) is an architecture of cognition (Anderson 2007, 2004). It describes the process of acquiring and reproducing knowledge based on prior practice. The architecture is extensive, but this research focuses only on the equations for storing facts to memory and retrieving facts from memory, e.g., the basis of the declarative memory. Anderson and Schooler (1991) proposed the first equations based on characteristics found after analysis of human memory behaviour. One can, they theorized, calculate the activation with these equations. This activation represents the strength of an item in memory. The main characteristics found were that this strength is a distribution that can vary within an infinite domain (1). The strength of each individual encounter or presentation decays as a power function over time (2) and summed together produce a total strength, e.g., its activation (3). To account for the spacing effect, they proposed separate decays for each encounter since the previous presentation as a function of time. A big interval between the presentations would then lead to a low decay, so that the presentation has a greater effect on the strength in memory than a presentation after a short interval.

The activation function as it is nowadays in ACT-R (1) corresponds with assumptions 1-3. The function returns an activation value of an item for the current time (t). A high activation means that recall of an item will be accurate and fast. A low activation means that the recall will be slow or might not happen at all. The activation is based on the summation of the individual practice events. Each practice event is based on the difference between the current time and moment of practice (t_j) in seconds. The strength of these practice events decreases as this interval grows with a rate of forgetting, the decay $-d_{i,j}$, where i stands for the item and j for the number of repetitions. The total activation is then the natural logarithm over the summation of these individual practice events.

$$m_i(t) = \ln\left(\sum_{j=1}^{n;t_j<t} \left(t - t_j\right)^{-d_{i,j}}\right) \tag{1.1}$$

In 2005, Pavlik and Anderson introduced the decay function of (2) to create unique decay values for the individual encounters and dependent on the time interval between the encounters.

$$d_{i,j} = ce^{m_i(t_j)} + \alpha \tag{1.2}$$

It determines the decay of an item (i) for a practice event (j) on its activation at the moment of the event (t_j), scaled with the scaling parameter c and added to a constant parameter, α. The constant α also designates the minimum decay value; the scaling factor determines the

spacing effect. By using individual decays, a distinction can be made between the impact on learning brought about by the individual encounters. As a result, repetition in short intervals gives repetition on high activations, which leads to higher decays. This implies that many rehearsals or practice events of an item over a short period lead to high decay rates for the individual encounters. Over time, this leads to relatively fast forgetting. However, a longer retention interval, or more spacing, leads to a lower decay, which creates benefit in the long term.
Both the models of Van Woudenberg (2008) and of Anderson and Pavlik (2008) are based on these two equations and achieved promising retention results. Before introducing the model used in this research, the next section will compare these models as to how they adapt to users and determine their word order.

2.2 Comparison of two ACT-R models

Both Van Woudenberg (2008) and Pavlik and Anderson (2008) have created ACT-R-based models which adapt to users in different ways. In Van Woudenberg, pupils learned the Dutch translation of twenty French words during class in a fifteen-minute training session. One day later, all participants received a pencil-and-paper test to see which words they still remembered. Pavlik and Anderson taught participants 180 Japanese-English word pairs, in three learning sessions of one hour each, on days one, three and five. An assessment session was done a week later. Although the participants in both studies learned word pairs, the studies are not comparable because of the differences in the learning sessions, retention intervals, session lengths and the numbers of words used. Furthermore, it is doubtful whether the tasks are comparable because it tends to be more difficult for native English speakers to learn Japanese words than it is for native Dutch speakers to learn French words. The French words which are learned come from a language with which there is some acquaintance. This will bring more foreknowledge and background information that can be used during the learning. And when participants have foreknowledge, differences in individual performances will be greater, because besides the learning capacity some will have more benefit from their foreknowledge than others. However, both models are based on the ACT-R equation of memory strength including the decay function as discussed earlier. Therefore, it is possible to compare these models with each other.
In the next section, we will compare the different types of adaptations for optimizing the word order and the consequences of these adaptations. Based on this analysis and the results of these studies, we have determined the adaptation method for the model used in this research. And, because the activation and decay formulas do not directly determine which word should be rehearsed next, the section ‘Determining the word order’ describes and analyzes the different ways of modifying the word order.

2.2.1 User adaptation/model

When no prior knowledge about a user is known, adaptation to the user’s performance is only possible by using the correctness or reaction time of the response. There are no other relevant measurements that lead to adaptation in the ACT-R formulas that discriminate between the capacity of the users. Pavlik and Anderson only adapt for correctness, whereas Van Woudenberg implements a condition that adapts for correctness and a condition that adapts for both correctness and reaction time. This section compares the different ways in which these measurements are processed.
Pavlik and Anderson (2008) introduce three new parameters in the activation function to capture item, participant and item-participant differences. The three parameters βi (item difference), βs (participant difference) and βsi (participant-item difference) are additional values in the activation equation, see 2.1.

$$m_n(t_{..n}) = \beta_s + \beta_i + \beta_{si} + \ln\left(\sum_{k=1}^{n} {t_k}^{-d_k}\right) \quad (2.1)$$

The βs and βi are updated every 300 trials while the βsi is updated after every response based on the success or failure performance with a Bayesian algorithm. Unfortunately, exactly how this is done remains unclear. However, the parameters are additional to the old activation formula (1) and this creates, as noted earlier (Van Rijn, 2009), a minimum activation of an item. As a result, and independent of how much time passes, the activation will never drop below this level. Also, since decay and activation are recursive functions, the chosen β values will influence the decay positively. A higher decay will lead to faster forgetting and will influence the activation both indirectly and negatively over time. In their paper, only the standard deviations of βs and βsi are given, but it is unclear what the range of these parameters is. Therefore, it is difficult to estimate the real influence. However, the question still remains of whether these side effects are desirable.

In the work of Van Woudenberg (2008), only the alpha in the decay is adapted for each participant-item combination. By influencing the rate of forgetting separately, the recursive side effect on the longer term has less impact. Of course, the activation for the next encounter is influenced, but this time not in a counter-productive manner. In addition, no extra parameters are required. The only disadvantage is that no foreknowledge regarding the item or participant is used. However, a concrete value judgment cannot be made since it is unclear how this is done in the work of Pavlik and Anderson, and which benefits are reached.

Van Woudenberg (2008) decided to adapt the alpha parameter of the decay function in a robust way, with a maximum of 0.01 per trial. The adjustment made is based either on the correctness of the response or a combination of correctness and reaction time. The condition that adjusts on correctness is called the dynamic spacing - response condition. When the model predicts a correct response but an incorrect response is given, the alpha value increases by 0.01. Or, when the subject unexpectedly answers correctly, the alpha decreases by 0.01. The adjustment based on correctness and latency is called the dynamic spacing - reaction time condition and adapts the alpha after each response, depending on correctness or the difference between expected and observed reaction time. After an incorrect answer, the alpha is raised by 0.01. The difference between the measured reaction time and the expected reaction time is calculated after an incorrect answer. If this difference is larger than 0.5s, the alpha is adjusted according to the difference, with a maximum adjustment of 0.01.

Although both Van Woudenberg's adaptive conditions scored better then a non-adaptive condition, only the reaction-time condition scored better then the control condition, which was a flashcard method that will be explained later. Pavlik and Anderson (2008) had already stated that adaptation based on latency might be useful to describe learning processes, but chose otherwise because failure latencies do not correlate with learning. Given this argument, it seems that a combination of correctness and latency is not considered as an option for their model. However, the Van Woudenberg results indicate that the combination of latency and correctness could be a good measure to which to adapt.

2.2.2 Determining the word order

As noted in the introduction of this chapter, activation values and decays do not directly determine which word should be repeated next. Determining the next word pair directly by the lowest activation is not a realistic option. This method is naive because the application would offer all words once before rehearsal takes place. After this, irrespective of the adaptation, there is no real optimization process. Adaptations made can increase or decrease activation values, but this does not by definition lead to a situation in which all activations

will be above a forgetting threshold at the end of a learning session. Finding the optimal word order thus takes more thoughts.

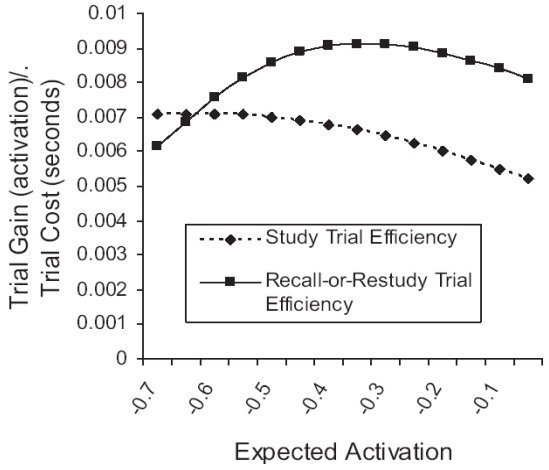


Figure 1.1: Efficiency functions for recall-or-restudy trials and study-only trials as a function of current activation for a retention interval of 9 days (the mean expected retention interval in the current experiment). (Pavlik & Anderson, 2008)

Pavlik and Anderson chose to optimize according to their self-defined learning rate. This learning rate is the gain in activation over nine days when rehearsing a word now divided by the time it took to practise the item, see formula 2.2:

$$learning\ rate_n = \frac{activation\ gain\ at\ retention\ test\ for\ item_n}{time\ cost\ now\ to\ practise\ item_n} \qquad (2.2)$$

The time cost in a recall-or-restudy is defined as: the time cost for latency times the chance of a correct response added to the time cost of a failed trial times the chance of an incorrect answer. They plotted the results of the gain depending on the present activation, see Figure 1.1. The black line denotes the gain for a recall trial, the dotted line for a study trial. From these results they concluded that when the activation is below -0.63, there is more gain in offering a study trial. Optimum benefit can be reached when the activation is at -0.33. For the algorithm, see Figure 1.2. When determining which word to recall or restudy next, these values are used as reference. Pavlik and Anderson claim that this model aims at a global optimum policy instead of local optimum policy, because it is not the maximum learning rate of each item on each trial that is used, but the overall maximum learning gain that can be attained by the independent items. To optimize for gain at retention is an excellent consideration, because performance at the moment of retention is used to measure model performance. However, the straightforward implementation raises some questions.
Firstly, calculating activation (gain) over a longer time interval causes loss of accuracy. For the intersection interval, a psychological time parameter is used. In between sessions and over longer intervals, when the subject is engaged in other activities and the working memory is not being trained continuously, less forgetting takes place.

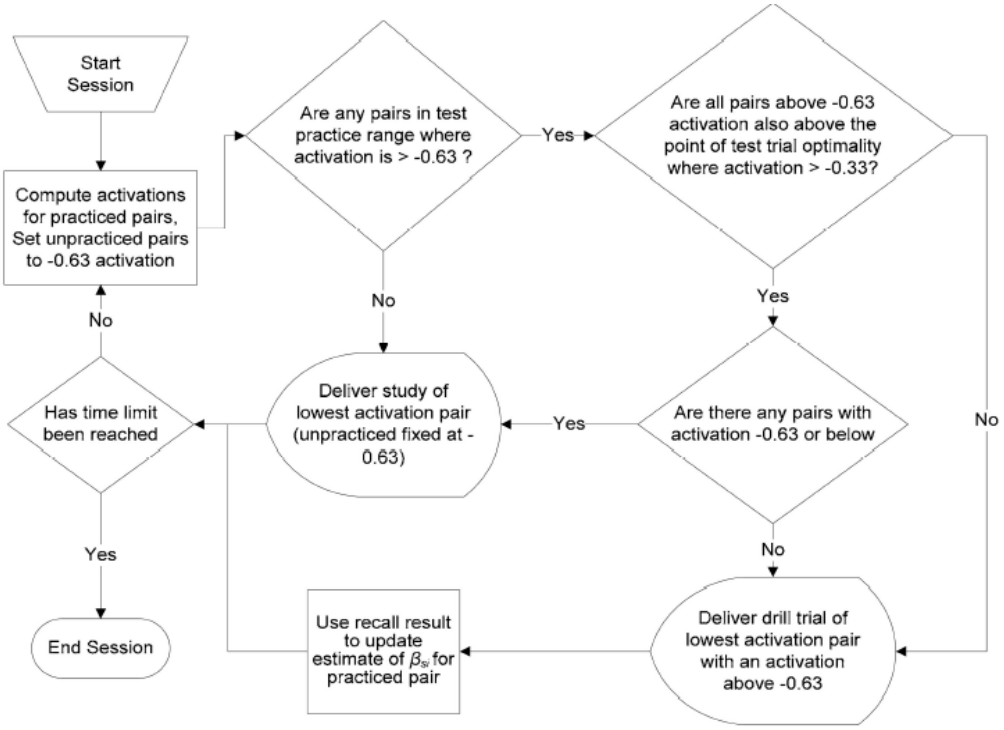


Figure 1.2: Schedule optimization algorithm flowchart. (Pavlik & Anderson, 2008)

Therefore, the time in between these intervals is scaled down with a scaling parameter *h.* In this way, time is scaled so that it proceeds at a slower rate compared to real time or the time taken during training sessions. Even though Pavlik and Anderson use a psychological time parameter for the intersection time, noise influence will increase and predictability will decrease over time. Hence, one can ask how plausible or realistic it is to optimize for small activation differences over a time interval of seven, nine, or eleven days. Pavlik and Anderson already average these different retention intervals because of the small differences. Thus, to optimize for a study trial instead of a recall because the gain in nine days will be one thousandth higher lacks real persuasiveness. Thereby, the activation gain in the learning rate is calculated based on the derived activation values generalized over a complete data set without accounting for different decays or number of rehearsals of the items learned. This means that the found range for optimal learning rate is also generalized and does not account for individual differences in, for example, the rate of forgetting. Nonetheless, these differences can also lead to other optimal ranges in learning rate.

Secondly, Pavlik and Anderson state that this results in global optimization instead of finding the maximum gain for the individual items. However, by directly implementing activation values for modifying word order in their formulas, they did not account for any rate of forgetting belonging to the individual items. However, the decay influences the activation over time and determines how much influence the separate rehearsals have over time. So, by generalizing the maximum gain on activation values, the influence of a forgetting rate is not considered. Thus, the statement that this algorithm optimizes globally is very general because it does not take into account influences of item-specific decays or user characteristics of retention.

Nevertheless, with the choice of rehearsing a word when the activation is in a certain range, Pavlik and Anderson's method does not differ much from the method used in Van Woudenberg (2008). Van Woudenberg uses a look-ahead time, which means they calculate activation not on the current time, but on a fixed interval in the future to predict whether a word pair would be below the threshold several seconds after a given moment. When

comparing the algorithms, one sees that both Pavlik and Anderson's algorithm and that of Van Woudenberg offer the word before it reaches a certain activation value. For Van Woudenberg, this is before the threshold of -0.5 is reached. Pavlik and Anderson found an optimum gain at an activation of -0.33, but waited until the activation dropped below this value and tried to rehearse the items before they reach -0.63. In a sense, this could be interpreted as a threshold.
Or, it can be interpreted as a threshold in combination with a look-ahead time. Pavlik and Anderson used a threshold of -0.704 and found the optimum learning rate at -0.33. Thus, recalling a word between -0.63 and -0.33 is comparable with a threshold of -0.73 and a non-fixed look-ahead time. Pavlik adds an additional nuance to his work by presenting a study trial instead of a rehearsal because calculation shows a more effective time investment. This means that the chance of an incorrect answer – and the possible extra time costs – cannot compensate for the smaller impact of the short duration of a single study trial. This trade-off is not present in Van Woudenberg's (2008) implementation. However, Van Woudenberg does take decay into account when calculating activation with a look-ahead time. In this way, the rehearsed word pair is not chosen based only on the activation, but on a combination of activation value and decay. When more items drop below the threshold in fifteen seconds, the one with the lowest activation in fifteen seconds is rehearsed. It is likely that this is the item with the highest decay. The fifteen-second period might look like a local optimum, but over a longer period of time, rehearsal of the items with the highest decay has a higher priority. As Pavlik and Anderson's graph indicates, this surpasses the gain that can be achieved. Or, as reasoned in Van Woudenberg and in this paper, this will increase the chance of forgetting. This, in turn, would lead to incorrect trials, extra-time costs and a possible decrease in efficiency. In conclusion, the method used by Van Woudenberg, in which the word order is determined by using a combination of a threshold and a look-ahead time has similarities with the determination based on optimal gain as indicated by Pavlik and Anderson, but takes into account the rate of forgetting of the individual items.

2.3 Implementation of latency-based spacing

With a maximum adjustment of 0.01 made according to only one response, Van Woudenberg's type of adjustment is unsubtle and slow. This means, firstly, that more repetitions are needed to create a real distinction between the participant-item combinations, and secondly, due to adaptation being only on a local scale, the influence of a single alpha is limited. The revised model also adjusts alpha values to minimize the difference between the measured reaction time and the calculated reaction time based on activation. The revised model is more subtle, because it optimizes every time for all encounters. It is faster because there is no maximum adjustment value. In this section, we will explain how we implemented adaptation of the alpha value. Next, we will explain how the word order is determined. This section explains how the alpha value is determined after repetition based on the reaction times of the user. After measuring the reaction time, some manipulations are made before the optimization process starts. We will explain these manipulations first. Thereafter, the optimization is explained.
Unfortunately, reaction times contain a great deal of noise. A small distraction can influence the reaction time by seconds. The noise is asymmetric because it will always influence the reaction time in a positive way. Therefore, we maximized the reaction times. Although it is uncertain whether these responses are slow due to noise, these reaction times influence the adaptation heavily. For example, a reaction time of 11 seconds corresponds to an activation of -2.37 in the latency formula, which is far below the threshold. This suggests that these reaction times are not solely determined by the activations of the related chunks. The

maximum reaction time used is determined with the latency formula 2.3. For consistency, latency is from now on called 'reaction time' and refers to the first key press of the user, so $L_{i,j}$ is now $RT_{i,j}$ – see formula 5. The *i* and *j* refer to the item and the number of repetitions.

$$L_{i,j} = Fe^{-m_{i,}(t_j)} + fixed\ time\ cost$$

$$RT_{i,j} = Fe^{-m_{i,}(t_j)} + fixed\ time\ cost \quad (2.3)$$

This formula consists of a fixed time and a part depending on the activation at the time of item *i* for encounter *j*. The part depending on the activation determines how quickly something can be retrieved from memory. A highly activated item will be recalled faster than an item with low activation. For example, your first name is highly activated in memory, so recall should be fast. On the other hand, the name of your old neighbour is less activated and recall should take more time. The F parameter scales this. The fixed time cost refers to the time of perception and motor actions needed to respond. The F and fixed time values are fixed at 1 and 300ms. The values are taken from Van Woudenberg (2008). Normally, this equation predicts the reaction time based on the activation of an item. We determined the fixed maximum reaction time at 3788ms, which corresponds to one-and-a-half times the reaction time of the threshold value, see equation 2.4, where τ is the threshold and *i* is the item and *j* the encounter. The threshold value is changed compared to Van Woudenberg from -0.5 to -0.8, because a pilot study showed that a threshold of around -0.8 still gave an 80% chance of a correct answer.

$$RT_{i,j} = \mathrm{Min}\left(RT_{max}, RT_{i,j}\right) \quad (2.4)$$

with:

$$RT_{max} = 1.5\ (Fe^{-\tau} + fixed\ time\ cost)$$

Next, when an incorrect response is given, the measured reaction time does not necessarily give information about the strength of the item in memory. Therefore, we did not derive the activation from the reaction time of these trials directly. However, since we adapted the alpha based on reaction times, we replaced the reaction times of incorrect answers with the maximum reaction time. In this way, the model can also adapt to incorrect answers irrespective of the applicable response time.

Before calculating the optimal alpha we used the manipulated reaction times to make an estimation of the strength of an item in memory. Again, we used the equation for latency retrieval to derive an activation from a measured reaction time.

So, recalling the latency formula:

$$RT_{i,j} = Fe^{-m_{i,}(tj)} + fixed\ time\ cost$$

Now instead of deriving a reaction time from this activation, we derived the activation from the reaction times. For this, we rewrote the latency formula. The first step is subtracting the fixed-time cost from the observed reaction time:

$$RT_{i,j} - fixed\ time\ cost = Fe^{-m_{i,}(tj)}$$

Then divide this value by the F and swap:

$$e^{-m_{i,}(tj)} = \left(\frac{RT_{i,j} - fixed\ time\ cost}{F}\right)$$

Then, to get the activation value, take the natural log and multiply it by -1:

$$m_{i(tj),} = -\ln\left(\frac{RT_{i,j} - fixed\ time\ cost}{F}\right)$$

Now we can derive an activation, or strength of an item in memory, corresponding with a measured reaction time.
To prevent confusion, in future we will call the activation derived from the latency formula the derived activation, $m_{\_derived\_i,j}$.
So:

$$m_{derived_{i,j}} , = -\ln\left(\frac{RT_{i,j} - \ fixed\ time\ cost}{F}\right) \tag{2.5}$$

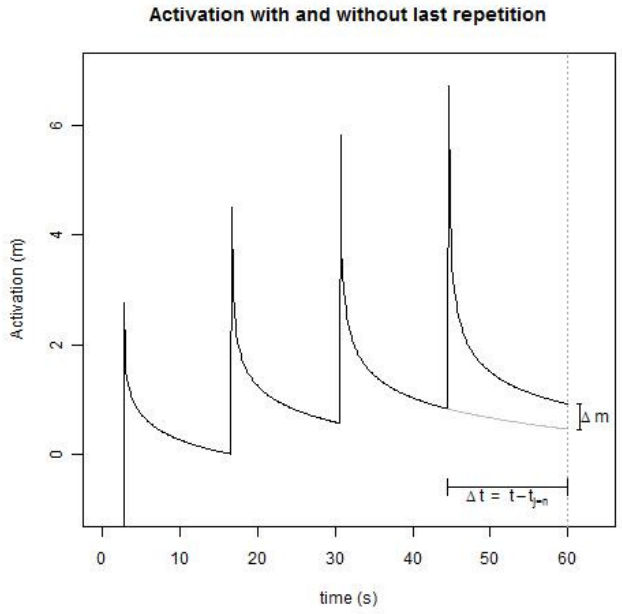


Figure 2.1: Example of activation after four encounters. The gray line denotes the activation without the fourth encounter.

Now we have measured strengths of items in memory. This derived activation is used to calculate an optimum alpha that minimizes the difference in measured and calculated reaction times, e.g. an alpha that minimizes the differences in derived and calculated activation. This optimization is done by first fitting the alpha of the latest decay. Figure 2.1 shows an example to clarify this optimization step. The figure shows the activation after four rehearsals. The fifth rehearsal takes place at t=60s. At this point a new reaction time is observed. Now, a new alpha is calculated that best fits all derived activations. To do so, first an alpha is calculated fitting only the latest decay, this is the decay belonging with the activation from t=45 till t=60 but without the activation of the first three encounters. The activation of these previous rehearsals is the dotted line, the difference between the black and gray line denotes the activation difference belonging with the fourth rehearsal. Notice that the picture is simplified because all decays are equal and optimize according to minimal differences. This means that activations derived for each encounter can be below or above these plotted activations. The alpha fitting this decay is the alpha in the decay of item i and encounter j=n at the time $t_{j=n}$. We can rewrite the activation formula (2.6) and separate decay $d_{i,j=n}$ in the formula.

$$m_i(t_j) \ = \ln\left(\left(\sum_{j=1}^{n-1;t_j<t} \left(t - t_j\right)^{-d_{i,j}}\right) + \left(t - t_{j=n}\right)^{-\boldsymbol{d_{i,j=n}}}\right) \tag{2.6}$$

Now the first part is the summation of all encounters except the last, e.g. the activation of the first three rehearsals or the dotted line in the example. The second part will create the delta activation in the figure. This decay can be mathematically derived. How this is done is explained in appendix B.
With this decay, we can rewrite the decay function already known,

$$d_{i,j} = 0.25e^{m_i(t_j)} + \alpha_{i,j}$$

To get the alpha:

$$\alpha_{i,j=n} = d_{i,j=n} - 0.25e^{m_i(t_j)}$$

This alpha is now optimized for the last encounter (j=n), thus only on the latest reaction time. To find an alpha that fits best for all encounters, we search between the last optimized alpha and the alpha just calculated and fitting the latest decay. Therefore, a binary search, explained in the flowchart in Figure 2.2. is used. After each trial, the algorithm starts at the top of this flowchart. First, the number of rehearsals is raised. Then, after the first rehearsal, which takes place directly after a study trial, a standard alpha of 0.3 is returned. After the second rehearsal, an alpha that fits the decays of the first and second rehearsal is searched for. This is done by taking an interval with a minimal alpha of 0.01, a maximal alpha of 0.5 and the mean 0.2505. Then, for both alphas and with the moment of rehearsals, the activations of these moments and corresponding reaction times are calculated. Next, the difference between these reaction

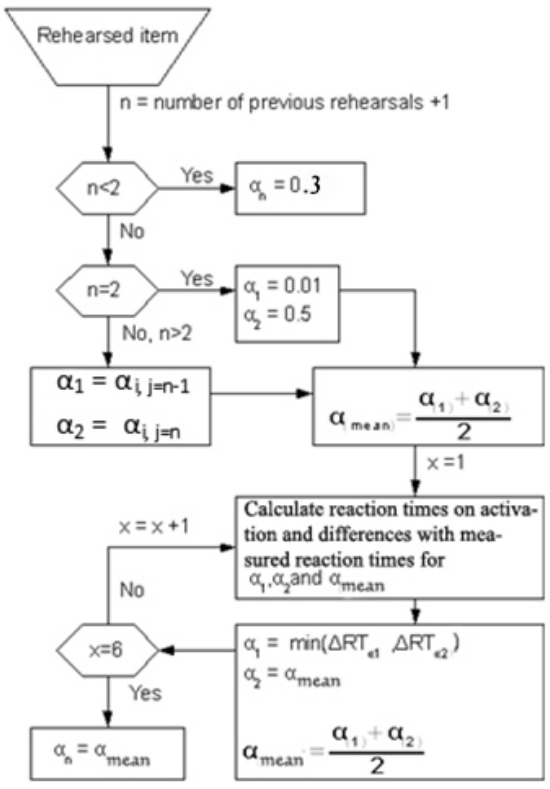


Figure 2.2: Flowchart of binary search to alpha

times and the measured reaction times are summed together. After that, the summed differences per alpha are compared with the mean. If the error of the smallest alpha is smaller than the error of the mean, the new interval will be between the small alpha of 0.01 and the mean alpha of 0.2505. If the error of the large alpha is smaller than the error of the mean, the

new interval will be between the mean alpha of 0.2505 and the 0.5. With this new interval, this process of calculating the summed differences is repeated. This is done six times. The difference in error decreases until all alpha values are very close together and the last mean value is returned. After the third rehearsal this process is done with the optimal alpha of the previous repetition and with the new calculated alpha.
Lastly, there is a difference between the model of this study and that used in Van Woudenberg (2008) by which one of the recommendation of Van Woudenberg (2008) is indirectly met. He argued that a distinction should be made between active and passive rehearsal, because '*active rehearsals are more beneficial and this will separate the difficult from the easy items. Increasing the activation of a correct rehearsed item more than a studied item, for example, will result in remembering this item for a longer time period. This does not only increase the spacing of that item, but it will also help to present new word pairs earlier in the sequence of repetitions.*' Although we did not distinguish between active and passive rehearsals, we did remove the second rehearsal of the study trial after an incorrect trial. Instead of one encounter at the attempt to recall and one at the rehearsal, we only implemented one rehearsal. More distinction can be made, but it is also reasonable to argue that a study rehearsal can consist of more (weak) repetitions in a row, particularly after an incorrect rehearsal. So, the decision was made not to change the implementation of Van Woudenberg any further.

Model parameters	α=0.3 c=0.25 τ=-0.8 Fixed time cost =0.3s F=1
Program parameters	Study time = 5s Rehearsal time = 0 to 15s Feedback time =2s Look-ahead time=15s

Table 2.1: Overview of model parameters

3. Experiment

In this study, we wanted to test whether we could develop a user-adaptive cognitive model that captures memory effects and modifies the word order for better retention results. This is done with an experiment. These retention results are compared with a standard learning method, e.g., the flashcard condition. This section has three subsections: the method, the model results and the test results. The method describes the experiment; the model results cover an analysis to see whether the model meets our expectations of the training session. It contains an analysis of the model's adaptability to individuals and its accuracy on the strength of items in memory. The results of the pencil-and-paper test of both conditions are compared in the test results. We were interested whether learning with the model improved retention, e.g., whether the group of participants in the spacing condition scored better at the test than the participants in the control condition.

3.1 Method

Participants

All participants were students from the dr. Aletta Jacobs College in Hoogezand. We tested two third-year classes – one *havo* and one *VWO* – and one fifth-year *VWO* class. We collected training and test data of 40 participants. Data from five participants were rejected for model and test results analysis. Three participants did not remember more than six words on the test set. Compared to the other participants, these participants were considered to have found the task too difficult to be representative. Two participants were rejected because of doubts as to the degree of serious effort they had invested as they came up with silly answers during study trials. A possible reason was a low level of concentration, since another lesson took place in the same room during one of the training sessions. This clearly led to problems in focusing on the task for some of the participants.

Materials

Two lists of twenty French-Dutch word pairs were compiled and approved by the teacher – one for the third-year students and one for the fifth-year students (see Appendix A). All words were selected from a chapter's word list that was to be discussed in the weeks after the experiment.

Design and procedure

The program let the students learn the Dutch translation of the French words. The learning program had two different type of trials: study or test. A study trial consisted of a five-second display of both French and Dutch words. A test trial only showed the French word and the participant had 15 seconds to type the Dutch translation. After a subject pressed Enter or the 15 seconds had passed, feedback appeared for two seconds. This was either *Correct*, *Almost Correct* if the Levensthein distance between the given and expected response was smaller than three, or *Not Correct* in all other incorrect cases. An *almost correct* or *incorrect* answer was always followed by a study or restudy trial of five seconds.

All students were randomly divided into two conditions: the spacing and the control condition. The order of the presentation of the words in the spacing condition was determined by the model described in section 2.2.2. The control condition was based on the flashcard method. This method is used more often as a control condition (Bahrick, 1797; Van Woudenberg, 2008; Pavlick & Anderson, 2007). In this condition, the word list is divided into sets of five cards. The first five words are stacked. If a word pair has never been studied before, a study trial takes place first, followed by a rehearsal. If the word has already been shown, the event is a rehearsal trial. If an item is recalled incorrectly, this item is put on the

bottom of the stack, so an extra rehearsal takes place following the rehearsal of the five words. When the stack is exhausted, it is filled with a new set of five word pairs. When the entire set has been rehearsed, the program starts over again with the first five word pairs.
After the fifteen minutes of learning had passed, the subjects were thanked for their participation, not knowing that another test moment was to follow later. On the following day, all words were tested with a traditional pencil-and-paper test. The subjects did not receive any feedback regarding these test results and were told that the results would not be graded by the school.

3.4 Results

Model results

This paragraph evaluates the predictions of the algorithms discussed in 2.3. The model creates representations of the strength of items in memory. These strengths give an indication of whether and how fast an item can be retrieved. When an item drops below the threshold, this implies that it cannot be retrieved and it is indicated as having been forgotten. The algorithm attempts to prevent this from happening. The model attempts to prevent items being forgotten by rehearsing the items before they drop below the threshold. This should lead to less incorrect responses and improvement of the predictions of the reaction time.
Firstly, we will address whether the model accurately adjusts itself to differences associated with participants and items by adjusting the alpha value. Figure and 3.1 show the mean and standard errors of the alpha values per participant and per word pair. The graphs indicate that the model adapts to the participants and words, because of the differences in alpha values and because the differences in alphas are bigger than their standard errors. Although this does not mean that the model fits the participants or words correctly, it does show that it discriminates between participants and between words. For example, the model can indicate a user as being a poor learner or a word as being harder to remember than others.
Secondly, the model tries to prevent that activation of items drops below the threshold with the look-ahead-time. Absolute prevention of incorrect responses is not possible, since (temporary) forgetting or typing errors can occur even if items are highly activated in memory. In addition, we did not define a concrete proportion of incorrect responses to be acceptable or not. Therefore, we used the control condition as the baseline.

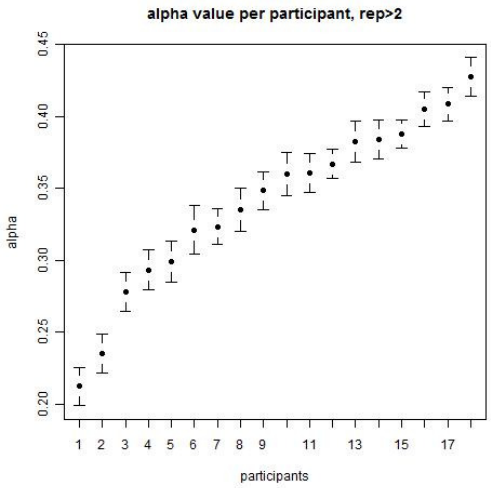


Figure 3.1: Mean and standard error of alpha values per participant. Values calculated for all encounters n>2 and sorted by means.

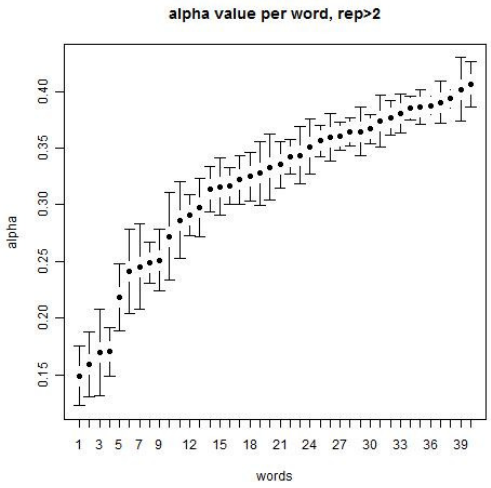
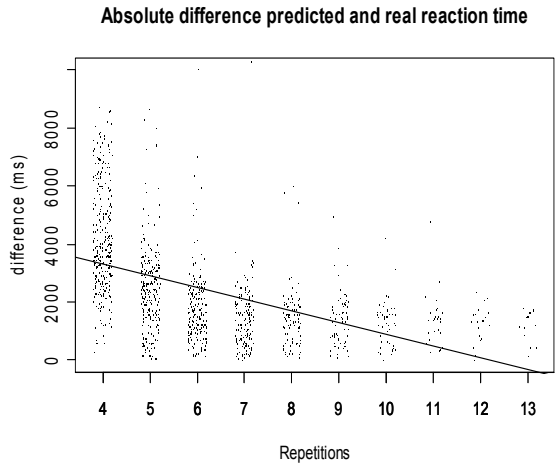
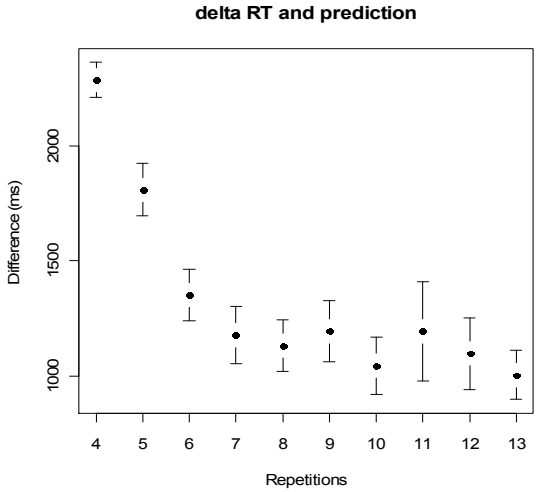


Figure 3.2: Mean and standard error of alpha value per word pair. Values calculated for all encounters n>2 and sorted by means.

As a result, it is possible to compare the number of errors with the control condition. A one-sided t test on the mean percentage of correct trials per participant confirms this difference with $t(25,81) = 3.3716$, $p < 0.01$. In addition, as noted earlier, preventing incorrect trials from happening yields a time advantage, because less study trials or repeated study trials are necessary (Pavlik & Anderson, 2008). This time can be invested in more test trials. Therefore, we also looked at the difference in the number of trials. Participants in the spacing condition had on average 122 trials; participants in the control condition 108. The one-sided t test also shows a significant difference with $t(30,40) = 2.9553$, $p<0.01$. Hence, the model meets our expectations in the prevention of forgetting, so less incorrect trials need take place. The time saved is used to study new items or rehearse previously presented items.
Thirdly, it is possible to examine the accuracy of the representations of the items in memory. The prediction is that the model's representation of the strength of items should become more accurate after increased repetition. Reaction times do contain noise, but when more trials of the same item have been presented, the alpha fitting the rehearsals will be less sensitive for noise in the separate trials. Therefore, the representation of the strengths in memory at any given moment in the training session should become more reliable. This should lead to more accurate predictions of the reaction times. The means and standard errors of the absolute difference between prediction and measured reaction times are plotted in figure 3.4. Five trials in which no response was given after 15 seconds were removed from this representation. It is unknown why no response was given in these trials, but they are not representative when testing the prediction of our model. Since the first adaptation of alpha takes place after the third encounter, prediction can only be made at the beginning of the fourth encounter. We also cut the graph at thirteen repetitions because fewer than ten items were presented more then thirteen times. The graph shows a decrease in the difference between prediction of the latency at the moment of retrieval and measured latency. Hence, the model's representations become more accurate after optimizing the alpha value, especially during the first four adaptations. After the seventh repetition, the mean absolute difference stagnates at around 1000ms. This is still a relatively large difference, because the mean reaction time over all trials was 2661ms. However, as figure 3.3 shows, the trials contain many high differences in reaction times, which makes more accurate prediction difficult.



Figuur 3.3: Absolute differences between predicted and measured reaction times. The line represents the regression line.



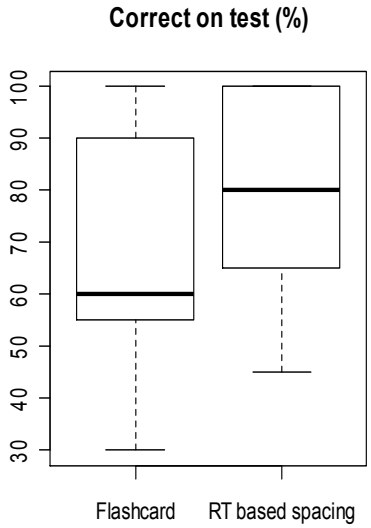
Figuur 3.4: Mean and standard error of the absolute differences between predicted and measured reaction time.

Test results

Although the results of the model of the training data look promising, the results of the post-test are the most important to evaluate. Figure 3.5 shows a box plot of the results of the two conditions. Both conditions lead to a wide spread in the results. To test whether there is a difference between the two conditions, we did a test with a binomial linear mixed effect model. We did not use an Anova test to test differences in the condition. This test cannot cope with the binary 'Correct'/ 'Incorrect' results, so only the score of the test can be included in the test. As a result, correction of word difficulties are not possible. However, there were differences in word difficulty that did influence test results. Because we wanted to calculate only the effect of the condition on the test results, we chose a binary model that could correct for this effect. When corrected for words and tested for better performance in the spacing condition, the test shows a significance level of $p<0.01$ (SE = 0.1736, z=-3.402). This means that the participants in the spacing condition performed better than in the flashcard condition.

3.5 Conclusion

As discussed above, the model matches the predictions. The alpha value is adjusted to each participant-item combination. With the look-ahead time to prevent forgetting, less incorrect trials take place compared to in the control condition. This leads to less time being required to restudy items, making it possible to do more test trials in the same session. Another advantage is that the adjustments of the alpha values improve the accuracy in the prediction of the reaction times. A more accurate representation of the strengths in memory thus develops. The post-test results indicate that latency-based spacing, as in our model, gives better retention than the flashcard method with a retention interval of one day.



Figuur 3.5: Boxplot of percentage correct trails on test per condition

4. Discussion

This thesis describes the study leading towards a user-adaptive system for optimizing the learning of word pairs. The system used is an activation-based model working with ACT-R equations. It adapts each participant-item combination by using the reaction time and the correctness of the responses. The results indicate that, in addition to correctness, reaction time is also a useful measurement to which to adapt the model. This was already demonstrated in Van Woudenberg (2008), but has been further refined in this study.

However, the present results do not show whether the success of the spacing model is due to more trials, the word order or a combination of both. Participants in the spacing condition did significantly more trials compared to the participants in the control condition. This means that more repetition of the items could occur, which enhances the strength of the items in memory. The number of trials is a consequence of the chosen algorithm, since users make less errors in the spacing condition, which yields time efficiency. By changing the design of the experiment, one can study whether better retention results can be made independent of the number of trials. For example, by equalizing the number of trials at 110. Nevertheless, the study time of participants in the control condition will most likely be higher. In addition, probably not all participants in the spacing condition trials will study all words, since the algorithm only allows a user to learn a new word pair after other words have been sufficiently studied to be remembered for at least fifteen seconds. One could also shorten the study time, e.g. the time to restudy an item after an incorrect response. This would mean that errors are less influenced by time, allowing more rehearsals even if many trials are answered incorrectly. The differences in the test results would then be less dependent on the difference in the number of trials, but more certainly in the order in which the items are presented.

The next question is whether one can generalize this model's results for retention over a longer time period, to other types of participants or other types of tasks. This research only had a retention interval of one day. The retention interval can be increased, but must not be lengthened by too much. When time passes, more forgetting takes place and, without rehearsal, less items will be remembered. To prove a difference in method, the retention interval must be limited to an interval that allows some remembering to still be possible. Otherwise, no differences will be found,. because when time increases, less items tend to be remembered, which makes it harder to annotate a difference in study method. As to the participants: we used a relatively homogenous group with regard to properties such as age and facility in learning. It is unknown how well the model will adapt to other types of learners. Nevertheless, it seems plausible to reason that this model is at least useful for the task in this study and for other tasks entailing the learning of facts, for example, history facts, since the spacing effect is proven to appear often in these simple tasks and in many different conditions (Capeda et al., 2006).

Before more research is done, further refinement of the model may be considered. Some suggestions would be:

- Currently, the model is optimized by minimizing the difference in measured and derived reaction time for all reaction times. However, the first reaction time is not representative because it appears directly after a study trial. By removing the difference in this calculation, the alpha is adjusted to more reliable reaction times and can therefore become more accurate.
- Secondly, the model only adapts by adjusting the alpha value per participant-item combination. Pavlik and Anderson (2005, 2008) used foreknowledge in specific participant variables in the activation formula. Adaptation according to specific participant features is plausible and can make adjustments to items faster and more accurately. But adaptation directly in the activation formula has some side effects, as is explained in 2.2.1. However, it is possible to make adaptations to account for foreknowledge, such as in the

fixed-time parameter. This is the minimum time needed for motor actions and visual processing to respond. The data of the experiment and data of an earlier pilot experiment indicated that the shortest response in fifteen minutes can differ over 500ms per participant. Some can respond in 300ms, others never respond faster than 1000ms. These relatively slow responses are interpreted by the model as low activation. The model would then adapt alpha as if the participant in question has more problems remembering items than a user who can respond more quickly. But this is not sound reasoning. Therefore, an adjustment of the fixed-time value would be more consistent.

- Thirdly, the threshold represents the boundary of forgetting. Of course there is no distinct boundary between remembering and forgetting and it is even more unlikely that this boundary is the same for everybody. Although our estimation proved to be a robust estimation, more examination can possibly refine this.
- Fourth, we used a maximum reaction time corresponding to 1.5 times the reaction time belonging to latency based on the activation of the threshold (see Formula 2.4). This value automatically changes when the fixed time becomes user-specific. However, when the fixed time is changed, an extra effect in the maximum reaction time is noticeable. For example, if somebody has a fixed time of 0.8s, the maximum reaction time becomes 4538ms. This will influence the activation and alpha values. The alpha values will increase and the activation values will decrease. This is a negative side effect of the adjustment of the fixed time, because a higher fixed time does not indicate that the participant forgets items faster. A more accurate situation is created by adapting only the part of the reaction time based on the activation value. Another possibility is to adapt the maximum reaction times personalized at a maximum of twice the standard deviation above the mean (for example).
- Lastly, in this research, the same algorithm in determining the next word pair is used as in Van Woudenberg (2008). However, when all words are still above the threshold value after fifteen seconds, the item with the longest spacing will be rehearsed. Although this situation does not appear often, it would be more consistent to increase the look-ahead time. Then optimization takes place based on retention over longer time. This is a more plausible solution.
- When implementing this algorithm for to use for actual practice purposes, some changes can be made to improve the didactic properties of the program. For example, as is also done in an often-used program (www.wrts.nl): when the answer is incorrect but the Levensthein distance is below a certain value, students are given the opportunity to correct their answer. This can improve motivation or save frustration compared to a condition in which these trials are assessed incorrectly and rehearsed more often.

A more general statement for further research is choice of a control condition. The flashcard method is an often-used method and seems to be a good control condition. It is not totally naive since it rehearses incorrect responses, but is very robust since it does not take into account any other measurements. Two variables can be chosen for this method. The first parameter is after how many trials the first repetition will take place and the second is after how many trials an incorrect item will be rehearsed. These parameters can greatly influence performance with this method. For example, if the first repetition of an item takes place only after 30 trials, as is done in Pavlik and Anderson (2008), one is unlikely to remember the time. This would lead to many incorrect answers, which makes the condition harder to perform with a normal learning curve. A consistent control method can offer a solution to finding a proper comparison of different methods over different studies.

To summarize, after the work of Van Woudenberg (2008) and Pavik and Anderson (2008), we created a more refined though robust cognitive model that captures memory effects (1) such as frequency, recency and spacing, that adapts to users (2) and uses this to modify the

word order in a session for better retention (3). Although some refinements are still possible and further research is necessary to show possible better retention over longer time intervals, for other types of users and different tasks, this study again shows that with a cognitive model one can enhance retention performance. With these models, application for actual practice purposes in class or for individuals becomes easy and practicable. There is no need for a specific schedule, so users can decide themselves how much time they spend learning and when they study and still benefit from these learning advantages compared to with a standard learning method.

References

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*,2 (6), 396–408.

Anderson, J. R., Bothell, D., Byrne, M., Douglass, D., Lebiere, C., and Qin, Y. (2004). An integrated theory of mind. *Psychological Review*, 111(4):1036–1060.

Anderson, J. R. (2007). How Can the Human Mind Occur in the Physical Universe? New York: *Oxford University Press.*

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

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, 245-248.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, 132 (3), 354

De Boer, V. (2003). Optimal learning and the spacing effect: Theory, application and experiments based on the Memory Chain model. *Unpublished master's thesis, University of Amsterdam*

Dempster, F. N. (1988). The spacing effect. *American Psychologist*, 43, 627–634.

Donovan, J. J., & Radosevich, D. J. (1999). A Meta-Analytic Review of the Distribution of Practice effect: Now You See It, Now You Don't. *Journal Of Applied Psychology*, 84, 795–805.

Ebbinghaus, H. (1964). Memory: A contribution to experimental psychology (H. A. Ruger, C. E. Bussenius, & E. R. Hilgard, Trans.). *New York: Dover Publications.* (Original work published 1885)

Mozer, M. C., Pashler, H., Lindsey, R. V., & Vul, E. Predicting the optimal spacing of study: A multiscale context model of memory. (Submitted)

Pavlik Jr, P., & Anderson, J.R. (2005). Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect. *Cognitive Science*, 29 (4), 559–586.

Pavlik, P. I., & Anderson, J. R. (2008). Using a model to compute the optimal schedule of practice. *Journal of Experimental Psychology: Applied*, 14 (2), 101–117.

Rickard, T.C., Sin-Heng Lau, J., Pashler, H. (2008). Spacing and the transition from calculation to retrieval. *Psychonomic bulletin & review* 15(3), 656-661.

Rohrer, D., & Taylor, K. (2007). The shuffling of mathematics problems improves learning. *Instructional Science, 35*(6).

Van Rijn, H., Van Maanen, L, & Van Woudenberg, M. (*submitted*) Optimizing short-session learning by individualizing schedules of practice.

Van Woudenberg, M. (2008). Optimal word pair learning in the short term: Using an activation based spacing model. *Unpublished master's thesis*, University of Groningen.

Appendix A: Word lists

French	Dutch
fier	trots
volontaire	vrijwilliger
une ampoule	een gloeilamp
environ	ongeveer
rater	missen
depuis	sinds
doué	begaafd
drôle	grappig
développer	ontwikkelen
néanmoin	toch
ailleurs	elders
l'amélioration	de verbetering
un nuage	een wolk
en revanche	daarentegen
le ciel	de lucht
mignon	schattig
une voix	een stem
dès que	zodra
coller	na laten blijven
le brouillard	de mist

Table A.1: Wordlist 3 Havo and 3 VWO

French	Dutch
feindre	doen alsof
épanouissement	ontwikkeling
gaieté	vrolijkheid
murmure	geruis
ininérant	rondtrekken
arbitrairement	willekeurig
négligeable	onbelangrijk
rabattre	verminderen
avidité	hebzucht
invresse	dronkenschap
faucher	jatten
aspirateur	stofzuiger
devoir	verschuldigd zijn
faire les frais de	opdraaien voor
franc	openhartig
tandis que	terwijl
vacarme	herrie
indulgence	toegeeflijkheid
esclavage	slavernij
équitablement	eerlijk

Table A.2: Wordlist 5 VWO

Appendix B: Derivation of decay

To calculate the decay of $d_{i,j=n}$ we can split the summation in a part with the already optimized decays and the latest decay:

$$m_i(t_j) = \ln\left(\left(\sum_{j=1}^{n-1;t_j<t} (t-t_j)^{-d_{i,j}}\right) + (t-t_{j=n})^{-\boldsymbol{d_{i,j=n}}}\right)$$

Then, we remove the natural logarithm by using the inversed exponential function on the activation:

$$e^{m_i(t_j)} = \left(\left(\sum_{j=1}^{n-1;t_j<t} (t-t_j)^{-d_{i,j}}\right) + (t-t_{j=n})^{-\boldsymbol{d_{i,j=n}}}\right)$$

Then subtract the summation since the latest repetition from the exponential function:

$$e^{m_i(t_j)} - \left(\sum_{j=1}^{n-1;t_j<t} (t-t_j)^{-d_{i,j}}\right) = \left((t-t_{j=n})^{-\boldsymbol{d_{i,j=n}}}\right)$$

Now the decay can be derived (formula swapped):

$$\boldsymbol{d_{i,j=n}} = -\left(^{t-t_{j=n}}\log\left(e^{m_i(t)} - \left(\sum_{j=1}^{n-1;t_j<t} (t-t_j)^{-d_{i,j}}\right)\right)\right)$$

With this decay, we can derive an alpha using the decay equation:

$$d_{i,j} = 0.25e^{m_i(t_j)} + \alpha_{i,j}$$

Appendix C: More analyses

Correctness on activation

The figure shows the mean percentage correctness based on activation ranges of 0.15, calculated with the previous alpha. The new alpha is already optimized on the corrected alpha and therefore not representative. This activation is the activation at the moment of rehearsal. The graph shows a strong rise in correctness between the third and fourth range. The vertical line indicates the threshold value. The drop at the -0.5 activation is because of the second rehearsal. This always takes place around the same time interval because all items still have the same alpha then.

Frequency effect

In this figure, the frequency effect is visible. More repetition means more chance of a correct answer. The first repetition is a study trial, so correctness can not be given. Again, a drop at the second rehearsal is visible, but after that an increase in percentage of correctness is visible. Because the frequency effect creates an increase in correctness over more repetitions, this can lead to lower alpha values. This means that the alpha value does not have to be a static value per combination of item and participant, but can change according to the number of repetitions, even though the activation formula accounts for the number of repetitions.

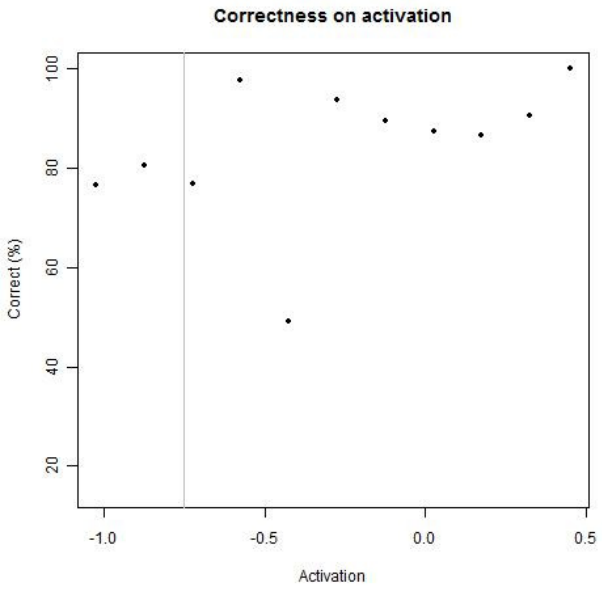


Figure C.1: Percentage correct of 0.15 activation ranges

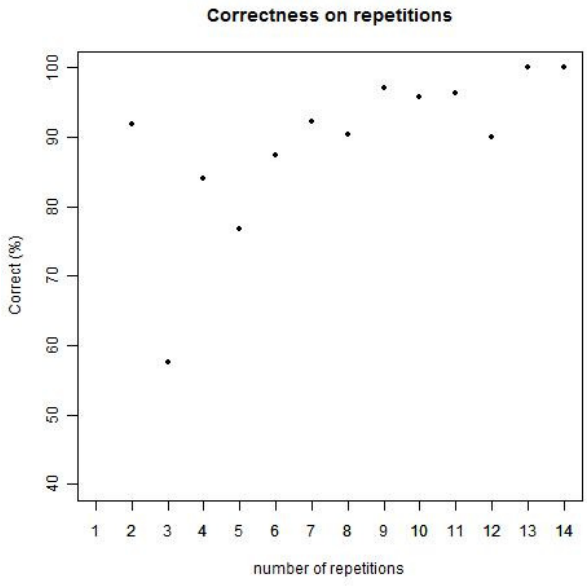


Figure C.2: percentage correct based on the number of rehearsals. Only the experimental data is used for this plot.

Number of words seen

The flashcard condition starts repeating items after an incorrect trial or when all items have been seen. The spacing condition only presents new items when all words seen in fifteen seconds stay above the threshold. This means that it is possible that not all words will be seen. It is a consequence of the method that when it adapts to users it will be better for some users to learn less words better instead of all words without remembering them. Table B.1 shows the number of words seen by the participants.

Participant	Number of words seen
1	**11**
2	**16**
3	**17**
4	**20**
5	**20**
6	**20**
7	**20**
8	**20**
9	**20**
10	**20**
11	**20**
12	**20**
13	**20**
14	**20**
15	**20**
16	**20**
17	**20**
18	**20**

Table C.1.: Number of different words seen in fifteen minutes per participant