Optimal Fact Learning: Applying Presentation Scheduling to Realistic Conditions

MENNO NIJBOER

UNIVERSITY OF GRONINGEN

Department of Artificial Intelligence
Human-Machine Communication

MASTER THESIS

SUPERVISOR: HEDDERIK VAN RIJN
INTERNAL SUPERVISOR: JELMER BORST

MAY 2011

Optimal Fact Learning: Applying Presentation Scheduling to Realistic Conditions

Abstract: Learning facts is a large part of any education. Earlier research has shown that learning can be improved by carefully scheduling the order of fact presentations. Concretely, these schedules balance two important memory effects: spacing and the testing effect. In this thesis the latest evolution of a scheduling algorithm, which is based on a cognitive model, is tested under realistic (classroom) conditions. These conditions include regular, fixed length study moments. The model is a refinement of earlier work done by Van Thiel (2010) and Van Woudenberg (2008), and uses the memory equations found in ACT-R (Anderson et al., 2004) to model how participants forget information. This model is continuously adjusted based on reaction time feedback. Two studies were performed: the first experiment had a single study session, while the second experiment had three sessions spread over an entire week. Both studies were concluded with a test. The model was compared against other scheduling algorithms and a standard learning method, and was found to increase performance significantly. Learning behaviour of the participants was analysed, and it was discovered that test performance can be predicted from the model, which opens up possibilities for tailoring the length of study sessions. Finally, analysis of the test results showed that participants using the model performed significantly better than participants in all other conditions. Thus, this work shows that scheduling using a cognitive model improves learning when used in typical educational circumstances.

Keywords: Spacing, testing effect, memory, forgetting, ACT-R, practice scheduling, computer assisted learning

Acknowledgements

I would like to thank my parents for giving me the support and opportunity to create this work. I would also like to thank Hedderik van Rijn for giving a helping hand whenever it was needed.

Contents

1 Introduction **1**

1.1 A History of the Spacing and Testing Effects . . . 1
1.2 Spacing Divided . . . 2
1.3 Effective Spacing . . . 3
1.4 Models of Spacing . . . 4
1.5 Optimizing Learning . . . 5
1.6 Other Learning Strategies . . . 6
1.7 Spacing and Testing in Practice . . . 7
1.8 Thesis Overview . . . 7

2 Revisiting the ACT-R Model for Spacing **9**

2.1 The Model Explained . . . 9
2.1.1 Origins: Spacing in ACT-R . . . 9
2.1.2 Current State of the Model . . . 11
2.2 New Improvements . . . 15
2.2.1 Estimating Reading Times . . . 15
2.2.2 Adapting the Model for Multiple Sessions . . . 18
2.2.3 Model Overview . . . 21

3 Pilot Study **23**

3.1 Another Model: Scheduling for Efficiency . . . 23
3.1.1 Practice Efficiency as Selection Criteria . . . 23
3.1.2 Dynamic Adaptation . . . 25
3.1.3 Item Scheduling . . . 26
3.1.4 Model Overview . . . 26
3.2 Experiment . . . 28
3.2.1 Method . . . 28
3.2.2 Results . . . 30
3.2.3 Conclusion . . . 40

4 Final Study **43**
4.1 Model Changes . . . 43
4.2 Method . . . 43
4.2.1 Participants . . . 43
4.2.2 Materials . . . 44
4.2.3 Design . . . 45
4.3 Results . . . 45
4.3.1 Differences Between Methods . . . 45
4.3.2 Item Difficulty . . . 46
4.3.3 Change Over Item Presentations . . . 46
4.3.4 Change Between Study Sessions . . . 49
4.3.5 Item List Selection . . . 51
4.3.6 Test Score Prediction . . . 51
4.4 Conclusion . . . 53

5 Discussion **55**
5.1 Item Scheduling Under Practical Circumstances . . . 55
5.1.1 The Studies in Short . . . 55
5.1.2 The Bottom Line: Which Method Works Best? . . . 56
5.1.3 Computer Aided Learning . . . 57
5.2 Describing the Learning Process . . . 58
5.3 Assessing the Study Difficulty of Individuals . . . 59
5.4 Future Research Directions . . . 59
5.5 Final Thoughts . . . 60

A Pilot Experiment Fact List **65**

B Final Experiment Fact List **69**

CHAPTER 1

INTRODUCTION

Whether it is historical dates or capital cities, everyone will have had to learn facts at some point during their life. Preferred learning strategies differ from person to person, given differences in learning ability and exposure to such strategies. However, one quite ubiquitous method is called 'cramming'. This method consists of rehearsing material as often as possible in a short timespan, often just before a test. However, this type of *massed practice* has been found to lead to poor long term memory retention (Ebbinghaus, 1885). Despite efforts to develop an optimal learning style, the sheer exposure to cramming will likely have resulted in many people using it as their dominant approach to studying. This would mean many people are using a method that does not deliver the maximal benefit. As such, many people are not living up to their full learning potential. Although efforts are being made (Pashler et al., 2007), many study methods (such as the repeated testing of study material) have seen no widespread adoption by educational institutes, and as such few students are aware of them. This is particularly true for two effects which improve memory and decrease the total time spent learning: the spacing and testing effects. This thesis will explore what the results are of applying the ideas of the spacing and testing effects to real classroom conditions. This exploration will be done through a model, which seeks to use both effects by presenting facts in a specific order. Two studies were performed to investigate the effect the model had on test scores, compared to other study methods. Before moving on to an explanation of the model, an account of the spacing and testing effects will be given to explain how they affect memory.

1.1 A History of the Spacing and Testing Effects

A history of spacing must start with Ebbinghaus, who first described the phenomenon over a century ago (1885). While training himself to remember lists of nonwords[1], much like one would learn vocabulary of a foreign language. Ebbinghaus discovered that his performance improved when he spread the repetitions over a larger period of time. This 'spaced' learning is often contrasted with massed learning, where many repetitions are performed over a short time span. Spacing (a term used interchangeably with distributed learning) has been shown to be relevant for many different memory tasks, including vocabulary, fact, and

[1] Nonwords are words that have no meaning, or are not known to exist

motor learning (Donovan and Radosevich, 1999; Pashler et al., 2007). The testing effect was first presented by Carrier and Pashler (1992): it states that retrieval from memory results in increased retention. They support two possible explanations for the effect: retrieval strengthens existing retrieval routes, or it causes the creation of new routes.

In contrast, a variety of theories have been proposed to explain the spacing effect. An early attempt at explaining the spacing effect is the deficient-processing theory by Hintzman (1974), which states that during massed presentations the second presentation is insufficiently processed. More plainly put, the second presentation is given less attention. One of the most popular accounts of the spacing effect is the variability theory (Glenberg, 1979). According to this theory, contextual fluctuation plays a large role in the effects of practice and forgetting. Each practice results in an encoding of the stimulus and its context. Over time, the context will change. As the time between repetitions increases, the overlap in context becomes smaller, leading to more unique retrieval cues for the stimulus, resulting in greater retrieval probability.

Aside from encoding variability, Janiszewski et al. (2003) identified several other theories: the attention, rehearsal, retrieval and reconstruction hypothesis. They compared predictions of these theories with results from spacing literature, and found that no theory could predict all outcomes correctly. This would imply that a model based on these theories would be incomplete, and unable to account for the spacing effect in some cases. Despite this, attempts to model the spacing effect have been fairly successful at fitting data from a variety of spacing experiments (Raaijmakers, 2003; Pavlik and Anderson, 2005; Lindsey et al., 2009), and have proven useful in practical learning applications (Van Rijn et al., 2009). So far spacing research has been presented as having a singular focus. This, however, is not the case.

1.2 Spacing Divided

A distinction within spacing research must be made. On one hand there is research into the effects of the ISI (InterStudy Interval) and RI (Retention Interval) lengths on performance, which could be considered between-session spacing. On the other hand there is research into the effects of within-session spacing, where spacing is applied to the item order during a study session. Most spacing research so far has been done into between-session effects. Recently, an important finding for between-session spacing was done by Cepeda et al. (2008), who performed a large meta study of ISI and RI combinations to find the best between-session spacing interval. They found that the optimal ISI duration increases non-monotonically as the length of the RI increases: the optimal ISI declined from 20 to 40% of a one week RI to about 5 to 10% of a one year RI. Their analysis of older spacing studies also revealed that an ISI that is too short is more

harmful to retention than an ISI that is too long, as the accuracy of recall decreases much more gradually with larger intervals. In practice this means it is better to over-estimate a spacing interval than to under-estimate it. An issue with this type of spacing is motivation: studying using this approach requires a long-term investment, and has a relatively low short-term benefit.

Within-session spacing applies the effect on a much smaller timescale. Study sessions are typically kept between twenty minutes and one hour. The time between sessions is kept constant, and does not necessarily depend on the RI. Recent advances in computer models of the spacing effect are probably an important factor in the increased interest into within-session spacing (evident by the rise of publications on the subject: Pavlik and Anderson (2008); Pavlik et al. (2008); Van Rijn et al. (2009)), as they have made it practical to do realtime predictions for short-term spacing effects on individual items. Consequently, these models open up the possibility of creating a time-constrained learning schedule where each item rehearsal is optimally spaced within a relatively short time span in order to maximize long term retention. What is considered optimal depends on the model, but the general idea is to give an additional practice just before the item has been forgotten by the participant.

Between and within-session spacing are not two mutually exclusive types of spacing. Rather, the two could be combined: the content of a study session can be spaced according to within session spacing, while the sessions themselves are scheduled using between-session spacing. This thesis will cover this setup to some extent by investigating a combination of within-session spacing and evenly spaced sessions. However, the between session spacing interval is a result of practicality, and not the focus of this work. As explained, the two types of spacing are quite different. However, the requirements to be effective are the same for both.

1.3 Effective Spacing

The effectiveness of spacing (in a practical setting) is partially dependent on the practice method that is used. Most experimental setups for spacing only use drill trials (where the participant has to provide the answer to a question) after the initial presentation, and with good reason. Drill trials require memory recall, which has been found to be critical to learning. They are far more effective than study trials (where the participant is shown both the question and the answer), which provide almost no benefit (Pashler et al., 2007; Karpicke and Roediger, 2008). This is consistent with the testing effect. Another important element of practice is feedback. Pashler et al. (2007) found that, during study, presenting the correct answer if participants were wrong improved successful recall on the final test around fivefold. Giving feedback after a correct recall made no difference. Neither did giving more elementary feedback, such as ‘Correct’ or ‘Incor-

rect'. Spacing, testing, and feedback will form the basis of the models described in this thesis. All of these models find their origins in earlier attempts at modelling the effect itself.

1.4 Models of Spacing

Several models that attempt to describe the spacing effect have been proposed. Raaijmakers (2003) presented a model based on the variability theory (Glenberg, 1979). This model is called SAM (Search of Associative Memory), and is is successful in creating qualitative fits for most of the datasets that were examined. In SAM, information is represented by memory images, and retrieval is based on retrieval cues. Memory strength is calculated by multiplying the strengths of all individual cues linked to a particular image. The probability of successful retrieval is calculated by dividing the memory strength of the image by the sum of strengths of all other images. Contextual fluctuation is used to account for forgetting. This means that the retrieval probability decreases over time because the context is becoming increasingly different from the context in which the image was created. The biggest concern with the model is that it uses a large number of free parameters, which is undesirable for practical applications, as estimating these values could be costly, or simply not possible. This is also the reason why an implementation of SAM was not considered for this thesis.

Another influential model is that of Pavlik and Anderson (2005), based on the ACT-R (Adaptive Control of Thought - Rational) architecture (Anderson et al., 2004). The model uses a modified version of the standard ACT-R declarative memory equations to account for spacing. These equations express the strength of an item in memory as an activation value. Over time, the strength (or activation) of an individual presentation decays as a power function. Summed together, these presentations represent the activation of a particular item. The modification, as proposed by Anderson et al. (2004), lies in the use of separate decays for each presentation, instead of a fixed decay parameter. These decay values cause large intervals between presentations to cause low decay, and short interval to cause high decay. While effective at describing the spacing effect over a short period, such as a single study session, the model needs a slight modification to work with large time gaps between sessions: the model predicts too much forgetting in such cases. To account for this, psychological time is used; the time between sessions is scaled by some value to diminish forgetting during this period. With this adjustment, the model forms a basis on which a practical implementation can be built.

1.5 Optimizing Learning

One very specific implementation of spacing models is the creation of *personalized learning schedules*. These schedules take into account the differences in forgetting between individuals and the items being learned. This information is used to determine the optimal amount of spacing between rehearsals of an item, for that particular individual. Essentially, such schedules try to combine spacing with the testing effect: items are presented just before they are forgotten so that the spacing between repetitions is maximized, while still retaining the benefit of the testing effect. Presentations are in the form of study trials where the item and the answer are shown, and test trials where only the item is shown and the answer must be recalled. Recently, several models that seek to optimize item scheduling within a session have been proposed. These models are based in varying degrees on the Pavlik and Anderson (2005) ACT-R model. The model by Pavlik and Anderson (2008) was used to teach participants 180 Japanese-English word pairs. The model produces a presentation sequence for items which maximizes retention for a specific retention interval, by calculating the learning rate for each item. The learning rate is the gain in activation at the time of testing, divided by the time cost incurred when studying the item now. This value is computed separately for drill trials and study trials, and the most effective of the two is chosen. The model was compared against a control condition (the Flashcard method) and a model based on an earlier attempt to optimize practice schedules by Atkinson (1972). The new model outperformed both the control and the Atkinson model.

Of particular interest to this thesis is a model by Pavlik et al. (2008), which is a simplified version of the model by Pavlik and Anderson (2008). Although the stimuli used with the experiments are word pairs again, the model is presented as a more general approach for learning facts of any kind. Perhaps the largest difference between this simpler version and its more complex counterpart is the lack of study trials beyond the initial presentation. The simpler model only presents drill trials, and therefore only computes the learning rate (or efficiency, as it is called in this case) for drills. The pilot experiment, which will be discussed later, uses this model as one of the study conditions.

The MCM (Multiscale Context Model) by Mozer et al. (2009) is a combination of SAM (Raaijmakers, 2003) and the MTS (Multiple Time-scale) model by Staddon et al. (2002), which was designed to explain rate-sensitive habituation in animals. MTS uses a cascade of leaky integrators[1] for each item. The integrators are ordered from short to long according to their individual time constants. Traces for items that are repeatedly presented with short ISI can be represented by integrators with short time constants, which will lead to faster decay of that trace. The MCM model takes the contextual fluctuation from the SAM model and combines it with the multiscale representation of the MTS model. The model is able

[1] Leaky integrators are differential equations that can describe components which slowly 'leak' input over time

to fit optimal ISI/RI combination data from earlier experiments (Cepeda et al., 2008) quite accurately, and can make predictions for an arbitrary RI. However, this model cannot adapt to individual differences, and its computational complexity makes it less suitable for realtime purposes.

Finally, there is the model of Koelewijn (2010), which builds upon the work of Van Thiel (2010) and Van Woudenberg (2008). While technically without a name, the model is called '*Improve Retention by Spacing*' in this work. For brevity it will be referred to as the IRS model. In essence, this model adjusts the decay of the memory strength of an item to make the predicted reaction time (calculated from the memory strength) for that item fit the actually observed RT (reaction time) as closely as possible, for each trial. The RT is defined as the time from the item presentation until the first keypress. The adjustment is done using the difference between obtained and predicted RT values. Individual differences between participants are taken into account by determining the standard reaction time cost per participant, through a quick reaction test before the study session. The IRS model was used for learning Dutch-French word pairs in a 15 minute study session, followed by a test the next day. Test scores from the adaptive spacing condition were significantly higher than the control condition (Van Thiel, 2010).

1.6 Other Learning Strategies

To properly judge the value of spacing in a practical setting, it must be compared against a control condition: a different learning strategy. One such strategy is the Flashcard method, which is widely used as a study tool. In the Flashcard method, the items to be studied are ordered in small batches, usually around 5 items each. Each batch is practiced in turn. One cycles through each item of the batch, until all items have been recalled correctly at least once. When this state has been reached, the next batch will be practiced. When all batches have been successfully practiced, the process starts anew from the first batch. This method resembles learning factual information with decks of flashcards: small cards with the question on one side, and the answer on the other. This fairly simple method is easy to use with limited resources, and even provides some spacing as practice rounds of the same batch are separated by the entire set of batches. This method also makes use of the testing effect: more difficult items will be presented more often, which keeps them from being forgotten. Unfortunately, this inherent spacing makes it more difficult to find a meaningful difference between the flashcard method and spacing schedules, which will be discussed later. Another popular learning method is to drop an item from further practice once it has been learned. However, Karpicke and Roediger (2008) showed that this method results in significantly less recall compared to conditions where no items were dropped. In a sense this method uses someone's own estimation of how well they have learned. Due to the large number of facts to learn, and short study sessions, the

flashcard method used in the two experiments combines the Flashcard method with the option of dropping groups of items from study. In later sessions, this solves the problem of having to go through all the old items in order to start learning new ones.

1.7 Spacing and Testing in Practice

Given that spacing and testing increase memory retention, it would seem to be desirable to use them in a practical application. However, despite the proven advantage of both, there has not been a widespread adoption of strategies based on the effects in educational systems. Although there is relatively little research into the use of spacing in the classroom, there is evidence that the effect is not applied much: study material is rarely handled more than once, even though this is essential to spacing. This also means that the material is never practiced through, for instance, tests with feedback. Typically, the only type of test performed is one to assess knowledge retention (Dempster, 1989). Therefore, it seems there are still many opportunities for insights gained from study of spacing to improve learning practices. Dempster (1988) examines possible reasons why application of the spacing effect in this context has failed. He presents nine plausible reasons, of which the four most likely are: discontinuities and confusion within the literature on the spacing effect, failures in obtaining the spacing effect in school-like tasks, an absence of classroom demonstrations of the effect, and a lack of understanding of the spacing effect itself. Of these four reasons the absence of classroom demonstrations is considered to be the most serious problem. While Dempster focused on between-session spacing, his points are also valid for within-session spacing. In fact, within-session spacing would be a better fit for the current educational system. Currently, as a consequence of course scheduling and weekly workloads, students typically start learning very shortly before a quiz. This means there is limited time for study, and no room for between-session spacing. Within-session spacing, however, is not affected by such restraints. Furthermore, because within-session spacing fits so well within the current educational system it could be used as a first step in the gradual introduction of the spacing effect. This might lead to more fundamental changes within educational practices which allow for the use of between-session spacing as well. As such, the type of scheduling methods presented in this work would be a logical choice for this first step.

1.8 Thesis Overview

As mentioned, the research on spacing and testing in realistic educational conditions is very thin. The goal of this thesis is to gain more insight into the practical benefits of within-session spacing; the type of spacing that is best suited for

the current educational environment. This environment advocates regular, fixed length study sessions (as dictated by the scheduling of classes each week), using a related and organized body of study work. Furthermore, learning material is divided into modules that are only handled once. Learning the same material over several years, as advocated with between-session spacing, would not fit within this structure. In much of the research a central question has been '*what is the effect of item scheduling on learning?*'. However, for spacing research with practical uses it should be '*what is the effect of item scheduling on learning, given practical, realistic constraints?*': a question which is explored in this thesis.

This is realized by testing the IRS model under real world circumstances: a full and existing body of coursework, studied over several sessions. The main focus of this thesis is a longitudinal study over multiple days to investigate the effect of the within-session spacing schedule predicted by the IRS model, and to ascertain whether the spacing effect occurs in this practical, realistic setting.

Most research into spacing has been done using one or two study sessions (separated by the ISI), followed by the final test (taken by the participant after the RI has passed). This does not deliver an account of spacing in regular study sessions over a longer period. By scheduling a greater number of study sessions, insight into spacing on a more practical (or educationally relevant) level can be gained.

The remainder of this thesis is structured in the following way:

- Chapter 2 provides a detailed explanation of the IRS spacing model used in the experiments. It introduces the most important concepts that are required to understanding the model and how it is used. The chapter also contains the changes made to the model.
- Chapter 3 is an account of the pilot study and its results. The study consists of a shorter version of the final experiment. It also includes additional information about the model proposed by Pavlik et al. (2008), which was used as an extra condition during the experiment.
- Chapter 4 presents the final experiment, methods, results and analysis of the results. The experiment is in the form of a longitudinal study comparing spacing against a flashcard control condition.
- Chapter 5 contains a general discussion of all the results, and their implications. It also touches upon research directions for future projects.

CHAPTER 2

REVISITING THE ACT-R MODEL FOR SPACING

This chapter describes the IRS model, which is used to compute an optimized item order during a study session. The model uses key parts of ACT-R, which are used to account for memory effects. The origins and evolution of the model will be explained, followed by the changes made to the model as it is used in the experiments in this thesis.

2.1 The Model Explained

2.1.1 Origins: Spacing in ACT-R

The origins of the model lie in the ACT-R architecture (Anderson et al., 2004). In short, ACT-R describes a part of human cognition; the acquisition and production of declarative and procedural knowledge. Internally, knowledge is represented by chunks. Each chunk is an individual memory item, such as a fact or a word. ACT-R has a variety of components (called modules), which describe different functions of the brain, such as the perception of the outside world, or the storage of facts. The IRS model uses only a small part of ACT-R: a variation of the equations from the declarative memory module.

To retrieve items, a measure of how strongly an item is present in memory is used: more strongly present items are more likely to be recalled. This is represented by the activation level of an item, which is used to determine the likelihood and speed of recall. A high activation means fast and correct recall, while low activation means slow recall, or none at all. Activation can be calculated using the *base-level activation* equation, as shown in Equation 2.1.

$$m_i(t) = ln\left(\sum_{j=1}^{n;t_j<t}(t-t_j)^{-d_{i,j}}\right) \tag{2.1}$$

The function returns the activation m_i for an item i at time t, and is the summation of the individual practice events for that particular item. The time at which a practice event occurred is represented by t_j. In earlier versions of the equation, the rate at which each practice event is forgotten was given by the implied decay

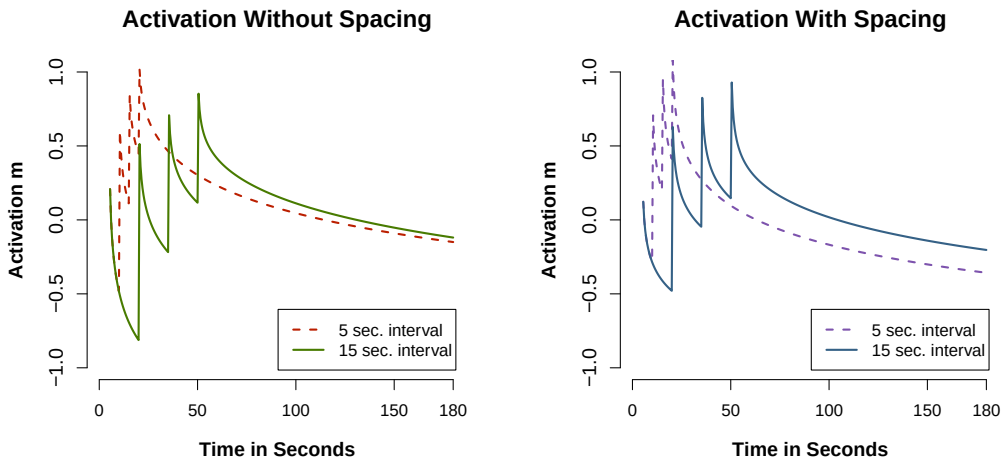


Figure 2.1: The left plot shows the activation of two times four presentations using the default ACT-R equation. It is clear that long term retention is almost equal, regardless of the spacing between presentations. The right-hand plot uses the activation equation that accounts for spacing. Here, there is a clear effect of the spacing of presentations.

constant d, which was set at 0.5. A higher decay means quicker forgetting. However, as the decay is the same for each memory trace, the effect that the spread of presentations has on the decay of memory strength cannot be modelled with this approach. A modification was required: individual decay values $d_{i,j}$ for each encounter (Pavlik and Anderson, 2005). These individual decay values are calculated with Equation 2.2.

$$d_{i,j} = ce^{m_i(t_j)} + \alpha \tag{2.2}$$

Here, c is a scaling parameter that determines the strength of the spacing effect and α is a constant which represents the minimal activation. Furthermore, $m_i(t_j)$ is the item activation at the previous encounter. This dependence on (the time since) the last encounter of the item is what accounts for the spacing effect: if the item was recently encountered, the activation will still be high which results in a larger decay value. This causes the activation of the new encounter to decay more rapidly. Looking at the activation equation as a whole, this means that many rehearsals over a short timespan result in high decay, which leads to faster forgetting. Conversely, more spaced encounters lead to lower decay, which promotes long-term retention. Both of these effects can be seen in the right-hand plot of Figure 2.1, where activation is plotted for two sets of presentations, using the activation equation that accounts for spacing. The left-hand plot shows that the default ACT-R activation equation shows no spacing effect.

2.1.2 Current State of the Model

With the ACT-R equations described earlier it is feasible to create personalized learning schedules. The overall concept of the model is to present an item as late as possible; just before it falls below the threshold of forgetting. This retrieval threshold τ is a certain activation value, usually between -0.5 and -0.8 (Van Woudenberg, 2008; Van Thiel, 2010). This gives the maximum spacing, while retaining the benefit of the testing effect (Carrier and Pashler, 1992). Given the equations discussed so far, combined with the threshold value, a model can be constructed that tracks the change in activation for each participant/item combination.

By creating a set of selection rules, an item scheduling algorithm can now be devised. First, determine which item, of the items that have been presented at least once, has the lowest activation 15 seconds from now. If the item is below the threshold τ at that time, it will be selected for the next trial. A 15 second look-ahead time is used to try to select items before they are forgotten: if an item is below the threshold at the current time, it is unlikely that it can be recalled. If there are no items with an activation value below the threshold, a new item is selected which has had no presentations yet. If no such item exists, the item that has the lowest activation is selected (as it is closest to being forgotten). These rules are used in all subsequent models covered in this section.

2.1.2.1 Van Woudenberg: Dynamic Spacing

Van Woudenberg (2008) implemented this model, which he called *Dynamic Spacing*. However, this model did not increase performance significantly over the flashcard control condition. The greatest shortcoming of the *Dynamic Spacing* model is that it does not account for the difference between participants and item difficulty. Van Woudenberg created two variations of the model: *Dynamic Spacing - Response* and *Dynamic Spacing - Reaction Time* to account for this. Only the *Reaction Time* variation will be discussed, as it is the only one that performed significantly better than the control condition. To realize this model, a way to estimate the expected reaction time (that is, the time from the presentation until the first key press) is required:

$$RT = Fe^{-m} + f \tag{2.3}$$

In Equation 2.3, the RT depends on the activation m, and a scaling parameter F. Additionally, there is a fixed time cost f. This fixed cost is an estimate of the time it takes to process the visual information of the stimuli, and the motor control required to press a key. Typically, this value is estimated around 300ms. This formula (apart from the f variable) is also a part of ACT-R.

The *Reaction Time* model is able to adapt itself to the participant by having the value of the α parameter, or decay intercept, in Equation 2.2 depend on the recall

speed and correctness of a trial, giving each item its own personal α value. The α is adjusted in the following way: when the observed reaction time is smaller than the expected reaction time (calculated using Equation 2.3), the model activation value must have been too low, which means the decay for that item is too high. Therefore, the decay parameter is adjusted by changing the α. If the observed time is less than the expected time, and the difference between observed and predicted RTs is larger than 0.5s, the α is decreased by this difference divided by 1000. The opposite also holds: if the observed time is larger than the expected time, the α is increased. The reaction time difference is limited to a maximum of 10 seconds, and an incorrect answer is treated as a 10 second difference, leading to an intercept change of 0.01:

$$\Delta\alpha = max\left(0.01, \frac{RT_{observed} - RT_{expected}}{1000}\right) \tag{2.4}$$

When using this model to schedule item trials, participants scored significantly higher on a Dutch-French word pair test, compared to the flashcard condition (Van Woudenberg, 2008; Van Rijn et al., 2009). The main criticism of this model is that the maximum step size of 0.01 causes the adaptation to converge very slowly. In short learning sessions, convergence might not occur at all. Furthermore, the change is based solely on the last trial, and as such does not use the information of previous trials to calculate a more accurate estimate. Van Thiel (2010) addresses these issues in a revision of the model.

2.1.2.2 Van Thiel: Refined Alpha Estimation

In the method of Van Thiel, the α is calculated in a different manner: there is no upper-bound to the change in alpha, so convergence should happen faster than in the previous model. The first step is to find the α that best fits the decay of the last encounter. This search uses the difference between observed and predicted activation: calculating the observed activation can be done by rewriting Equation 2.3 so that it converts RT into activation, and using observed RTs as input:

$$m_{obs}(t_j) = -ln\left(\frac{RT - f}{F}\right) \tag{2.5}$$

The observed activation should be equal to the activation calculated by the model (Equation 2.6). To see how much activation was added by the last encounter, it is split out from the summation as shown in Equation 2.7.

$$m_{obs}(t_j) = ln\left(\sum_{j=1}^{n}(t - t_j)^{-d_{i,j}}\right) \tag{2.6}$$

$$m_{obs}(t_j) = ln\left(\left(\sum_{j=1}^{n-1}(t-t_j)^{-d_{i,j}}\right) + (t-t_j)^{-d_{i,j=n}}\right) \tag{2.7}$$

The next step is to remove the logarithm by taking the exponent on both sides, and rearranging the terms to separate the last encounter from the rest:

$$(t-t_j)^{-d_{i,j=n}} = e^{m_{obs}(t_j)} - \left(\sum_{j=1}^{n-1}(t-t_j)^{-d_{i,j}}\right) \tag{2.8}$$

Finally, the decay value d can be calculated by taking the $(t-t_j)$th logarithm (and negative) of both sides of Equation 2.8:

$$d_{i,j=n} = -^{(t-t_j)}log\left(e^{m_{obs}(t_j)} - \left(\sum_{j=1}^{n-1}(t-t_j)^{-d_{i,j}}\right)\right) \tag{2.9}$$

By rewriting the decay equation (Equation 2.2) and using the activation of the previous encounter the α that corresponds to that encounter can be calculated:

$$\alpha_{i,j=n} = d_{i,j=n} - ce^{m_i(t_j)} \tag{2.10}$$

Now that the α_{new} for the previous encounter is known, and assuming there is some α_{fit} which best fits all previous encounters, a new α can be computed using these two values. This is done using a simple binary search between α_{fit} and α_{new}. First, compute the error between observed and calculated reaction times, for both α_{fit} and α_{new}. Next, take the value $\alpha_c = (\alpha_{new} + \alpha_{fit})/2$, and again compute the error. If α_{new} is a better fit than α_{fit}, repeated the previous steps with α_{new} and α_c as boundary values. If α_{fit} is a better fit, use α_{fit} and α_c instead. After each repeat of this procedure the search interval is effectively cut in half, creating a convergence towards the optimal α. This process is repeated six times, which is considered an acceptable compromise between speed and precision.

One drawback to this approach is that all previous encounters are considered for the α fit. This creates significant inertia: if a participant suddenly switches from being consistently bad to consistently good at recalling a certain item, the bad results will still play a large role in determining a new α, even though these results are no longer relevant. Consequently, the α value will be overestimated. This issue will be addressed in the final section of this chapter.

There is another notable differences with the original model. Instead of a fixed 10 second reaction time cut-off, a *maximum reaction time* is calculated:

$$RT_{max} = 1.5(Fe^{-\tau} + f) \tag{2.11}$$

and

$$RT_{i,j} = min(RT_{max}, RT_{i,j}) \tag{2.12}$$

The reasoning behind this choice is that large reaction times can contain a great deal of noise: small changes in participant behaviour, such as a distraction, can lead to large increases in reaction time. Large RTs lead to low activation, which in this case would lead to a strong and unjustified effect on the α adaptation, as processes other than memory retrieval are likely involved.

Van Thiel performed an experiment using French-Dutch word pairs to test the model against a flashcard control condition. Participants performed significantly better in the spacing condition, using the improved model. However, it was unclear whether this was due to the item sequencing, or because participants in the model condition had a larger number of trials on average. Van Thiel mentions several further improvements to the model, some of which Koelewijn (2010) implemented.

2.1.2.3 Koelewijn: Personalized Fixed Cost

So far, the fixed cost in Equation 2.3 has remained the same for each participant. Of course, reaction time and motor function differs from person to person. Koelewijn (2010) used a small reaction time test before the study session to estimate this personal fixed cost. During this reaction test participants were presented with two identical Dutch words. One of the words would disappear after 3 seconds, at which time the participant would have to retype the word. After 10 trials, an average of the time between the disappearance of the word and the first key press was calculated to serve as fixed cost value.

As recommended by Van Thiel (2010), the first test trial is no longer used when calculating the α fit. This test trial immediately follows the initial presentation of the item. Considering the short timespan between presentation and test, the item should still be present in the working memory of the participant, and as such no (declarative) memory retrieval is required to access the information (Baddeley, 2003; Cowan, 2000). Therefore, this trial contains no useful information for the model.

As a consequence of adding a personal fixed time, calculation of the *maximum reaction time* (Equation 2.11) was changed:

$$RT_{max} = Fe^{-1.5\tau} + f \tag{2.13}$$

In the original model, the personal fixed time would have been multiplied by 1.5 as well, which would lead to an overestimation of the difference between people (as this fixed time is separate from the activation based latency). Therefore, in Equation 2.13 only the retrieval threshold is multiplied.

Koelewijn (2010) conducted two experiments using Dutch-English word pairs: the first to test the addition of the personal fixed time, and the second to test the effect of a personalized initial α, obtained from the data of the first experiment.

Both experiments tested the model against a flashcard control condition. No significant performance difference was found in either experiment. Koelewijn still argues for the use of both a personal fixed cost and a personal initial α, as they did not have a negative impact on the final test scores.

2.2 New Improvements

The first major change that was made to the model lies in the method used to estimate the new α. Using the estimation of the previous decay to limit the search interval adds a great deal of complexity. However, the impact of careful selection of the initial values has only a minor impact on the result of a binary search. Instead, a combination of the approaches presented by Van Woudenberg (2008) and Van Thiel (2010) was used: if the observed RT was faster than the expected RT, α_{new} is set to $\alpha_{fit} - 0.05$. Vice versa, if the observed RT was slower, $\alpha_{fit} + 0.05$ is used instead. Next, a binary search between α_{fit} and α_{new} is performed to find the best fit.

This leads to the second change, which addresses the concern raised earlier: currently all previous trials are used when calculating the fit of an α. This means unsuccessful trials still have an impact on the α calculation, even if the participant has since had a streak of successful trials (for a particular item). For example, the participant may have suddenly realized a mnemonic device[1] that results in perfect recall. A solution would be to take a sliding window approach: only use the last n trials for the fit calculation. The question then becomes how many encounters should be used. RT values are very noisy, so using too few is likely worse than using too many, as the noise would not be sufficiently smoothed out. A window size of 4 is used, as this is the minimal number of encounters needed to get a reasonable estimate of α, according to Van Thiel (2010).

2.2.1 Estimating Reading Times

For the experiments in which the model has been used so far, reading time has not been an issue: all items have been in the form of a one word cue. The time to read this word can be accounted for in the (personalized) fixed time quite well, and will not fluctuate much between items. However, the items used during the final experiment are full sentences. One consequence is that the personalized fixed time used by Koelewijn (2010) cannot be used: there is too much difference between item lengths. Instead, a function of sentence length is used to determine the average reading time f_s:

$$f_s = (1180 + 44.3C^{count})s^r \tag{2.14}$$

[1] A learning technique which aids memory.

Sentence Metrics	
Character count	**0.62**
Average number of syllables	0.09
Textual Difficulty Measures	
Flesch-Kincaid reading ease (Flesch, 1948)	-0.20
Gunning Fog index (Gunning, 1969)	0.32
SMOG index (McLaughlin, 1969)	0.42
Coleman-Liau index (Coleman and Liau, 1975)	0.07
Automated readability index (Smith and Kincaid, 1970)	0.41
Empirical Measures	
Summed log of word-frequency	0.56
Spache method (Spache, 1953)	0.59

Table 2.1: Correlation between predictor variables and observed RTs.

Where C^{count} is the number of letters (including spaces) in the sentence, and s^r is a scaling factor (which was set at 0.3). The solution, of which the origin will be explained shortly, is quite ad-hoc, and models which can estimate reading time have been proposed, such as *E-Z Reader* (Reichle et al., 2006). However, the *E-Z Reader* model is far too complex for the needs of the spacing model, as it was designed to account for more than reading time alone.

Equation 2.14 was based on the RT data from a small experiment where participants were asked to perform a *comprehensive reading task*: The objective was to rate the answers of definitions on familiarity. Participants were presented with one definition at a time, which they had to read carefully. When done, they had to press the space bar or click the left mouse button. This caused a Likert[2] scale ranging from one to seven to appear, with one being 'completely unfamiliar' and seven being 'known by heart'. After picking a number on the scale, the next definition was presented. The stimuli were sampled from the set of definitions used for both experiments. Therefore, the observed RTs are representative for the whole set. 50 sentences of varying length were picked, of which a random selection of 25 was shown to a particular participant. Data was gathered for a total of 18 participants: 12 male, 6 female.

Several predictors were considered to estimate the reading times of the definitions. The correlation between each predictor and the RTs is shown in Table 2.1. Interestingly, the most simple predictor is also the most effective: the character count has a 0.62 correlation with the reading time. Two other predictors are

[2]The Likert scale is a numerical rating commonly used in survey research. Participants specify in how far they agree with a particular statement, where the ends of the scale are usually opposites of each other (e.g. 'Agree' and 'Disagree').

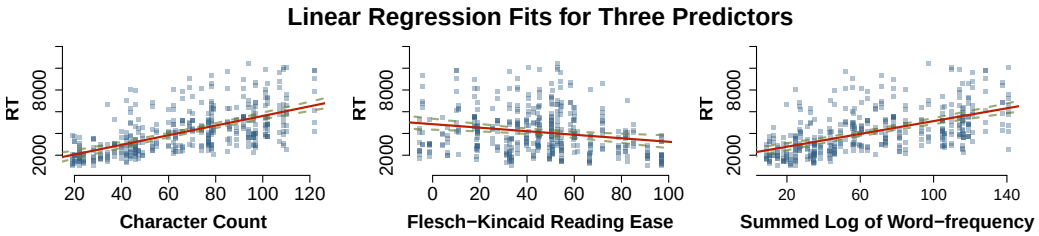


Figure 2.2: Linear regression fits for three predictors. The blue circles are the observed RTs. The solid orange line is the linear regression fit. Observations of all the participants were combined to calculate these fits.

considered in detail: the *Flesch-Kincaid reading ease* and the *summed log word-frequency*. These three predictors represent three very different approaches. The *character count* is a simple sentence metric. The *Flesch-Kincaid* scale represents the group of measures for textual difficulty, which use a combination of word count, sentence length and syllables per word to calculate the reading difficulty (or 'readability'). While the *Flesch-Kincaid reading ease* does not give the best fit for the data out of the methods in this group, it is a method that has been widely used and validated (D'Alessandro et al., 2001). The *summed log of word-frequency* represents the set of empirical predictors: the word frequency table of the CELEX English lexical database was used as an estimate for word difficulty: less familiar words (words with a lower frequency) should be more difficult to process. The log of each word frequency in a sentence was taken and summed, which also adds the factor of sentence length to the predictor. The log for each frequency was taken to incorporate a kind of soft ceiling effect for familiarity. Otherwise, often used words words such as 'a' and 'for' might skew the results.

Figure 2.2 shows the observation distributions and the resulting linear regressions. The results for the *character count* and *summed log word-frequency* look similar, with higher RTs as the value of the predictors increase. The *Flesch-Kincaid reading ease* shows a negative correlation. This is expected as a higher value means the sentence is easier to read, which should lead to a lower RT.

Merging the observations of all the participants into one set might obscure significant effects and patterns present in the individual data sets. Therefore, the fits for each participant were examined. The resulting plots for five participants can be seen in Figure 2.3. The graphs show that while the average correlation might not be high, the average fits of all three predictors tend to be near the individual fit. Notable exceptions occur in the *Flesch-Kincaid* fits, where, against expectation, the individual slope is sometimes positive. Given the simplicity and good correlation of the *character count* predictor, it was used to calculate the average reading time as previously shown in Equation 2.14. Early tests showed that reading speed increased significantly after the first presentation. Recognition of the questions likely plays a large role in this. Therefore, the scaling factor s^r was introduced to approximate this change.

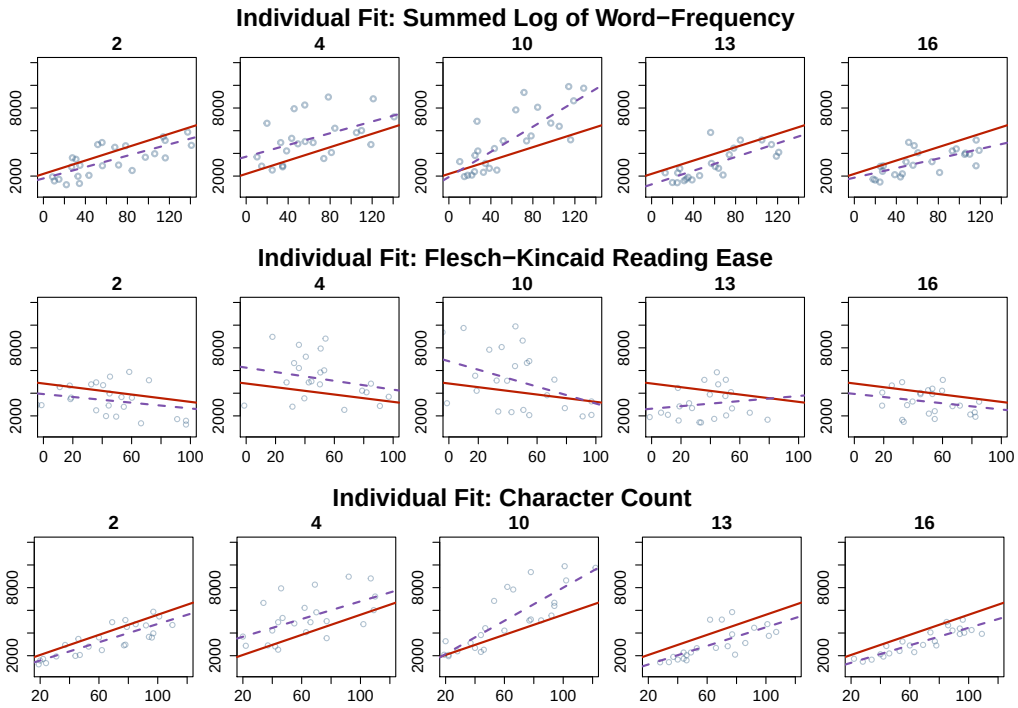


Figure 2.3: Individual fits for five participants and three different predictors. The blue circles are the observed RTs. The dashed purple lines are individual fits, and the solid orange lines are the linear regression fits for the corresponding predictors as shown in Figure 2.2.

Overall, the textual difficulty measures have a much lower correlation than both the simple and empirical measures. Possible explanations are that the textual difficulty measures are optimized for texts instead of individual sentences, that the reading ability of the participants exceeds the difficulty of the sentences (therefore reducing the effect of difficulty on reading speed), or that the reading speed is not influenced significantly by the readability of a sentence.

2.2.2 Adapting the Model for Multiple Sessions

It has been shown that the rate of forgetting is different between study sessions (Anderson et al., 1999). There is less destructive interference during this time than during a study session (Pavlik and Anderson, 2005; Pavlik et al., 2008). This is an important difference between the two types of spacing: within-session spacing has no concept of sessions, and does not consider the difference in forgetting rates. However, because each participant of the final experiment will have several sessions, a way to deal with this difference in the rate of forgetting is required. A common solution is to scale the time between sessions by some constant. This constant is called the *psychological time*. To use this, the total amount of time spent outside practice sessions is multiplied by the *psychological time* constant

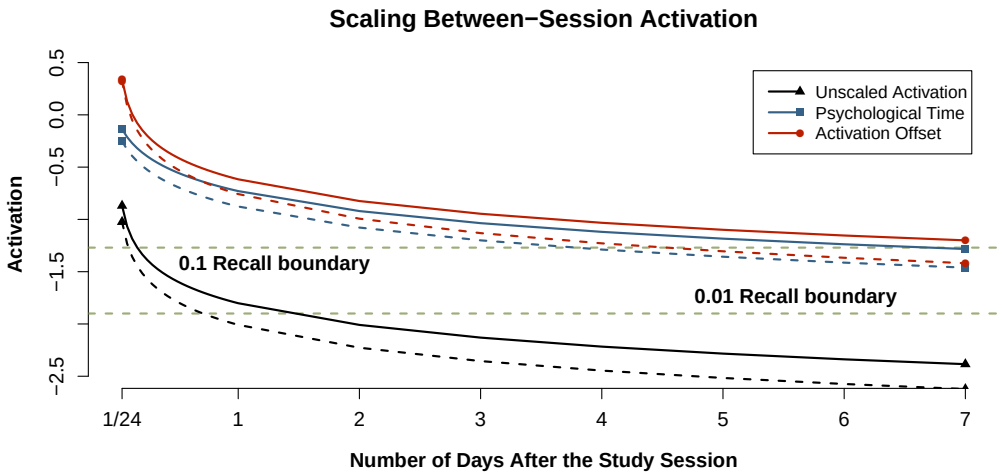


Figure 2.4: Different types of between session activation scaling. The change in activation for two items with $\alpha = 0.3$ (solid lines) and $\alpha = 0.34$ (dashed lines) were computed using three methods. The 'unscaled activation' method simply uses the regular activation formula (Equation 2.1), while 'psychological time' and 'activation offset' uses the methods explained in Section 2.2.2. The recall boundaries were calculated using the recall probability equation which will be introduced in Chapter 3 (Equation 3.4). The psychological time was computed using a δ of 0.025.

δ, and this value is added to the previous session duration:

$$t_{current} = t_{eos} + \delta t_{out} \tag{2.15}$$

Here $t_{current}$ is the new total study duration, t_{eos} is the time at the end of the previous session, t_{out} is the total time spent outside of the study session, and δ is the between session scaling factor. While scaling the time between sessions seems like an inelegant way to correct for a shortcoming in the memory and spacing equations, it has been used successfully to fit experimental data (Anderson et al., 1999; Pavlik and Anderson, 2003). Unfortunately, the value for δ seems to vary considerably between experiments (ranging from 0.00046 to 0.025). This means that the value must be estimated for a particular setup, which reduces its practical use. It seems that typically the amount of forgetting between sessions can only be roughly approximated. In earlier research, Anderson et al. (1999) present a more comprehensive method to account for intersession forgetting. However, this method assumes a fixed decay value d, and therefore cannot be used with the spacing model.

The approach used in this thesis is different. The *psychological time* solution requires that all previous encounter times must be stored in order to compute the activation at the start of a new session. This might be impractical or inefficient (for instance in web-based applications, where bandwidth may be limited), so a

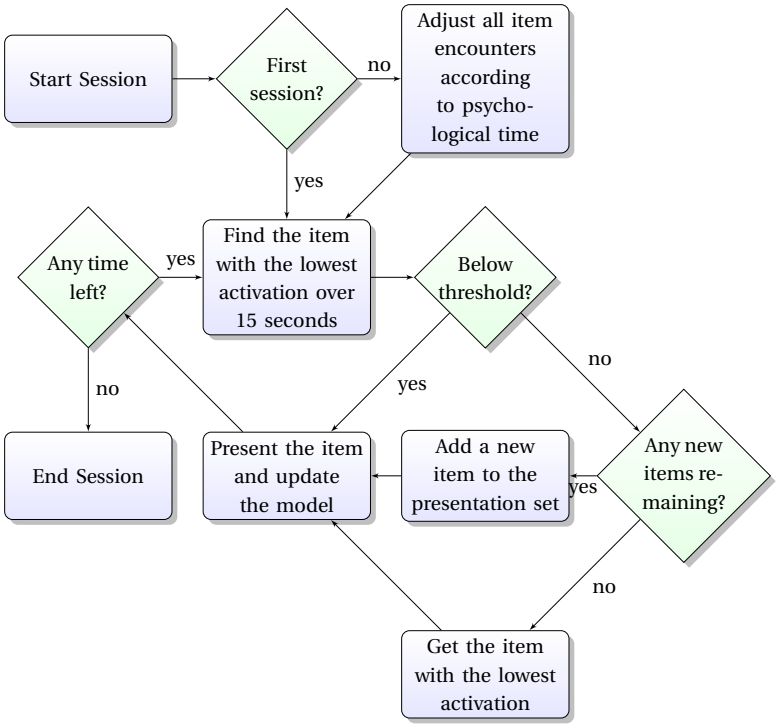


Figure 2.5: Flow chart of the model.

simpler solution was developed. For each item, the estimated activation at the end of the session is stored. At the start of a new session, this value is used to calculate a new activation offset value m^{os} for item i:

$$m_i^{os} = m_i^f - \alpha_i^{last} log(t_i^{now}) \tag{2.16}$$

In Equation 2.16 m_i^f is the 'future' activation, α_i^{last} is the last calculated α, and t_i^{now} is the time in days since the last presentation. Forgetting is scaled by α to capture the difference in item difficulty, which should still be a factor between sessions. For very short intervals the 'Activation Offset' becomes less accurate. Therefore the upper bound of the activation is set to the item activation at the end of the session. The exact function of between-session forgetting is unknown: one that contains a forgetting curve which is scaled by the rate of forgetting is sufficient for its purpose in this thesis. In Figure 2.4 the resulting between-session forgetting curves are very similar to those obtained by using the *psychological time* method. Because the between-session forgetting scaling is a rough guess at best, there should be little practical difference between both approaches. In a new session, the m_i^{os} value is used in the activation equation (Equation 2.1) as a summation of all the previous encounters:

$$m_i(t) = ln\left(e^{m_i^{os}} + \sum_{j=1}^{n;t_j<t}(t-t_j)^{-d_{i,j}}\right) \tag{2.17}$$

2.2.3 Model Overview

A flow chart of the model in its final form can be seen in Figure 2.5. It illustrates all the steps that are taken during a study session, focusing on the item selection logic. The model as it has been described in this chapter (including improvements) is what will be used for the spacing condition in both experiments.

CHAPTER 3

PILOT STUDY

This chapter describes the first of the two experiments: the pilot study which consists of a single session test of the different presentation scheduling methods. Two of the methods have already been discussed in detail (Flashcard and IRS). Before moving on to the study itself, the third method that was included will be described, which was created by Pavlik et al. (2008).

3.1 Another Model: Scheduling for Efficiency

The IRS model is very similar to the model for optimized scheduling of practice introduced by Pavlik and Anderson (2008). Compared to the IRS approach, their method can be seen as using a forgetting threshold with a non-fixed lookahead time for determining the activation. The model has a fair degree of complexity, including three personalized parameters which have to be estimated beforehand: item difference β_i, participant difference β_s, and participant-item difference β_{si}. Pavlik et al. (2008) created a simpler variation of the model, meant for basic fact learning. It is that version of the model which was implemented as an extra condition for the pilot experiment. The motivation behind adding this extra condition is to see whether the conceptual differences between both models lead to significantly different performance.

3.1.1 Practice Efficiency as Selection Criteria

The model of Pavlik et al., which will be referred to as the TBF (Train Basic Facts) model, uses a variation of the ACT-R activation equation:

$$m_n(t_{1..n}) = \beta_s + ln\left(\sum_{i=1}^{n} b_i(t-t_i)^{-d_i}\right) \tag{3.1}$$

There are a few new parameters compared to the activation equations that have been shown so far: β_s accounts for the difference between participants. This value is initialized at 0, and is estimated every 50 trials to improve the fit of the model to the observed results. This is done by incrementally adjusting the value, as will be shown later in this chapter. The b_i values represent the effect of correctness during practice. According to Pavlik et al., b_i is high for a successful recall, and low for a failure. No specific values are mentioned. For the experiment 0.3

was chosen as b_{low} and 1.0 as b_{high}, based on information obtained from Pavlik et al.. In essence this is how the model adjusts the expected activation per item: more failures equals a more difficult item, which will have less activation because the encounter is penalized through b_{low}. The decay values d_i are calculated as shown in Equation 2.2.

In order to determine which item to practice next, the TBF model predicts the time when practicing said item is most efficient. Here, efficiency is the *long-term learning gain* (where long-term is defined as the RI until the test) divided by the *expected time cost*. The *long-term learning gain* (which is the predicted gain in activation at the moment of the test, compared to the current time) is made up of two components: the expected increase in activation for a successful trial plus the expected activation increase for an unsuccessful trial. Similarly, the *expected time cost* (the predicted time length of the trial) consists of the time cost for a successful trial (which is dependent on latency) plus the time cost of an unsuccessful trial:

$$eff_m = \frac{p_m b_{suc} r^{-d_m} + (1-p_m) b_{fail} r^{-d_m}}{p_m(l(m) + fsc) + (1-p_m) ffc} \tag{3.2}$$

In Equation 3.2, r is the retention interval in seconds. Both b_{suc} and b_{fail} are scaling parameters. The value of b_{fail} is always 0 after the first trial: failures are given no credit beyond the first practice. This is consistent with the implementation of Pavlik et al. (2008). The latency is scaled by F:

$$l(m) = Fe^{-m} \tag{3.3}$$

F is fit every 40 trials using an incremental adjustment, which will be explained later in this chapter. The parameters *fsc* and *ffc* are the '*fixed success cost*' and '*fixed failure cost*', respectively. The value of d_m is calculated with the decay formula (Equation 2.2) from the current activation of the item. Finally, p_m is calculated using the recall probability equation:

$$p(m) = \frac{1}{1 + e^{\frac{\tau - m}{s}}} \tag{3.4}$$

Here, m is the activation, τ the threshold value, and s explains the noise in activation: the transition from no recall to perfect recall becomes more fuzzy as s increases. The values for all the parameters (as used in the Pavlik et al. (2008) paper) can be seen in Table 3.1.

Plotting efficiency against activation using Equation 3.2 shows that there is an optimal point; a level of activation that produces the maximal efficiency. In Figure 3.1, this point is around -0.04 activation. Any activation level lower than this point results in more failed trials, which are time consuming. A higher activation means that the spacing effect is not used optimally, as trials become too closely grouped together.

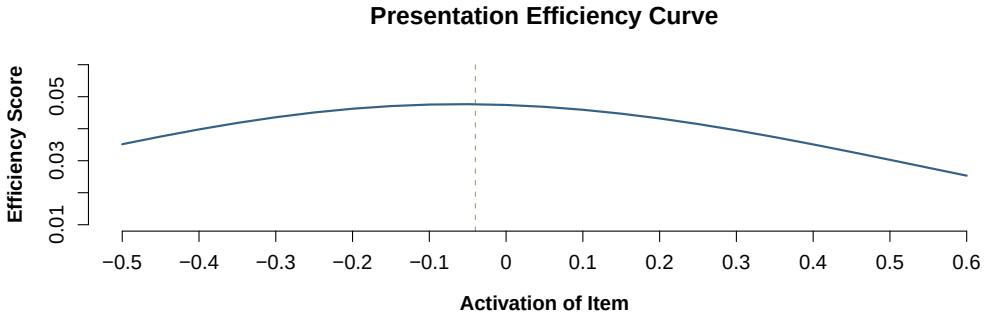


Figure 3.1: Activation plotted against efficiency.The value of F was set at 2.322 to create this graph.

Parameter	Value	Parameter	Value
a	0.17	*ffc*	9s
b_{suc}	2.497	*fsc*	0.63s
b_{fail}	1.526	τ	-0.7
c	0.21	r	1191
s	0.261		

Table 3.1: Parameter values used during the experiment.

3.1.2 Dynamic Adaptation

In order to customize the TBF model for individual participants, two parameters are re-estimated after a certain number of trials: b_i and F. The value for β_s is adjusted incrementally: a fit is computed for $\beta_+ = \beta_s + \beta_\Delta$, as well as $\beta_- = \beta_s - \beta_\Delta$, and the current β_s. The step-size $\beta_\Delta = 0.05$ was estimated from Pavlik (2005). There are two ways in which a fit could be calculated: the first option is to calculate the recall probabilities for the last 50 trials (which is the interval used in Pavlik et al. (2008)), using the activation values as predicted by the TBF model. Summing these probabilities would give an estimate of how many times a correct answer was given. Comparing this value against the observed number of correct recalls produces a measure of error for the new estimates. This is approximately what Pavlik et al. (2008) did. The second option is to compare the predicted activation against the observed activation and select the estimate with the smallest error. This procedure would be similar to the one used in the IRS model, using Equation 2.5 to retrieve observed activation from latency (with $f = 0$). Both approaches have caveats: using probabilities is not very precise, as the probability equation includes noise. The second approach uses F to compute the fit, which itself is re-estimated periodically. This should not be a large issue however, as F should converge to an optimal value, and small fluctuations in F should not effect the fitting process much. Therefore, the second approach was chosen for

the TBF model implementation in this thesis.

As mentioned, F is fit dynamically as well. This process is very similar to the adaptation of β_s: the expected latency is compared against the observed latency. This is done for $F_+ = F + F_\Delta$, as well as $F_- = F - F_\Delta$ (with $F_\Delta = 0.05$), and the current value for F, using the last 40 trials (again, this is the interval which was used in Pavlik et al. (2008)). The value that produces the smallest error becomes the new F.

3.1.3 Item Scheduling

One way to use Equation 3.2 would be to estimate all parameters a priori, find the activation value that gives maximum efficiency, and compare item activations against this value to determine which one to present. However, because F is estimated every 40 trials, this optimal value can shift. A better approach is to take the first derivative eff'_m of eff_m, and present the item when eff'_m (the rate of change in efficiency) is 0. If eff'_m is positive, the optimal point has not yet been reached, and if it is negative, the point has already been passed. While it is possible to derive eff'_m algebraically, it is more efficient to compute the difference quotient of eff_m:

$$\frac{\mathit{eff}(m+h) - \mathit{eff}(m)}{h} \tag{3.5}$$

with $h = 0.001$, or some other small step-size.

Item scheduling can now be performed by practicing each item at its optimal point. Before each trial, eff'_m is computed for each item. When the change in efficiency approaches 0, that particular item is selected for the upcoming trial. If none of the items have an efficiency change close to 0, a new item is introduced. If all items have already been introduced, the item with the efficiency change closest to 0 is picked.

In one of the experiments performed by Pavlik et al. (2008), the model is compared against a random item selection control condition. Chinese vocabulary was used as study material. Instead of a direct comparison of the final tests between both conditions, time on task was used: because the study time was short, it is unlikely there would be a significant difference in the final test. Participants were free to choose a condition, and were allowed to switch at any time. Results showed a preference for using the described model.

3.1.4 Model Overview

The flow chart of the model as shown in Figure 3.2 is very similar to the one of the IRS model as presented in the previous chapter (Figure 2.5). This shows how

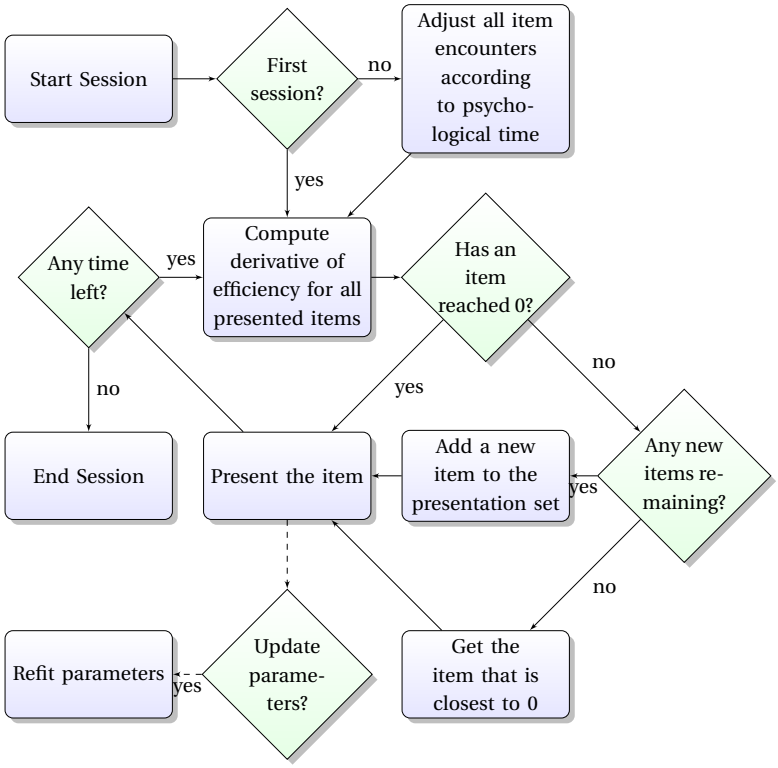


Figure 3.2: Flow chart of the TBF model.

similar the scheduling methods of both models are. The main differences are the item selection criteria and the parameter adaptation. In terms of selection criteria, the IRS model uses a fixed threshold to determine what item should be presented. The TBF model uses a variable threshold which is adapted through the values of F and β_s. The difference could also be explained in terms of the lookahead time: the IRS model always looks 15 seconds ahead, while the TBF model does not look ahead explicitly, but rather uses the efficiency equation to determine how fast an item is approaching the critical point.

An important difference is the method used to adjust the parameters. The adjustments made by the TBF model could be considered *discrete*, as parameters are always changed by a fixed amount. The IRS model uses *continuous* adjustment, as the amount of change depends on observed latency. Finally, there is a distinct difference in which parameters are adapted: the IRS model adjusts the α to change the rate of decay for each item, while the TBF model penalizes the activation gain from incorrect trials and offsets the minimal activation by β_s. Furthermore, F is adjusted to more accurately predict the RT of a participant, which is required for the efficiency computation. The rate of adjustment also differs between models: while the IRS model performs adjustments each trial, the TBF model does so every 40 or 50 trials.

Before presenting the pilot experiment, one issue should be discussed. Even

though three different methods were tested, the results show four conditions. Halfway through the pilot the collected data showed that the TBF model was performing below expectations. The cause was found to be the models absence of a fixed time component in the activation equations. As explained earlier, the stimuli are sentences instead of single words, which has a huge impact on the trial RT. This effect has to be accounted for by the activation equations. Therefore, the TBF model was modified to use the fixed time component. To avoid confusion, a distinction will be made between the '*Original TBF*' (OTBF) model and the '*Alternate TBF*' (ATBF) model.

3.2 Experiment

Considering the scale of the final experiment (four sessions spread over a week), it was considered prudent to first perform a smaller version of the setup. This pilot experiment would be helpful to determine possible problems, and to obtain data that could be used to improve the final experiment. Despite the exploratory nature of the pilot, the setup was used to investigate several issues. Specifically, a comparison was performed between the two spacing models that have been discussed in Chapters 2 and 3. Furthermore, it was investigated whether all available items are presented at least once in the spacing conditions, as earlier research shows that this is not necessarily the case (Van Woudenberg, 2008; Van Thiel, 2010).

3.2.1 Method

3.2.1.1 Participants

All participants were first year Psychology students from the University of Groningen. Curriculum requirements obligate the students to participate in a number of experiments. In total, data for 65 participants was collected (10 Male, 55 Female), with an average age of 21 (SD: 2.2).

3.2.1.2 Materials

As stimulus, 48 definitions from Biopsychology were used. In this instance, Biopsychology refers to anything physically related to the mind, including (but not limited to) cell biology, head trauma, and sensory perception. The items had varying definition lengths and difficulty. The rather large number of facts to be learned was intentional; it removed the possibility of a ceiling effect on the number of items learned, thus keeping the task sufficiently difficult. A complete overview of the stimulus used in the pilot experiment can be found in Appendix A.

3.2.1.3 Design

The setup consists of one study session of 30 minutes and a final test. Between the session and the test participants had to perform a distraction task for 10 minutes, in order to remove the possibility of rehearsing the items. Participants were randomly assigned to one of three conditions: the *free selection* Flashcard control, IRS spacing, and TBF spacing. The IRS spacing condition uses the model from Chapter 2 to determine the order of items. The TBF spacing condition uses the model described earlier in this chapter.

In the *free selection* flashcard method the items are divided into four selectable lists. Participants were allowed to choose which lists they want to practice, and the set of active lists could be changed at any point during the experiment. This approach was chosen as it is a fair approximation of a real learning strategy using flashcards: it creates the option to drop a group of items from practice once they have been learned, at the discretion of the participant (Kornell and Bjork, 2008). Items are presented in batches of four items. Once each item has been recalled correctly at least once, the next batch of items is loaded. Once all batches have been presented, the cycle repeats with the first batch.

Both spacing conditions do not allow participants to chose which lists to practice. Instead, the models can select from all items that are available at that point. In all three conditions the first presentation of an item is a study trial, and each subsequent trial is a test. Study trials consisted of both the question and answer displayed on screen, and require the participant to type over the answer. Test trials only provide the question and an answer prompt.

After each test trial the participant receives feedback. Feedback is presented for 1.5 seconds: for a successful trial, the word '*Correct*' is displayed. For responses within the maximum Damerau-Levenshtein[1] distance, '*Almost Correct*' is displayed on screen (Levenshtein, 1966; Damerau, 1964). The distance is a function of response length: a Damerau-Levenshtein distance of 1 accounts for roughly 80% of all spelling mistakes. However, for short words such as acronyms even a distance of 1 is quite large. Hence, for items with a character length of four or less, no errors are allowed. For answers with five characters or more a distance of 1 is allowed. Internally, an almost correct answer is registered as correct, as perfect spelling is not required for the type of material being studied. In the case of a wrong answer the text '*Incorrect*' is displayed, followed by a 4 second presentation of the correct response.

[1]The Levenshtein distance is a measurement of the difference between two sequences. It is defined as the minimum number of changes needed to transform the first sequence into the second. Valid changes are insertion, deletion, or substitution of a single character. The Damerau-Levenshtein distance also allows the transposition of two characters. Combined with the other operations, this explains most of the spelling mistakes made by people.

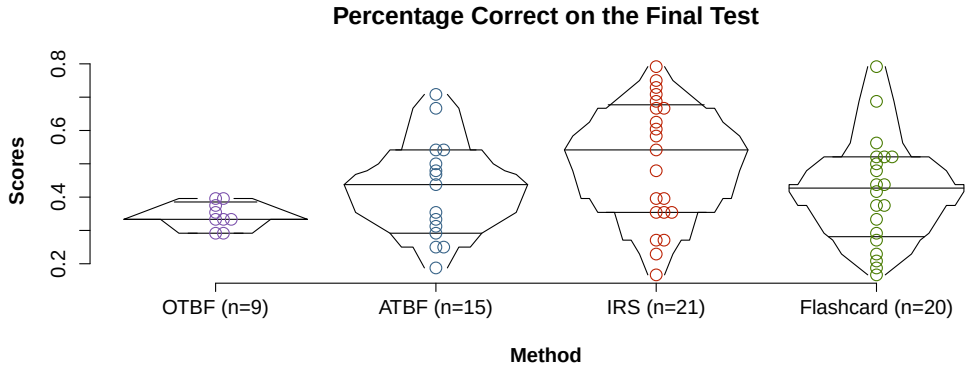


Figure 3.3: Difference in score on the final test, per condition. These so called box-percentile plots convey the same information as the standard box plots, but show more of the underlying distribution. For that reason, this type of plot will be used instead of the boxplots.

3.2.2 Results

3.2.2.1 Differences Between Methods

The results on the final test for each condition can be seen in Figure 3.3. Overall, the scores are quite low. This is a consequence of the design, as the pilot was used to find out how many items can be learned in half an hour. To avoid ceiling effects the study session contained a substantial amount of items to learn. The IRS model has lead to the highest overall test scores (n:21, M:0.51, SD:0.19), followed by the ATBF (n:15, M:0.42, SD:0.16), Flashcard (n:20, M:0.42, SD:0.17), and OTBF (n:9, M:0.34, SD:0.04) methods. A closer look at the data does show some differences in the way the scores are distributed: in the IRS condition there is a close grouping of the top half of the scores, while the lower half are more spread out. The Flashcard condition shows the opposite pattern. The ATBF condition looks very similar to the flashcard, albeit with a slightly higher 50$^{\text{th}}$ percentile. The addition of the fixed time component has a remarkable effect on the OTBF condition, increasing the average grade with almost a full point (0.34 vs. 0.42). The spread of the scores has increased quite a bit as well, in both directions. When looking at which scores lead to a passing grade (55% and above in the Netherlands), the IRS condition produced 10 passing scores (M:0.68, SD:0.07), while the Flashcard condition only has 3 passing scores (M:0.68, SD:0.11). For both TBF conditions combined, 2 scores were above 55% (M:0.69, SD:0.03).

A linear mixed model was constructed from the *logit*[2] transformed data to analyse the difference between groups. The best and worst participant in each condition were removed: these are either very good or very bad learners, and

[2]The *logit* is the logarithm of $p/(1-p)$, where p is a ratio. This transformation is used to map probabilities or scores onto the scale of the linear predictor used in logistic regression.

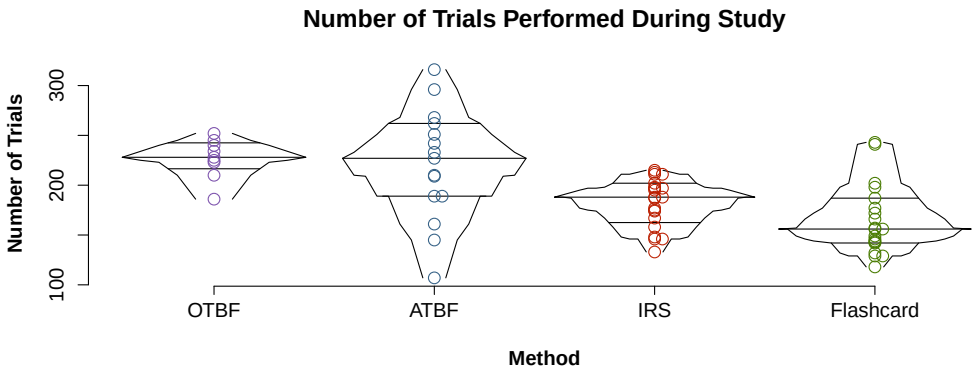


Figure 3.4: Number of trials performed during the study session, per condition.

would likely perform the same regardless of the learning method used. Removing more than two outliers per condition did not change the analysis qualitatively. An ANOVA showed that the difference between conditions was significant ($p = 0.037$). Further comparisons between groups were performed, and the IRS condition was found to be significantly different from the other three ($p = 0.007$). The difference remains significant when removing the condition with the lowest performance (OTBF, $p = 0.027$). The only significant individual difference is between the OTBF and IRS conditions ($p = 0.019$).

Despite not being significant, the difference between the performance of both TBF methods is quite striking, and strong evidence that taking into account a fixed time component is important for activation or RT estimation. But even with the fixed time component the model does not surpass Flashcard in terms of final test performance. The number of passing grades further supports the difference between conditions: almost half (48%) of all participants in the IRS condition score a sufficient mark, while in the Flashcard condition only 15% pass. The TBF conditions fare even worse, with only 8.3% passing for both conditions combined. Furthermore, in the IRS condition the average score is almost a full point higher compared to the other conditions. Overall, the evidence suggests that the IRS model is the best choice when one is allowed to pick a learning method.

Several aspects of the different conditions, which might impact test performance, were examined. The first of these is the total number of trials performed during the study session. The results can be seen in Figure 3.4. With an average of 220 (SD:56.5), the ATBF condition manages to fit significantly more trials in half an hour than the other two conditions (ANOVA: $p = 0.018$ vs. IRS and $p < 0.001$ vs. Flashcard). The difference between IRS (M:182, SD:24.7) and Flashcard (M:164, SD:34.9) is much less defined, although IRS seems to fit slightly more trials into half an hour on average. This might be caused by the time it takes to change the module selection (during study) in the Flashcard condition. Thus, two groups can be distinguished in terms of trials: high (ATBF) and low (IRS and Flashcard).

Closely related to the number of trials is the number of presentations per item.

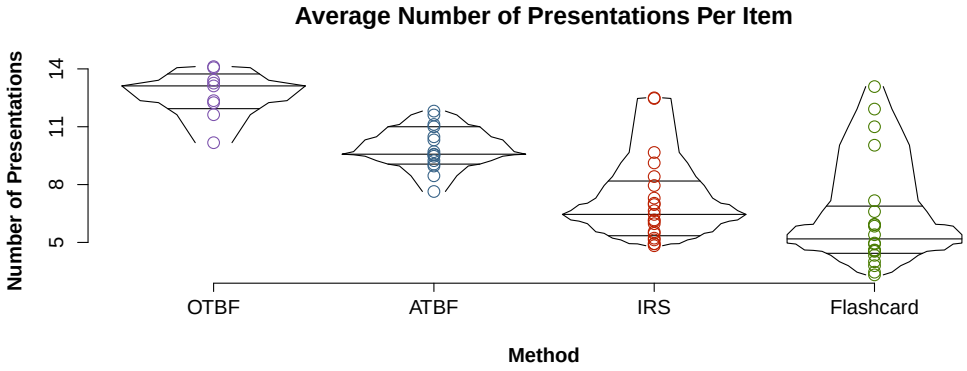


Figure 3.5: Average number of item presentations for a given participant, per condition.

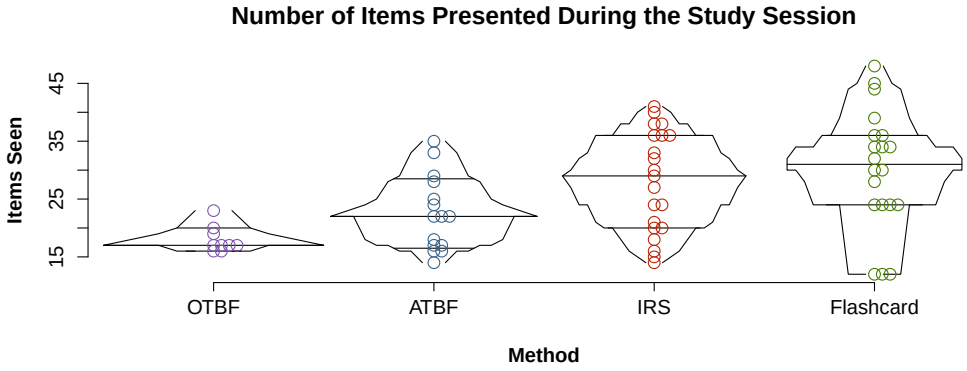


Figure 3.6: The number of items seen during the study session, per condition.

A larger number of trials should lead to more presentations per items. The result is quite similar to that seen in the difference between the number of trials (Figure 3.5). A notable change is that the difference between both TBF conditions (M:12.7, SD:1.3 and M:9.9, SD:1.2) is now significant as well ($p = 0.017$). The IRS (M:7.1, SD:2.3) and Flashcard (M:6.3, SD:2.9) conditions remain very similar.

Looking back at Figure 3.3, and taking in account the number of trials and presentations data, it seems more trials (and more presentations per item) does not result in a higher score. This is evident from the ATBF condition, which produces the largest number of trials on average, but does not score higher than either of the other two conditions. One could say that this is evidence against massed practice, but that would not be accurate in this case. The reason for the comparatively bad performance becomes clear when looking at the total number of items seen during study. Pavlik and Anderson (2008) argue that more practice should lead to better performance. However, better performance is not just about the number of trials, but also the number of items that are practiced.

More trials should not only lead to more presentations per item, but also more items seen: there are more opportunities to present new items (something which

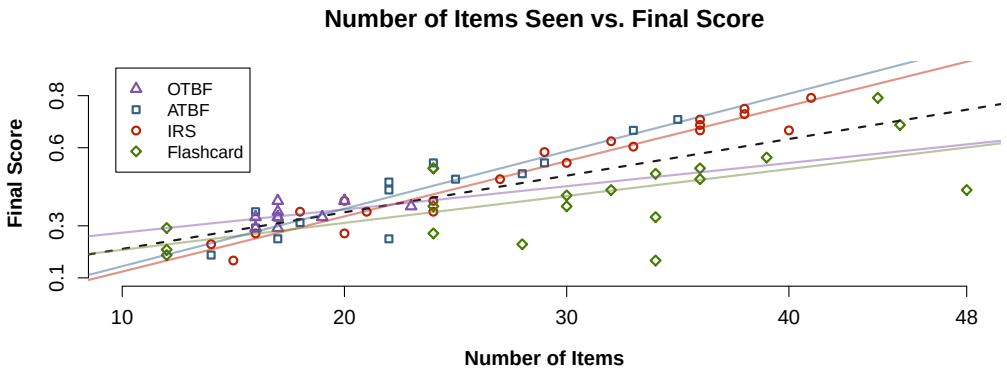


Figure 3.7: The relationship between the final score and the total number of items seen. Two groups or clusters can be distinguished: the spacing models, and the flashcard method.

is more relevant for the spacing models, as the number of items seen in the Flashcard condition is controlled by the participant). Figure 3.6 shows an inverse relation between the number of different items seen and the two previous measures. Rather than leading to more items seen, a higher number of trials seems to lead to the same items being rehearsed much more often. This is confirmed when looking at the average number of presentations shown earlier. The effects are not as strong as in the previous two comparisons, and only two differences are significant: OTBF (M:18, SD:2.3) vs. IRS (M:28, SD:8.9) ($p = 0.02$), and ATBF (M:22.5, SD:6.5) vs. Flashcard (M:30.1, SD:10.4) ($p = 0.049$). Still, Figure 3.6 shows clear trends. Again, the difference between both TBF models is present: almost half the participants in the ATBF condition have seen more items than all participants in the original model. Unfortunately the opposite is also true, although to a much smaller degree.

The number of different items seen might be a large factor in the low final scores produced by the ATBF method. The balance between learning new items and rehearsing old ones does not appear to be optimal. Furthermore, it seems very likely that the low number of items seen and the high trial counts are linked. This will be confirmed later on in this chapter. The average number of items seen is slightly lower than the Flashcard condition, although the distributions are too close to conclude (from this data alone) that the IRS model provides better retention (or higher efficiency) per item than the Flashcard method. To further investigate this, the relation between score and the number of items seen was examined.

One would expect the total number of items seen to have a strong effect on the final scores. Figure 3.7 shows the relation between score and items seen. As expected, this relation is quite explicit, especially for both spacing models ($r = 0.97$ for IRS, $r = 0.92$ for ATBF, and $r = 0.65$ for Flashcard). This seems to imply that the performance of an item practiced in a spacing condition is more predictable

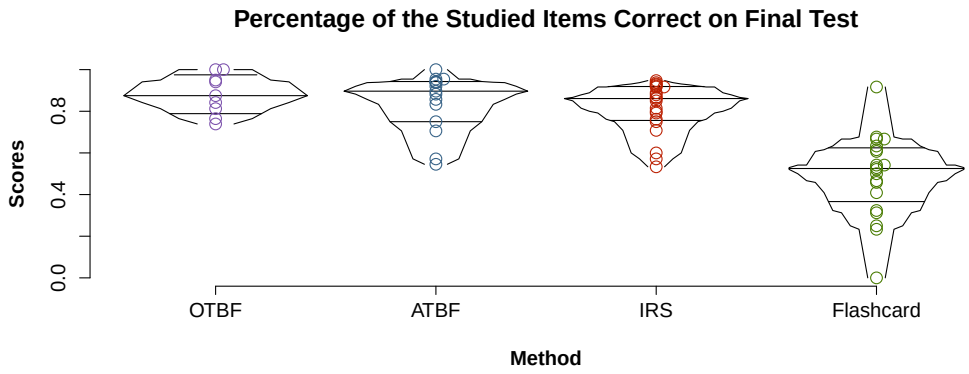


Figure 3.8: Adjusted scores (percentage correct of the items seen during the study session), per condition.

than one practiced in the Flashcard condition. A possible explanation is that strategies for free list selection in the Flashcard condition vary greatly between participants. Some of these strategies might produce very good results, others very poor results. In contrast, the scheduling methods use the same strategy with each participant (albeit tailored to their specific learning ability). From these results it seems that the 'items seen' measure combined with a scheduling method could be used for score prediction. Later on in this chapter this point will be examined again when another measure for score prediction is introduced, which is even stronger than the number of items seen. The scatter plot reveals something else; with equal numbers of items seen, the ATBF and IRS models result in higher final scores than the Flashcard condition. To further investigate this, adjusted scores were calculated and compared between conditions. The adjusted scores are the number of items correct on the test, out of all the items actually seen during study. This removes the effect caused by the difference in the number of items seen. The comparison between conditions can be seen in Figure 3.8.

An ANOVA shows a significant difference between the Spacing and Flashcard methods ($p < 0.001$ for all three spacing conditions), with the Flashcard method resulting in lower performance. Just as with the normal scores a *logit* transformation on the data was performed prior to the ANOVA. This indicates that items are not remembered as well in the Flashcard condition compared to the spacing models. All three spacing approaches have more or less equal performance, with both TBF methods on top. This difference can likely be explained by the larger average number of presentations per item seen in the TBF conditions (Figure 3.5). However, there is a ceiling effect in the scores, and this difference is not significant. In this instance the larger number of rehearsals per item does pay off, although the gain over the IRS condition is minimal and not significant. This implies that learning efficiency in the TBF method is not optimal, and more trials should be used to learn new items.

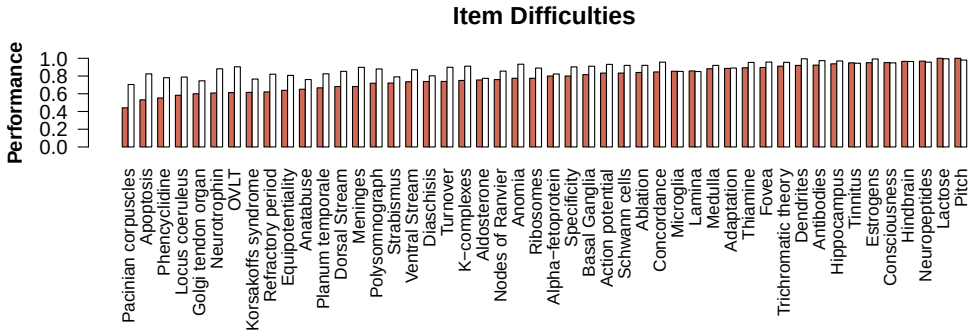


Figure 3.9: Item difficulty. Orange bars show the average correctness on the final test, and white bars show average correctness during study.

3.2.2.2 Item Difficulty

Item difficulty was examined to make sure it has no unexpected impact on test performance. The difference in item difficulty can be seen in Figure 3.9. The most difficult items have an average correctness around 50%, while the easiest items are near 100%. The spread of difficulty is fairly even, and difficulty increases almost linearly. No items were extremely difficult. The reason for the difference in difficulty remains elusive, as no interesting correlations between difficulty and possible predictors were found (i.e. sentence length).

3.2.2.3 Change Over Item Presentations

To examine the effect (and the differences per condition) of repeated item presentation, the average performance and RT were plotted per presentation, as seen in Figure 3.10. Very similar patterns can be seen for the IRS and Flashcard conditions: after the first presentation, there is a characteristic dip in performance at the second presentation. This is caused by the intervening trials of other items: with the Flashcard method the three other items in the batch are shown before the next presentation of the same item. The dip is less severe for the IRS condition, as there is only one intervening trial before the next item presentation. Hence, IRS performance remains higher by a fair amount. For both conditions, the increase in performance seems to level off after the third presentation. The TBF conditions does not have a dip at the second presentation, because the second presentation is always directly after the first one. Overall, performance in the ATBF condition is high and essentially constant across presentations.

In a sense, much of the same information can be seen in the RT graph. Again, the IRS and Flashcard methods show very similar curves, both levelling off after the third presentation. Here, the Flashcard method has somewhat higher RT values compared to the IRS condition. There is a clear mirroring (or reversal) compared to the performance graph: the lines are in opposite order, and now there is a clear

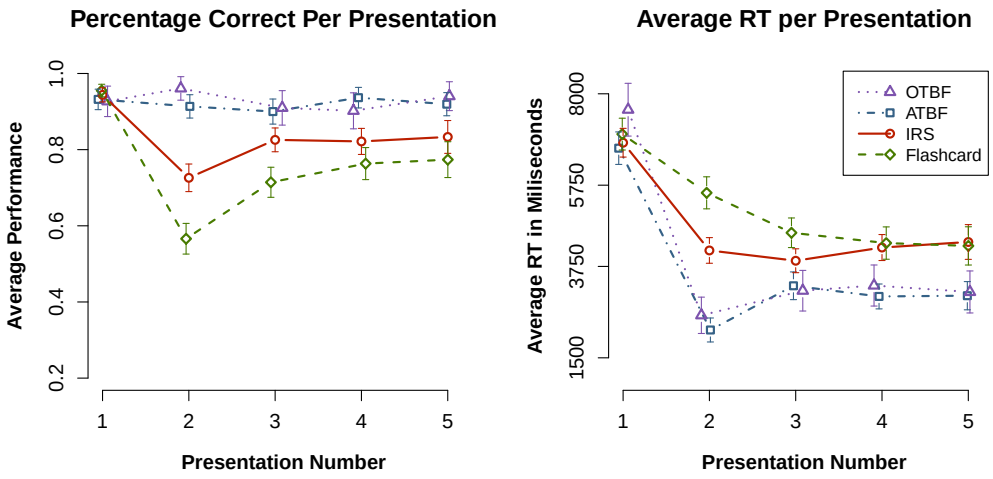


Figure 3.10: Change per presentation (with 95% confidence intervals). The left-hand graph shows the percentage of correct responses, and the right-hand graph shows the average reaction time. The values have been averaged over all participants per condition.

dip at the second performance for the TBF methods. This is a consequence of the second presentation of an item following the first. Participants can respond very quickly to this second presentation, as the answer is still readily available from STM.

The graphs in Figure 3.10 seem to confirm that the high number of trials in the TBF conditions is a result of the low number of items seen: having fewer items to rehearse leads to higher accuracy and shorter RTs, which results in more trials. Regarding the Flashcard condition, the low performance and high RTs might explain why the method leads to less efficient learning, as shown earlier. The trials might be too difficult: too many intervening items have led to too much forgetting. This could reduce the testing effect, which would have a negative impact on retention (Carrier and Pashler, 1992).

The average spacing between presentations can be seen in the left-hand graph of Figure 3.11. The IRS and Flashcard methods show large variances and a much steeper increase in spacing. The spacing in the Flashcard conditions flattens out after the first presentation, and decreases slightly. This is a result of the difference between the first round of a Flashcard batch (where each item is presented in order) and subsequent study rounds where the items are studied that have not yet been answered correctly. This also explains the variance in spacing for these subsequent rounds (shown in the right-hand graph of Figure 3.11), as this is dependent on rehearsal performance. The steadily increasing variance in the IRS condition is because the model personalizes each item, which leads to different presentation rates per item. The sudden jump in spacing distance between the third and fourth presentation is curious; a density estimation of the spacing at this point was cal-

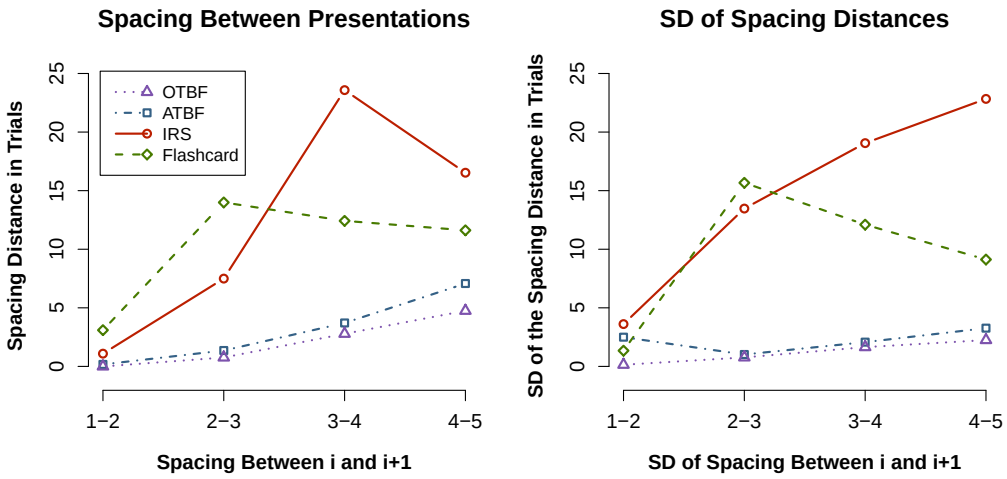


Figure 3.11: *Left-hand:* Spacing between subsequent presentations. Error bars were not included due to their great length. In stead, SD values were presented in a seperate graph. *Right-hand:* The SD values for the corresponding points in the left-hand graph: these indicate how different the items are spaced compared to each other.

culated (Figure 3.12) to investigate this occurrence. The density estimation shows three peaks in spacing: one for difficult items (which have low spacing), one for easier items (higher spacing), and one small peak for very easy items. This result does not indicate anything out of the ordinary is happing, thus the high value is likely caused by the large variance in spacing at this point. The spacing densities between later presentations show that the peaks disappear as they are smoothed into a single curve. This indicates that the spacing adjustment to each item is very granular. Finally, the ATBF condition has a slow but steady increase in spacing between rehearsals. This might indicate that the model will perform better with longer study sessions. The graph indicates that beyond the fifth presentation the average spacing will quite quickly be in the same range as the other conditions. The amount of variance in spacing distance is quite low, which indicates that the personalization per item (and participant) does not cause any large changes in presentation rate. This could be because almost every trial is answered correctly, due to the short spacing intervals.

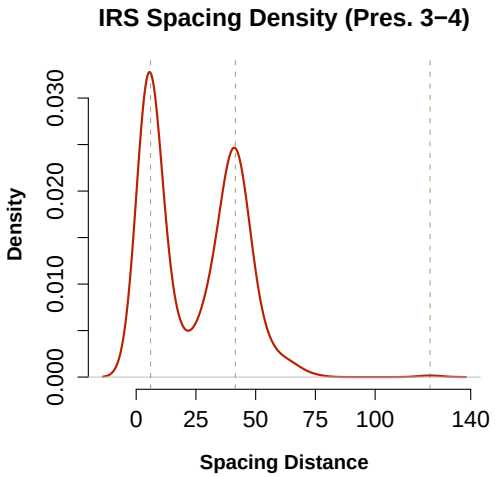


Figure 3.12: The density estimation of the IRS spacing between presentation 3 and 4.

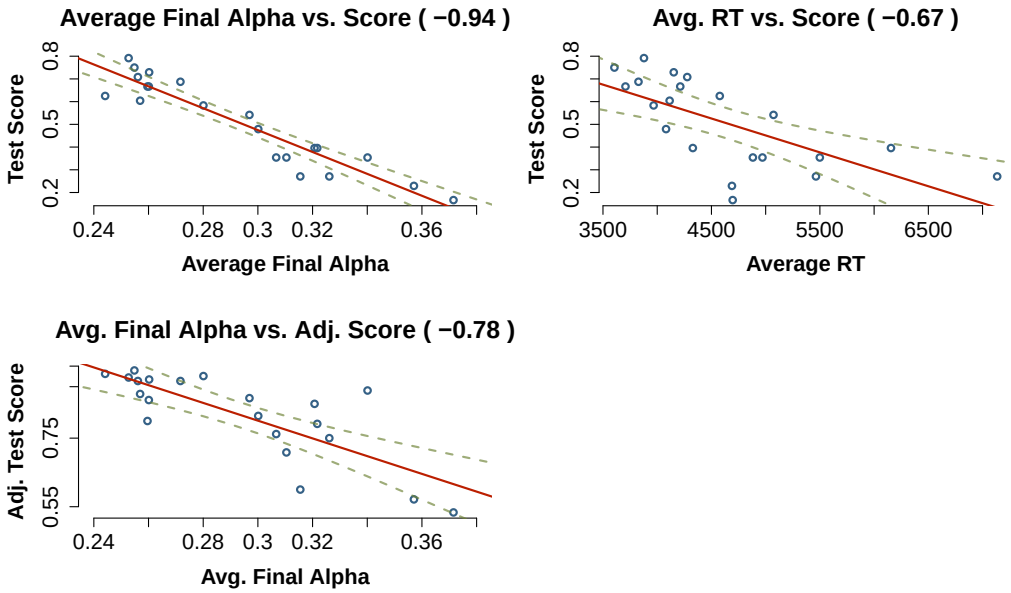


Figure 3.13: Predictors for final scores from participant input, with 95% confidence intervals. The average RT is the average calculated over all study trials. The average final alpha is calculated by taking the last calculated alpha for each studied item, and taking the average over all these values.

3.2.2.4 Predicting Test Scores

Earlier it was shown that the number of items seen during study is a good predictor for the final score. This line of investigation was continued to see if there are other predictors, obtained from participant input, which can lead to accurate score predictions. This might give some insight into the qualities found in good learners. Many different predictors were examined (although only the IRS data was considered), and the best two are shown in Figure 3.13. The *average RT* shows a reasonable correlation with score, but the *average final alpha* is an excellent predictor for both the regular and adjusted final scores ($r = -0.94$ and $r = -0.78$). In fact, for the adjusted score it is an even better predictor than the *number of items seen* ($|r| = 0.78$ vs. $|r| = 0.73$).

The idea that scores can be predicted from a single variable is a very interesting one. It could open up the possibility to estimate study lengths required to obtain a certain grade. However, the correlations presented here are specific to the pilot experiment. The real test is whether the score estimations generalize to a different experimental setup. Therefore, this idea will be revisited in the final experiment, where score estimation will be performed for a final test following multiple study session. What makes the *final alpha* predictor especially interesting is that it contains no information about the number of items seen, yet its relationship with the final score is almost as strong as the number of items seen, and for the adjusted score it is actually higher.

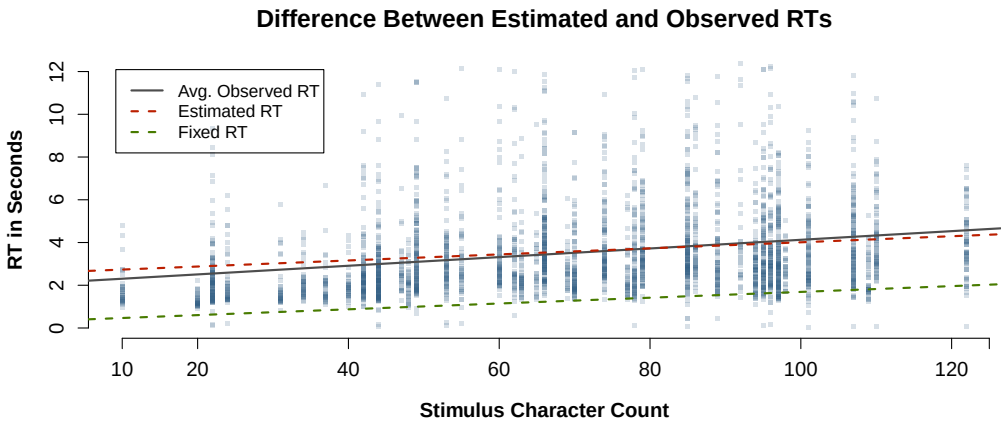


Figure 3.14: The difference between estimated and observed RTs found in the IRS data.

Strangely, even though the opposite is true, the intuitive expectation is that the *final alpha* should be better at predicting the adjusted score than the normal score. This is because a participant only has α values for items that have been presented, and there is no information in these values about the total number of items that will be tested. The adjusted score is just that; the score over all the items for which the participant has an α value. A possible explanation is that the adjusted scores is simply less robust: someone who has seen only five items and answers one of these incorrectly will have a 20% reduction in adjusted score, while someone who has seen 10 items will only receive a 10% reduction for one wrong answer. In short, the effect of an incorrect answer can differ quite a lot between participants, making it more difficult to predict the scores. In comparison, an incorrect answer in the regular score will always result in (only) a 2.1% reduction in performance, independent of the number of items the participant has seen. A different explanation would be the nature of the α value itself: smaller α values means less forgetting, which results in less time being spent rehearsing old items. With a fixed study session duration this frees up time to learn more items. Thus, the α value does not only represent the rate of forgetting, but also the rate at which new items can be learned.

3.2.2.5 Reaction Time Estimation

To see how well the fixed time estimation works in practice, the estimated RTs were compared against the observed values. The result can be seen in Figure 3.14. It is immediately apparent that there is a great deal of noise in the observed values, which is to be expected from RTs. The estimated values do not exhibit such variance, as the dominant part of these values are the sentence fixed times, which do not take noise in account. However, when looking at the difference in averages, the estimated values are quite close to the observed RTs. On average

the difference between estimation and observation is 0.59 seconds ($min = 0.01$, $max = 1.69$). This relatively small difference means that the original estimate was close to the mark, even though Equation 2.14 was based on RTs which represent the careful reading of sentences for the first time. For the final experiment, the fixed time estimation will be updated to better reflect the observed values. This might lead to an improvement in the models performance and efficiency.

3.2.3 Conclusion

Comparing different learning methods is no easy feat; From the data is appears that effective learning is a careful balance of rehearsing old and presenting new items, combined with a rehearsal order optimized for spacing (and the testing effect). Overall, the IRS model seems to deliver the best performance, followed by the Flashcard method. The difference is quite significant: almost a full grade point on average. The low number of unique items seen in the ATBF condition likely plays a large role in its low performance. This, combined with the large number of presentations per item indicates that the model is less efficient compared to the other two methods. In short, the ATBF model seems too conservative when it comes to presenting new items. However, the model seems to be optimized for very long study sessions (at least an hour; Pavlik and Anderson; Pavlik et al.) so perhaps a 30 minute study session is too short for the model to be effective. This is further supported by Figure 3.11, which shows that the average spacing between rehearsals slowly but steadily increases for both TBF models.

A problem with this type of experimental setup is that the effect of spacing is not very strongly present in the final test, as the retention interval is quite short. This could diminish the difference between the models. This problem should be eliminated in the final experiment. Related to this is the spacing present in the Flashcard condition. This means the difference between the scheduling methods and the control condition is relatively small (although clearly present), and therefore difficult to find in a between-subject setup. This problem will not be addressed in the final experiment, but is certainly something that should be investigated in future research.

Of all the possible participant input to use for performance prediction, the RT seems to show the strongest relation. This supports the use of RT to estimate activation, as performance could be considered a different way to represent forgetting. More surprising however, is the quality of the average final alpha as a score predictor. It seems this value combines the information of the average RT with the total number of items seen, resulting in strong relationship with both the regular and adjusted scores.

Finally, the pilot study was used to see how many items can be studied in 30 minutes. The data shows a large variance between participants in this regard. When only considering the IRS and Flashcard data, the average appears to be 29

items. The number of items per session in the final experiment will be around this value, but optimized for the particular setup used.

CHAPTER 4

FINAL STUDY

The final experiment tries to answer the questions posed at the start of this thesis: does intelligent item scheduling provide a benefit to memory retention under realistic circumstances? These circumstances refer to the short, fixed length study opportunities that educational institutes seem to encourage. The question will be answered using a longitudinal study: participants learn facts in three equally spaced sessions, followed by a (pen and paper) test some time later.

4.1 Model Changes

Based on the results from the pilot study, the reading time calculation was updated to better represent the data:

$$f_s = max(-157.9 + 19.5C^{count}, 300) \tag{4.1}$$

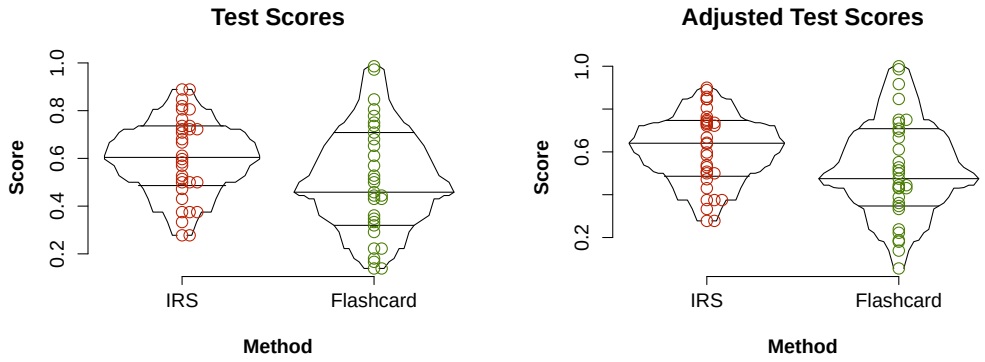
The scaling factor s^r was omitted, as the calculation is based on reading times after the first presentation. The lower boundary was set to the typical 'one word' reading time of 300 milliseconds.

While no other changes were made, this experiment will mark the first real use of the psychological time replacement discussed at the end of Chapter 2. Instead of scaling the encounter times, as *psychological time* does, the new approach calculates an activation offset for each item, which is used to estimate the activation at the start of the next study session. Because the pilot experiment only consisted of one study session, there was no opportunity to test this part of the model earlier.

4.2 Method

4.2.1 Participants

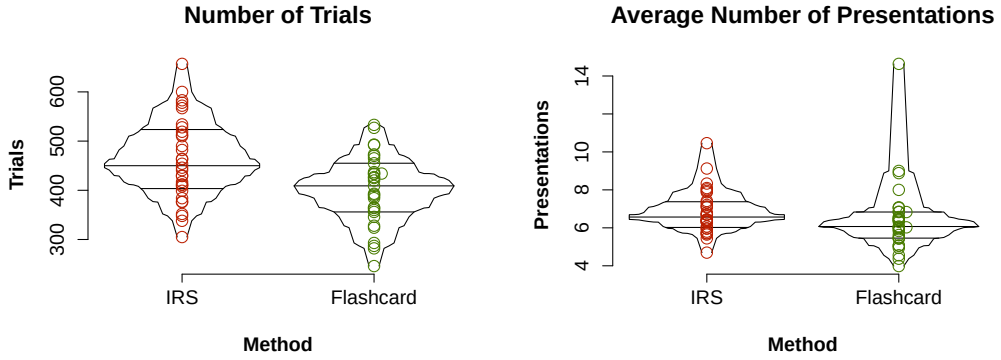
Participants were taken from the same pool as the pilot study: first year Psychology students from the University of Groningen. A total of 63 participants were tested (17 male, 46 female), with an average age of 20.6 (SD:1.6).



(a) The regular scores, taking all items in consideration.

(b) Scores computed from the items the participant actually studied.

Figure 4.1: The scores on the final test.



(a) The total number of trials performed, summed over all three study sessions.

(b) The average number of presentations per item, considering all three study sessions.

Figure 4.2: Additional measurements between conditions.

4.2.2 Materials

For this experiment, 72 facts from the Biopsychology course were adapted as stimuli. No facts from the previous experiment were used. The complete list of facts can be seen in Appendix B. The choice for the number of facts is based on the results of the pilot study, which showed that on average a participant was able to learn approximately 1 fact per minute. For this study the sessions are slightly shorter, so the number of items that could be learned per session was estimated at 24.

4.2.3 Design

The study consisted of four sessions: three study sessions, followed by a test. The study sessions were spaced with one day in between, and the test was taken three days after the third study session (ISI=1 and RI=3). Each study session lasted 25 minutes. The test was performed with pen and paper, and the participants essentially had unlimited time to fill in the answers. Participants were assigned to one of two conditions: (free selection) Flashcard or IRS. Aside from the changes mentioned at the beginning of the chapter, the models are identical to the ones used in the pilot study. At the beginning of each session 24 new facts become available for study. So during the first session there are 24 facts, and during the third there are 72. In the Flashcard condition the facts were subdivided into lists of 12, from which participants could choose freely at any time during the study session. In the IRS condition participants could not choose which lists they wanted to learn.

The study session itself is the same as in the previous experiment: after each test trial feedback is presented in the form of the word '*Correct*' for a successful trial, or the word '*Incorrect*', followed by a four second presentation of the correct answer otherwise.

4.3 Results

4.3.1 Differences Between Methods

The distribution of the final scores can be seen in Figure 4.1(a). The scores are in the same range as the earlier study, which means the study task was sufficiently difficult. The results are very similar to the earlier study as well: The IRS condition has higher scores on average (n:32, M:0.60, SD:0.18) compared to the Flashcard condition (n:31, M:0.50, SD:0.24). Aside from the differing average, the Flashcard condition also seems to exhibit a larger spread of scores, evidenced by the larger SD value. The number of passing grades (a score above 55%) differ as well: 19 passing scores for IRS vs. 12 for Flashcard. For the sake of completeness, the adjusted test scores were calculated as well (Figure 4.1(b)). However, the adjusted measure is less interesting as there is a ceiling effect in the number of items seen per participant. Furthermore, as participants see less items, it is more likely they will have a higher adjusted score, as the study task becomes easier. In an extreme case, someone might see only one item, which would very likely result in a perfect score. However, this score reveals nothing about the learning ability of that individual.

Because the earlier study showed that the IRS condition was significantly better, a one-sided t-test was performed to see if it still outperforms the Flashcard condition. Again, the IRS condition was found to be significantly better ($t = 1.883$, $p = 0.033$, and $df = 55.303$). Given that the average difference is a full grade point

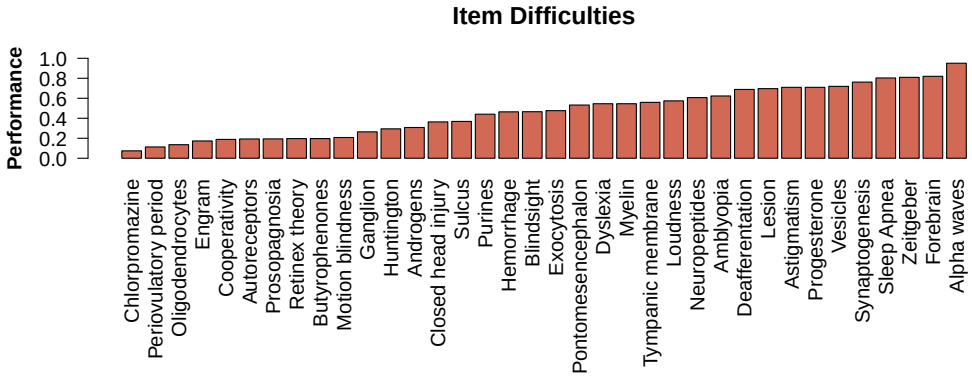


Figure 4.3: Item difficulty on the final test. For the sake of readability, only every second item is displayed.

(60% vs. 50%), this significance is not trivial. Removing the best and worst participant per condition did not change the analysis in a qualitative manner, but did make the difference more significant. Furthermore, the adjusted test scores were significant as well ($t = 1.848$, $p = 0.035$, and $df = 55.496$). The difference in adjusted scores is not as large as in the pilot study, however.

Other measurements between conditions were revisited as well: the number of trials performed during study, and the average number of presentations per item. Due to a ceiling effect, the number of items presented was omitted. As shown in Figure 4.2, the IRS condition is somewhat higher for both measures: the difference in the number of trials is quite pronounced (M:462 vs. M:401). Of the two measures, only the difference in the total number of trials is significant ($t = 3.047$, $p = 0.003$, and $df = 59.831$). Removing the outliers causes the difference in the average number of presentations to become significant as well (ANOVA, $p = 0.029$), with IRS having half a presentation more on average (M:6.83 vs. M:6.17).

4.3.2 Item Difficulty

The item difficulty was examined again, and shows a very regular curve (Figure 4.3). One difference from the earlier study is that the range in difficult is much larger, with the most difficult items being answered correctly only around 10% of the time.

4.3.3 Change Over Item Presentations

As in the pilot study, the change in performance over presentations was investigated. The result can be seen in Figure 4.4, and is quite different from the graphs presented in the previous chapter (Figure 3.10). There is a substantial difference in setup however: a subsequent item presentation may happen in a different session. Furthermore, as the total study duration was much

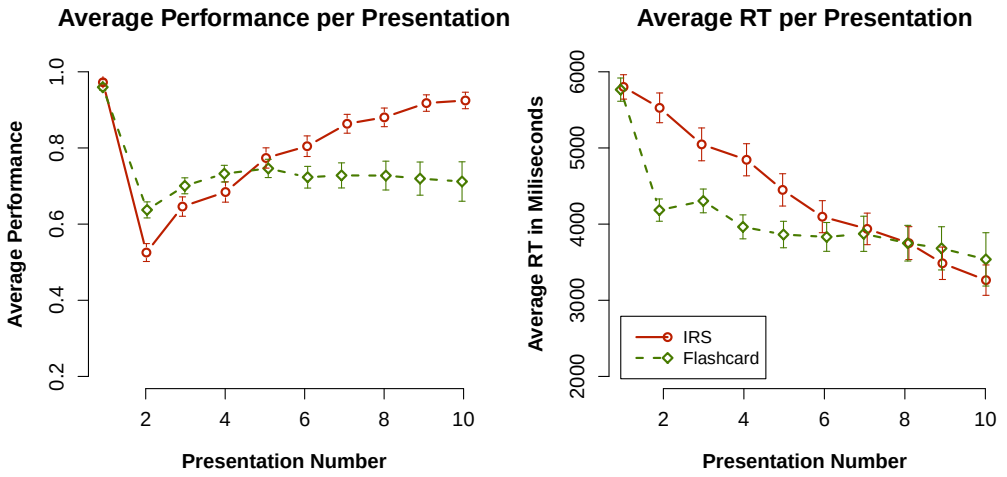


Figure 4.4: change per presentation (with 95% confidence intervals). the left-hand graph shows the percentage of correct responses, and the right-hand graph shows the average reaction time. the values have been averaged over all participants per condition.

longer, the average number of presentations per item was higher. As a result, data from presentations 6 through 10 was added this time. At first glance, the scores start out much lower for both conditions than previously. This might be a result of the higher item difficulty. In the IRS condition the performance per presentation increases in a smooth, almost linear way. This is complemented by the RTs, which decrease in a similar fashion. The sudden drop in RT at the second presentation is gone; likely caused by the greater variance in spacing between the first two presentations (which will be discussed later). The Flashcard condition shows a very different pattern. After a large drop in both performance and RT at the second presentation, the change levels out. Performance is almost constant after the fourth presentation, and the RT shows a slow but steady drop.

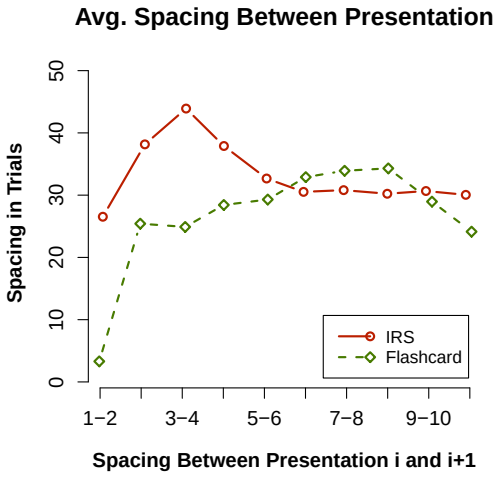


Figure 4.5: The spacing distance between subsequent presentations.

Both graphs show an intersection point. Initially, performance and RT is better in the Flashcard condition. After a number of presentations, however, this reverses. Study performance becomes higher in the IRS condition around the fifth presentation. The change in RT comes a bit later, around the eight presentation. The discrepancy between the score and RT inter-

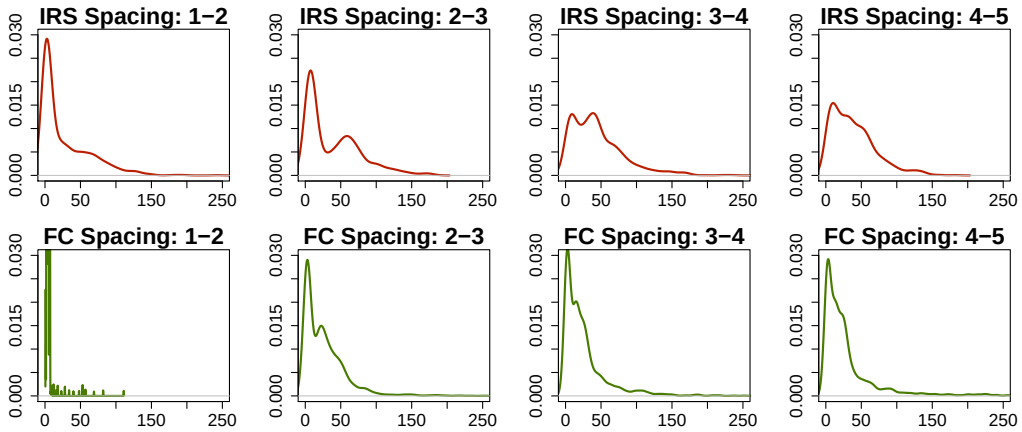
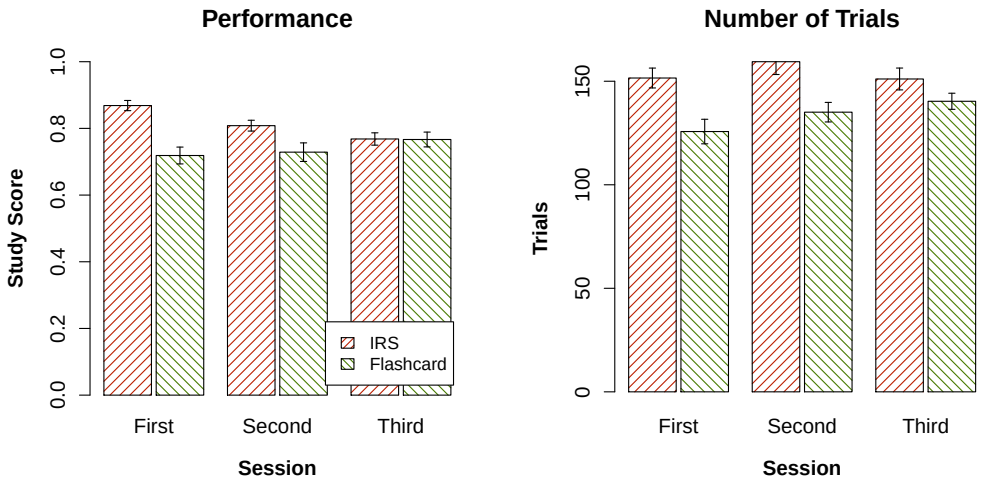


Figure 4.6: Density estimation of the spacing intervals between trials. The x-axis is the spacing in trials, and the y-axis is the density. The top row shows the first four estimates for the IRS condition, the bottom row shows the estimates for the Flashcard condition.

section points seems to indicate that while closely related, one does not automatically infer the other.

The change in spacing over presentations was also revisited. Figure 4.5 shows a picture quite different from the first study. Spacing in the Flashcard condition is similar, but much higher; even at the second presentation. It still seems to level out quite quickly, and after the eighth presentation it even goes down somewhat. The IRS condition shows a similar shape to the one found in the earlier study, but also with much higher average spacing values. A possible cause is that some items are very easy (or already known to the participant), causing them to have very large spacing intervals. This was already apparent in the pilot study, but with more items and study time this effect may have become more pronounced. This could also explain why the spacing levels off after the fifth or sixth presentation; the spread of spacing distances has become so diverse between items that the average interval no longer gives any insight into the spacing within the condition. To investigate this, the density estimations of the first four presentations were plotted for both conditions.

The densities can be seen in Figure 4.6. Even between the first and second presentation there is already a considerable difference in spacing, indicating that some items are indeed very easy. The second interval shows two distinct bumps: a group of difficult and a group of easier items. The third and fourth interval show that the bump is smoothed out until a single slope remains: the measurement of item difficulty has become more fine-grained. This means the spacing intervals are very diverse, which fits with the earlier remark that the average spacing interval has lost much of its meaning. The Flashcard condition shows a similar progression, but much more concentrated on the lower side of the interval scale.



(a) The average performance during study.

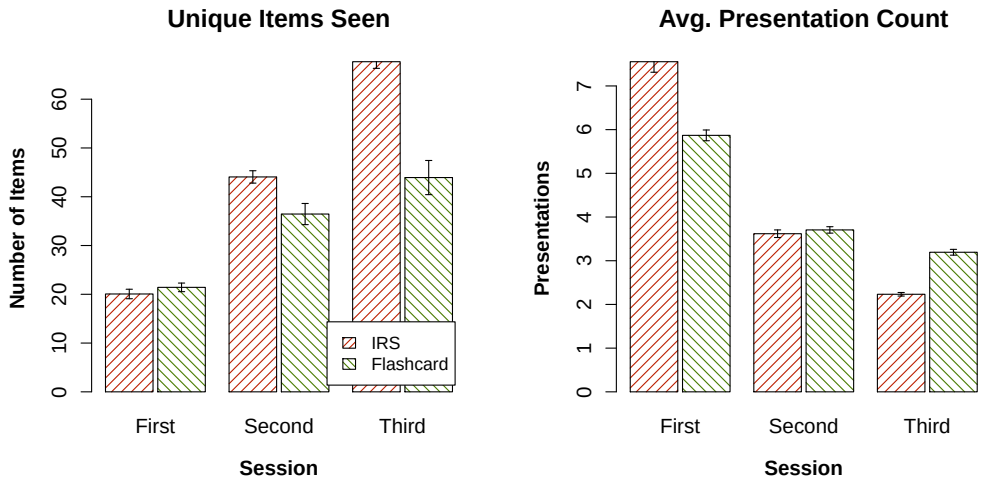
(b) The average number of study trials.

Figure 4.7: Measurements per session and method.

4.3.4 Change Between Study Sessions

To investigate the effect of multiple sessions on the learning methods, a number of measures were compared between sessions and conditions. Figure 4.7(a) shows that the study performance decreases over sessions for the IRS method, and increases for Flashcard. The interaction between session and method was found to be significant (ANOVA, $p = 0.003$). Looking at the confidence intervals, this seems to be caused by the larger intervals between presentations in the Flashcard condition. The change in score fits with the results from the earlier study: the larger initial spacing makes the Flashcard method more difficult. Older lists become increasingly easier to learn in later sessions (when not 'mixed' with other lists). In contrast, the increase in the total number of available items makes learning new items more difficult in the IRS condition, as the average spacing increases. For the number of trials (shown in Figure 4.7(b)) there is no effect for session, nor any interaction between session and method. Still, there seems to be a slight increase in the number of trials for the Flashcard condition. A possible explanation is that participants can rehearse the lists from previous sessions quicker as they have seen the items before, which could add a small amount of extra trials.

The change in the number of unique items seen per session is shown in Figure 4.8(a). To clarify, this measurement was taken by counting the unique items seen by a participant during a session, regardless of whether the item was seen before in a previous session. Both the main effect of session and the interaction between session and method are significant (ANOVA, $p < 0.001$ for both). During the first session, the number of items seen is more or less the same. The number rises over sessions, but much faster in the IRS condition.



(a) The number of unique items seen.

(b) The average number of presentations over all items.

Figure 4.8: Measurements per session and method.

This can be explained: in the Flashcard condition participants are more likely to rehearse the new lists and forgo the lists of the previous session. This will be examined in more detail later in this chapter. The opposite pattern is seen in the average number of presentations per item (Figure 4.8(b)): The number of presentations drops for both conditions, but more quickly for IRS. This makes sense; more items seen means less time to rehearse each item within the time limit. Again, both the session main effect and the interaction are significant (ANOVA, $p < 0.001$ for both). Interesting is that there is a difference in average presentation for the first session, despite the number of items seen being the same. This is explained by the higher average trial count in the IRS condition.

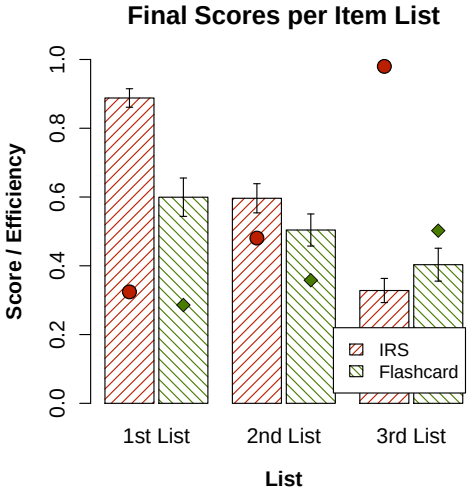


Figure 4.10: Comparison between average performance on each list during the test. Study efficiency (the gain in test score per second spent learning, in this case normalized in regard to the highest efficiency) is shown as points centred at the bar of each list. These values are normalized to the highest found efficiency.

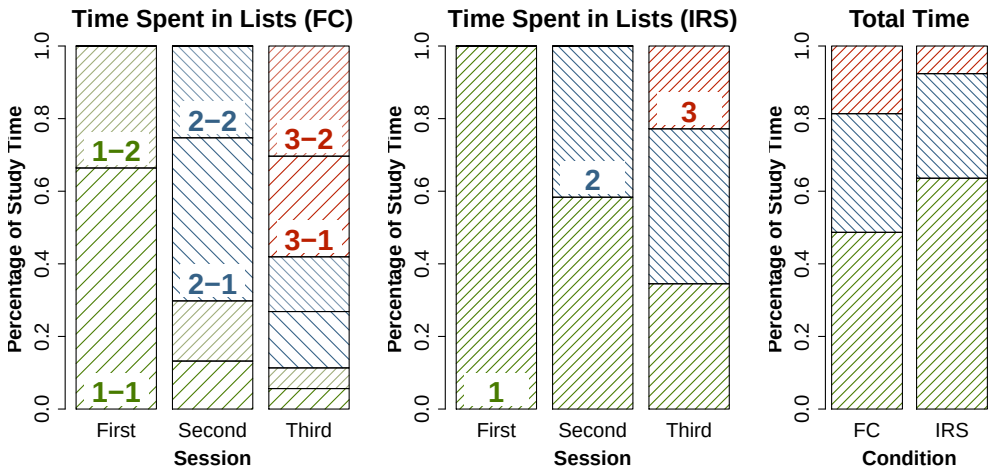


Figure 4.9: Distribution of study time over the different lists, per session. The rightmost graph compares the total time spent in each 'list' over all three session between both conditions.

4.3.5 Item List Selection

Given that the experiment is now spread over three study sessions, list selection becomes much more important for performance. The time spent learning a particular list was plotted per session in Figure 4.9. Immediately apparent is that the IRS condition spends a lot of time in the first list, even at the third session. In the Flashcard condition, participants spent most of the time in the newly available lists for that particular session. This leads to the expectation that participants in the Flashcard condition had substantially more correct answers on the test for items found in the last list. Figure 4.10 confirms this: the main effects of list and method, as well as the interaction between list and method are significant (ANOVA, respectively $p < 0.001$, $p = 0.004$, and $p < 0.001$). The difference in performance on the first list is quite dramatic. Even for the second list the IRS condition performs better, even though more time was spent in that list within the Flashcard condition (33% vs. 29%). This presents additional evidence that the scheduling in the IRS condition is more optimal. This can be taken one step further by computing the study efficiency of both conditions, as the ratio between *time spent studying* and *final score*. These are plotted as points in Figure 4.10. These points reveal several things: learning is subject to diminishing returns (ie. each additional trial spent learning the same material yields less retention gain), and learning in the IRS condition is more efficient.

4.3.6 Test Score Prediction

Finally, score prediction was investigated once more. The two most successful measurements of the pilot study were applied to the re-

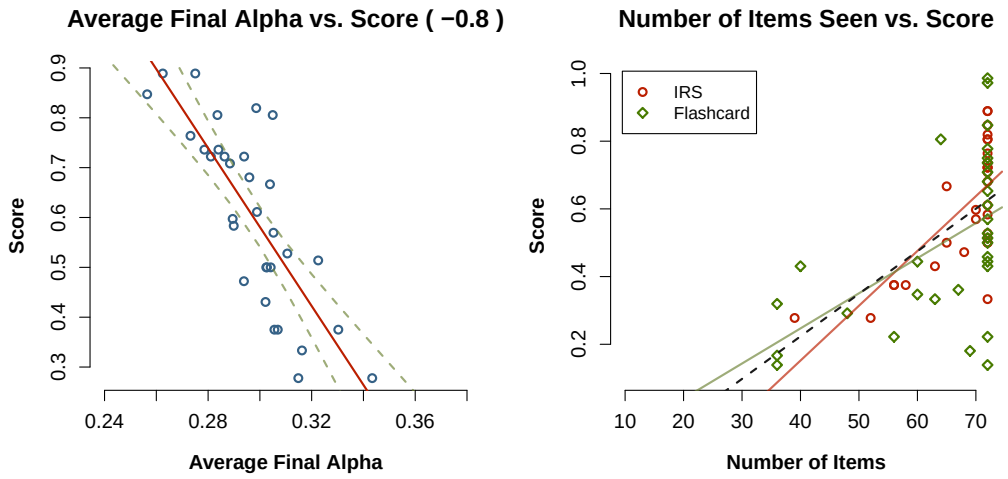


Figure 4.11: The two best score predictors of the pilot study, now plotted for the final study data.

sults of the final experiment, and the results are shown in Figure 4.11. The correlations are somewhat lower this time: $|r| = 0.8$ for *final alpha*, and $|r| = 0.7$ for *items seen* in the IRS condition (for *items seen* in the Flashcard condition the correlation $|r|$ is 0.52, and over the entire set it is 0.59). The lower correlation for the *number of items seen* is due to a ceiling effect; many participants saw all the items at least once. Hence, the measurement cannot be used as a score predictor in such cases. The *average final alpha* does not suffer from this problem, and remains a very good indicator of the final score. To summarize the score prediction efforts, Figure 4.12 contains the three most promising indicators in one graph. It clearly describes the relation between the different predictors and the final score: participants who scored well saw most items, and had low alphas and RTs. Conversely, participants who scored poorly did not always see all items, and had large alphas and RTs. The graph also shows that not everybody fits the mold; some participants had large RTs but still scored well. This is not surprising, as the analysis of the pilot study had already confirmed that RT is less suitable as a predictor.

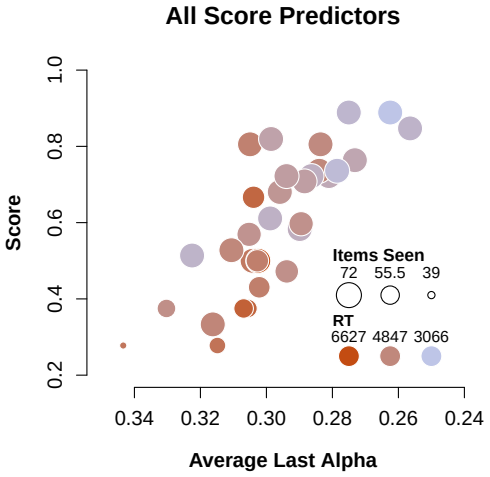


Figure 4.12: A bubble plot combining the three best predictors: final alpha, items seen, and RT. Each point represents a participant.

4.4 Conclusion

Despite a very different setup, the final study confirms many of the findings from the pilot. There is an average difference of 0.1 (which represents one grade point) in favour of the IRS method. This is remarkably similar to the difference between both methods found in the earlier study, which was 0.09. Therefore, it seems that the difference between methods is constant, and is not affected by the amount of study time (given that both methods receive equal attention). This cannot be stated as fact, however. It is possible that three study sessions within the span of a week is still too short to reveal any interaction between the average score difference and total study time. It could also be that more exposure to the free selection Flashcard method made participants better at list selection, therefore increasing the performance of the condition. Still, it seems that the increased room for spacing presentations in the IRS condition does not have much of an effect, which invalidates the prediction that the difference in test scores between the conditions would become larger. This could mean that the increase in spacing distances is not enough to affect the test scores, or that the maximal benefit of within-session spacing was already achieved in the pilot experiment. Having tested only two different spacing setups, the experiments done in this work are not enough to provide conclusive results on the matter.

One clear difference between the pilot and final studies is the slight ceiling effect in the total number of different items that were seen by the participants. In the pilot, virtually no participant saw all 48 items. In the current study, however, many participants saw all 72 facts. This was true for both study methods. This did not seem to affect the difficulty of the task much, as only a few participants came even close to a perfect score. This is to be expected; seeing an item is not that same as rehearsing it often enough to consider it readily available knowledge. Difficulty of the task clearly increased for each session: studying 24 facts in 25 minutes was well within the capabilities of most participants. Rehearsing 72 facts in 25 minutes was quite a challenge, even if most of the facts had already been seen before.

Overall, there are quite clear differences between the IRS and Flashcard conditions, ranging from simple measures (also seen in the pilot study), to complex differences across sessions. For one, the IRS condition seems more efficient, as it is able to fit more trials in the same amount of time. Performance over presentations keeps increasing when using IRS, but levels off in the Flashcard condition. A similar difference is seen for the decrease in reaction time. When looking at spacing, the IRS condition adapts itself much more to the differences between items: the spread in spacing is much larger.

In terms of time distribution over all of the lists, two very different strategies become apparent: if Flashcard were to be described as 'progressive' than IRS would most certainly be 'conservative'. IRS puts the focus on rehearsing facts that have already been presented, while participants in the Flashcard condition gave pref-

erence to studying new facts. Subdividing the final test scores over each list shows that the more conservative approach results in better scores. Given the differences in test scores, two observations can be made: people are too quick to drop items from study (Karpicke and Roediger, 2008), and focussing on only a part of the material is a viable learning strategy. It is better to know a portion by heart than the whole set superficially. Given all these differences, it is difficult to say how much each of these contributes to the overall difference in test performance between both methods.

The data from the final experiment also validates the idea of score predictions, as evidence suggest the phenomenon was not confined to the particular conditions of the pilot study. The effects seem slightly less powerful in this study. However, the interval between study sessions and a much longer RI could have introduced additional noise. Taking this in mind, the quality of final alpha as a score predictor is quite remarkable. The total number of items seen still works quite well, as long as a participant has not seen all the items. This limitation is quite severe, and therefore the total items seen cannot be considered a good predictor: in real study situations it is quite likely that an individual will have seen all the items.

CHAPTER 5

DISCUSSION

In this chapter the implications of the results found in the previous two chapters are discussed. Afterwards, the thesis is concluded with a look at possible directions future research could take.

5.1 Item Scheduling Under Practical Circumstances

Simple fact learning is an integral part of education. Facts are the building blocks on which complex rules and inferences are built. However, the time people have available for study is often limited. Therefore, it is important to optimize the study process and maximize the memory retention while minimizing the time spent rehearsing. Proper use of two powerful phenomenon can help in reaching this goal: spacing and the testing effect. These two phenomenon imply that long term memory retention can be improved by increasing the time between rehearsals and by making the memory retrievals more difficult. Thus, the optimal time to schedule an item is just before it will be forgotten, as this optimizes the gain from both effects. The scheduling algorithm presented in this thesis is the IRS model, which is a continuation of earlier work by Van Woudenberg (2008); Van Thiel (2010), and optimizing the long term retention gain is exactly what it tries to do. Thus, the central question posed in this thesis is whether item scheduling determined by a cognitive model leads to an increase in memory performance, when used under practical circumstances. To answer this question two studies were conducted.

5.1.1 The Studies in Short

In the pilot study participants received a thirty minute study session, in which they had to learn 48 facts. This session was followed by a ten minute distraction task, and concluded with a final test over all the facts. In total, there were four different study conditions: two versions of the Pavlik TBF model (Pavlik et al., 2008), the IRS model, and as control condition the Flashcard method. The difference between the two TBF conditions was the addition of an estimator for item reading time in the model. The final study had a more elaborate setup, with each participant completing three 25 minute study sessions, with a one day interval between sessions. The test was performed three days after the final study, and

Not at all - 1	**2**	**3**	**4**	**5 - A lot**
Did you enjoy using the study system?				
6%	0%	17%	44%	33%
Would you use the system to learn facts?				
0%	6%	11%	22%	61%
Do you think the system was helpful?				
6%	0%	22%	17%	56%
Do you think you would have learned better without the system?				
17%	44%	17%	17%	6%

Table 5.1: Survey taken after the studies. Results of both the IRS and Flashcard methods were merged, as no difference in answer distribution could be found between the two conditions.

was done with pen and paper this time. The facts were evenly distributed, with 24 extra facts becoming available at the start of each session, totalling in 72 facts that could be studied. Only two conditions were tested this time: IRS and Flashcard. These conditions were the same as in the pilot study.

5.1.2 The Bottom Line: Which Method Works Best?

The final test performance per condition paints a clear picture: participants in the IRS condition scored higher on average than any other condition, in both studies. In the pilot study the only significant difference was found between IRS and the other three conditions as a group. Given the difficulty in evaluating study methods and the somewhat low number of participants per condition, this was a very encouraging result. Furthermore, the difference in participants with a passing grade was quite dramatic: Three times as many participants passed in the IRS condition compared to the second best method (Flashcard). The difference between the performance of both Pavlik models validated the investigation into reading times performed in this thesis, as the reading time estimation increased the average score by 30% within that condition.

The final study provided definite evidence: the IRS condition performed significantly better than the Flashcard control. The difference is almost a full grade point, which is roughly the same difference between these two conditions which was found in the pilot study. Again there is a large difference between the number of passing grades, with the IRS condition containing almost twice as many. Therefore, it seems the IRS method is quite robust, performing better than other methods in two very different, yet realistic, circumstances. The IRS model has another advantage over free selection Flashcard: it does not require the participant to manually select lists. This saves both time and allows the participant to

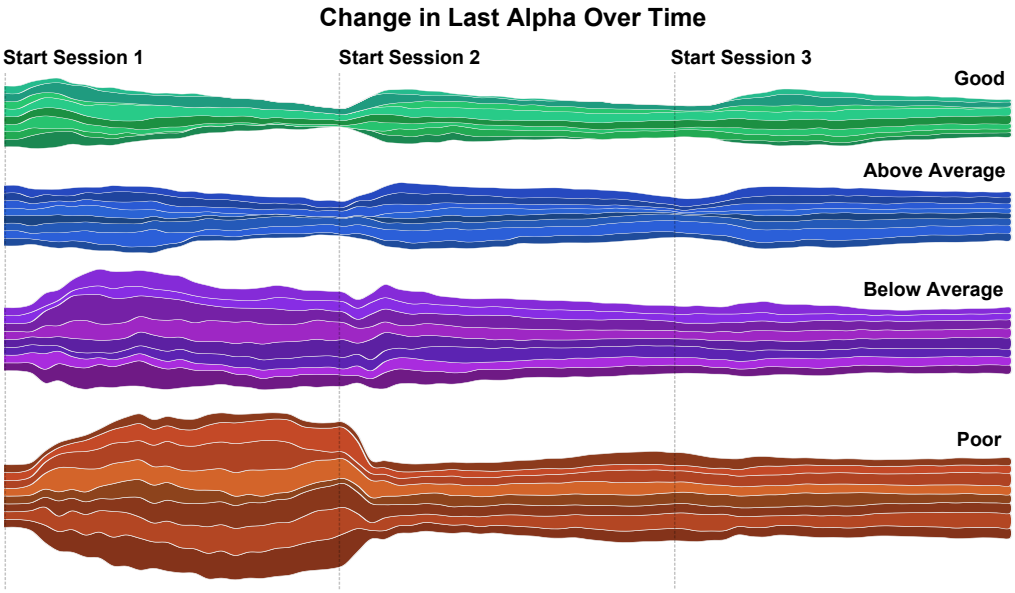


Figure 5.1: A streamgraph (Byron and Wattenberg, 2008) visualizing 'forgetting' or 'study difficulty'. The quantification used for this is the *average final alpha* (as discussed in earlier chapters), plotted over time: the value was recomputed every thirty seconds. Each colored line represents one participant, and the thickness of the line represents the *final alpha* at that point. A thicker line means a higher *alpha*, which in turn means a greater study difficulty. The plot shows how this difficulty changes over all three study sessions. Participants have been divided into four groups, based on performance on the final test. Study time over all three sessions has been graphed as a continuous line.

fully concentrate on studying. Considering all the results, the central research question has been answered: item scheduling using a cognitive model does lead to an increase in memory performance, when used in realistic circumstances.

5.1.3 Computer Aided Learning

Taking the particular method out of the equation, there is something to be said for learning facts using a simple computer interface. A follow-up survey with the participants was performed, which focused on their experience with the system itself. As such, results were not grouped per condition. Overall, the reaction is very positive (Table 5.1). Only a few expressed that they could have learned better using their own methods. Comments by participants were enthusiastic, and some even asked if they could use the system for their own study purposes. This goes to show that even the simple act of automating what is normally considered an unpleasant task can have a positive impact.

5.2 Describing the Learning Process

Apart from providing evidence to what the best method is, the studies have supplied an interesting look at behaviour during learning. The final study showed that people seem to focus on new material over old: participants in the Flashcard condition allocated the majority of study time in each session for the newly available facts. The test performance of earlier material suffers significantly from this strategy, so it would seem that spending time on this material in later sessions is important. This is consistent with earlier findings (Karpicke and Roediger, 2008).

The data obtained from the study sessions allows for one step further, and can be used to visualize the amount of forgetting (or study difficulty) over time. This is shown in Figure 5.1. There is a very clear difference in behaviour between the four groups. Those with a sufficient test mark show a clear pattern: when new items are introduced the difficulty increases, and by the end of the session the new material has been 'mastered' and difficulty becomes less. The time between sessions does not seem to cause much forgetting: at the beginning of each session the items from previous sessions are rehearsed. However, there is no immediate spike in difficulty at the start of the second and third sessions. Instead, there is a kind of lag before a spike which signals the introduction of new items. Furthermore. it can be clearly seen that the good learners have less difficulty over the entire study time than the above average learners: average learners need more study time to reach the same level. Below average learners show a very different pattern. After a very difficult first session, the difficulty values seem to stabilize, or at most decline very slowly. Gone are the typical spikes seen for the good and above average learners. This would indicate that these participants are still struggling with the old items, and as a result new facts are only introduced sparingly. While more similar to 'below average' than any other type of learner, the poor learners show an unexpected pattern; In the second session the difficulty reduces significantly, only to rise slowly during the remaining two sessions. It seems the study behaviour of these learners lead to almost no new items being introduced.

Good learning methods must take in account this fundamental difference if they are to be successful. The cognitive models described in this thesis all handle this problem in their own way, through parameters that offset activation, or by changing the rate of forgetting. These adjustments optimize the time spent learning, but do not indicate how much time is actually needed to learn the whole set. This seems like a logical extension of the scheduling algorithm, and would require a prediction of test performance based on the current 'study difficulty' an individual is experiencing.

5.3 Assessing the Study Difficulty of Individuals

More out of curiosity than anything else, learning behaviour was studied to see if it is possible to estimate a participants final score using some quantification obtained from study trials. Somewhat surprising, the pilot study yielded two measures that could estimate scores in the IRS method with great accuracy: the *number of items seen* and the *average final alpha*. In a way, the learning capacity was predicted: the amount of information an individual can learn in a fixed time period. Despite a very different setup, the two predictors were still very effective in the final study. However, one problem emerged with the *number of items seen*; many participants saw all 72 items at least once. This led to a significant ceiling effect, greatly reducing the predictive power of the *items seen* measure. In a real learning situation it is quite likely that many of the learners will reach this ceiling. The *final average alpha* was not affected by this problem. The correlations for the final experiment were somewhat lower than in the pilot (as well as having a slightly different regression line). However, the more elaborate setup of the study introduces a lot more variability, so this is to be expected. Given the evidence, this thesis has shown that the study difficulty experienced by an individual can be predicted quite accurately.

This creates some exciting possibilities. For one, the amount of time spent learning could be tailored to students: when the *final alpha* drops low enough the long term retention is sufficient to pass a test. Perhaps tests might even be obsolete; judging a student by his or her *final alpha* (and thus study time) corresponds almost directly to judging by test performance. Of course values that lie around the pass threshold would require special attention, but this might significantly lower the workload of educators. However, the two studies show that the interpretation of the predictor changes under different circumstances. So before the earlier mentioned claims can be substantiated, more research would need to be done into the relationship between the *final alpha* and other factors, such as the number of sessions and the type of material that is learned. As it stands, the two studies do not provide enough information to adequately interpret the predictor under such varying circumstances.

5.4 Future Research Directions

One way to further explore learning capacity is an experiment where the predictor is used to determine the study session length. An example of such a study would be a setup with two study sessions: for the first session the study time would be equal for all participants, while for the second session total study time is estimated per participant. The expectation is that the scores on the items of the second session would be closer together, if learning capacity can be adequately assessed.

Another possible follow-up to the work in this thesis would be to combine aspects of free selection Flashcard and IRS spacing: giving participants the option to drop old items from study at the start of (or during) each new session, after they have answered the item correctly. If one knows the item at that point, it is unlikely to be forgotten during the study session. This would make the model somewhat less conservative in presenting new items, without being over confident in the learning ability of the participant, as seen with free selection Flashcard.

5.5 Final Thoughts

What is left to be said? The goal set out in this thesis has not only been achieved, but exceeded: the value of cognitive modelling in learning has been shown not only in performance gain, but also in the new insights provided through it. A digital learning environment which can estimate individual behaviour in order to optimize learning seems more important than ever, in an age where information is crucial and time is short. However, educational systems have largely been ignoring (psychological) effects which are beneficial for memory. Such effects are perceived as incompatible with typical educational structures. Hopefully the work presented here has shown that this need not be the case.

BIBLIOGRAPHY

Anderson, J. R., D. Bothell, M. D. Byrne, S. Douglass, C. Lebiere, and Y. Qin (2004). An integrated theory of the mind. Psychological Review 111(4), 1036–1060. (Cited on pages 1, 4, and 9.)

Anderson, J. R., J. M. Fincham, and S. Douglass (1999). Practice and retention: A unifying analysis. Journal of Experimental Psychology: Learning, Memory, and Cognition 25, 1120–1136. (Cited on pages 18 and 19.)

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

Atkinson, R. C. (1972). Optimizing the learning of a second language vocabulary. Journal of Experimental Psychology 96(1), 124–129. (Cited on page 5.)

Baddeley, A. (2003). Working memory and language: An overview. Journal of Communication Disorders 36, 189–208. (Cited on page 14.)

Byron, L. and M. Wattenberg (2008). Stacked graphs – geometry and aesthetics. http://www.leebyron.com/else/streamgraph/. (Cited on page 57.)

Carrier, M. and H. Pashler (1992). The influence of retrieval on retention. Memory and Cognition 20, 632–642. (Cited on pages 2, 11, and 36.)

Cepeda, N. J., E. Vul, D. Rohrer, J. T. Wixted, and H. Pashler (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. Psychological Science 19, 1095–1102. (Cited on pages 2 and 6.)

Coleman, M. and T. Liau (1975). A computer readability formula designed for machine scoring. Journal of Applied Psychology 60, 283–284. (Cited on page 16.)

Cowan, M. (2000). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. Behavioural and Brain Sciences 24, 87–185. (Cited on page 14.)

D'Alessandro, D., P. Kingsley, and J. Johnson-West (2001). The readability of pediatric patient education materials on the world wide web. Archive of Pediatric Adolescent Medicine 155, 807–812. (Cited on page 17.)

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. Commun. ACM 7(3), 171–176. (Cited on page 29.)

Dempster, F. (1989). Spacing effects and their implications for theory and practice. Educational Psychology Review 1, 309–330. 10.1007/BF01320097. (Cited on page 7.)

Dempster, F. N. (1988). The spacing effect. American Psychologist 43, 627–634. (Cited on page 7.)

Donovan, J. J. and D. J. Radosevich (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), 795–805. (Cited on page 2.)

Ebbinghaus, H. (1885). Memory: A contribution to experimental psychology. Translated by Henry A. Ruger and Clara E. Bussenius (1913). (Cited on page 1.)

Flesch, R. (1948). A new readability yardstick. Journal of Applied Psychology 32, 221–233. (Cited on page 16.)

Glenberg, A. M. (1979). Component-levels theory of the effects of spacing of repetitions on recall and recognition. Memory & Cognition 7(2), 95–112. (Cited on pages 2 and 4.)

Gunning, R. (1969). The fog index after twenty years. Journal of Business Communication 6(2), 3–13. (Cited on page 16.)

Hintzman, D. L. (1974). Theoretical impications of the spacing effect. In Theories in cognitive psychology: the Loyola symposium, pp. 77–99. (Cited on page 2.)

Janiszewski, C., H. Noel, and A. G. Sawyer (2003). A meta-analysis of the spacing effect in verbal learning: Implications for research on advertising repetition and consumer memory. Journal of Consumer Research 30(1), 138–149. (Cited on page 2.)

Karpicke, J. and H. Roediger (2008). The critical importance of retrieval for learning. Science 319(5865), 966–968. (Cited on pages 3, 6, 54, and 58.)

Koelewijn, L. (2010). Optimizing fact learning gains: Using personal parameter settings to improve the learning schedule. Master's thesis, University of Groningen. (Cited on pages 6, 14, and 15.)

Kornell, N. and R. A. Bjork (2008). Optimising self-regulated study: The benefits-and costs-of dropping flashcards. Memory 16(2), 125–136. (Cited on page 29.)

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady 10, 707–710. (Cited on page 29.)

Lindsey, R., M. Mozer, N. J. Cepeda, and H. Pashler (2009). Optimizing memory retention with cognitive models. In Proceedings of the 9th International Conference on Cognitive Modeling. (Cited on page 2.)

McLaughlin, G. (1969). Smog grading: a new readability formula. Journal of Reading 12, 639–646. (Cited on page 16.)

Mozer, M. C., H. Pashler, N. Cepeda, R. Lindsey, and E. Vul (2009). Predicting the optimal spacing of study: A multiscale context model of memory. Advances in Neural Information Processing Systems 22, 1321–1329. (Cited on page 5.)

Pashler, H., P. Bain, B. Bottge, A. Graesser, K. Koedinger, M. McDaniel, and J. Metcalfe (2007). Organizing Instruction and Study to Improve Student Learning (NCER 2007-2004). Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education. (Cited on page 1.)

Pashler, H., D. Rohrer, N. Cepeda, and S. Carpenter (2007). Enhancing learning and retarding forgetting: Choices and consequences. Psychonomic Bulletin & Review 14(2), 187–193. (Cited on pages 2 and 3.)

Pavlik, P. I. (2005). The microeconomics of learning: Optimizing paired-associate memory. Ph. D. thesis, Carnegie Mellon University. (Cited on page 25.)

Pavlik, P. I. and J. R. Anderson (2003). An act-r model of the spacing effect. In Proceedings of the Fifth International Conference of Cognitive Modeling, pp. 177–182. (Cited on page 19.)

Pavlik, P. I. and J. R. Anderson (2005). Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect. Cognitive Science 29(4), 559–586. (Cited on pages 2, 4, 5, 10, and 18.)

Pavlik, P. I. and J. R. Anderson (2008). Using a model to compute the optimal schedule of practice. Journal of Experimental Psychology: Applied 14(2), 101–117. (Cited on pages 3, 5, 23, 32, and 40.)

Pavlik, P. I., T. Bolster, S. Wu, K. R. Koedinger, and B. MacWhinney (2008). Using optimally selected drill practice to train basic facts. In Proceedings of the 9th International Conference on Intelligent Tutoring Systems, pp. 593–602. (Cited on pages 3, 5, 8, 18, 23, 24, 25, 26, 40, and 55.)

Raaijmakers, J. G. W. (2003). Spacing and repetition effects in human memory: application of the sam model. Cognitive Science 27(3), 431–452. (Cited on pages 2, 4, and 5.)

Reichle, E. D., K. Rayner, and A. Pollatsek (2006). E-z reader: A cognitive-control, serial-attention model of eye-movement control during reading. Cognitive Systems Research 7, 4–22. (Cited on page 16.)

Smith, E. A. and J. P. Kincaid (1970). Derivation and validation of the automated readability index for use with technical materials. Human Factors: The Journal of the Human Factors and Ergonomics Society 12(8), 457–464. (Cited on page 16.)

Spache, G. (1953). A new readability formula for primary-grade reading materials. The Elementary School Journal 53(7), 410–413. (Cited on page 16.)

Staddon, J. E. R., I. M. Chelaru, and J. J. Higa (2002). Habituation, memory and the brain: the dynamics of interval timing. Behavioural Processes 57(2-3), 71–88. (Cited on page 5.)

Van Rijn, H., L. Van Maanen, and M. Van Woudenberg (2009). Passing the test: Improving learning gains by balancing spacing and testing effects. In A. Howes, D. Peebles, and R. Cooper (Eds.), Proceedings of the 9th International Conference on Cognitive Modeling. (Cited on pages 2, 3, and 12.)

Van Thiel, W. (2010). Optimize learning with reaction time based spacing: By modifying the order of items in a learning session. Master's thesis, University of Groningen. (Cited on pages 1, 6, 11, 12, 14, 15, 28, and 55.)

Van Woudenberg, M. (2008). Optimal word pair learning in the short term: Using an activation based spacing model. Master's thesis, University of Groningen. (Cited on pages 1, 6, 11, 12, 15, 28, and 55.)

APPENDIX A

PILOT EXPERIMENT FACT LIST

Table A.1: Stimuli used in the pilot experiment.

Question	**Answer**
The site of protein synthesis in the cell.	Ribosomes
Very small cells that remove waste material.	Microglia
A drug that blocks the effects of the enzyme acetaldehyde dehydrogenase by binding to its copper ion.	Anatabuse
Brain structure that controls breathing, heart rate, vomiting, coughing, and other vital reflexes.	Medulla
A row or layer of cell bodies separated from other cell bodies by a layer of axons and dendrites.	Lamina
A group of subcortical structures including the caudate, putamen, and globus pallidus.	Basal Ganglia
Decreased activity of surviving neurons after other neurons are destroyed.	Diaschisis
Theory which states that humans have three different types of cones, each sensitive to different wavelengths.	Trichromatic theory
Condition in which the eyes do not point in the same direction.	Strabismus
The perception of frequency of a sound wave.	Pitch
Frequent or constant ringing in the ear.	Tinnitus
A hormone which causes the kidneys, salivary glands, and sweat glands to conserve sodium.	Aldosterone
Estradiol and several other steroids.	Estrogens
Y-shaped proteins which circulate in the blood and attach specifically to one kind of antigen.	Antibodies
The notion that all parts of the cortex contribute equally to complex behaviors like learning.	Equipotentiality
The mostly magnocellular pathway that leads to the parietal cortex and is associated with integrating vision and movement.	Dorsal Stream
The rough equivalent to attention.	Consciousness
Agreement.	Concordance
A drug that blocks NMDA glutamate receptors.	Phencyclidine
Difficulty recalling the names of objects.	Anomia
A section of the temporal cortex that is larger in the left hemisphere in approximately 65% of the population.	Planum temporale

Continued on the next page.

Continued from the previous page.

Question	Answer
The effect that only activated synapses become strengthened.	Specificity
Brain damage caused by prolonged thiamine deficiency.	Korsakoffs syndrome
Amount of release and resynthesis of a neurotransmitter by presynaptic neurons.	Turnover
A protein that binds with estrogen and keeps it from entering cells during the prenatal period.	Alpha-fetoprotein
Sugar found in milk.	Lactose
Area located around the third ventricle which is responsible for detecting osmotic pressure.	OVLT
A combination of EEG and eye-movement records, used in sleep research.	Polysomnograph
Proprioceptors located in the tendons at opposite ends of muscles.	Golgi tendon organ
Sharp, high-amplitude waves followed by a smaller, positive wave.	K-complexes
A structure in the pons that is inactive at most times but emits impulses in response to meaningful events.	Locus coeruleus
Decreased response to a stimulus as a result of recent exposure to it.	Adaptation
Receptors which detect sudden displacements or high-frequency vibrations on the skin.	Pacinian corpuscles
Central portion of the retina specialized for acute, detailed vision.	Fovea
The parvocellular and magnocellular pathways sensitive to shape, movement, and color-brightness.	Ventral Stream
Programmed cell death.	Apoptosis
A chemical that promotes the survival and activity of neurons.	Neurotrophin
Removal of a brain area.	Ablation
Thin membranes that surround the brain and spinal cord.	Meninges
A large structure between the thalamus and the cerebral cortex, important for new memory storage.	Hippocampus
Posterior part of the brain, consisting of the medulla, pons, and cerebellum.	Hindbrain
A rapid depolarization and slight reversal of the usual membrane polarization.	Action potential
Type of glia that builds the myelin sheaths around certain neurons in the periphery of the body.	Schwann cells
Chains of amino acids.	Neuropeptides

Continued on the next page.

Continued from the previous page.

Question	**Answer**
Short unmyelinated sections on a myelinated axon.	Nodes of Ranvier
Different name for vitamine B1.	Thiamine
A period immediately after an action potential.	Refractory period
Branching fibers that extend from the cell body.	Dendrites

APPENDIX B

FINAL EXPERIMENT FACT LIST

Table B.1: Stimuli used in the pilot experiment.

Question	**Answer**
Formation of synapses.	Synaptogenesis
Number of compressions per second.	Frequency
A predominantly female hormone.	Progesterone
Inability to breathe during sleep.	Sleep Apnea
A protuberance on the surface of the brain.	Gyrus
Precursors of the male reproductive organs.	Wolffian ducts
A blurring of vision for lines in one direction.	Astigmatism
Condition in which a child ignores the vision in one eye.	Amblyopia
Comprised of the superior colliculus and inferior colliculus.	Tectum
Ballistic eye movements from one fixation point to another.	Saccades
Stimulus that is necessary for resetting the circadian rhythm.	Zeitgeber
An opening in the center of the iris in which light enters the eye.	Pupil
The area that surrounds the immediate damaged area of the brain.	Penumbra
A hormone released by the pineal gland, which increases sleepiness.	Melatonin
Brain waves with a frequency of about 8-12 brain waves per second.	Alpha waves
The most anterior portion and most prominent part of the human brain.	Forebrain
Individuals whose development is intermediate between male and female.	Intersexes
Decreased response to a stimulus as a result of recent exposure to it.	Adaptation
Brain structure that organizes sensory information that guides movement.	Cerebellum
Cells which recieve information from bipolar cells and relay it to others.	Amacrine cells
A part of the reticular formation that contributes to cortical arousal.	Pontomesencephalon

Continued on the next page.

Continued from the previous page.

Question	Answer
A disorder characterized by frequent unexpected periods of sleepiness during the day.	Narcolepsy
The elimination of sensory nerve impulses by destroying the sensory nerve fibers.	Deafferentation
Nerve which carries information to the brain regarding the stretching of stomach walls.	Vagus nerve
The sense of smell.	Olfaction
Damage to a brain area.	Lesion
Chains of amino acids.	Neuropeptides
Adenosine and several of its derivatives.	Purines
The process of neurotransmitter release.	Exocytosis
Precursor to female reproductive organs.	Mullerian ducts
Internal rhythms that last about a day.	Circadian rhythm
Tiny nearly spherical packets storing neurotransmitters.	Vesicles
Midbrain structure that contains dopamine neurons.	Substantia Nigra
Rear surface of the eye which is lined with visual receptors.	Retina
A class of neuroleptic drugs that includes haloperidol.	Butyrophenones
A type of stroke; bleeding due to the rupture of an artery.	Hemorrhage
The belief that mind and body are different kinds of substances.	Dualism
A type of glia that helps synchronize the activity of neurons.	Astrocytes
Substance which increases dopamine release from presynaptic terminals.	Amphetamine
Change over generations in the frequencies of various genes in a population.	Evolution
The inability to recognize faces without an overall loss of vision or memory.	Prosopagnosia
A fold or groove that separates one protuberance on the surface of the brain from another.	Sulcus
A technique that measures changes in the blood's hemoglobin molecules as they release oxygen.	fMRI
Structure which vibrates at the same frequency as the sound waves that strike it.	Tympanic membrane
Making up an answer to a question and accepting the invented answer as if it were true.	Confabulation
The ability to respond in some way to visual information after extensive damage to area V1.	Blindsight
A set of axons that allows the two hemispheres to exchange information with one another.	Corpus callosum
Type of glia that builds the myelin sheaths around certain neurons in the brain and spinal cord.	Oligodendrocytes
White blood cells.	Leukocytes

Continued on the next page.

Continued from the previous page.	
Question	**Answer**
Specialized gaps between neurons.	Synapses
Physical representation of learning.	Engram
Testosterone and several other steroids.	Androgens
The perception of intensity of a sound wave.	Loudness
Theory proposed to account for color constancy.	Retinex theory
Brain structure that lies anterior and ventral to the medulla.	Pons
Inability to read despite adequate vision and intelligence.	Dyslexia
A long, thin fiber which is the information-sending part of the neuron.	Axon
First drug used successfully for the treatment of schizophrenia.	Chlorpromazine
A receptor that detects the position or movement of a part of the body.	Proprioceptor
A cluster of neuron cell bodies, usually outside the Central Nervous System.	Ganglion
A sharp blow to the head that does not actually puncture the brain.	Closed head injury
Insulating material composed of fats and proteins found on some vertebrate axons.	Myelin
Presynaptic receptors sensitive to the same neurotransmitter they release.	Autoreceptors
Midway point of the menstrual cycle when sexual interest increases.	Periovulatory period
Inability to determine where objects are moving or even if they are moving.	Motion blindness
Part of the cortex which responds to sensory signals that lead to movements.	Prefrontal cortex
A device that records electrical activity of the brain through electrodes.	Electroencephalograph
Individuals whose genitals do not match the normal development for their genetic sex.	Hermaphrodites
Severe neurological disorder with symptoms that include twitches, tremors, and writhing movements.	Huntington
Effect that pairing a weak input with a strong input enhances later responses to the weak input.	Associativity
Effect that nearly simultaneous stimulation by two or more axons produces long-term potentiation.	Cooperativity
Area which is a small part of the frontal lobe of the left cerebral cortex that when damaged leads to language impairments.	Broca