Optimizing fact learning gains

Using personal parameter settings to improve the learning schedule

Laurens Koelewijn

June 2010

Master Thesis

Human-Machine Communication
Dept. of Artificial Intelligence
University of Groningen, The Netherlands

Supervisor:
Dr. Hedderik van Rijn (Experimental Psychology, University of Groningen)

Internal supervisor:
Prof. dr. Niels Taatgen (Artificial Intelligence, University of Groningen)

Abstract

Learning a list of facts in an efficient way is not as simple as it might appear. The spacing effect and time costs of the learning trials are amongst other aspects that have to be taken into consideration. This thesis is about creating an algorithm which produces learning schedules that maximize the retention of a learned item set on a test. This has been attempted before (Pavlik & Anderson, 2008; Van Rijn, Van Maanen & Van Woudenberg, submitted for publication; Van Thiel, 2010), but none of these studies account for the large differences in individual learning abilities that exist between people. I present an adaptation of the latency-based ACT-R spacing algorithm used by Van Thiel (2010) and in addition introduce the personalization of two important parameters to account for these individual differences. A series of experiments is performed in a laboratory setup as well as in a more realistic real-world setting to test the algorithm's performance. Analysis of the results shows no significant increase in retention on a test of the learned items when using personal parameter settings. All data do indicate however that the use of personal parameter settings does not hurt retention on a test. The analysis also shows personalization is potentially more important in a real-world setting. Including personal parameter settings thus seems to be justified.

Acknowledgements

I would like to thank Hedderik van Rijn for supervising my research. He always took the time to provide feedback and advice and really helped the project to run smoothly from the start. Next to that I would like to thank Wendy van Thiel for providing me with the raw data of her experiment as well as the source code of the spacing algorithm she used. I would also like to thank Willie Lek for letting me conduct the experiments at the ID College. And finally I would like to thank Ada Koelewijn, Duncan Hulleman and Esther Kuilema for giving me the opportunity to conduct an experiment at the Dirk van Dijkschool.

Contents

Introduction

Almost everybody who received basic education has had to learn a list of facts at some point. If you are one of these people, you have probably experienced that you remembered these facts better when the time spent on learning them was divided in a few separate sessions. This effect is called the spacing effect (or distributed practice effect). Leaving some time in between rehearsals of a fact allows for better recall of that fact at a later point than when you leave no time in between. There seems to be an optimum though. Leaving too much time in between rehearsals seems to not aid recall and even hurt recall (Cepeda, Pashler, Vul, Wixted & Rohrer, 2006). My thesis will be about creating an algorithm which produces learning schedules that, after a period of learning, maximize the facts' strength in memory. This has been attempted before (Pavlik & Anderson, 2008; Van Rijn, Van Maanen & Van Woudenberg, submitted for publication; Van Thiel, 2010), but although mentioning the importance and influence of individual differences, none of these studies incorporate them into their algorithm for producing learning schedules.

The importance of taking individual differences into account when designing interactive systems has been known for quite a while (Atkinson & Paulson, 1972; Rich, 1983) and its application is widespread (Kobsa, 1993). This is also true for the area of e-learning, which is involved with developing methods for computer (and internet) supported learning, where ways of personalizing learning aids are being developed (Conlan, Dagger & Wade, 2002; Chen, Lee & Chen, 2005). Intelligent tutoring systems are another kind of interactive learning aid that have proven to be quite successful in helping learners solve problems in fields such as mathematics, science and technology (Graesser, Van Lehn, Rose, Jordan & Harter, 2001). Most of them use a model of the individual student in order to personalize the content and help offered (Murray, 1999). Some researchers even argue such a model is necessary for an intelligent tutor (Self, 1990).

A learning aid for creating optimal learning schedules will also have to include personal differences. There are multiple aspects of human cognition involved with building a learning aid that differ from person to person. It has been found that there are big differences in learning capabilities between people (Jonassen & Grabowski, 1993) including explicit learning in complex (Reber, Walkenfeld & Hernstadt, 1991) as well as simple tasks (Kliegel & Altgassen, 2006). Next to that, because the learning aid will most likely be computer based, individual differences relating to computer use are relevant as well. Czaja & Sharit (1993) found significant influence of age and computer experience on response times and errors in a data entry task. This task is part of nearly all computer use so the performance differences will be relevant for a learning aid as well.

In this thesis I will present an extension to the latency-based ACT-R spacing model as proposed by Van Rijn, Van Maanen and Van Woudenberg (submitted for publication) and later adapted by Van Thiel (2010) to implement the spacing effect and predict when the optimal interval has passed and thus when it is time to rehearse a fact. In addition I will introduce the personalization of two important parameters of the spacing model to try and improve performance on a test of the learned facts. This personalization is based on data gathered on the participants prior to the learning of the tested set of facts. The question I like to answer is whether learning a list of facts using a learning schedule created by an algorithm that is personalized in advance of the learning will lead to better performance on a test. In other words: will adjusting the parameters of a latency-based spacing model to prior learning data, improve the learning schedules produced by the model for that person?

Background

The spacing effect has been a popular research topic for more than a century. Ebbinghaus (1913/1885) is cited as the first to describe the phenomenon and research on the topic has been extensive ever since. It can be defined as the superior retention realized by spaced practice as opposed to massed practice. Spaced practice in this sense is practice in which trials of an item are separated by a time gap (or trials of other items which are essentially a time gap as well). Massed practice on the other hand consists of consecutive trials of an item without any time gaps. Subsequent research on the spacing effect shows it is widespread and relevant for many memory tasks, such as vocabulary learning (Bloom & Shuell, 1981), skill acquisition (Wisner, Lombardo & Catalano, 1988) and mathematics (Rohrer & Taylor, 2006). It has even been shown that the spacing effect applies to learning by certain animals (Carew, Pinsker & Kandel, 1972; Beck, Schroeder & Davis, 2000; Menzel, Manz, Menzel & Greggers, 2001). Most research however, has been done on verbal recall tasks and an extensive review of this can be found in Cepeda, Pashler, Vul, Wixted and Rohrer (2006).

So the effect itself is widely recognized, but the nature of the spacing effect is still a matter for debate. Janiszewski, Noel and Sawyer (2003) state five theories that have been proposed as possible explanations, namely the attention, rehearsal, encoding variability, retrieval and reconstruction hypothesis. They conducted ten tests in which they compared the prediction of the different theories on a certain topic (as far as these could be derived) with the outcomes of meta-analyses of the spacing literature. No theory predicted all outcomes correctly. The encoding variability hypothesis (Glenberg, 1979) however, is still one of the most popular explanations of the spacing effect, despite other criticism and evidence against it (Dempster, 1987; Dempster, 1989). It states that the effects of spacing are caused by the differences in environmental cues while storing and retrieving memories. If one leaves some time in between repetitions while learning, the environment during a repetition will have changed since the last repetition, giving you a greater set of retrieval cues. This will facilitate retrieval, because it is likely the environment during retrieval has drifted away from the one during learning, so a larger amount of retrieval cues increases the chance that some of them match the retrieval environment.

In recent years, with the rise of cognitive modeling, Raaijmakers (2003) implemented the contextual fluctuation hypothesis in a model called SAM (Search of Associative Memory). This model is quite a direct implementation of the hypothesis and fairly successful at fitting data from three different experiments. The main drawback of the model is the large amount of free parameters, which make the good fits less impressive. The model is nonetheless capable of explaining the data based on the contextual fluctuation hypothesis.

Pavlik and Anderson (2005) then propose a different explanation for the spacing effect based on the ACT-R cognitive modeling architecture (Anderson, Bothell, Byrne, Douglass, Lebiere & Qin, 2004) that has been evolving since the first ideas were proposed by Anderson & Schooler (1991). They implemented an activation-based spacing model that calculates the activation of each fact in memory. The total activation of an item *i* is calculated by summing over the activation generated by every encounter of the item:

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

As can be seen in the equation, the activation generated by an encounter *j* depends on the decay *d* for that encounter. This decay for this encounter depends on the activation of the item at the present moment:

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

We can now see that the decay will be great if the activation of that item is high. This accounts for the spacing effect, because it is now not beneficial to present an item right after the previous encounter, since the activation will be very high and the activation added by this encounter after a little while thus very low. Leaving some time between encounters leads to a higher activation in the long run and thus to a higher learning gain. For a more detailed explanation of the model, the reader is referred to Pavlik and Anderson (2005).

Pavlik and Anderson conducted an experiment in which they let participants learn Japanese - English word pairs while varying the number of test trials and intervening trials. They then compared the predictions of their ACT-R model and the predictions of the SAM model as proposed by Raaijmakers (2003) with the generated data. Both models fit the data reasonably well, although the ACT-R model resulted in a slightly better fit. Comparison of the predictions of both models on datasets from 'classic' experiments in the spacing literature shows similar results, both models produce good fits. An important thing to note though is that the ACT-R model uses less free parameters to achieve this.

So models prove to be useful in trying to explain the mechanisms underlying the spacing effect. An additional benefit is that due to their predictive power they can also be used to create efficient learning schedules. Since the 1960's researchers have tried to do this and a good example of these modeling efforts in the early days is Atkinson (1972). He created a Markov Model that was fairly successful at producing efficient learning schedules. However interest faded and only recently, models of the spacing effect are again being applied to create efficient learning schedules for fact learning (Pavlik & Anderson, 2008; Van Rijn, Van Maanen & Van Woudenberg, submitted for publication; Van Thiel, 2010).

Pavlik and Anderson (2008) use an extension of their spacing model (Pavlik, 2007) in an algorithm for creating learning schedules. This model takes the differences in the effectiveness of study and test trials (b_j) into account and uses an extension of the earlier activation formula to also adjust for variation in individual learning ability (β_s), item difficulty (β_i) and individual learning ability for an item ($\beta_{s,i}$) during learning:

(3) $$m_i'(t_j) = \beta_s + \beta_i + \beta_{s,i} + \ln\left(\sum_{j=1}^{n} b_j (t - t_j)^{-d_{i,j}}\right)$$

These new β parameters make the model adapt to the individual and matter studied during learning, but they are not adjusted to the individual prior to the learning. The algorithm uses this extended activation equation to produce a presentation sequence for a list of facts that will maximize retention on a test. It does this by calculating the learning rate for each item, which is the gain in activation at the time of testing, divided by the time cost now to study the item. Items are not scheduled for practice until their learning rate is maximal. The model outperforms a control flash card algorithm and an implementation of the Atkinson model (Atkinson, 1972) in producing optimal learning schedules for a test, but Van Rijn, Van Maanen and Van Woudenberg (submitted for publication) argue that these results are not very useful when considering how learning takes place in a real world classroom setting. The three learning sessions of an hour used in the test are much longer than the usual time spent studying for a test. In addition to this, the type and amount of learning materials used are claimed to be unrealistic. This is an important point and it has longer been argued that the spacing community should focus more on realistic classroom applications (Dempster, 1989).

To see whether a model of the spacing effect could be used to create optimal learning schedules for pre-university students learning vocabulary in a classroom setting, Van Rijn, Van Maanen and Van Woudenberg *(*submitted for publication) also adjust the original Pavlik and Anderson (2005) model to make their model adaptive. They do however not use the extended activation formula as in Pavlik and Anderson (2008), but instead make the α parameter in the decay formula (2) dependent on the recall speed of a fact. This α parameter acts as the baseline value of d representing the difficulty of remembering a particular item. Increasing or decreasing it directly influences d and can be used to account for an over- or underestimation of the activation. The recall speed that the adaptation is based on is defined as the time between the presentation onset of the request for the fact and the moment the answer is given.

The activation formula can be used to predict the time a participant will need to recall a fact when prompted to do so. A high activation will lead to a short recall time and a low activation to a long recall time or even to no recall at all if the activation has fallen below the recall threshold. If a participant needs more time than predicted or fails to recall the fact, the α parameter is increased. This is done because apparently the predicted activation was too high. Increasing the α parameter will lead to a higher estimate of the decay and in turn to a lower estimate of the activation of that fact. The next presentation will thus be sooner, because the activation will fall below the retrieval threshold sooner. Adaptation however, is done with steps of 0.01 at a time, so convergence can be very slow and potentially not even be possible during shorter learning sessions.

To improve this, Van Thiel (2010) changed the adaptation algorithm of the model to increase the speed of the adaptation. This adaptation is based on the latency of the responses during rehearse trials. This model also significantly improves the performance in recall over a control flashcard model when used to create a learning schedule for a test. The model however, still does not take individual differences in learning ability into account. Every participant starts with the same initial α parameter for every item, while it is possible to gather, and use, data on the individual learning ability to personalize the α parameter prior to learning.

Personalizing the initial α parameter seems to be a good method. The research by Van Thiel (2010) revealed that the α parameter shows great variability between subjects, but not too much variability within subjects. A personal estimate of the initial α parameter for a person should thus be quite representative for most items. The research also revealed there is convergence to an appropriate α parameter for a fact, but that this can now take several encounters. If the initial α value is a better estimate of the real value, convergence will be quicker and the model can produce a more optimal learning schedule.

The adjustment of the α value itself is based on the discrepancy between the activation as predicted by the model and the activation deduced from the response latency at the moment of rehearsal. This response latency is the time between the onset of the presentation of a word and the first key stroke of the participant. The response latency (*RT*) corresponds to an observed activation according to Equation 4:

(4) $$RT = Fe^{-m} + f$$

The response latency consists of two parts. The first part represents the time it took the participant to retrieve the fact from memory. As can be seen it depends on the activation of the fact m and a scaling parameter F. The second part is a standard reaction time cost (f) corresponding to the processing of the stimulus on the screen and the act of pressing the key on the keyboard. This parameter thus represents the part of the response latency that is not involved with memory retrieval. In previous research then this parameter had been kept constant for every participant (Pavlik & Anderson, 2008; Van Thiel, 2010). There is however great variability in the speed with

which people can process a cue and respond to it. Children and older people for example are found to be slower than young adults (Kail & Salthouse, 1994). The fixed value can thus differ greatly from the real standard reaction time for a participant. This introduces an error into the estimation of the time it took a participant to retrieve a fact from memory, because this is the time that is left after one subtracts f from the response latency. If the f value is too big, the memory retrieval time is underestimated, if the value is too small, this item is overestimated. Because the adaptation of the α parameter is based on the estimation of the memory retrieval time, it also introduces an error in the adaptation of the α parameter.

I will examine the effects of including a personal value for the standard reaction time parameter f. This is expected to provide a better fit between the model and the observations and next to that should allow for more accurate adaptations of the α parameter. This because the optimization of the α parameter does not need to account for the error in the f parameter, or at least not for as big of an error.

The latency adaptive algorithm

To optimize the word order during a learning session I will use a latency adaptive algorithm that is a slightly modified version of the one used by Van Thiel (2010). As its name reveals, this algorithm adapts to the latencies of responses by the participant to adjust the amount of spacing between words. The algorithm is based on an earlier spacing model by Pavlik and Anderson (2005) which works by calculating the activation of each fact in memory. So the selection of the next word pair is based on the activations of the different pairs. The algorithm for selection is the following:

1. First it is determined which of the presented word pairs has the lowest activation at 15 seconds from now. Note that only word pairs already presented earlier are taken into consideration here. If this word pair's activation is below the retrieval threshold (τ) of -0.8, it is selected as the next word pair to present. By using the 15 second look ahead an attempt is made to select word pairs for presentation before they are forgotten.
2. If no word pairs have an activation that will fall below the retrieval threshold, the next new word on the list is selected for presentation. New word pairs are thus only presented when the word pairs already presented have activations above the retrieval threshold.
3. If all words have been presented, the word with the biggest interval between its last presentation an now, is selected for presentation. This to keep the spacing of the word pairs maximal.

Word pairs are initially presented in a study only trial immediately followed by a test trial of the same word pair. This to encourage conscious processing of the word pair. Subsequent presentations are test trials only. This because testing is found to be a more effective way of learning than additional studying (Roediger & Karpicke, 2006). If the test trial is responded to incorrectly however, it is followed by a study trial. This to remind the person of what was forgotten. The duration of a study trial is 5 seconds which is in line with research by Metcalfe & Kornell (2003) who found that information gain per second declines strongly beyond 4 seconds after presentation onset. They also found that too short an interval can hurt the spacing effect, so a 5 second study trial duration seems safe. The duration of a test trial is of course variable but has a maximal length of 15 seconds after which the test trial is judged as incorrect. The length of the feedback on whether the response was correct or not is 2 seconds.

Whether a word pair is selected for presentation is dependent on the activation of the word pair. As mentioned before this activation is calculated by Equation 1 (repeated below). The total activation of the word pair *i* is calculated by summing over the activation associated with every encounter of the word pair. As can be seen in the formula, the activation generated by an encounter *j* depends on the time that has passed since the encounter ($t - t_j$) and the decay *d* for that encounter. This decay depends on the activation of the item at the time of the encounter as presented in Equation 2.

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

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

So the decay for an encounter will be large if the activation of that item at the moment of the encounter is high. This accounts for the spacing effect, because the added activation from encounters that take place while activation is high will now decay quickly. After a certain period of

time the activation these encounters add to the total activation has decayed more than the activation added by spaced encounters. Spacing encounters apart will thus lead to less decay over a longer period and consequently better retention.

As said before, the decay also depends on the α parameter. This is the baseline of the decay function. When the activation is very low, the first part of Equation 2 will be close to 0 and the decay will be equal to α. Because a person will not find every item to be equally difficult to remember, this parameter can be adjusted for every item to obtain a good fit between the calculated activation and the 'real' activation of the word pair in the participant's brain. If for example an item is more difficult to remember than expected we can lower the α value to make the activation decay more quickly, because a larger α value leads to larger decay value. This adjustment of the α value is based on the discrepancy between the activation as predicted by the model and the activation deduced from the response latency at the moment of rehearsal according to Equation 4.

(4) $$RT = Fe^{-m} + f$$

Here we introduce a difference with the method of Van Thiel (2010). Instead of using a standard reaction time (f) of 300 ms, this time cost is participant specific and obtained from a small reaction time test conducted before the learning session. During this test, ten word pairs are presented consisting of two identical Dutch words (e.g. schoen – schoen). After three seconds one of the words disappears and the participant has to re-enter the word. The smallest latency obtained from this test is used as the personal standard reaction time cost for this participant. The participants are primed for what they have to type in since the words are already on the screen after which one disappears. This is done, because during learning the study trials followed by a test trial create the same situation. Not priming participants during this test leads to an overestimation of the standard reaction time for these situations. This has undesirable consequences that I will get back to later.

Because we subtract the standard reaction time (f) from the total latency of the response to leave the part representative of memory retrieval (Equation 4), having a better estimate of f also gives us a better estimate of the memory retrieval time. This in turn allows us to better estimate the observed activation of an item, because the time to retrieve an item from memory directly depends on m. The difference between the observed activation and the activation predicted by our model then indicates an error in the decay values for this word pair. Apparently the activation of the word pair decayed more, or less than we predicted and we can adjust the α to try and close the gap between predicted and observed activation. This adjustment of the α parameter is done according to the method proposed by Van Thiel (2010). First the α that produces the best fit with the decay of the last encounter is deduced. After the optimal α value for this last encounter has been deduced, a greedy search is performed on the interval between this new value for α and the previous value to find the value providing the best fit with all observations. This means the value for α resulting in the best fit between the predicted response latencies and the observed latencies is selected as the new α value for this word. For a step by step explanation of the deduction of α and the greedy search algorithm, the reader is referred to *Appendix A*.

Another difference with the method of Van Thiel is that the first test trial is not taken into account when calculating the mismatch between the observed and predicted response latencies. This first test trial always takes place immediately after an initial study trial. At this point the information from this study trial is very likely to be still available from working memory. Different accounts have been proposed as to what the nature (Baddeley & Hitch, 1974; Oberauer, Süß, Wilhelm & Wittman, 2003; Baddeley, 2003) or capacity (Cowan, 2001) of working memory might be and the scientific community is yet to achieve consensus on these matters. It is however widely agreed that

working memory, in whatever form, allows one to keep information readily available and eliminates the need for declarative memory retrieval. Raaijmakers (2003) and Pavlik and Anderson (2005) also find the need to include a mechanism in their model of the spacing effect to cope with working memory influences at very short lags. It seems therefore reasonable to treat the first test trial of a word as if it contains no information representative of declarative memory retrieval and ignore these rehearsals when calculating the mismatch. Because the first test trial cannot be used to calculate the mismatch, α is not adjusted on this test trial. Adjustment of the α parameter thus starts on the second test trial.

The last difference is the handling of incorrect responses and responses with very long latencies (more than 1.5 times the latency corresponding to an activation around the retrieval threshold (τ) of -0.8). Because the latencies for these responses contain a lot of noise and cause undesirably strong adaptation there needs to be a cutoff value. Response latencies larger than the cutoff value are replaced by the cutoff value itself. Van Thiel (2010) uses a cutoff latency (RT_{co}) that corresponds to 1.5 times the response latency associated with the activation around the retrieval threshold. This is the latency calculated with Equation 4 for $m = \tau$ with the outcome multiplied by 1.5. Equation 5 illustrates this:

(5) $$RT_{co} = 1.5 * (Fe^{-m} + f)$$

As this equation includes the now variable standard reaction time for a retrieval, people with a higher personal standard reaction time also have a higher cutoff value. This is not a problem and even desirable because we now also account for individual differences in the cutoff value. The problem is that the entire outcome of Equation 4 is multiplied by 1.5 as in Equation 5, also the standard reaction time part. This will magnify the differences between people more than is justified, so instead of multiplying the entire response latency with 1.5 we will multiply the activation corresponding to the retrieval threshold by 1.5 as in Equation 6. This will not cause an unjustified magnification of the differences, because the f value is not multiplied.

(6) $$RT_{co} = Fe^{-(1.5*m)} + f$$

Parameter	Value
c	0.25
F	1
τ	-0.8

Table 1: Values for the different fixed parameters used in the adaptive spacing algorithm.

Laboratory experiments

Method

Two laboratory experiments were conducted. The first one to test whether this implementation of a latency adaptive spacing algorithm performs better than a baseline flashcard algorithm. Others have already shown this is the case for similar algorithms (Van Rijn, Van Maanen & Van Woudenberg, submitted for publication; Van Thiel 2010), but used a between subject setup. I used a within subject setup, to make sure there is no bias caused by differences between the groups of participants. The two conditions here were a standard initial α condition and a flashcard condition.

In the standard initial α condition, the latency adaptive algorithm is used with an initial value for α of 0.32. This value is based on Van Thiel (2010), who used a value of 0.30 and found that about 57% of the responses on the second rehearse were incorrect. A slightly higher value thus seems appropriate (a higher value causes the words to be spaced closer together, so the percentage correct on the first rehearse is expected to be higher).

In the flashcard condition a flashcard algorithm determines the order of the word pairs. This algorithm is very similar to flashcard algorithms used in earlier research (Pavlik & Anderson, 2008; Van Rijn, Van Maanen & Van Woudenberg, submitted for publication; Van Thiel, 2010) and serves as a control condition. The algorithm takes all the word pairs and divides them into decks of 5 word pairs. The word pairs in this deck are presented one by one and after the initial presentation a word pair is rehearsed immediately with a test trial. If a word pair is not recalled correctly it is placed back at the bottom of the deck. If a word pair is recalled correctly it is put aside. After the whole deck has been recalled correctly, it is rehearsed a second time with another test trial, starting with the first word pair in the deck. As soon as the whole deck is recalled correctly the second time, a new deck of 5 word pairs is selected. When all decks have been presented, the algorithm starts again with deck 1. From now on the cycle through a deck consists on one test trial per word only.

This algorithm resembles a more traditional method of learning word pairs. It is however different from the flash card algorithm used by Van Thiel (2010). In this study there is no second test trial for the word pairs in a deck before moving on to the next deck. The first test trials are not very useful however, because they occur immediately after the initial presentations, so there is no spacing between the initial study trial and the test trial. By adding the second test trial in the first cycle through the decks, the algorithm includes more spacing at the start of the learning session, while still being very simple. It thus provides us with a better and fairer baseline to compare the latency adaptive algorithm to.

The second experiment was conducted to test whether using a personalized default α parameter improves performance over the usage of a standard value for the α parameter as in earlier research. The two conditions in this experiment were the personal initial α condition and the fixed initial α condition.

In the personal initial α condition, the median of the final α values gathered in the first experiment is used as the initial α value. Only the α values of words that had a minimum of 4 encounters were used (word pairs with less encounters have not had their α value adjusted yet, or the adjustment is based on only one observation). For words with more than 7 encounters, the α value as obtained at the seventh encounter is used. This is done because data from Van Thiel (2010) show that α stabilizes around the seventh encounter. Since there is always the risk of overlearning, which will drop the α value due to very quick responses, it is best to not use the value obtained at later encounters. In addition to this, if the median of the final α values was less than 0.28, this value of 0.28 was used as the default α value. This is done, because low values for α will lead to very wide

initial spacing. An α value however cannot be adjusted before the second test trial. If the value turns out to be a poor estimate for a particular item, correcting it takes a very long time. Because of this very low initial values were expected to do more damage than good and the minimum was set at 0.28.

In the fixed initial α condition, a value of 0.30 is used as the initial value for α. This is the value used by Van Thiel (2010) and will allow us to compare the performance of the model using Van Thiel's fixed initial α with the performance of the model using the personal initial α. It has to be noted that as mentioned earlier there are some differences between my model and the one used by Van Thiel (2010), so a direct comparison of the models themselves is not possible in this case.

A total of 20 first year psychology students participated in the experiments (6 males, average age 21). Word pairs were presented on MacBooks in the Safari web browser using a self-made web application. This application also logged all the data collected during learning for later analysis. A maximum of three participants were tested at the same time. The experimental setup was within subject and the experiments were conducted together during three separate events spread out over three days. Both experiments used a 2 x 2 setup in which word lists and order of conditions were counterbalanced between different participants. This to eliminate any bias caused by fatigue or difference in word list difficulty.

The first event on day one started off with a small test to obtain a personal fixed time cost as mentioned earlier. After this small initial test, two 15 minute learning sessions were scheduled right after each other. During each of these sessions the participants had to learn a list of words pairs in the standard initial α condition or the flashcard condition. The word lists consisted of 15 difficult English – Dutch word pairs that first year psychology students generally have no knowledge of (*Appendix C*). The number of word pairs was set at 15, because a pilot study using 20 word pairs showed the participants had great difficulty retaining even half the pairs. To keep participants motivated 12 word pairs were used in a first attempt of this study (see *Appendix B*), but this caused a considerable ceiling effect rendering the data useless. 15 word pairs were chosen then for this series of experiments to try and prevent a ceiling effect while keeping the amount of words manageable as to keep the participants motivated.

The second event took place the next day. It started off with a quiz of the words learned on the first day. Participants were given a maximum of 10 minutes to complete the quiz. After this quiz there were again two learning sessions in which the participants had to learn new lists of 15 difficult English – Dutch word pairs in the personal initial α condition or the fixed initial α condition. During the third event, the day after that, the participants were quizzed on their knowledge of the 30 words learned on the second day. Again participants were given a maximum of 10 minutes to complete the quiz.

Results

Latency adaptive versus flashcard

The test results of the first laboratory experiment were analyzed to see whether the participants' performance in the latency adaptive α condition was better than their performance in the flashcard control condition. A one sided paired *t* test however, shows that the number of correctly recalled words in the latency adaptive α condition is not significantly greater than in the flashcard control condition, $t(19) = 0$. Correct in this case is defined as conceptual correctness. Participants sometimes answered on the test with a synonym of the learned word, which was judged as correct. A boxplot of the percentages correct in both conditions is shown in Figure 1. We can see the median for the latency adaptive condition is a little higher, but there is also more variability.

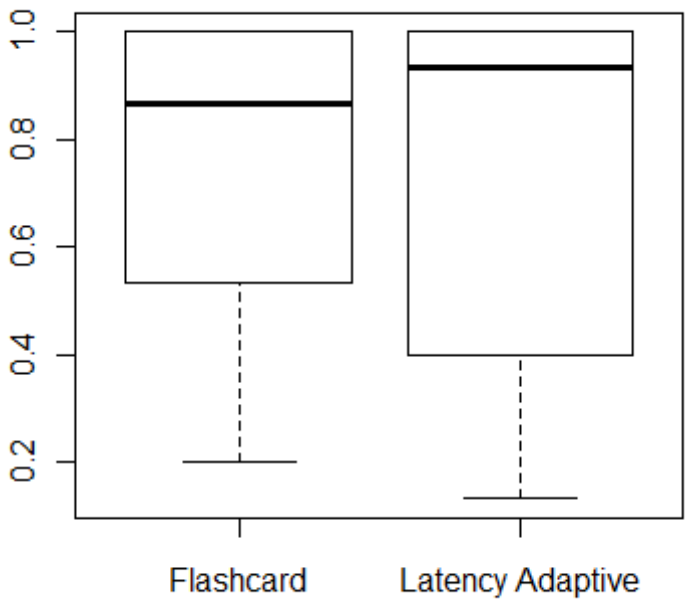

Figure 1: Boxplot showing the percentage correct on the test for the flashcard and latency adaptive conditions.

Because of the within subject setup of the experiment, there could be a difference in word list difficulty. And although the setup was balanced, it is interesting to look at this. Using linear mixed-effect models the answers on the test were analyzed for an effect of word list difficulty or effect of condition. No main effects of word list or condition were found. Including condition for every participant separately however, shows there is a significant effect of condition per participant ($\chi^2 <$ 0.001), but not in one direction, so some participants perform significantly better in the latency adaptive condition and others in the flashcard condition. The analysis also shows there is no significant effect of difference in word list difficulty per participant.

The effect size of condition per participant is plotted against the total performance of the participant (percentage correct in both conditions combined) in Figure 2. As can be seen, condition has no effect on participants that perform very well. Indicating that if you are very good at learning facts, it does not matter which method you use. Unfortunately Figure 2 also shows a substantial amount of our participants belongs to this group which might explain absence of an effect.

Looking at the learning data reveals that there are differences in the percentages of correct answers on the different test trials (Figure 3). Note that one participant is not included here, because due to technical failure the learning data of the flashcard condition were not saved. Because of the nature of the flashcard algorithm, one would expect the latency adaptive condition to give a higher percentage of correct recalls on the third test trial. This because (given that someone does not make incorrect responses) the first two test trials are spaced fairly closely together when using the flashcard algorithm. The third test trial however does not take place until all the other decks have been rehearsed. One would expect a reasonable amount of incorrect responses on the third test trial and thus the latency adaptive condition to perform better. We can see that this is the case, but also that the latency adaptive condition keeps performing better on the next three test trials, after which its percentage correct drops below the percentage correct in the flashcard condition. The percentages for the flashcard condition on these later test trials though are based on a very small number of trials making them noisy and a comparison unreliable. Another aspect to note here is the fact that in the latency adaptive condition there were 2190 rehearsals (study and test trials together) in total as opposed to a total of 2096 in the flashcard condition. This corresponds to around 5% more rehearsals per participant, but the difference is not significant, $t(18) = 0.78$.

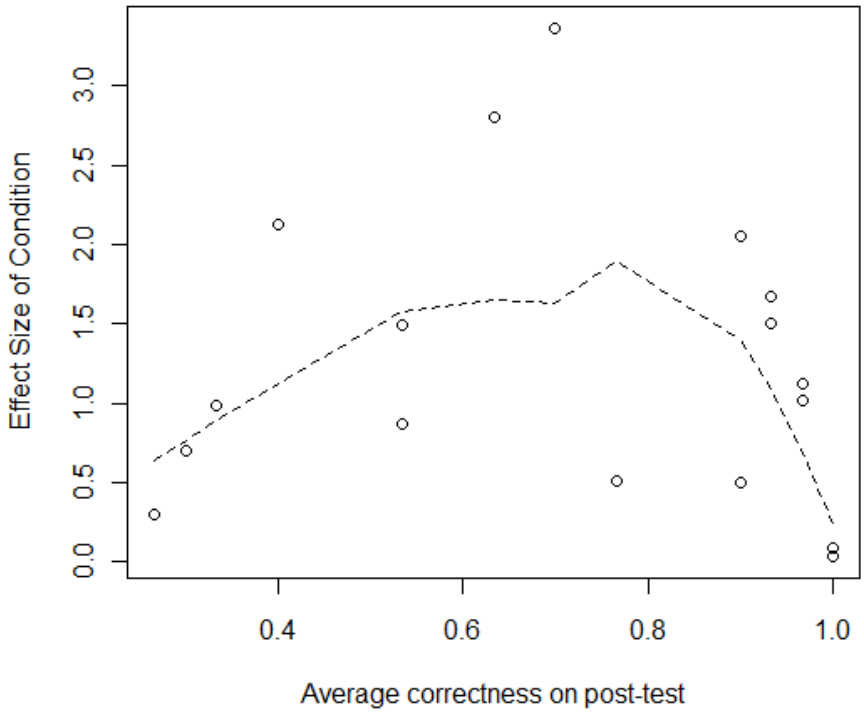

Figure 2: The effect size of condition per participant.

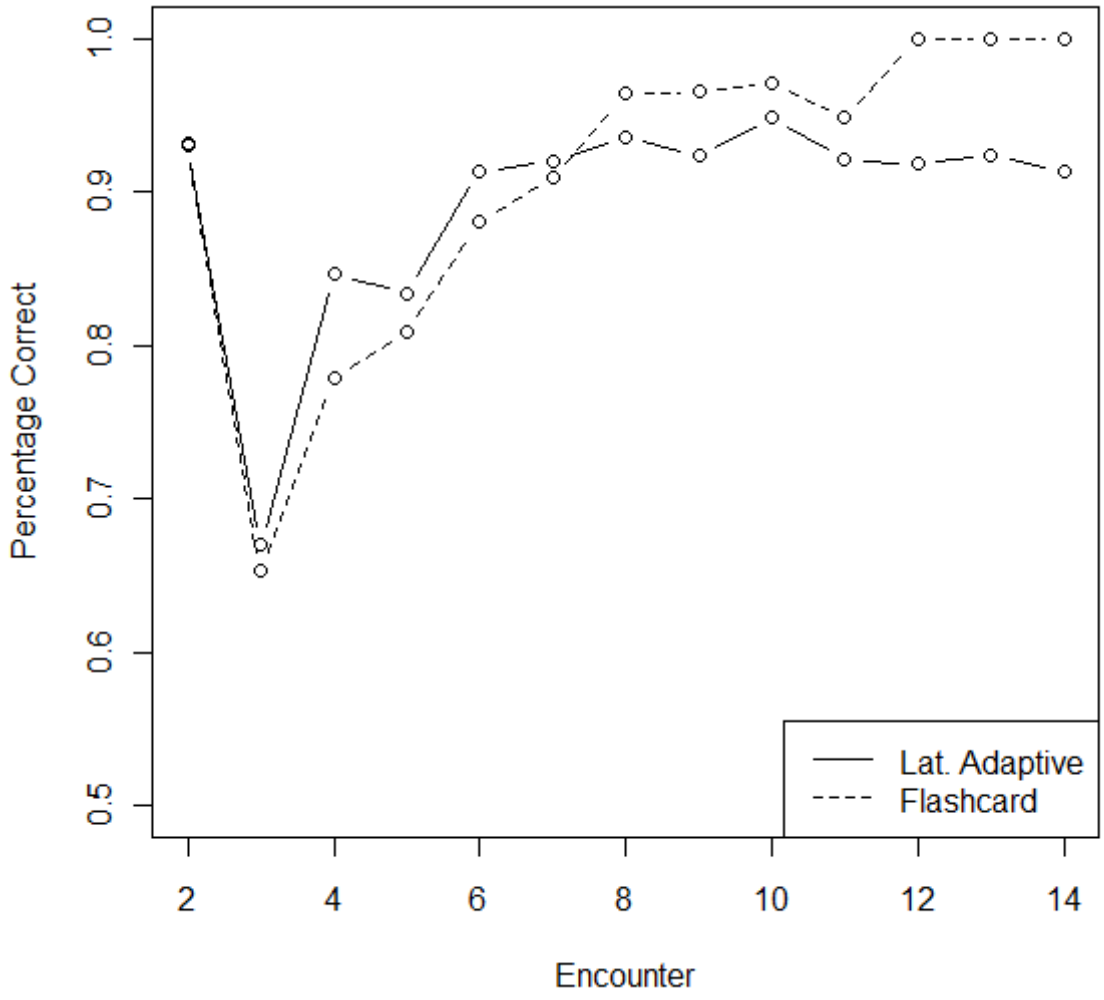

Figure 3: Percentage correct per encounter for the latency adaptive and flashcard conditions. Note the graph starts at encounter 2 because this is the first test trial.

Personal initial α versus fixed initial α

The test results of the second laboratory experiment were analyzed to see whether participants' performance in the personal initial α condition was better than their performance in the fixed initial α condition. The initial α values are shown in Figure 4. A one sided paired *t* test shows that the number of correctly recalled words in the personal initial condition α is not significantly greater than in the fixed initial α condition, t(19) = 0.75. Correct again defined as conceptually correct, so a response with a synonym was judged as correct. A boxplot of the percentages correct in both conditions is shown in Figure 5. We can see the distributions are fairly similar. Further analysis of the test data using linear mixed-effects models reveals no effect of condition and no effect of a difference in word list difficulty.

Analysis of the learning data reveals the percentages of correct answers on the different test trials are higher for the personal initial α condition (Figure 6). This was expected for the first few test trials, because a personal initial value for α is of influence there, but the difference remains at later test trials as well. To test whether this lasting difference is significant a one sided paired *t* test was conducted for the average percentages correct on encounter 5 to 9 of all participants. This showed the effect is not significant, t(19) = 1.29, p = 0.106. Looking at the total number of rehearsals (study and test trials together), we can see that there were 2585 rehearsals in the personal initial α condition and 2468 rehearsals in the fixed initial α condition, again around 5% more rehearsals per participant, which in this case is significant, t(19) = 1.77, p = 0.046. A better estimate of the initial α seems to lead to a higher percentage of correct answers, and thus shorter trials. This means we can fit more trials in our learning session.

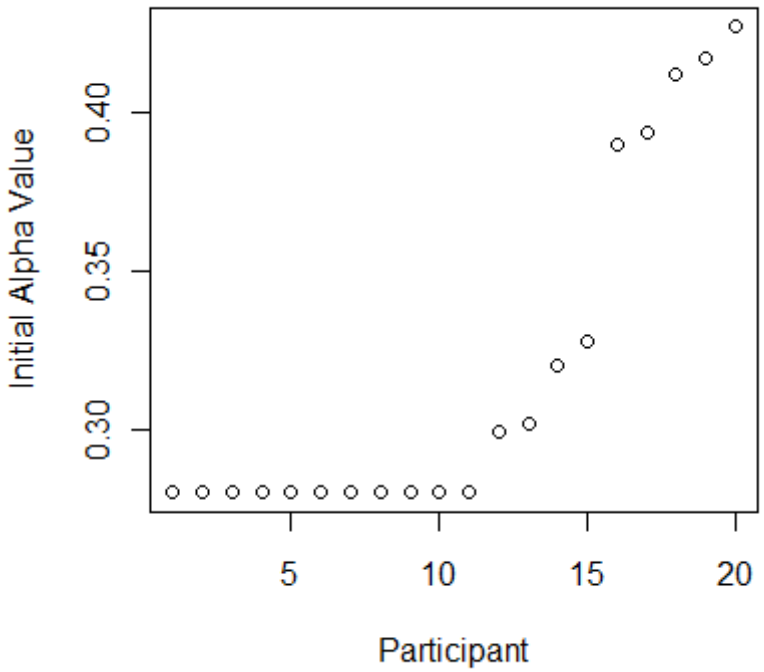

Figure 4: Initial α values in the personal initial alpha condition.

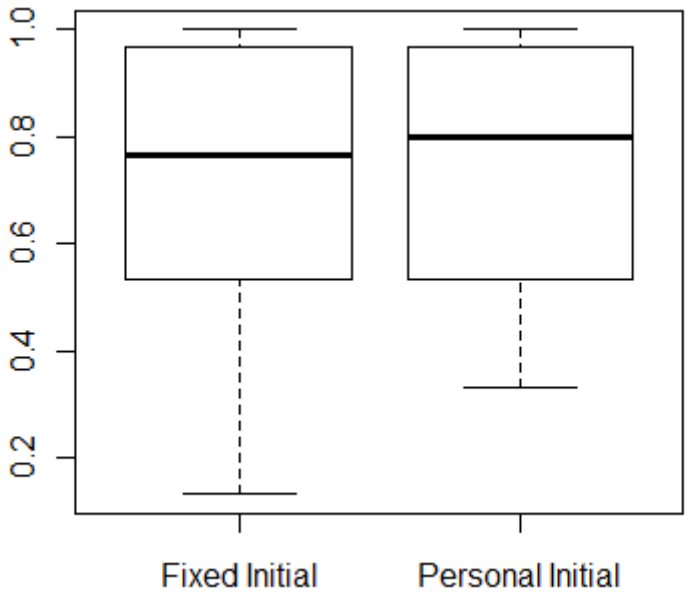

Figure 5: Boxplot showing the percentage correct on the test for the fixed initial α and personal initial α conditions.

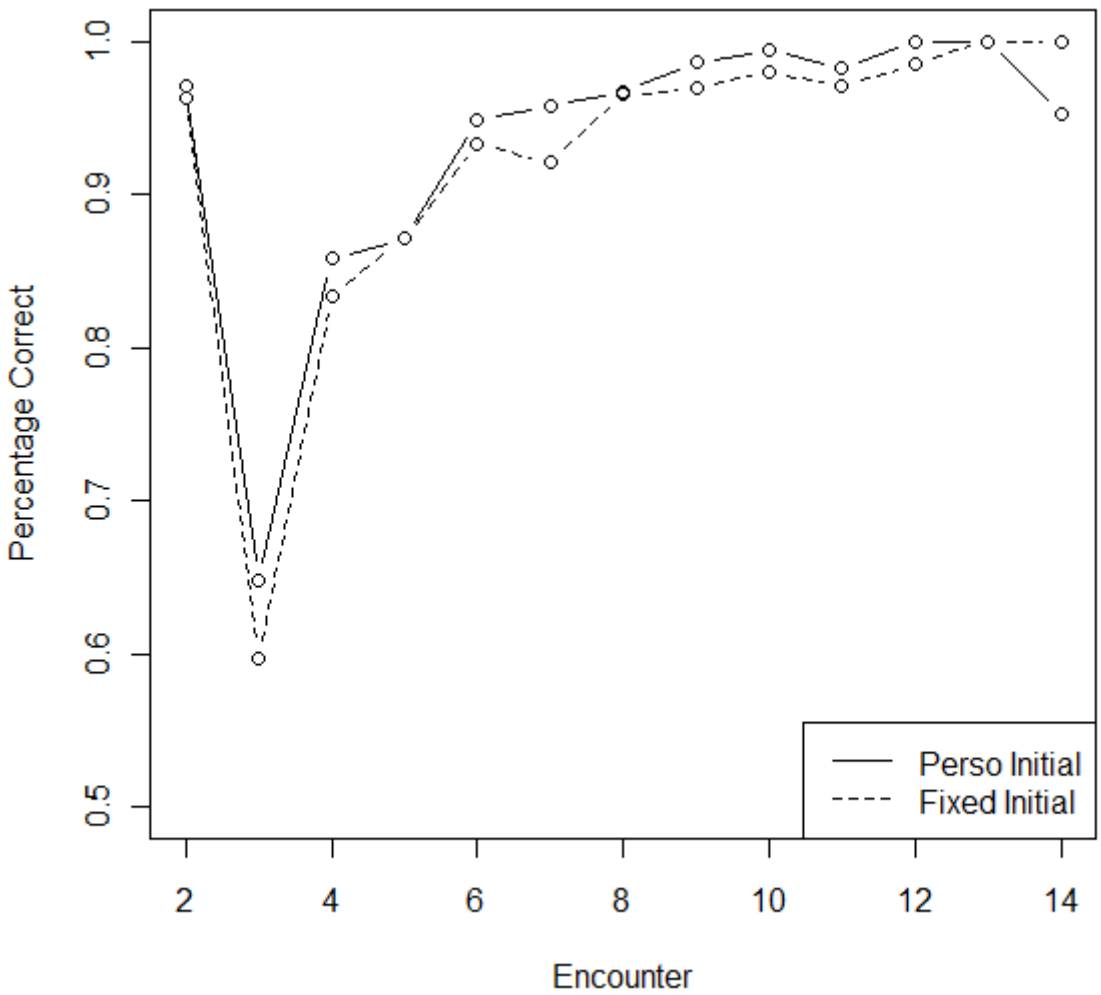

Figure 6: Percentage correct per encounter for the personal initial α and fixed initial α conditions. Note the graph starts at encounter 2 because this is the first test trial.

So there are effects in the learning data, but only the number of rehearsals is significantly greater. This is not too disturbing, because we have to consider that only the timing of the second test trial was affected by the use of personal initial values for α. On later test trials the effect was quickly diminished by the very sensitive α optimization algorithm. If one would use a more robust and thus a probably more slowly converging algorithm, the benefits from using personal initial α values can be greater. To illustrate this I have taken all values for α between 0.0 and 0.6 (steps of 0.005) as the initial value and for each of these values calculated the mean of the absolute error between the observed reaction times and the reaction times predicted by the model for the first three test trials of every word (not counting the very first test trial that takes place right after the initial study trial). While doing this α was kept constant to simulate a very slowly converging algorithm. Next to this, the median of the final α values gathered during the first experiment is plotted as a vertical line. This to verify whether the method used produces initial values for α that are close to the optimal value. The graph for every participant is shown in Figure 7. The personal initial α values are based on the data of the first experiment so the graph is only based on the data of the second experiment. The initial α values are most likely close to the value that produces the minimal error for the data of experiment one, because they are based on these data. Including the data from the first experiment will thus bias the results of the applicability of the initial α values on a new data set.

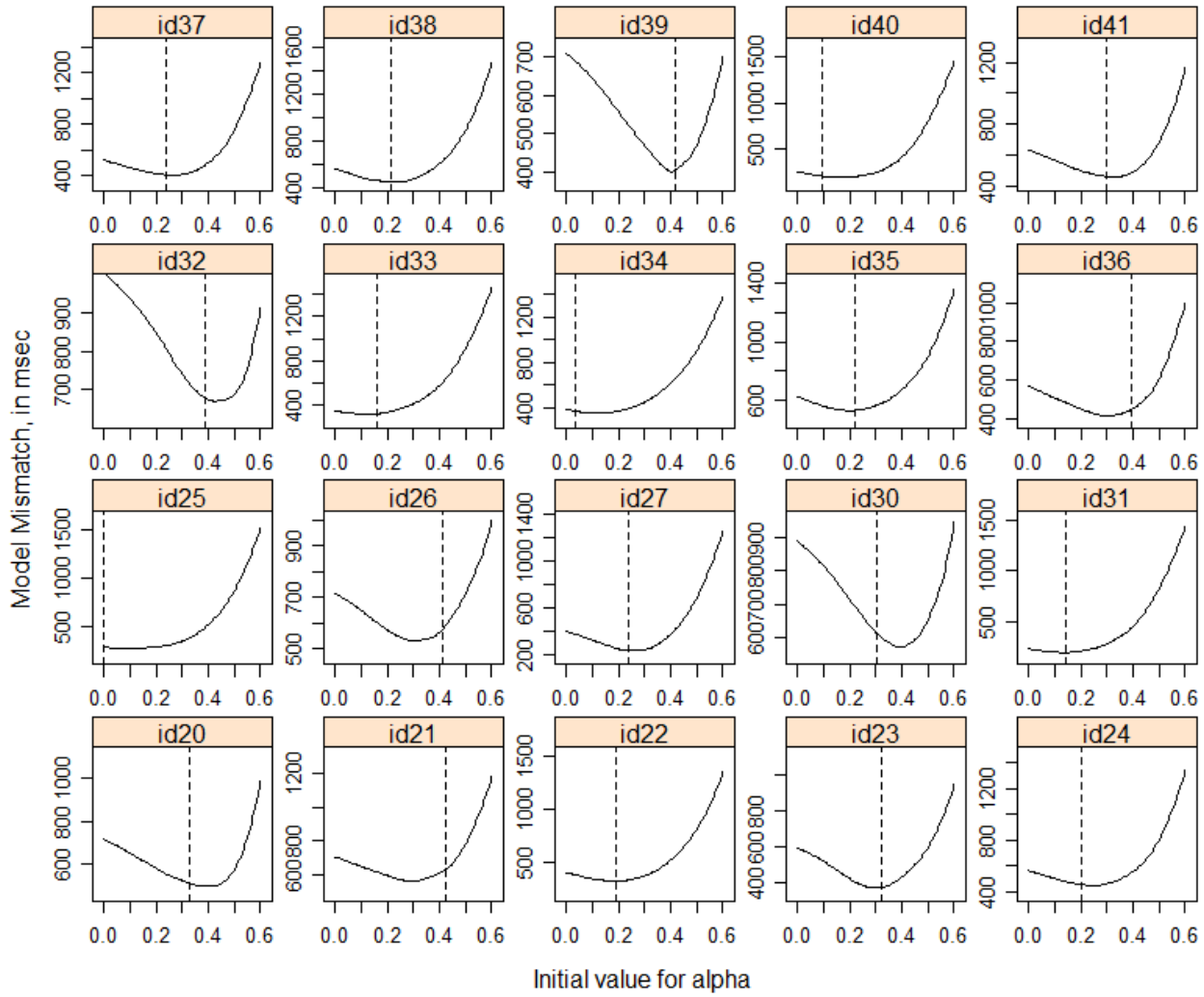

Figure 7: Mismatch between model and observations, given different values for the initial value of α. Note that the scale on the y-axis differs for each plot.

We can see there is an initial value for α that indeed minimizes the mismatch between the predicted and observed reaction times, because the curves in the graphs have a lowest point. We can also see the mean of the final α values in the first experiment is a reasonable approximation of this value, because the vertical lines representing the initial α values intersect with the error graphs very close to the minimum. This is good news, because it indicates the α value for new words is generally the same as for previous words and we can actually use a personal initial value as a better starting point.

A one sided paired *t* test confirms the mean absolute error between the model predictions and the observations for the first three test trials is significantly smaller, $t(19) = -2.13$, $p = 0.023$, when using personal initial values for α as opposed to using a standard value of 0.3, as used in earlier research (Van Thiel, 2010). This is the extreme case however, in which α is not adjusted at all during these trials, so the effect is going to be smaller when an adaptive algorithm is used.

Standard reaction time

I also conducted an exploratory analysis of the learning data of the laboratory experiments to examine the effects of the personal standard reaction time (f). As mentioned previously, this personal f value for every participant is obtained from a small test at the start of the first session. The f values for the participants of the laboratory experiments are shown in Table 2. This f value is used instead of the standard value of 300 ms and should be closer to the optimal value for f. The optimal value thus needs to be determined first to verify whether this is indeed the case.

In determining this optimal value of f, only the first three α adjustments are taken into consideration. This because the number of words that had their α adjusted 4 or more times is less than 75%. These are the harder words, because they are rehearsed relatively often. The response times for test trials of these words are relatively long, because they are forgotten more often or take a longer time to be retrieved from memory. This will bias the optimal value towards greater values for f. The cause of this is the optimization of α that is performed in parallel. If there are a lot of long response times it might be the case that a high value for f and a small value for α produce the best fit, while these might not at all reflect the real values for f and α. By considering only the first three α adjustments this effect is reduced. It still means the optimal value for f is not necessarily the true value for f. It is merely the value that minimizes the error.

To find the optimal value, all values for f between 0 and 2000 ms (50 ms steps) were taken and for each of these values the mean of the absolute error between the observed reaction times and the reaction times calculated by the model was determined for the first three optimized values of α for every word. These mean errors are plotted against f for every participant in Figure 8. We can see that for every participant, there is a value for the standard reaction time that minimizes the error between the model and the observations. The solid vertical lines then, represent the personal value of f the test produced for this participant and the dashed line the f value of 300 ms.

Participant	*f* value	Participant	*f* value
id20	351	id32	399
id21	247	id33	391
id22	303	id34	303
id23	158	id35	439
id24	318	id36	295
id25	880	id37	383
id26	169	id38	351
id27	68	id39	170
id30	735	id40	336
id31	423	id41	383

Table 2: f values for the participants of the laboratory experiment.

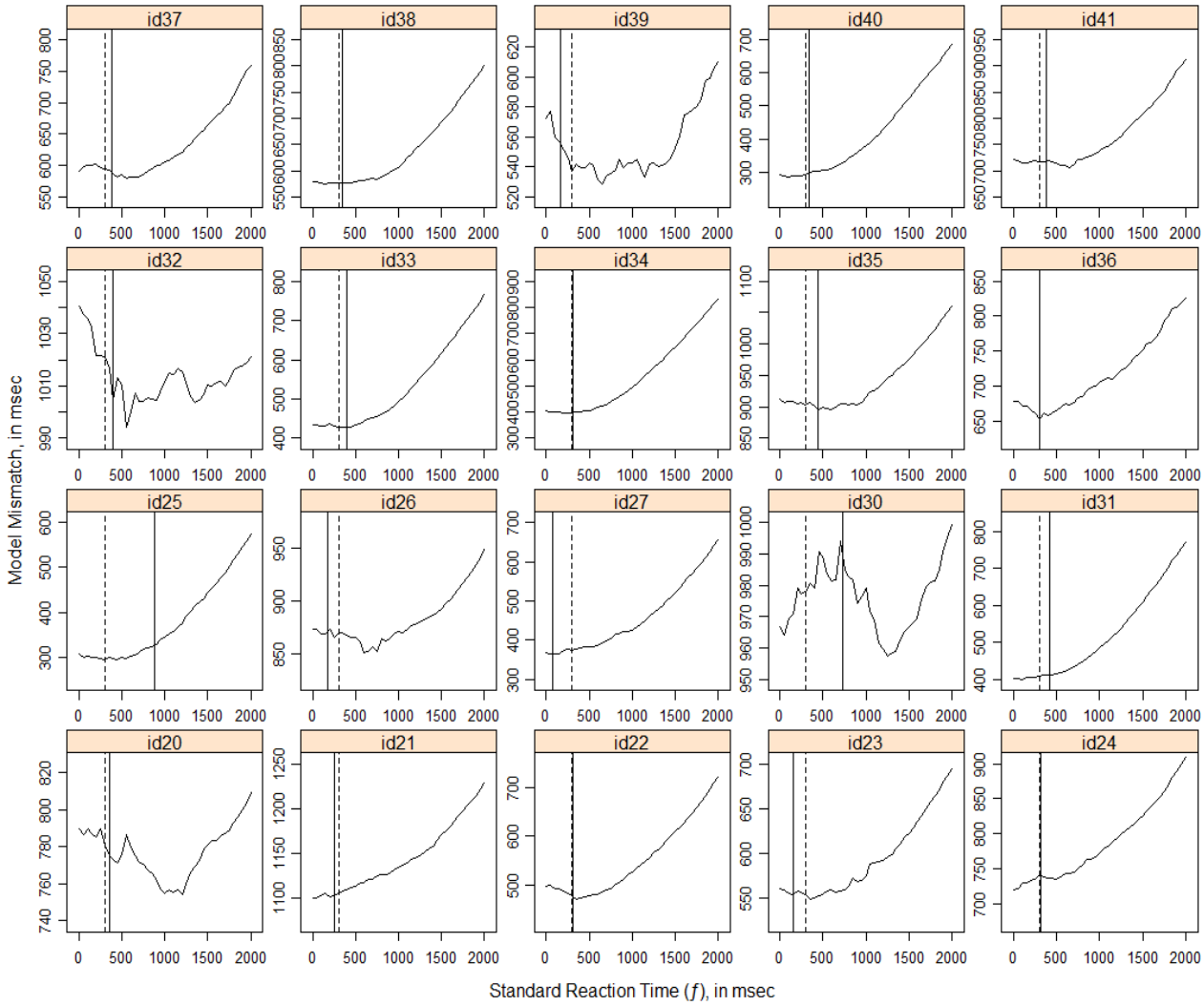

Figure 8: Mismatch between model and observations, given different values for the standard reaction time (f). Note that the scale on the y-axis differs for each plot..

We can see the included test does not always find a value for f that is optimal. However it does usually produce a value for f that is closer to the optimal value than the standard value of 300 ms is. With respect to whether the estimated personal values for f lead to a better fit of the model, a one sided paired t test shows the mean absolute error of the model fit for the first three optimized values of α is not significantly smaller ($t(19) = 0.55$) when using personal values for f than when using a standard value of 300 ms. It is even bigger. This is unfortunate, but could be caused by the noise in the data. The amount of noise in the data often causes the error using 300 ms to be smaller than the error using our personal value, while one can see in the graphs of Figure 8 that the personal value lies closer to the optimal value. A very discrete test like this, might thus not be the best way to look at the data.

Another consequence of a value of f that is closer to the real value of f should be more variation in the value of α. If the α value does not need to compensate for an incorrect value of f, it has more freedom and thus should show more variation. If for example the value for f is too small, the value for α will always need to be higher than it should be, because it needs to account for the longer response times. If the f value is too large, the α value will always be too small, because it needs to compensate for the shorter response times. Again this implicates there should be an optimal value of f for which the standard deviation of α is maximal. To find out whether this is the case, all values for f between 0 and 2000 ms (50 ms steps) were taken and for every one of these values the standard deviation of α was calculated. The results for every participant are plotted in Figure 9.

As we can see, the graphs for the different participants show a large variety in their values for *f* where the standard deviation of α is maximized. For many participants, the optimal value for *f* is very large or even larger than 2000 ms and thus not even visible in the graph. This is the case, because the standard deviation of α is very sensitive to the length of the observed response latencies. This causes the maximum to be skewed towards higher values for *f* when there are a lot of slow responses, making it a poor measure for the quality of the *f* values produced by the test. These graphs are nonetheless interesting, they show the value for *f* does indeed influence the variability of α. We also see that the initial direction of the curve is up towards higher standard deviations. This indicates that using too low a value for *f* needs to be compensated for during the optimization of α, which leads to this lower variability and biases the values for α.

The consequences of using too high a value for *f* are a little more complicated. One would expect it to again cause a decrease in the variability of α, because α will now have to compensate for this high value. But what we see is that initially it can lead to more variability in α values. Figure 10 shows the logarithmic nature of α lies at the base of this. One should ignore the absolute values in this graph, since they are highly dependable on the chosen parameters. As can be see, in this case reaction times ranging from 0 to 4000 ms will lead to more variability in α values than reaction

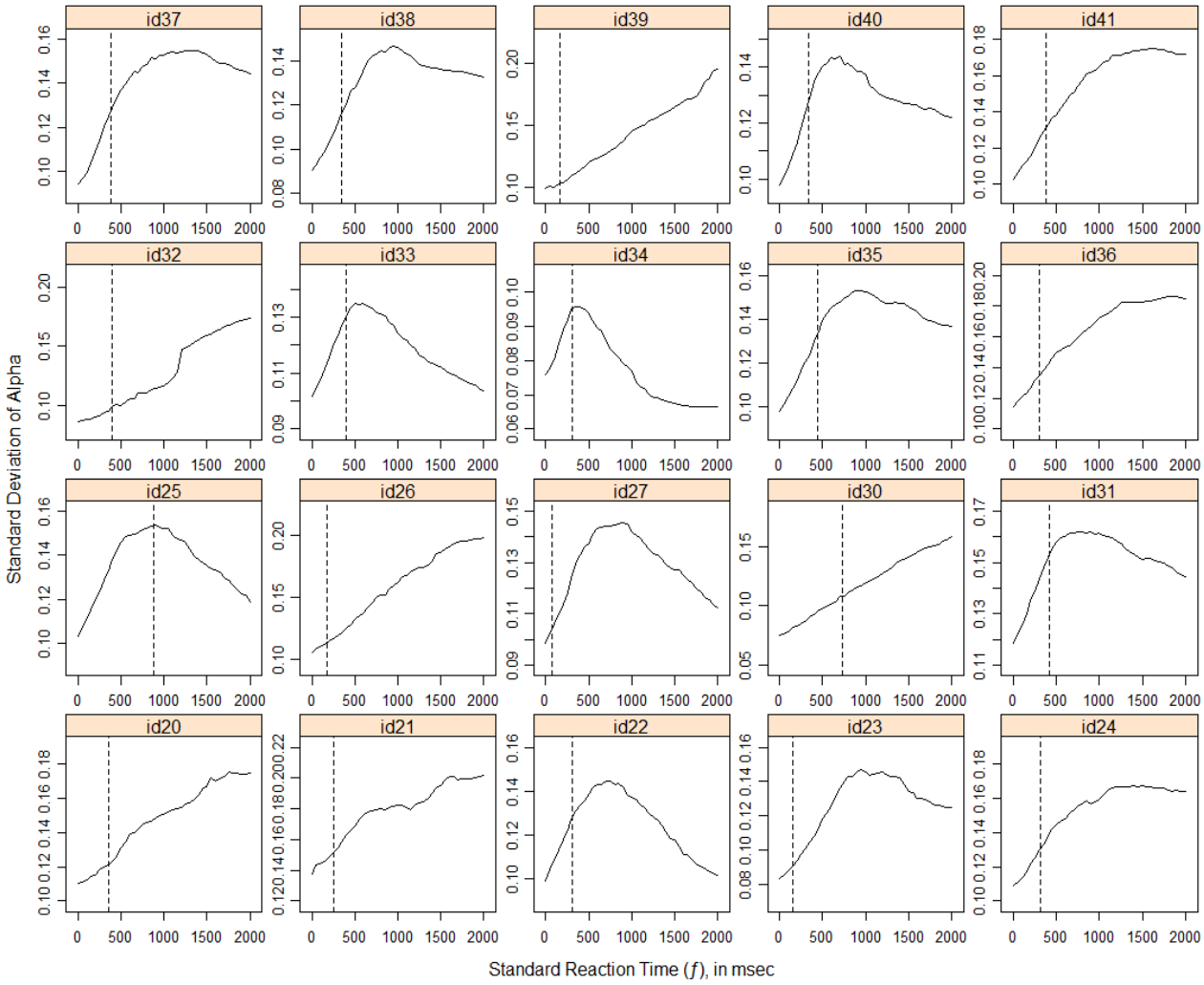

Figure 9: The standard deviation of α given the given different values for the standard reaction time (f). Note that the scale on the y-axis differs for each plot.

times ranging from 4000 to 8000 ms. If most of the responses are thus in the 4000 - 8000 ms range, subtracting a larger value for f will scale down the reaction times to a range that causes more variability in α. The expected decrease in variability of α beyond the real value of f is thus mixed with this effect and shifted towards greater values for f. Although in the graphs of participants with mostly short reaction times and thus without this shift (or at least a very small one) such as "id33" and "id34", we can already see the drop at much lower values of f. This indicates overestimating can also lead to compensation during α optimization and thus biased values for α. This causes spacing to be too wide or too close leading to a learning schedule that is not optimal.

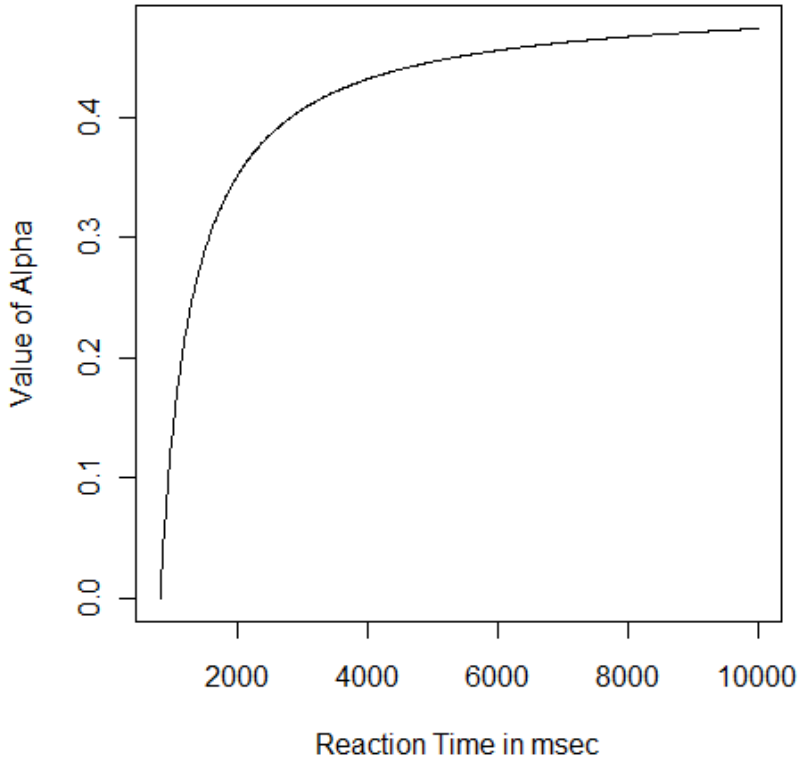

Figure 10: Distribution of α given different observed reaction times with a fixed decay of 0.5 and a standard reaction time of 300 ms. The absolute values on the axes should be ignored. since they are highly dependable on these chosen parameters.

Real world experiment I

Method

In order to test whether the results are applicable in a real world situation I also conducted an experiment at the ID College in Gouda (The Netherlands). From the students following a Dutch language course here, 7 participated in our study. The experiment consisted of two sessions, spread one week apart. During the first session participants had to learn word pairs in the fixed initial α condition, with a fixed initial α value of 0.32 in this case. During the second session word pairs were learned in the personal initial α condition using the median of the α values collected during the first session as the initial α value.

During the first session the participants had to learn two word lists containing 22 word pairs each. These word pairs consisted of the Dutch word and it's translation in the participant's native language. The languages included French, English and Polish. Words were presented in the participant's native language and had to be translated to Dutch. All words were taken from a chapter in the book the students use for their Dutch course. The participants had 15 minutes to learn each word list, so a total of 30 minutes to learn 44 word pairs. The same self-made application was used over the internet to learn the word pairs. This means no experimenter was present at the location, but only the participant's teacher to assist with possible problems. No problems were reported however.

During the second session participants had to learn two new word lists containing 21 word pairs each. Again a word was presented in the participant's native language and had to be translated to Dutch. The participants, again, had 15 minutes to learn each word list for a total of 30 minutes to learn 42 word pairs. The personal initial α values used for learning the second word list were updated before learning, based on the learning data obtained from learning the first word list. Words were taken from the same chapter in the course book. Because the participants were only available once a week and there was no time in their schedule for testing, only the learning data were collected.

Results

Personal initial α versus fixed initial α

We can look at the learning data to see whether we can find the same effects in this more real life setting. Only the personal and fixed initial α conditions were tested here. Again we see the percentages of correct answers on the different encounters are higher for the personal initial α condition (Figure 11). We have to keep in mind that there were only 7 participants here. There were also some early terminations, restarts and skipping of parts of the learning sessions by some participants. I chose not to discard participants that did this, because the data are already very noisy given the lack of experimental control and the experiment is only conducted to search for the same patterns in this real world learning situation as found in the laboratory experiment. Because of these restarts, skips and terminations the total learning time in both conditions differs. This means comparison of the number of rehearsals in both conditions is not possible. To test whether the long term differences are significant a one sided paired *t* test was conducted for the average percentages correct on encounter 5 to 9 of all participants. This showed there is no significant effect, $t(6) = 0.83$. An interesting thing to note is that the personal initial α values were updated before learning the second list based on the learning data obtained during the learning of the first list in the personal initial α condition. When we look at the personal initial α values (Table 3) we can see the updated values to not differ much from the original values, indicating two 15 minute learning sessions might be enough to produce a reasonable estimate of the personal initial α value.

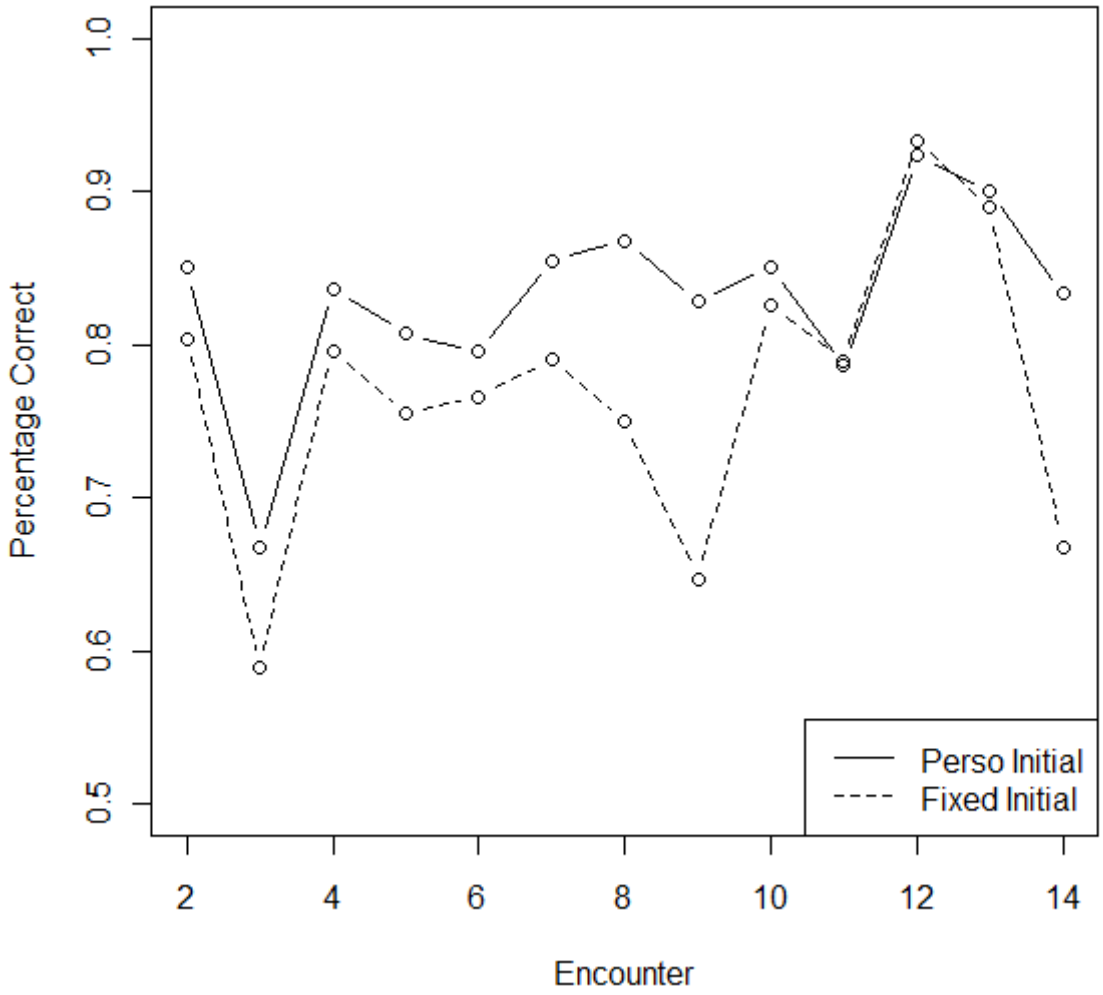

Figure 11: Percentage correct per encounter for the personal initial α and fixed initial α conditions.

Participant	First α	Second α
Feest	0.38	0.34
Huis	0.43	-
Koffie	0.42	0.41
Oma	0.35	0.34
Poes	0.28	0.29
Taart	0.36	0.35
Zee	0.28	0.28

Table 3: Initial α values for the participants of the ID College experiment. The second value for participant "Huis" is missing, because this participant only learned one wordlist.

Standard reaction time

Regarding the effects of the personal standard reaction time (f), the same analysis was performed on the learning data of the ID College experiment. To find the optimal value, all values for f between 0 and 2000 ms (50 ms steps) were taken and for each of these values the mean of the absolute error between the observed reaction times and the reaction times calculated by the model was determined for the first three optimized values of α for every word. These mean errors are plotted against f for every participant in Figure 12.

Again we can see that for every participant, there is a value for the standard reaction time that minimizes the error between the model and the observations. The vertical solid lines again represent the personal value of *f* the test produced for this participant and the dashed lines the *f* value of 300 ms. We can see the included test usually produces a value for *f* that is closer to the optimal value than the standard value of 300 ms is. A one sided paired *t* test shows the mean absolute error of the model fit for the first three optimized values of α is significantly smaller ($t(6) = -2.36$, $p = 0.028$) when using personal values for *f* than when using a standard value of 300 ms. As we can also see, the values produced by the test seem to differ more from 300 ms than the ones produced during the laboratory experiments. Table 2 shows the values for the laboratory experiment have a median of 344 ms. Table 4 shows the values for the ID College experiment have a median of 563 ms. The greater difference with 300 ms in the ID college experiment indicates personalizing *f* might have a greater influence in this case.

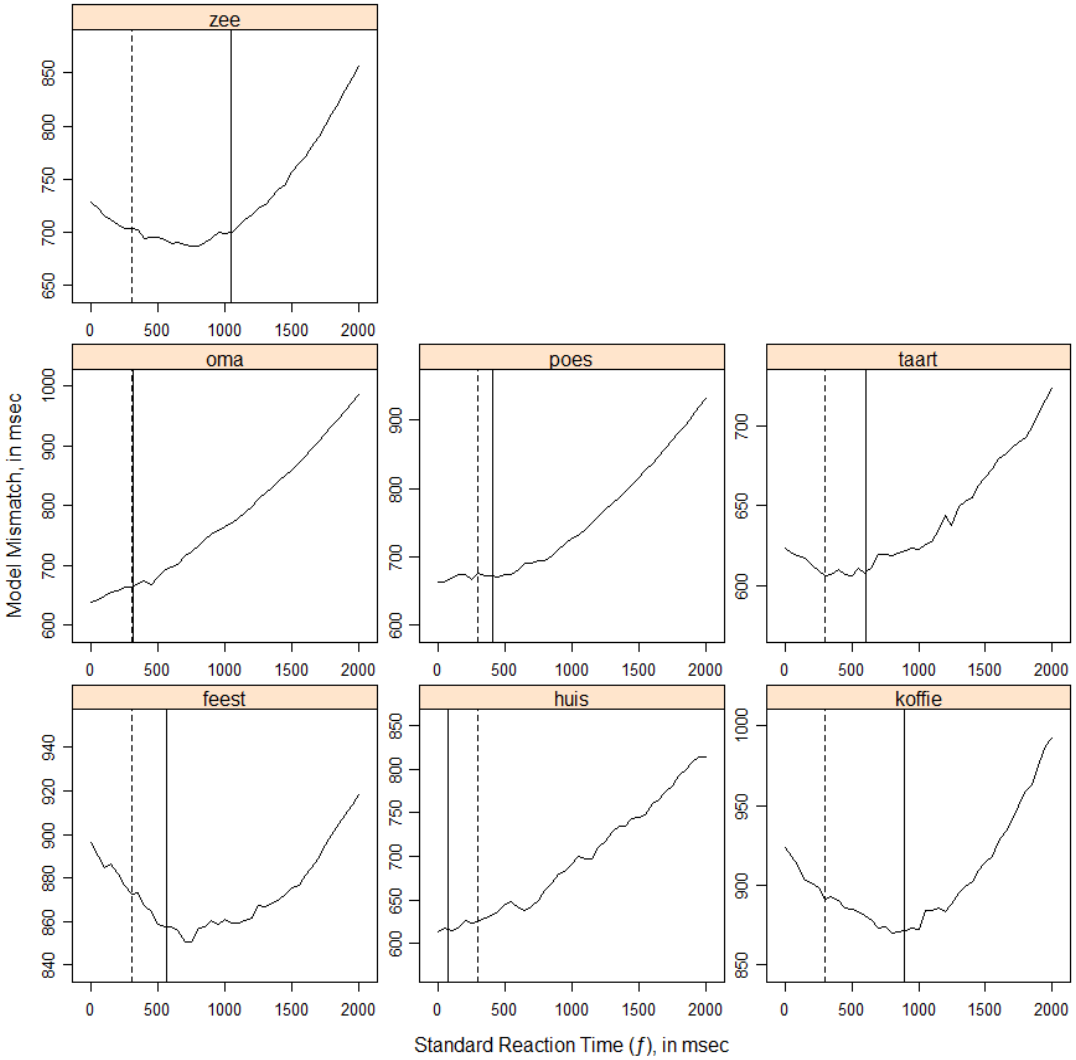

Figure 12: Mismatch between model and observations for the ID College experiment, given different values for the standard reaction time (f). Note that the scale on the y-axis differs for each plot.

Participant	*f* value
zee	1047
oma	312
poes	406
taart	609
koffie	891
huis	78
feest	563

Table 4: f values for the participants of the ID College experiment.

Real world experiment II

Method

A final experiment was conducted at the Dirk van Dijkschool, a primary school in Kampen, The Netherlands, to test whether incorporating the personal standard reaction time parameter (f) leads to better retention on a test as compared to a fixed value. In the previously described experiments the personal standard reaction time is used, but not compared to the use of a fixed standard reaction time. The analysis of the learning data gives some clues as to what the influence of the personal standard reaction time parameter might be, but a direct comparison of the performance on a post-test is needed to find out how useful the personalization really is. The pupils of a primary school are chosen as the participants, because there are expected to be clear differences between the personal f values of primary school pupils. They are also expected to generally have personal f values greater than 300 ms because of their fairly limited experience with computers (at least less than the average psychology student).

The participants were two groups of pupils, adding up to a total of 43 participants. Their age ranges from 11 to 13 years old with a median of 12. A total number of 21 participants were male. A between subject setup was used and pupils were evenly distributed over two conditions: the personal standard reaction time (f) condition and the fixed standard reaction time (f) condition. In the personal f condition a personal value for f is used as obtained by the reaction time test described earlier. In the fixed f condition a fixed value is used that is set at 300 ms. All other parameters were kept the same in both conditions including the initial α value which was fixed at a value of 0.32. A between subject setup was used in this case, because the time each participant was available was limited.

The experiment was conducted at the Dirk van Dijkschool itself. A room was outfitted with four laptops to allow 4 pupils to participate at the same time. 3 of the laptops were MacBooks and one an Asus K50IJ series. The use of this Asus laptop was counterbalanced between the two conditions. The self-made application was again run in the Safari web browser on all laptops. The word list consisted of 20 English – Dutch word pairs (*Appendix C*) that were selected by the pupils´ teacher as being unfamiliar to them. The participants were informed they would be tested on their knowledge of the words the next day, but that their scores would not be part of their grade in English.

The experiment was spread out over two days. The first day consisted of a 15 minute learning session in one of the two conditions preceded by the reaction time test. All participants conducted the reaction time test, also the participants in the fixed standard reaction time condition. The second day consisted of a pen and paper test of the words learned on the first day. The time limit for this test was 10 minutes. 2 participants in the personal f condition were not present during this pen and paper test leaving 21 participants in the fixed f condition and 20 in the personal f condition.

Results

The post-test results of the Dirk van Dijkschool experiment were analyzed with an one sided t test and show no significant increase in number of correctly recalled items on the post-test for the personal f as opposed to the fixed f condition, $t(39.0) = 0.22$. A boxplot of the percentages correct in both conditions is presented in Figure 13. We can see there is quite a wide spread in the test scores. The mean number of recalled items in the fixed f condition is 9.8 with a standard deviation of 5.0 while the mean number of recalled items in the personal f condition is 10.2 with a standard deviation of 4.8. So there is an absolute difference between the conditions, but the standard deviations are very high making it very difficult to find a significant difference between the two.

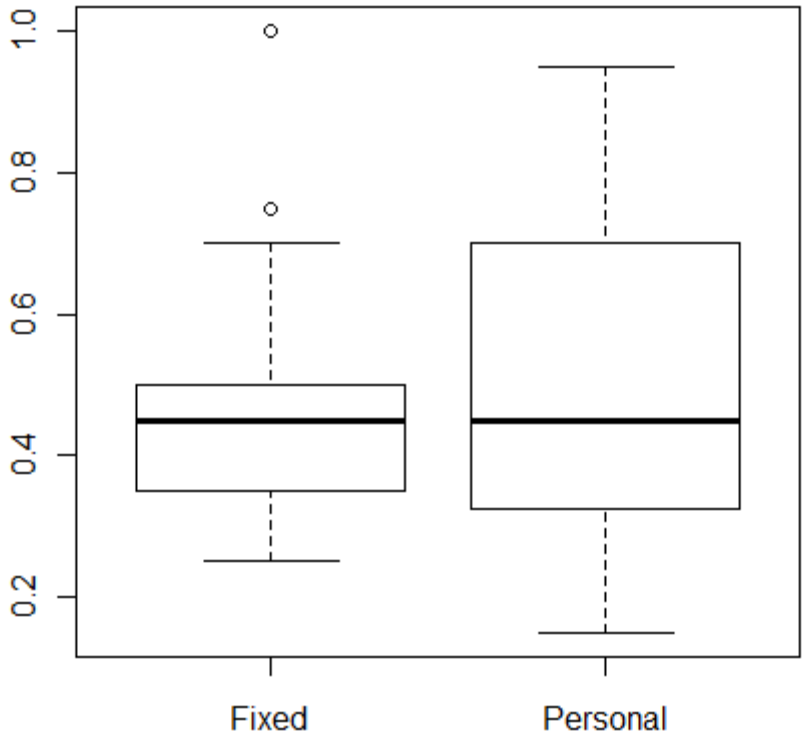

Figure 13: Boxplot showing the percentage correct on the test for the fixed initial α and personal initial α conditions.

To reduce the within group variance the 5 best scoring and 5 worst scoring participants were removed from both groups and another one sided *t* test was performed to test for a significantly higher number of recalled words in the personal *f* condition. The effect is now stronger, but still not significant, $t(15.6) = 1.31$, $p = 0.104$.

There is also a difference between both conditions in the number of distinct words encountered during learning. In the fixed *f* condition the mean number of distinct words encountered is 12.2 with a standard deviation of 4.0 and in the personal *f* condition this is 14.0 with a standard deviation of 4.6. An one sided *t* test however cannot show a significantly higher number of words encountered in the personal *f* condition, $t(37.7) = 1.31$, $p < 0.100$. Again removing the 5 best and worse scoring participants leaves two groups where the number of words encountered during learning is significantly greater for the personal *f* condition, $t(14.9) = 3.88$, $p < 0.001$. This could explain the slightly higher test scores in the personal *f* condition, because words that were not encountered during learning are most probably not recalled during the test. Having encountered more words during learning gives the participants in the personal *f* condition an advantage.

Looking at the percentages correct on every encounter during learning we can see the percentages are higher for the fixed *f* condition beyond the fourth encounter (Figure 14). Because the number of words seen during learning was higher for this condition while the average total number of encounters was practically the same for both conditions (96.9 encounters for the fixed and 97.7 encounters for the personal *f* condition) the spacing of the encounters for every word was wider which could explain why the percentages correct are lower for the personal *f* condition. This is in line with what one would expect given that most personal *f* values are greater than 300 ms (Table 5). This leads to a smaller part of the response latencies to be treated as representative of memory retrieval and thus greater estimated activation values. These on their part lead to smaller estimations in α values and thus wider spacing. A *t* test shows the percentages correct on the 5th until the 14th encounter are indeed significantly higher in the fixed *f* condition, $t(23.1) = 1.74$, $p = 0.048$. It is interesting that despite the fact that the participants in the personal *f* condition answer correctly in

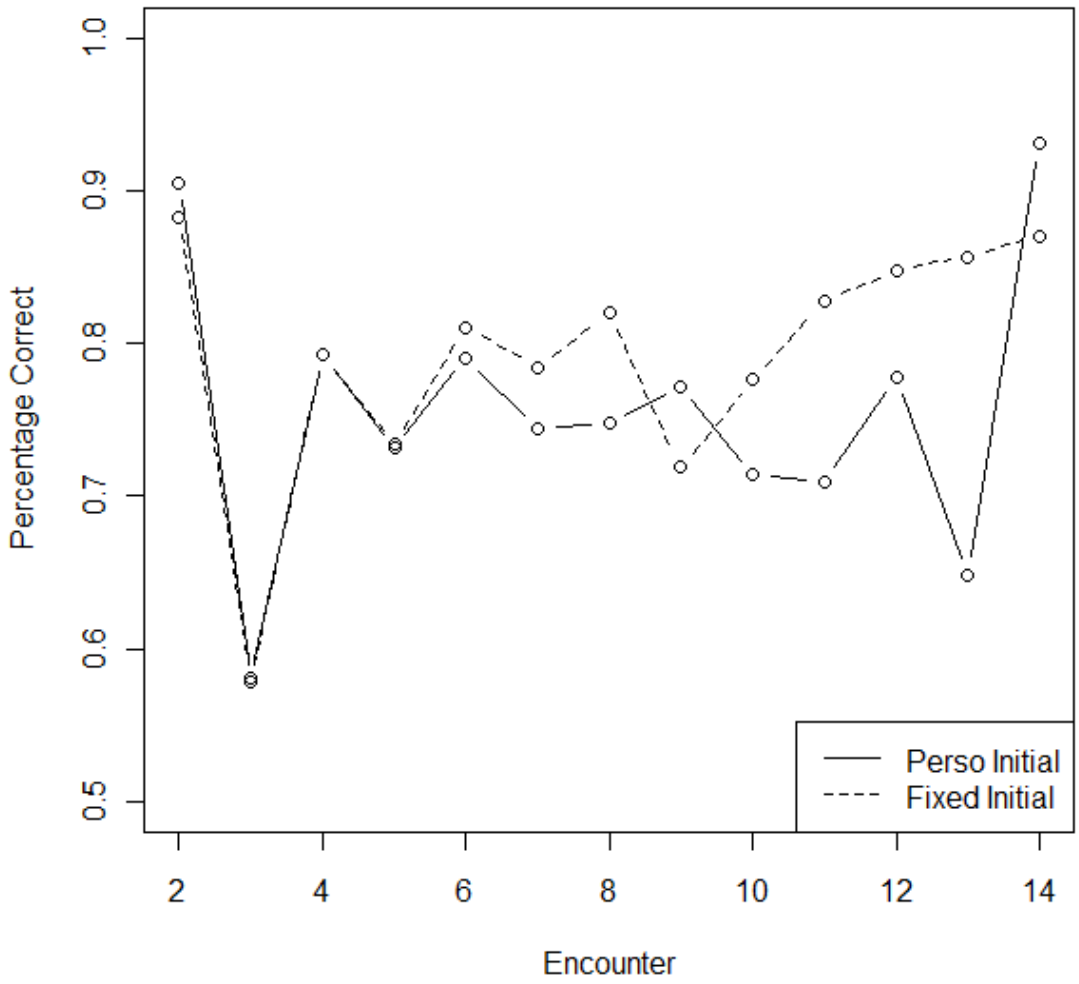

Figure 14: Percentage correct per encounter for the fixed ƒ value condition and personal ƒ value condition in the Dirk van Dijkschool experiment.

less trials during learning this does not show itself in their scores on the post-test.

The *ƒ* values of the participants are shown in Table 5. Some values seem to have been estimated incorrectly because they are close to 0 ms. No person is expected to be that quick. To check whether the found personal values for *ƒ* are generally an improvement though, all values for *ƒ* between 0 and 2000 ms (50 ms steps) were again taken and for each of these values the mean of the absolute error between the observed reaction times and the reaction times calculated by the model was determined for the first three optimized values of α for every word. These mean errors are plotted against *ƒ* in Figure 15 and Figure 16 for every participant in the fixed *ƒ* and personal *ƒ* condition respectively. Remember that in the fixed *ƒ* condition these values were not really used by the algorithm but replaced with the standard value of 300 ms.

As can be seen in the graphs, the personal *ƒ* values produced by the reaction time test are usually closer to the *ƒ* value that realizes the smallest error than the 300 ms standard is. We can also see the reaction time test might have been too conservative in this case. The *ƒ* value seems to have been underestimated substantially for quite a few participants, although it is still true that the *ƒ* value than minimizes the error in these graphs is not necessarily the real *ƒ* value. Because the is a substantial amount of slow responses in these data it is very well possible that the optimal *ƒ* value for some participants is bias towards higher values. The data also appear more noisy than in the previous experiments because the dataset for each participant is smaller. In the previous experiments the graphs are based on at least two 15 minute learning sessions. In this case only one session. It is therefore harder to say something about the quality of the personal *ƒ* value for these participants.

Fixed *f* condition				Personal *f* condition			
Participant	***f* value**	**Participant**	***f* value**	**Participant**	***f* value**	**Participant**	***f* value**
id103	471	id121	479	id101	423	id124	559
id102	384	id127	327	id104	694	id125	408
id105	479	id126	377	id106	62	id128	471
id107	487	id131	391	id108	319	id130	484
id111	295	id129	319	id109	479	id132	335
id110	292	id135	487	id112	439	id133	16
id115	471	id134	649	id114	502	id136	343
id113	367	id138	352	id116	487	id139	295
id118	2	id137	407	id117	31	id141	141
id119	296	id142	448	id120	127	id143	423
id123	15	id140	510	id122	664		

Table 5: ƒ values for the participants in the Dirk van Dijkschool experiment.

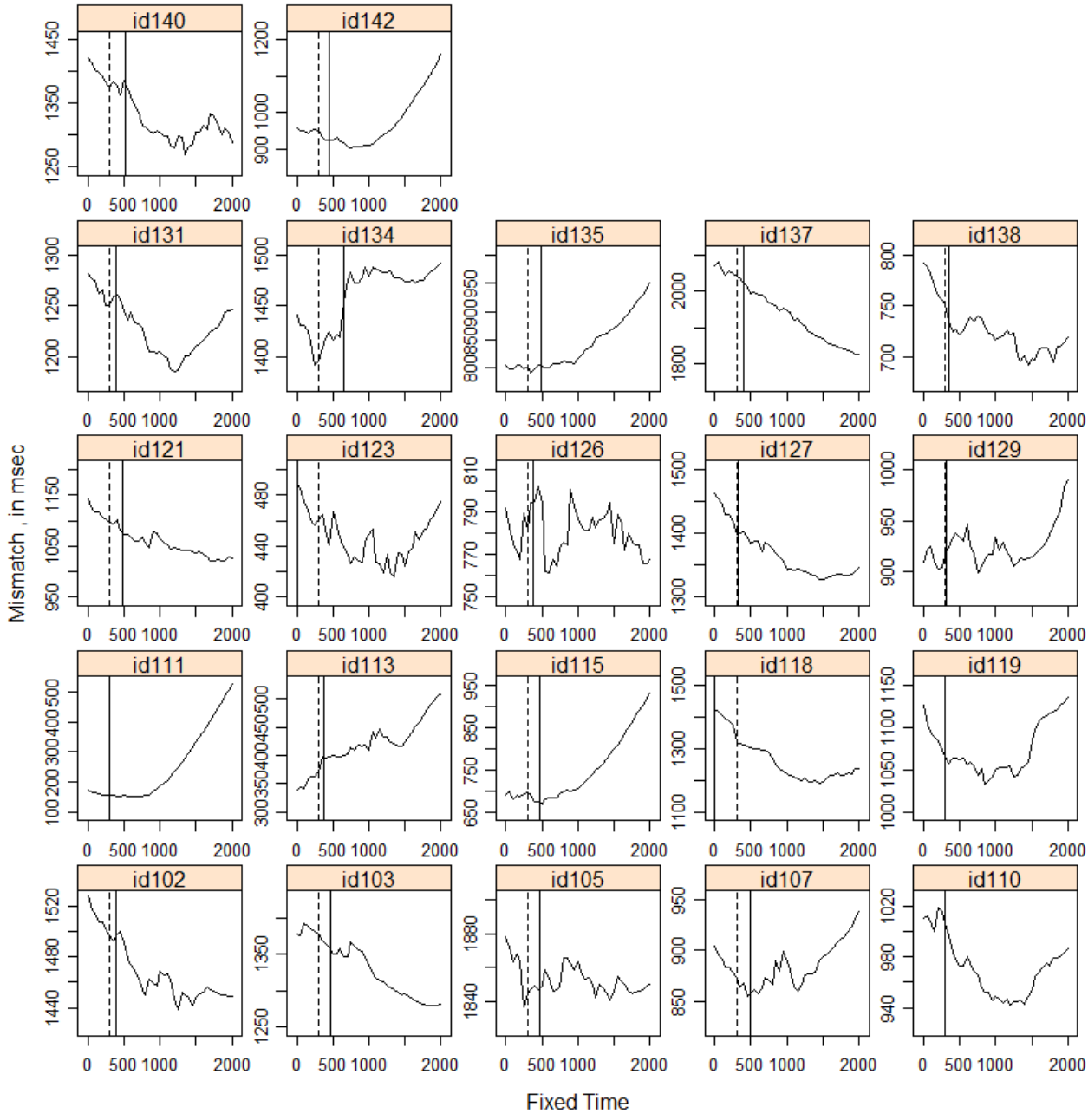

Figure 15: Mismatch between model and observations for the participants in the fixed ƒ value condition of the Dirk van Dijkschool experiment, given different values for the standard reaction time (ƒ). Note that the scale on the y-axis differs for each plot.

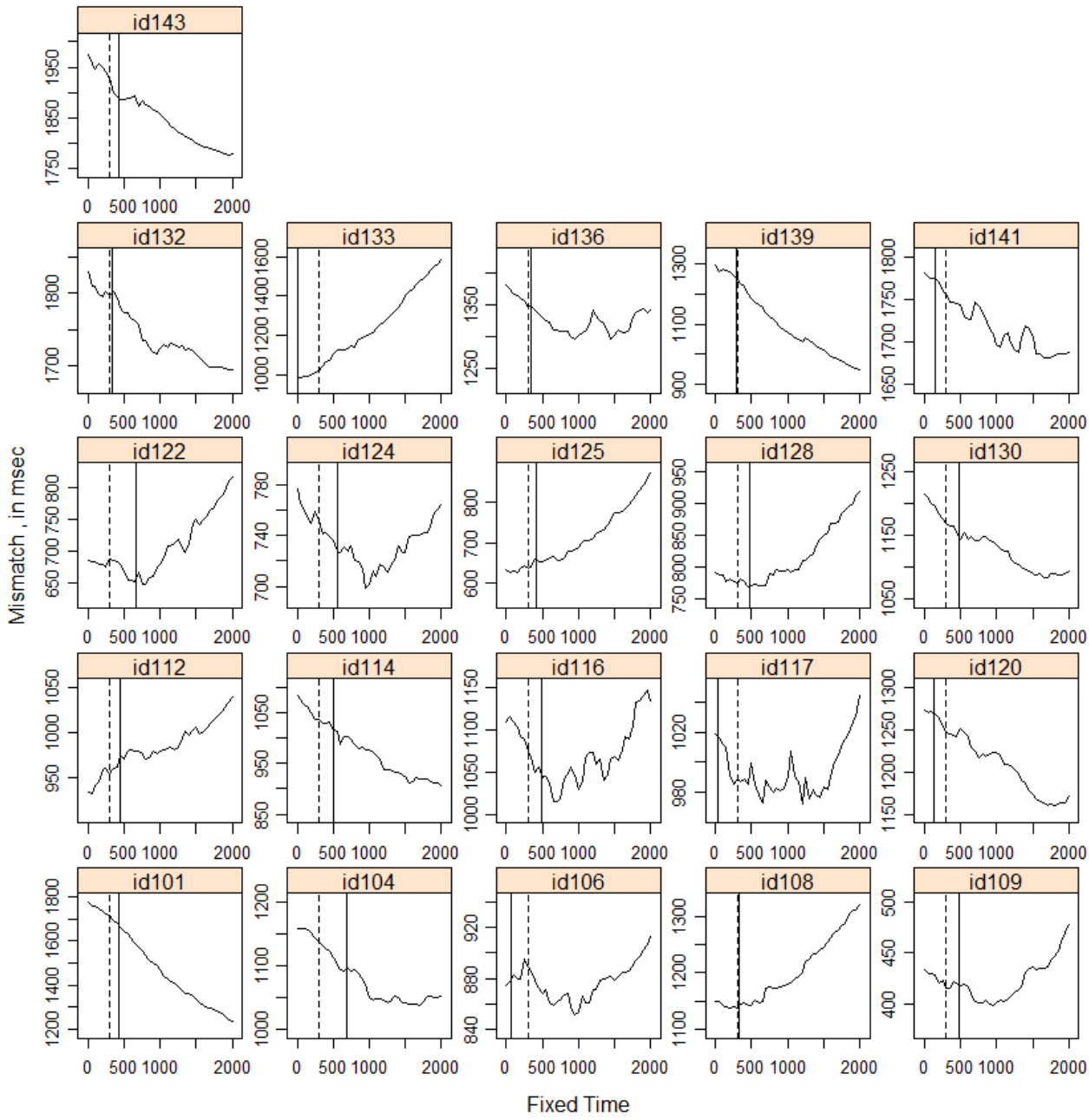

Figure 16: Mismatch between model and observations for the participants in the personal ƒ value condition of the Dirk van Dijkschool experiment, given different values for the standard reaction time (ƒ). Note that the scale on the y-axis differs for each plot.

Discussion

Personal α parameters

What stands out the most in the results is the lack of a significant difference in test retention between the flashcard condition and the standard initial α condition in the laboratory experiments. This is surprising given earlier results by others (Van Rijn, Van Maanen & Van Woudenberg, submitted for publication; Van Thiel, 2010), but not unexplainable. First, the flashcard algorithm I used is more similar to the latency adaptive algorithm than in these previous studies. In Van Thiel (2010), there is no second round of test trials of a deck after the initial study trials, causing much wider initial spacing. Next to that, in Van Thiel (2010) as well as in Van Rijn, Van Maanen and Van Woudenberg (submitted for publication) word lists consisting of 20 words are used. I used lists of 15 words because of the greater word difficulty, but this caused the spacing in the flashcard algorithm to be closer than when using 20 words, since the gap between decks is now 10 instead of 15 words (excluding additional rehearsals due to errors). So we can say the flashcard algorithm produces more comparable spacing in this short time interval, resulting in a lack of differences between the conditions. In the extreme cases of bad and good performance, the flashcard and latency adaptive algorithms produce fairly similar spacing anyway. If one keeps answering incorrectly, the flashcard algorithm will keep presenting the same list of 5 word pairs. Which is quite similar to what the latency adaptive algorithm would do, present a small set of words many times. If one answers mostly correctly, there is a gap of around 14 trials between the trials of one particular word for the flashcard algorithm. This is quite comparable to the latency adaptive algorithm, which would produce wide spacing of more than 10 trials as well. These are the extreme cases of course, but we have to consider the possibility that for these short learning periods with a relatively small number of words, the spacing produced by the flashcard algorithm is not that bad and comparable to the spacing produced by the much more complicated latency adaptive algorithm.

Given the lack of a significant difference in test retention between the flashcard condition and the standard initial α condition, it is not very surprising that the difference between the personal initial α and fixed initial α conditions is not significant either. Although a lasting effect is observed between the two conditions in the percentages of correct answers during learning, this is also not significant. As shown however, there is much more potential for this method when a more slowly adapting algorithm is used, as in Pavlik and Anderson (2008) and Van Rijn, Van Maanen and Van Woudenberg (submitted for publication). Because in the latency adaptive algorithm used in this study only the second rehearsal is directly affected by the personal α value, it's influence is small. To increase this influence, keeping α set at the personal value for a few more rehearsals and only start adjusting after that might be an option. But this had the disadvantage that a wrongly estimated personal α value has longer lasting and more serious consequences. Maybe a hybrid approach could be used in which the α value calculated according to the response to a rehearsal and the personal α value are combined in a weighted average to form the new α value.

Next to this, it has to be noted that 11 (more than half) of the participants had their personal α value initialized at 0.28, because their actual value was lower. As mentioned earlier, this bottom value of 0.28 is used , because lower values lead to very wide initial spacing. But because not all words are found equally difficult, some might need closer initial spacing. Adjusting the α value to allow for this however, can only be done after a test trial, but the number of test trials is very sparse when the value of α is very low, consequently adjustment will be very slow or not even possible. To prevent this the 0.28 bottom value for α is used. This means sometimes the value is too high, but if it is, a few test trials will cause α to drop to the correct value again. In the end, having a few extra rehearsals is not nearly as bad as having a few too little. Returning to the point raised at the start of this paragraph, 11 participants had 0.28 as their personal initial α value, which is very close to the 0.30 in the fixed initial α condition. Another 2 participants had a value around 0.30 and an additional 2 had a value a little bit above 0.30, leaving only 5 participants with a value clearly above

the fixed value. The lack of a significant difference in test performance could be attributed to this fact. The personalization of the initial α value will mostly benefit poorer learners and there were not many of them amongst the participants.

With regard to the results of the first real world experiment at the ID College, it is encouraging to see the same differences between the personal initial α and fixed initial α conditions in this real life setting. Although there is no significant effect, this can very well be caused by the small number of participants in this experiment because the absolute differences are greater than in the laboratory experiment. So this also points into the direction that although the results of the laboratory experiment indicate that the use of personalized initial α values has no added value, it might be the case that this is due to the homogeneous group of good learners that participated. A follow up study on this topic should ideally recruit participants from an ID College like institution. They will probably benefit more from personalization and it might be possible to find a significant improvement in retention on a test.

In general, the lack of significant differences is striking and it is very well possible that the conditions of the experiments lie outside the boundaries in between which the spacing effect operates. This would make the order of the words in these short learning sessions less important and would diminish differences between conditions. That these boundaries exist has been recognized in recent years (Donovan & Radosevich, 1999). Especially the relation between the size of the gaps between presentations and the retention interval appears to be nontrivial, as becomes apparent in Rohrer and Pashler (2007) and Cepeda, Coburn, Rohrer, Wixted, Mozer and Pashler (2009). In this last study they found that the optimal gap between presentations can grow as big as 28 days, when the retention interval becomes very large (6 months). They also claim that the optimal ratio between gap and retention interval for retention intervals in the order of minutes is close to 1, while this ratio is closer to 0.1 for retention intervals of multiple days. I used a retention interval of 1 day and gaps in the order of minutes. So even if the 0.1 ratio applies in this case, according to Cepeda *et al.* (2009) the gaps should still have been in the order of hours. Now of course this would be the optimal situation but the discrepancy between minutes and hours is quite large. It is therefore possible that in this case the effect of spacing is diminished after one day, hence the lack of differences between conditions.

Donovan and Radosevich (1999) on the other hand find that the optimal gap size would be between 1 and 10 minutes for a task such as paired associate learning. This is more or less the same range as the gaps in my experiment, although spacing at the start of the learning sessions is shorter than this to prevent forgetting. Maybe these very short learning sessions of 15 minutes fail to capture enough repetitions with gaps at the high end of this 1 to 10 minute interval. From Donovan and Radosevich (1999) it becomes clear that for gaps of less than 1 minute, spacing has significantly less effect than for gaps of 1 to 10 minutes. The 1 tot 10 minute gap is quite big though and gaps at the lower end of the 1 to 10 minute interval might also not be optimal indicating gaps at the higher end might be preferred. All in all it is still the case Van Rijn *et al.* (submitted for publication) and Van Thiel (2010) did find significant differences using roughly the same gap size and retention interval. They both used a larger number of words though, which by definition leads to wider average spacing, but less time spent learning every separate word. Maybe the extra practice in my experiments was ineffective given the lack of spacing. It is clear however, that the interactions are subtle and the boundaries in between which the spacing effect operates require more exploration.

Personal f parameters

The exploratory analysis of the personal values for the standard reaction time (f) in the laboratory experiment shows potential for its usefulness. There is indeed an f value for each participant that minimizes the mismatch between the model predictions and the observations and these values can vary quite a bit between participants. The absolute differences in the mismatch, however, are not

very big and at first sight one might conclude that the influence is small and insignificant. We must keep in mind however that the α optimization is performed in parallel that can make up for a wrongly estimated *f* value and thus diminishes the mismatch. As the data in Figure 9 indicate, over- or underestimation of *f* can indeed bias the optimization of α and restrict the freedom of the α value leading to inefficient learning schedules.

The second real world experiment was performed to prove the inclusion of personal *f* values indeed leads to higher scores on a post-test. This turns out not to be the case. Even if the best and worst performing participants are rejected from the analysis the differences between the fixed and personal *f* condition are still not significant. There is a difference however and although far from significant, the participants in the personal *f* condition performed a little better on the post-test than the participants in the fixed *f* condition and certainly not worse. Another thing is that the number of words seen during learning is greater when personal values are used. This might compensate for the lower percentages of correct responses to test trials during the learning session. It seems to be worth investigating whether the benefit from the increase in number of words seen outweighs the disadvantage of more incorrect responses in the test trials. In the laboratory experiments where almost all participants see all the words during learning this is not as much of an issue and maximizing the percentages correct seems to be the way to go. When however the number of words is large compared to the time available for learning them it might be better to adjust this strategy and try to keep the number of words seen at a reasonable percentage. It is not unlikely there are some easy to learn words amongst the ones not yet seen that need only a few encounters to be retained until the time of testing. This could be of more benefit than spending more time trying to learn the words already seen.

The quality of the personal *f* values produced by the reaction time test seems to be quite good. Most values are closer to the optimal value than the 300 ms standard value is. They are however not perfect. It has to be stressed again though that this optimal value is not necessarily the real value, but an estimate. There are a few examples of participants for whom the estimated *f* value seems to be a bit off (e.g "id25", id118"). This is unavoidable, the *f* value produced by the reaction time test is only a guess. It is therefore important to monitor the values and correct values when necessary. This is also argued by Rich (1983). In this case the value can be corrected after or during a learning session. The first test trial of a word is in principle identical to a trial in the reaction time test. It is preceded by a study trial so that one sees 2 words for 5 seconds and then one disappears after which it has to be typed in again. It is these trials where the design of the reaction time test is based on. One could thus take the reaction times on first test trials and use them to adjust the personal *f* value if these are very different. If for example a lot of reaction times on the first test trials are a lot quicker than the standard *f* value, this value is probably incorrect and needs to be lowered.

Another positive aspect of the reaction time test is that it usually produces an *f* value that is smaller than the optimal value. This means the test is quite conservative, which is a good thing. It is better to underestimate the value of *f* than to overestimate it. This because underestimation leads to α values that are too high, which is less harmful than α values that are too low. Too high a value means closer spacing than necessary, which is only a waste of time. Too low an α value means wider spacing and leads to a waste of time as well as less learning due to the increase in errors.

A negative aspect is that the priming in the reaction time test causes some people to prepare a little bit too well and produce very quick unrepresentative *f* values. The problem is that during learning sessions learning trials are sometimes immediately followed by test trials producing the same situation where participants can prepare for the response they have to give. This produces problems if the personal *f* value does not account for this. Preventing these situations from occurring is not desired. Using different personal *f* values for these situations might be a good solution.

Directions for future research

A next step in the personalization of the latency adaptive algorithm can be the inclusion of item difficulty as in Pavlik and Anderson (2008). They used initial item difficulty values (the β_i constant in Equation 3) based on earlier data. In a similar way, the initial value of α for a particular word can be based on both information of item difficulty and personal learning ability by for example combining the median of the α values of the participant and the median of the α values of different participants for that word.

Another improvement could be to use a more slowly adaptive algorithm because reaction times are very noisy. The current algorithm is very quick to adapt α to a value very close to the final value in one, maybe two, trials. As Van Thiel (2010) already found reaction times are noisy and converging to a near final value based on one or two reaction times can lead to wrong values. Again the underlying problem here is that once α is very small, spacing is very wide and rehearsals are very sparse, leaving little opportunity for correcting errors. And now with the introduction of personal parameters it might be possible to adjust α in a more robust way while maintaining quick convergence towards the appropriate value.

Looking at overlearning at the end of a learning session might be another thing. The current method sometimes causes the α value for a word to drop to a lower value at the end of a learning session when there have been lots of rehearsals of certain words. The words are then often strongly represented in memory and the activation formula does not seem to accurately capture what is going on at that point. Reactions times seem smaller than predicted by the algorithm. I will not go as far as to claim that a different mechanism is at work there, but it is worth looking into this and keeping it in mind for future attempts at creating optimal learning schedules.

Conclusion

It is hard to come up with a firm conclusion given the results of the various experiments. In the results of the laboratory experiment there is no significant difference in test retention between the flashcard condition and the standard initial α condition. Consequently, there is also no significant difference in test retention between the fixed initial α condition and personal initial α condition. The learning data in combination with the results of the ID College experiment however show there is potential for the use of personal initial α values. The absolute differences and effects in the data from the ID College experiment are greater than in the data from the laboratory experiments. This indicates there might be more benefit from personal initial α values in a real world application.

With respect to the personal standard reaction times (f) there also seems to be potential for their inclusion. Although this did not lead to a significant increase in test scores in the Dirk van Dijkschool experiment, it did lead to an increase and certainly not to a decrease in test scores. It is the same story as for the inclusion of personal initial α values. This research fails to show they significantly increase the retention on a test of the learned words. All the data however indicate that if they have an effect on retention, it will be positive.

It seems therefore reasonable to include the personal parameter settings for α and f in this latency adaptive scheduling algorithm while we wait for results from other research that hopefully produces more firm evidence of the benefits. People using an application that includes personal differences will not notice any difference and if it has an influence on the learning schedules produced there are only indications to assume it will be positive. This influence is maybe too small to measure at this stage, but it could mean the difference between a pass an a fail on a retention test the next day.

References

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

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

Atkinson, R. C. (1972). Optimizing the learning of a second language vocabulary. *Journal of Experimental Psychology.* 96 (1). pp. 124-129.

Atkinson, R. C., & Paulson, J. A. (1972). An approach to the psychology of instruction. *Psychological Bulletin,*78. pp. 49-61.

Baddeley, A. (2003). Working memory and language: An overview. *Journal of Communication Disorders,* 36. pp. 189–208.

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. A. Bower (Ed.), *Recent advances in learning and motivation,* 8. pp. 47–90. New York: Academic Press.

Beck, C. D. O., Schroeder, B., & Davis, R. L. (2000). Learning Performance of Normal and Mutant Drosophila after Repeated Conditioning Trials with Discrete Stimuli. *The Journal of Neuroscience,* 20 (8). pp. 2944-2953.

Bloom, K. F., & Shuell, T. J. (1981). Effects of massed and distributed practice on the learning and retention of second-language vocabulary. *Journal of Educational Research,* 74 (4). pp. 245-248.

Carew, T. J., Pinsker, H. M., & Kandel, E. R. (1972). Long-term habituation of a defensive withdrawal reflex in Aplysia. *Science,* 175. pp. 451-454.

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). pp. 354-380.

Cepeda, N. J., Coburn, N., Rohrer, D., Wixted, J. T., Mozer, M. C., & Pashler, H. (2009). Optimizing distributed practice: Theoretical analysis and practical implications. *Experimental Psychology,* 56 (4). pp. 236-246.

Chen, C. M., Lee, H. M., & Chen, Y. H. (2005). Personalized E-learning system using item response theory, *Computers and Education,* 44 (3). pp. 237–255.

Cowan, M. (2000). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioural and Brain Sciences,* 24. pp 87-185.

Czaja, S., & Sharit, J. (1993). Age differences in the performance of computer based work. *Psychology and Aging,* 8 (1). pp. 59-67.

Dempster, F. N. (1987). Effects of variable encoding and spaced presentations on vocabulary learning. *Journal of Educational Psychology,* 79 (2). pp. 162-170.

Dempster, F. N. (1989). Spacing effects and their implications for theory and practice. *Educational Psychology Review,* 1 (4). pp. 309-330.

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 (5). pp. 795-805.

Ebbinghaus, H. (1885/1913). *Memory: A contribution to experimental psychology.* Translated by Henry A. Ruger and Clara E. Bussenius (1913).

Glenberg, A. M. (1979). Component-levels theory of the effects of spacing of repetitions on recall and recognition. *Memory & Cognition,* 7 (2). pp. 95-112.

Graesser, A. C., Van Lehn, K., Rose, C. P., Jordan, P. W., & Harter, D. (2001), Intelligent tutoring systems with conversational dialogue. *AI Magazine,* 22 (4). pp. 39-51.

Janiszweski, C., Noel, H., & Sawyer A. G. (2003). A meta-analysis of the spacing effect in verbal learning: Implications for research on advertising repetition and consumer memory. *The Journal of Consumer Research,* 30 (1). pp. 138-149.

Jonassen, D. H., & Grabowski, B. L. (1993). *Handbook of individual differences, learning, and instruction.* Hillsdale, New Jersey: Erlbaum.

Kail, R., & Salthouse, T. A. (1994). Processing speed as a mental capacity. *Acta Psychologica,* 86. pp. 199-225.
Kliegel, M., & Altgassen, M. (2006). Interindividual differences in learning performance: The effects of age, intelligence, and strategic task approach. *Educational Gerontology,* 32. pp. 111-124.
Kobsa, A. (1993). User modeling: Recent Work, Prospects and Hazards. In: *Adaptive User Interfaces: Principles and Practice.* M. Schneider-Hufschmidt, T. Kühme, and U. Malinowski, (eds.). North-Holland: Amsterdam. pp. 111-128.
Menzel, R., Manz, G., Menzel, R., & Greggers, U. (2001). Massed and spaced learning in honeybees: The role of CS, US, the intertrial interval and the test interval. *Learning and Memory,* 8. pp. 198-208.
Metcalfe, J., & Kornell, N. (2003). The dynamics of learning and allocation of study time to a region of proximal learning. *Journal of Experimental Psychology: General,* 132 (4). pp. 530-542.
Murray, T. (1999). Authoring intelligent tutoring systems: An analysis of the state of the art. *International Journal of Artificial Intelligence in Education,* 10. pp. 98-129.
Oberauer, K., Süß, H. M., Wilhelm, O., & Wittmann, W. W. (2003). The multiple faces of working memory: Storage, processing, supervision, and coordination. *Intelligence*, 31. pp. 167-193.
Pavlik, P. I., Jr. (2007). Understanding and applying the dynamicsof test practice and study practice. *Instructional Science,* 35. pp. 407-441.
Pavlik, P. I., Jr., & Anderson, J. R. (2005). Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect. *Cognitive Science,* 29 (4). pp. 559–586.
Pavlik, P. I., Jr., & Anderson, J. R. (2008). Using a model to compute the optimal schedule of practice. *Journal of Experimental Psychology: Applied*, 14 (2). pp. 101-117.
Raaijmakers, J. G. W. (2003). Spacing and repetition effects in human memory: Application of the SAM model. *Cognitive Science,* 27. pp. 431-452.
Reber, A. S., Walkenfeld, F. E., & Hernstadt, R. (1991). Implicit and explicit learning: Individual differences and IQ. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 17. pp. 888-896.
Rich, E. (1983). Users are individuals: Individualizing user models. International Journal of Man-Machine *Studies,* 18. pp. 199–214.
Roediger, H. L., III, & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science,* 1. pp. 181-210.
Rohrer, D., & Pashler, H. (2007). Increasing retention without increasing study time. *Current Directions in Psychological Science*, 16 (4). pp. 183-186.
Rohrer, D., & Taylor, K. (2006).The effects of overlearning and distributed practice on the retention of mathematics knowledge. *Applied Cognitive Psychology,* 20 (9). pp. 1209-1224.
Self, J.A. (1990). Bypassing the intractable problem of student modelling. In Frasson, C., & Gauthier, G. (Eds.), *Intelligent Tutoring Systems: At the crossroads of artificial intelligence and education.* Ablex Publishing, Norwood. pp. 107-123.
Van Rijn, H., Van Maanen, L., & Van Woudenberg, M. (Submitted for publication). Optimizing short-session learning by individualizing schedules of practice.
Van Thiel, W. (2010). Optimize learning with reaction time based spacing: By modifying the order of items in a learning session. *Unpublished master's thesis*, University of Groningen.
Wisner, B. L., Lombardo, J. P., & Catalano, J. F. (1988). Rotary pursuit performance as a function of sex, sex-role, and intertrial interval. *Perceptual and Motor Skills,* 66. pp. 443-452.

Appendix A: Deduction of the new α value.

The deduction of the new α value is based on the difference between the reaction time belonging to the current activation of an item and the observed reaction time of recalling that item. As soon as a encounter has taken place an observed reaction time is available to adjust α with. Remember that a reaction time corresponds directly to an activation value through the equation:

(A.1) $$RT = Fe^{-m} + f$$

In this formula *m* is the activation, *RT* is the reaction time, *F* is a multiplier kept constant at a value of 1 and *f* is the standard reaction time (personal or fixed, this does not matter right now). By rewriting this formula we can thus convert the observed reaction time to the observed activation:

(A.2) $$m_i(t_j) = -\ln\left(\frac{RT - f}{F}\right)$$

So we now know what the actual activation of this item is. Remember that α is a measure of how quickly the activation added by the previous encounters has decayed. We can now calculate the α corresponding to the previous encounter, because the current encounter provides the reaction time with which we can measure how quickly the activation added by the previous encounter has decayed. *Equation A.2* provides us with the observed activation, this should be equal to the activation calculated by the model using *Equation A.3*.We can however separate the summation in *Equation A.3* into two parts: the activation added by all encounters except the last encounter, and the activation added by this last encounter. This will allow us to calculate how much activation was added by the last encounter. Combining all the above gives us *Equation A.4*, here you see the observed activation equals the summation of the activity added by all earlier encounters excluding the previous encounter, plus the activation added by the previous encounter.

(A.3) $$m_i(t_j) = \ln\left(\sum_{j=1}^{n} (t - t_j)^{-d_{i,j}}\right)$$

(A.4) $$m_i(t_j) = \ln\left(\left(\sum_{j=1}^{n-1} b_j (t - t_j)^{-d_{i,j}}\right) + (t - t_j)^{-d_{i,j=n}}\right)$$

If we remove the natural logarithm we can rewrite this to:

(A.5) $$(t - t_j)^{-d_{i,j=n}} = e^{m_i(t_j)} - \left(\sum_{j=1}^{n-1} (t - t_j)^{-d_{i,j}}\right)$$

We can then 'free' *d* by taking the $(t\text{-}t_j)$th logarithm which gives us:

(A.6) $$d_{i,j=n} = -^{(t-t_j)}\log\left(e^{m_i(t_j)} - \left(\sum_{j=1}^{n-1} (t - t_j)^{-d_{i,j}}\right)\right)$$

We now know the decay that corresponds to the previous encounter and we can calculate the α value that corresponds to the previous encounter, because the decay depends on the activation $m_{i,j}$ at the time the previous encounter took place and α (where *c* equals 0.25):

(A.7) $$d_{i,j} = ce^{m_i(t_j)} + \alpha_{i,j}$$

So we now have the α value corresponding to the previous encounter. We also have an α value corresponding to the earlier encounters. In the range between these two values we will now perform a greedy search to find the value for α that will produce the best fit between the reaction times predicted by the model and the observed reaction times for all encounters together. So one value for α that will produce the best overall fit.

The greedy search algorithm is taken directly from Van Thiel (2010) and is fairly straightforward. The old and the new found value for α are taken as the boundaries of the interval on which we have to search for the best overall fitting α. The reason behind this is that the old α value was the best fit for all encounters except the previous one. The new α value is the best fit for this previous encounter, so the best fitting α value for all encounters has to lay somewhere in between. The greedy search algorithm then consists of a couple of steps:

1. For the lower and upper α value, calculate the mismatch between the model's predicted reaction times for all encounters of this word (except the first two) and the observed reaction times.
2. Take the α value that is right in between the lower and upper value. If the lower value had the best match, this middle value is now the upper value. If the upper value had the best match, this middle value is now the lower value. This way we cut the search interval in half and throw away the part that was very likely to lead to a bigger mismatch than the α values in the remaining interval.
3. We repeat step 1. with these new lower and upper values and continue this process until we have done it six times in total. The value with the best match in the last run is the new α value. In practice the 6 repeats turn out to be enough to converge to a reliable estimate.

The value thus found is the new α value for this word pair.

Appendix B: The failed experiments

The similar version of the mentioned laboratory experiments was conducted earlier, but the results were only partially useful due to a ceiling effect. Of the 15 participants (6 males, average age 22), almost half still knew the correct translation of all the words in all conditions, making comparison of their performance in the different conditions impossible.

These experiments used the same setup as the experiments described above, but with three differences. The first difference was that the amount of word pairs in a word list was 12. This number was based on a pilot study amongst artificial intelligence students at the Rijksuniversiteit Groningen, The Netherlands, where a word list length of 20 was used, but the results on a retention test were so catastrophic it was decided to shorten the word list length to 12 word pairs. Because of the found ceiling effect amongst psychology students in this experiment, the amount of word pairs was increased again to 15 in the final experiments.

The second difference is the description of the experiment used to recruit participants. The initial description attracted participants very skilled in word pair learning. This contributed to the ceiling effect and in addition caused their final α values to be very low. This would mean a very small difference in default α values during the second experiment and little hope for a significant effect. The description for the later experiments was thus changed into something less appealing, in the hope of attracting less skilled participants.

The third difference was the order of the learning sessions. In these experiments the words were presented while iterating between two conditions, so for the first experiment, the first word was presented using the algorithm of condition one, the second using the algorithm of condition four, the third using the algorithm of condition one, etc. This method however adds extra spacing to the flashcard algorithm, because of the intervening presentations, so in the later experiments, the conditions were not iterated but a blocked presentation was used.

The rest of the experimental setup however was exactly the same.

Appendix C: Word lists

List 1	
English	**Dutch**
the farrago	de mengelmoes
to sunder	opdelen
to impugn	betwisten
profligate	verkwistend
to cosset	verwennen
to immure	opsluiten
turgid	gezwollen
obsequious	kruiperig
to propound	voorstellen
the prestidigitation	de goochelkunst
to adumbrate	afschetsen
somniferous	slaapverwekkend
the physiognomy	de gelaatstrekken
the anodyne	de pijnstiller
the tergiversation	de afvalligheid

List 2	
English	**Dutch**
to gainsay	tegenspreken
venial	vergeeflijk
the parsimony	de gierigheid
arcane	geheim
to inveigle	verleiden
hirsute	behaard
to obtrude	opdringen
the asperity	de ruwheid
rebarbative	afstotelijk
the argot	het jargon
factitious	kunstmatig
to undulate	golven
the harangue	de toespraak
the busker	de straatmuzikant
the quandary	het dilemma

List 3	
English	**Dutch**
to inveigh	schelden
the propinquity	de nabijheid
to collude	samenspannen
to expiate	boeten
the penchant	de voorliefde
to abrogate	intrekken
hoary	grijs
to expatiate	uitweiden
the compunction	het berouw
lascivious	wulps
the firmament	de hemel
ephemeral	kortstondig
to ruminate	herkauwen
the peregrination	de zwerftocht
to ululate	huilen

List 4	
English	**Dutch**
lachrymose	huilerig
diaphanous	doorschijnend
the augury	het voorteken
the calumny	de laster
to masticate	kauwen
to flummox	verwarren
jejune	schamel
to portend	voorspellen
the vicissitude	de wisselvalligheid
to acquiesce	berusten
condign	verdiend
portentous	onheilspellend
the admonition	de vermaning
to cozen	bedriegen
puissant	machtig

Table A.1: Word lists used for the laboratory experiments.

English	**Dutch**		**English**	**Dutch**		**English**	**Dutch**
offspring	afstammeling		temptation	verleiding		carriage	rijtuig
casual	oppervlakkig		to faint	flauwvallen		favourable	gunstig
limbs	ledematen		ordinary	gewoon		spectacle	bril
to perspire	zweten		scar	litteken		kerbstone	stoeprand
fatique	vermoeidheid		heavily	erg		to peel	schillen
giddy	duizelig		luggage	bagage		graceful	sierlijk
kidnev	nier		pinch	snufje			

Table A.2: Word list used for the Dirk van Dijkschool experiment.