Making the most of human memory

Sense, Florian

Making the most of human memory

Studies on personalized fact-learning and visual working memory

Supervisors
Prof. D. H. van Rijn
Prof. R. R. Meijer

Assessment committee
Prof. M.C. Mozer
Prof. P.C.J. Segers
Prof. N.A. Taatgen

Cover design : Florian Sense
Layout design : Briyan B Hendro *(briyan.2209@gmail.com)*
Printed by : Ipskamp Printing

ISBN (book) : 978-90-367-9719-1
ISBN (ebook) : 978-90-367-9718-4

university of groningen

Making the most of human memory

Studies on personalized fact-learning and visual working memory

PhD thesis

to obtain the degree of PhD at the
University of Groningen
on the authority of the
Rector Magnificus Prof. E. Sterken
and in accordance with
the decision by the College of Deans.

This thesis will be defended in public on

Thursday 20 April 2017 at 16.15 hours

by

Florian Sense

born on 24 January 1988
in Bremen, Germany

CONTENTS

1

General Introduction

The brain is often thought of and talked about as functioning like a computer. In this metaphor, memory is clearly the hard drive. Many interesting aspects of human memory are revealed by understanding why this metaphor is fundamentally flawed. Computers save files to a specific location on the drive such that storage space decrease with each file that is saved. In contrast, human memory actually gets bigger because information is stored in terms of relationships to existing knowledge. The more knowledge is already encoded, the easier it becomes to add more to it. That is why it is easier to learn a new word in a language one already speaks – such as *gorgonize* in English[1] – than a single word in a language one is not familiar with. However, the amount of information that can be encoded in a human brain is so large that it is practically infinite (e.g., Bartol et al., 2015). Furthermore, computers treat all files equally: how long it takes to open a file with a name in it does not depend on the name. Human memory, on the other hand, is subject to lots of biases: your mother's name will come to mind much easier than the name of the kid two rows ahead in fourth grade (LaBar & Cabeza, 2006).

One aspect in which human and computer memory are comparable is that we can distinguish between three distinct stages: encoding, storage, and retrieval (Tulving, 1995). Functionally, however, the three stages are fundamentally different. A computer would be useless if it did not encode, store, and retrieve data with perfect accuracy. And because of the location-specific and content-independent storage in a computer, saving a thousand copies of a French word is not different from saving a thousand different French words. Saving an additional copy has no effect on the existing copies and retrieving a copy does not alter any other copy. Not so in human memory. Storage is inherently interconnected and a thousand repetitions of the same French word would strengthen the neural representation of the concept, not create a thousand copies of it. Furthermore, retrieving information from memory changes that memory. A memory will be easier to retrieve if it has been retrieved recently (van den Broek et al., 2016). This is why retrieval practice is a more effective study strategy than re-reading (Karpicke & Roediger, 2008). Conversely, related memories that rely on the same cues might become less available – a phenomenon referred to as retrieval-induced forgetting (M. C. Anderson, Bjork, & Bjork, 2000, 1994). And even though the originally encoded information might become unavailable, human memory often fills in the gaps and supplies "false memories" (e.g., Brewin & Andrews, 2016; Otgaar, Merckelbach, Jelicic, & Smeets, 2016). And more worryingly, people often report inaccurately recalled information with high confidence (e.g., Busey, Tunnicliff, Loftus, & Loftus, 2000).

If a computer routinely provided wrong files without appropriate warning messages, we

[1] verb – gor·gon·ize – \ˈgȯr-gə-ˌnīz\ – to have a paralyzing or mesmerizing effect on

would deem it useless, throw it out, and replace it with reliable hardware. We do not have this privilege with our wetware, though. A key difference between the two systems is that computers are intelligently designed while memory evolved. As such, memory is adapted to the properties of the environment and while forgetting might seem like a nuisance, it can be adaptive (Kraemer & Golding, 1997). The world constantly changes and remembering all previous instances – for example, all the places you have ever parked your car – would greatly interfere with what is currently relevant (e.g., Unsworth, Brewer, & Spillers, 2013). Some aspects of the environment follow regularities, though, and it has been argued that memory evolved to mirror these regularities (an approached known as rational analysis; e.g., J. R. Anderson & Milson, 1989; J. R. Anderson & Schooler, 1991; Pachur, Schooler, & Stevens, 2014; Stevens, Marewski, Schooler, & Gilby, 2016).

Interestingly, memory has evolved to be seemingly infinite yet severely limited. The amount of autobiographical events we can recount in vivid detail is astonishing, we can sing along to hundreds of songs, can recall countless facts, and have learned hundreds of names and faces throughout our lives. And yet it is impossible for most of us to keep more than a couple of digits in working memory at the same time (Cowan et al., 2005). Making the most of human memory requires understanding "important peculiarities" of the storage and retrieval processes (Bjork & Bjork, 1992) and to be aware of its weaknesses and exploit its strengths (see Bjork, Dunlosky, & Kornell, 2013 for an excellent review).

Such understanding – just like memory – has many different facets. All investigations into the various parts of memory strive to develop theories that give coherent accounts of the complex behavioral patterns we observe (e.g., Gabrieli, 1998; Wixted, 2004). All theories have competitors and rigorously testing theories' assumptions and predictors is the core of the scientific method. The more tests a theory passes, the more confidence we put in the theory's predictions. And even though the brain-as-a-computer metaphor is not helpful for memory, implementing theories as computational models can be an excellent way to test assumptions and generate predictions.

The first part of this thesis applies the predictions of a prominent theory of declarative memory to improve learning and the second part of this thesis tests a specific assumption prevalent in the visual working memory literature.

Part 1: Studies on personalized fact-learning

An adaptive fact-learning system is a computerized learning environment in which learners can study a set of facts in a way that is tailored towards their particular strengths and weaknesses. Ideally, such a system can estimate – in real time – how well each learner has

mastered each fact in the set and adjust the order in which facts are presented for review dynamically. The goal is to adapt to the learners' needs and provide a personalized learning experience that yields better long-term retention of the studied material than a one-size-fits-all approach would. The crux, of course, is the way the computerized system probes the current state of the learner's memory to learn how well individual facts have been mastered and how that information is used to devise a personalized learning schedule on the fly.

Chapter 2 opens with the main motivation behind developing an adaptive fact-learning system that incorporates response latency instead of relying solely on the accuracy of responses: If only accuracy is used to gauge how well each fact has been learned, a fact-learning system has to ensure a sufficient number of errors to assess how well facts are known. This need for errors is greatly reduced if an alternative measure of memory strength can be used. Response latency can serve as such a measure. After outlining the adaptive fact-learning system that incorporates both accuracy and response latency, this chapter presents the results from an experiment demonstrating that the model outperforms a study method often used by students: flashcards. More specifically, the model is pitched against a "smart" flashcard system – essentially a control condition with high face validity – in a within-subject design: each participant studied a set of foreign vocabulary with both methods. The statistical analyses confirm that many participants benefit from studying with the adaptive system relative to the flashcards, especially low-performing learners.

Chapter 3 zooms in on the model parameter that makes the adaptive learning system adaptive. The theoretical framework is presented again to highlight the crucial role that the model's decay parameter plays in tracking the memory strength of each fact during learning. The reported experiment was designed to shed light on a specific question with regard to the model's parameter: if a parameter is estimated for a given participant while learning one type of material, how reliable is that estimate? This question has two important dimensions: how stable is the parameter over time – that is, when the parameter is estimated while learning one type of material this week, how similar will a parameter estimate be when estimated from the same type of material a week later? – and how stable is the parameter over materials – that is, how similar will parameter estimates for the same person be between different types of materials. The main finding is summarized in the title of the article that this chapter was published as: "An individual's rate of forgetting is stable over time but differs across materials".

In previous studies, we have observed a very high correlation between the parameters extracted from the model at the end of the learning session and delayed recall performance on the studied material. This is both interesting and potentially very useful because it implies that we can predict how well a learner will perform on a delayed test purely based on their

performance while studying the material. We were wondering whether the model parameter we estimated during learning truly reflected a marker of someone's ability to encode and retrieve the studied material or whether the parameter was merely an artifact of a more general property of the learner. Therefore, we set out to test whether the estimated *rate of forgetting* of a learner is related to two commonly used and validated measures of cognitive functioning: working memory capacity and general cognitive ability. That is, while Chapter 3 focuses on the "internal consistency" of the estimated model parameter, Chapter 4 looks at the relationship between the model's parameter with two "external" constructs. Two important conclusions can be drawn from the data presented in Chapter 4: first, we replicate earlier observations and find a very high, negative correlation between the estimated *rate of forgetting* and subsequent delayed recall performance; both with a delay of 80 minutes and three days. This implies that the model parameter is a good indicator of future test performance. And second, we show that the estimated *rate of forgetting* is not significantly correlated with either working memory capacity or general cognitive ability. Furthermore, the analyses presented in Chapter 4 show convincingly that in the tested sample, neither of the two "external" measures predicts significant amounts of variance in delayed recall and that the model's parameter is the single best predictor of future test performance.

In summary, the first part of the thesis introduces an adaptive, theory-based fact-learning system. We show that the system works better than a method that many students report using: flashcards. Then the parameter underlying the adaptive mechanism at the core of the model investigated more closely and we find that it is stable over time within a learner, strongly related to delayed recall of studied material, and not related to either working memory capacity or general cognitive ability. Overall, the work report in Part 1 of the thesis implies that the *rate of forgetting* can be estimated reliably, can distinguish between learners that are likely to have mastered the studied material and those that are not, and that it is an independent, useful measure of individual differences in the context of an adaptive learning environment.

Part 2: A study on visual working memory

We perceive the world through different modalities and can keep incoming information in working memory for short durations. One on-going discussion in cognitive psychology is the nature of working memory as a whole and how to estimate specific parts of it. (Whether those parts even exist is also part of the discussion.) One potential problem we face if we want to study *visual* working memory is that participants in our experiments might see, say, a red square but instead of remembering the red square as a visual percept, they might *recode* it by rehearsing, quietly to themselves, "red square, red square, red square". A potential problem

with this approach – especially in the framework of the influential multi-component model of working memory – is that participants are effectively using both their visual and their verbal memory resources to store the stimulus. This makes it impossible to get a clean estimate of the part of someone's memory that is purely visual.

The way to fix this problem is usually to enforce what is known as articulatory suppression: the participant is required to repeat out loud a series of non-sense syllables so that their verbal resources cannot be recruited to store the percept of the red square and they have to rely solely on their visual working memory. To the dismay of both participants and researchers, the use of articulatory suppression is common practice – it is inconvenient and awkward for everyone involved.

Various studies cast doubt on whether the distinction between verbal and visual resources is meaningful and, by extension, whether articulatory suppression is necessary. In Chapter 5, we report findings from an experiment specifically designed to test this idea. We had participants perform a visual change detection task to assess their visual working memory capacity. Each participant performed the task under different conditions that were designed to make it either very easy to recruit potential verbal resources (stimuli were presented sequentially and no articulatory suppression was required) or very hard to recode the visual material (stimuli were presented simultaneously and articulator suppression *was* required).

The collected data was subjected to a Bayesian state-trace analysis and were found to provide very strong evidence in favor of the statistical model that assumes only a single underlying resource. That is, the data presented in Chapter 5 strongly suggest that articulatory suppression is not necessary to obtain an unbiased estimate of a participant's visual working memory capacity. By extension, this implies that – at least in the setup used here – it is unlikely that multiple working memory components are at play.

We hope that fewer researchers in the future will be required to listen to a participant mumbling "da da da da" for extended periods of time.

Taken together, the studies in Part 1 demonstrate that the adaptive fact-learning model outperforms a traditional flashcard method. The parameter extracted from the model is stable over time and is not merely an artifact of attentional control and cognitive functioning. In the final chapter, we will discuss in more detail how the estimated rate of forgetting relates to delayed recall performance, how useful it is as an individual differences measure, and potential future developments of the model. The work presented in Chapter 5 suggests that articulatory suppression is not necessary to obtain unbiased estimates of visual working memory through visual change detection tasks.

2

The Theoretical and Practical Advantage of an Adaptive Fact-Learning System

Acknowledgements

This chapter was written in collaboration with Hedderik van Rijn.

We thank Charlotte Schlüter for her help with data collection and Michael LeKander for his help with programming the experiment.

Supplementary materials for this chapter are available at www.osf.io/3g6pq

Abstract

To learn effectively, students should exploit techniques that are known to enhance retention, such as the spacing and retrieval practice. However, self-reported study behavior often does not include such techniques. Computerized learning systems can incorporate effective techniques and help students to use their study time efficiently by tailoring the session to their needs based on information gathered during study. To determine whether such adaptive methods should be used instead of those commonly employed by students – primarily flashcards – requires testing their effectiveness relative to current methods. Here, we present data from a within-subject experiment contrasting an adaptive method with a traditional flashcard method. Statistical analyses reveal that students generally benefit from using the adaptive method, both in their performance during study and on subsequent tests of the studied material. The differences in the methods resulting in this benefit are discussed in detail and possible extensions of the current work are proposed.

Ever since the earliest systematic investigations of learning, we have known of a phenomenon referred to as the spacing effect: we learn more material when study time is distributed over multiple sessions (see Dempster, 1988; Donovan & Radosevich, 1999; Rohrer, 2015 for reviews). The optimal schedule for multiple study sessions depends on the time between the last study session and the moment of the test (Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008; Mozer & Lindsey, 2016). Surveys showing that students gravitate towards massing their practice (e.g., Taraban, Maki, & Rynearson, 1999) suggest that the benefit of the spacing effects is not exploited by students (Dempster, 1988).

In most studies of the spacing effect, a test is administered to assess retention of studied material after a delay. However, it was discovered early on that testing itself improves learning (Gates, 1917; Spitzer, 1939), a phenomenon that is still a highly relevant topic of research today (Karpicke & Roediger, 2008). This testing-induced benefit on long-term retention is referred to as the testing effect (van den Broek et al., 2016). The testing effect is most pronounced if retrieval attempts are successful (Carrier & Pashler, 1992) but no benefit is observed for non-retrievable items (Jang, Wixted, Pecher, Zeelenberg, & Huber, 2012). Furthermore, testing is especially effective when retrieval is difficult (Gardiner, Craik, & Bleasdale, 1973; Karpicke & Roediger, 2007; Pyc & Rawson, 2009). These constraints make clear when the optimal time for a testing event would be: not too close to the initial learning (to maximize spacing and produce "desirable difficulties", Bjork, 1994) but not too long after initial learning either to ensure that items are still retrievable.

Determining the optimal moment for a repetition is therefore dependent on knowing how well something has been learned (i.e., how retrievable a certain information is). However, making accurate judgments of learning is difficult, especially if they involve predictions regarding the retrievability of learned information in the future (see Bjork et al., 2013 for a review). Exploiting our theoretical knowledge in practice would also require that students are aware of effective study methods and use them. However, they are often not aware of them (Hartwig & Dunlosky, 2012; Kornell & Son, 2009; McCabe, 2011) and do not use them (Carrier, 2003; Karpicke, 2009; Karpicke, Butler, & Roediger, 2009; Kornell & Bjork, 2008b). For example, McCabe (2011) reports that students believe re-reading is a more effective study strategy than self-testing and Kornell (2009) demonstrates that even though spacing produced better learning outcomes than massing for 90% of the participants, 72% of those participants believed the opposite when asked after learning.

One study method that many students report using is flashcards (Wissman, Rawson, & Pyc, 2012), cards containing small bits of information (such as the definition of a term) used as a study and memorization aid. However, flashcards – even if combined with self-testing –

are often not ideal because students tend to opt for spacing intervals that are too short (Pyc & Dunlosky, 2010; Taraban et al., 1999; Wissman et al., 2012). This not only means that they do not take optimal advantage of the spacing effect but it also produces stronger feelings of familiarity (or *retrieval fluency*) with the material (Benjamin & Bjork, 1996; Matvey, Dunlosky, & Guttentag, 2001), which is not necessarily a good indicator of having mastered the material (Jacoby, Kelley, & Dywan, 1989; Kelley & Lindsay, 1993). This will lead a student to forfeit additional repetitions due to familiarity-induced, overconfident judgments of learning (Bjork et al., 2013).

Therefore, it would be desirable to provide students with the opportunity to circumvent biased meta-memory judgments and teach them to self-regulate their learning. One way could be through educating students about effective study techniques (de Boer, Donker, & van der Werf, 2014; Donker, de Boer, Kostons, Dignath van Ewijk, & van der Werf, 2014; Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013). Additionally, students could be supplied with tools to help them schedule the introduction of new material and repetitions of old material. The most promising tool would be computerized learning environments that exploit our theoretical understanding of human memory to devise a learning schedule that is as close to optimal as possible. Achieving this goal requires potential learning environments to be adaptive because the needs of individual learners vary: A schedule that is optimal for a high-performing learner with prior knowledge is not optimal for a low-performing learner with no prior knowledge, for example.

Here, we will focus specifically on scheduling repetitions of items within a single study session because students often report studying the night before a test (e.g., Taraban et al., 1999). Ideally, students would space their learning over multiple session leading up to the test but since they typically do not, providing them with information about how to study effectively (Putnam, Sungkhasettee, & Roediger, 2016) and supplying them with tools that fit their needs will have to suffice.

A promising path towards an adaptive system is via *retrieval practice* (e.g., Roediger & Butler, 2011). Rather than passively re-reading the to-be-remembered material, each repetition during retrieval practice is a test event that requires actively retrieving the answer from memory. This not only boosts retention by exploiting the testing effect but has an additional benefit: The probability of a successful retrieval decreases as a function of time (Wixted & Carpenter, 2007) and through continuous testing, the learner's current knowledge state can be assessed (Mozer & Lindsey, 2016). That is, we can use an observable behavior – a learner's response to a cue – to learn about the (unobservable) memory strength of an item. Many models developed to optimize fact learning use this basic idea by making each trial a test and using the response

to schedule subsequent trials (Atkinson, 1972; Lewis, Lindsey, Pashler, & Mozer, 2010; Mettler, Massey, & Kellman, 2011; Pavlik & Anderson, 2008; Woźniak & Gorzelańczyk, 1994).

However, using only accuracy information from each trial to personalize the scheduling of a learning session poses a potential problem: To discriminate between easy and difficult items, the system has to ensure a sufficient number of errors. Systematically producing errors is not optimal, though, because learners do not benefit from the testing effect if they cannot retrieve the item (Jang et al., 2012) and an opportunity to re-study is necessary after retrieval failure to enable subsequent testing events. The optimal duration of such corrective feedback is about four seconds (de Jonge, Tabbers, Pecher, & Zeelenberg, 2012; Zeelenberg, de Jonge, Tabbers, & Pecher, 2015) – time that could otherwise be spent on additional, more effective testing events. Consequently, errors should be avoided if possible.

Luckily, each trial contains at least one additional piece of information besides accuracy: response latency. The assumption is that there is a link between how quickly a learner can respond to a prompt and how strong the memory trace for that particular item is at that point in time. Strength theory assumes such a link (Murdock, 1985; Norman & Wickelgren, 1969) but the idea is still present in more recent work. For example, Madigan, Neuse and Roeber (2000) report experimental data supporting the link, and converting observed reaction times to estimated memory strength (and estimated memory strength to predicted reaction times) is a core function of ACT-R's declarative memory (J. R. Anderson et al., 2004). With regards to fact learning, Mettler et al. (2011) have shown that an adaptive learning system based on both accuracy and latencies outperforms one that only incorporates accuracy, which suggests that latency provides useful information on a trial-by-trial basis (Mettler & Kellman, 2014). More generally, Van den Broek and colleagues suggest that testing strengthens the cue-response association more than re-studying (van den Broek, Takashima, Segers, Fernández, & Verhoeven, 2013) and, more specifically, that the response latency can serve as a proxy for the strength of that association (van den Broek, Segers, Takashima, & Verhoeven, 2014). Thus, a system that seeks to improve the computation of an optimal schedule of practice might be improved by incorporating both the accuracy and the response latency recorded for each retrieval attempt.

Taken together, it seems likely that theory-based adaptive fact-learning systems could have a number of advantages over methods typically employed by students, particularly flashcards. First, they could maximize retrieval practice. Students report creating flashcards to study but often use self-testing as an assessment tool, not as a study strategy (Hartwig & Dunlosky, 2012; Karpicke et al., 2009; Kornell & Son, 2009; McCabe, 2011). Second, they can keep accurate records of response-related measures such as accuracy and latency rather than

relying on students' subjective judgments of learning or feelings of familiarity. Third, using a theoretical framework that allows incorporating those recorded measures to devise personalize learning schedules circumvents biased judgments of learning and, ideally, can provide unbiased estimations of mastery.

Evaluating whether an adaptive fact-learning system that utilizes these principles actually produces better learning outcomes requires comparing it against a criterion. Relatively smart flashcard systems provide a good benchmark because they have high face validity. Demonstrating experimentally that a candidate system produces better results than the default study method of many students (e.g., Wissman et al., 2012) provides a natural threshold for calling an adaptive method effective.

Here, we present data from a within-subject experiment that directly compares an adaptive system that makes use of response latencies with a traditional flashcard system[1]. Specifically, we will test how studying with the adaptive system affects the proportion of correct responses during learning and on two subsequent tests. In other words, how using the adaptive system affects both the process and the outcome when studying with the adaptive system rather than flashcards.

METHODS

Study Methods

Participants studied word pairs with two methods: an adaptive method and digital flashcards. It is important to note that the two methods only differed in the scheduling of the items (discussed in the next sections) but otherwise looked the same. We will describe these methods in detail[2] because they determine the order in which participants studied and repeated items during the 20-minute learning sessions. Both methods used the same presentation procedures. The first encounter of an item was always a *study trial* that featured the Swahili word along with the English translation and an input field. The participant had to type in the English

[1] Technically, both of these systems could be called flashcard systems. The graphical interface does not differ between the systems and the presentation style is akin to digital flashcards. What differs is only the algorithm determining the order of repetitions. We refer to them as "the adaptive system" and "flashcard system" because the emphasis in the former is on the adaptive nature of the underlying algorithm, while the emphasis in the latter is emulating traditional flashcards as often used by students.

[2] This description is largely taken from a manuscript that is currently in preparation and will combine the experiment reported in this chapter along with the one reported in Chapter 4 and an additional study conducted under realistic conditions in high school classrooms.

translation and press Enter to proceed. All subsequent repetitions of that particular item were presented on *test trials* that presented the Swahili word and showed an input field but not of the English word. If a correct response was entered, the test trial was followed by a 600 ms feedback screen displaying "Correct!". If the response was incorrect, a study trial without the input field was presented for four seconds along with the feedback "Incorrect!".

The Adaptive Method. The adaptive method is based on the ACT-R theory's declarative memory (J. R. Anderson, 2007). In this framework, each studied word pair is encoded in memory and has a memory trace associated with it. The trace's activation changes as a function of usage: Activation will be high if the word pair has been retrieved recently and decays as a function of time and previous encounters. The decay ($d_{i,j}$) of a particular trace (j) of an item (i) depends on an item's activation (A_i) at the time of the trace's instantiation (t_j), modulated by a scaling parameter (c, fixed at .25) plus a constant intercept (α):

$$d_{i,j} = ce^{A_i(t_j)} + \alpha_i$$

On the first encounter of each word pair, the decay will be identical to the intercept α (set to 0.3) because the activation is set to -∞ initially. This equation implies that if a study item (i) is highly activated (i.e., has a relatively high A_i) when a presentation of that study item is encountered (t_j), the new trace (j) will be associated with a high decay value ($d_{i,j}$). On the other hand, if activation is low, decay will also be relatively low. As activation is assumed to decrease with time, increased spacing will result in lower decay values, potentially resulting in better later recall (Pavlik & Anderson, 2005). The combination of decay values and the distribution of previous presentations (n) of a study item over time (t_j) determine the activation value of that study item at time t:

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

As explained in the Introduction, the goal of an adaptive learning environment should be to estimate – for every learner, for every item they study – the optimal time when a memory query yields maximum benefits in terms of the spacing and testing effects. This activation-based framework provides an opportunity to do exactly that: The system keeps track of the estimated activations of all word pairs that have been presented in the current session. At the start of each trial, the system assesses if any word pairs have an activation below the retrieval threshold (τ, set to -0.8) or an activation that is estimated to drop below the retrieval threshold during the next presentation (assuming a presentation duration of 15 seconds). If so, the system selects the word pair with the lowest activation. If the activation of all word pairs is expected to be above the retrieval threshold, a new word pair is chosen randomly and presented to the participant.

Although the activation of a memory trace for a given word pair is a theoretical construct, activation can be mapped to an observed variable as Anderson et al. (2004) proposed a link between memory activation and response latency: The expected response latency of an item (L_i) at time t can be computed using:

$$L_i(t) = Fe^{-A_i(t)} + \text{fixed time cost}$$

in which F is a scaling parameter set to 1. As we did not model variations in perceptual and motor action required to perceive the stimulus and type the first word, we set the *fixed time cost* parameter to 300 ms for all items and participants – assuming 100 ms for perceptual encoding and 200 ms for the initial motor response (Byrne & Anderson, 1998) – for cues that contain only a single word (as in the experiment reported here). If the cue has more than two words, the *fixed time costs* are (-157.9 + 19.5x), where x is the number of characters in the cue. However, the *fixed time costs* cannot be less than 300 ms.

Every time a response is collected, the system calculates the difference between expected and observed response latency and a binary search is performed to minimize the absolute difference between the predicted and observed response latencies across the last five encounters of the item. The boundaries for the binary search are the current α and the current $\alpha \pm 0.05$ to dampen the effect of extreme differences in expected and observed latency and to ensure a gradual adaptation. If an incorrect or no response was given, the observed response latency is replaced with the latency that would be expected if the activation had been 1.5 times the activation at the retrieval threshold. This dynamic adjustment of the α parameter is done for each item individually.

In summary, the equations outlined here can be used to ensemble an adaptive learning system that estimates the strength of the memory trace associated with each studied item and updates its estimate based on the accuracy and latency of the learner's incoming responses. In practice, the system will start a study session by selecting a random word pair and presenting it on a study trial. On every subsequent trial, the system goes through the following steps: the activation 15 seconds from now is calculated for all items that have already been presented. If any item's estimated activation is expected to be below the retrieval threshold, that item (or the one with the lowest estimated activation if there are multiple) is presented on a test trial. The expected latency will be compared to the observed latency to update the model's parameters. If any time is left, the same procedure starts again. If the estimated activation of all items that have already been encountered is still above the retrieval threshold 15 seconds from now, the model will select a new item randomly and present it on a study trial. If there are no new items left, the model selects the item with the lowest estimated activation regardless of whether the activation is below the retrieval

threshold or not. As a result, the study time can expire before the participant has been introduced to all items in the set.

Flashcards. In the flashcard condition, the list of word pairs was divided into "stacks" of five. All items in a stack were presented a single time on study trials, one after another. Next, items from the same stack were presented again on test trials. If a response on a test trial was incorrect, the associated word pair was added to the end of that stack. Correctly answered items were removed from the current stack. The system only commenced to the next stack of five after all word pairs in the current stack had been correctly responded to once. After all 50 items (i.e., ten stacks) have been completed, the participant started with the first again until the time was up. Each time a new stack was presented, the order of the items in that stack was randomized but the items making up each stack remained the same. Note that this setup results in closely spaced repetitions when a response is incorrect (at most four responses before a repetition), and widely spaced repetitions if a response is correct (at least [total number of items] – [stack size] = 45 trials but probably more assuming that errors were made which would add additional intervening trials).

Participants

A total of 53 participants were recruited through a local social media group and were paid €10 for compensation. One participant was excluded because the wrong tests were administered in both sessions. Therefore, data on the test performance for this participant was not available and the participant was excluded from all analyses.

Of the remaining 52 participants, 36 were female (69%) and the median age was 22.5 (SD_{age} = 2.82; $range_{age}$ = [18, 29]). Most (37) indicated no familiarity with Swahili but 15 indicated minimal familiarity – most likely because they have participated in experiments that used a different sets of Swahili stimuli earlier in the semester. All participants gave informed consent in line with the approval of the study by the Ethics Committee Psychology (ID: 15161-NE+PPP).

Design, Stimuli, and Procedure

Each participant came in for two 90-minute sessions. In the first session, they studied with both the adaptive method and the flashcard method outlined above. The order of the study conditions was counterbalanced between participants using the parity of random participant IDs. The studied stimuli were the 100 Swahili-English word pairs reported in Nelson and Dunlosky (1994). For each participant, the 100 stimuli were randomly split in two sets, one assigned to the adaptive and one assigned to the flashcard method.

At the beginning of the first session, participants signed an informed consent form and completed a short questionnaire regarding demographic information. Then they began the first 20-minute study session of the first random subset of 50 Swahili-English word pairs. The study session was followed by a 15-minute break during which participants played Tetris. The break was followed by a test of the word pairs. The test was a list of all 50 items that participants could have encountered during the study session and they were asked to provide answers to as many as items as possible without a time limit. No feedback was provided. After completion of the test, the participants went through the same study-break-test cycle again, this time studying with the other method.

At the beginning of the second experimental session, exact copies of the two tests taken during the initial 90-minute session were administered. Thus, each participant was tested on all potential items twice. The retention intervals ranged from 3 to 17 days with a mean of 7.5 days (SD = 2.12; 78% of the retention intervals were between 6 and 8 days). Participants spent the rest of the second session completing an unrelated experiment based on a variation of an attentional blink task. The outcomes of this task will not be discussed here.

RESULTS

The main questions were (1) whether the performance on the tests varied as a function of the study method and (2) which method produced fewer errors during learning. Not all items were necessarily studied by every participant (especially in the adaptive condition) but for items that were, we have information regarding the proportion of correct responses to test trials during study. Regardless of whether items were studied, we have information about whether an item was answered correctly on the two tests. To address the two questions, we fit mixed-effects regression models to the item-level data (instead of aggregates per participant) and determined the best-fitting model for each of the two outcome measures. In both analyses, the effect of *study method* is most relevant to evaluate the research questions but other predictor variables and possible interactions were also added to test for their potential influence.

Bayesian statistics offer various conceptual and practical advantages over null-hypothesis testing (e.g., Wagenmakers, Morey, & Lee, 2016) but current software packages do not offer the ability to fit logistic regression models, yet. For the sake of consistency, we will fit both the logistic and the linear mixed-effects regression models using the lme4 package (Bates, Mächler, Bolker, &

Walker, 2015) in R (R Development Core Team, 2016)[3] and approximate Bayes factors for model comparison from the output they provide. The best-fitting models were determined by fitting the full model first (i.e., the model including all predictors and all their interactions) and then removing fixed effects in a step-wise procedure. When removing effects, we started with the highest-order interactions and proceeded to the lower-order interactions and finally the main effects (Gelman & Hill, 2007). If multiple interactions of the same order were present, the one with the lowest absolute *t*- or *z*-value was removed first. At every step of the procedure, the simpler model was evaluated against its more complex predecessor by comparing models' Bayesian information criteria (BIC) to approximate Bayes factors (as advocated by Wagenmakers, 2007; see equation 10 specifically). The Bayes factors express how likely one model is relative to another given the data and provide an intuitive interpretation of the strength of the evidence (Kass & Raftery, 1995). If the Bayes factor for a comparison was inconclusive, the AIC and Chi-squared tests were consulted to determine whether using the more complex model was justified.

Performance on the tests

All participants were asked to provide an answer to all items on both tests, independent of whether they practiced the item during the learning phase. The scores on *all* items were included in the following analysis. Alternatively, one could only include the items that participants actually studied. However, due to the way new items were introduced, this would bias the analysis strongly in favor of the adaptive method, which only introduces new items if previous items are estimated to be learned well, whereas participants are much more likely see all 50 items when studying with flashcards. Only introducing new items when previously introduced items have a sufficiently high activation is an inherent part of the algorithm (as outlined in the sub-section *The Adaptive Method* above; also see the Discussion). If introducing the possibility that some learners will not see all items is a good strategy, their performance should benefit from it in absolute terms. Therefore, we included *all* items in the test performance analysis in both conditions.

A visual overview of participants' performance is provided in Figure 2.1. Logistic mixed-effects regression was used to predict whether an item was answered correctly on the test. The *study method*, *order*, *session*, and *retention interval* were used as predictors. The predictors *session* and *order* were contrast-coded (specifically, *session* = {-0.5 = first session; 0.5 = second

[3] Alternatively, we could have fit the logistic mixed-effects regression using lme4 and the linear mixed-effects regression using BayesFactor (R. D. Morey & Rouder, 2015b). We verified that fitting the linear models using BayesFactor yields the same best-fitting model (see supplement at www.osf.io/3g6pq) but opted to report the results from lme4 for consistency's sake.

session} and *order* = {-0.5 = started with flashcards; 0.5 = started with adaptive method}) and the predictor *retention interval* was centered by subtracting the median retention interval of 7 days. *Study method* was coded as {0 = flashcards; 1 = adaptive method}. This coding makes the interpretation of the model's coefficients straightforward: The intercept will yield the predicted probability of answering an item correctly on the test for a participant that studied with flashcards, independent of the *order* and *session* for the median *retention interval*. The coefficient for *study method* expresses the main comparison of interested: the change in predicted probability when studying with the adaptive method rather than flashcards, independent of the other predictors.

All models also include random intercepts for participants and items to capture variance associated with differences between learners and item difficulty that are unrelated to the predictors included in the model. As we are only interested in the group-level effects, we will only discuss the fixed effect The best-fitting model was determined using the model selection procedure outlined above.

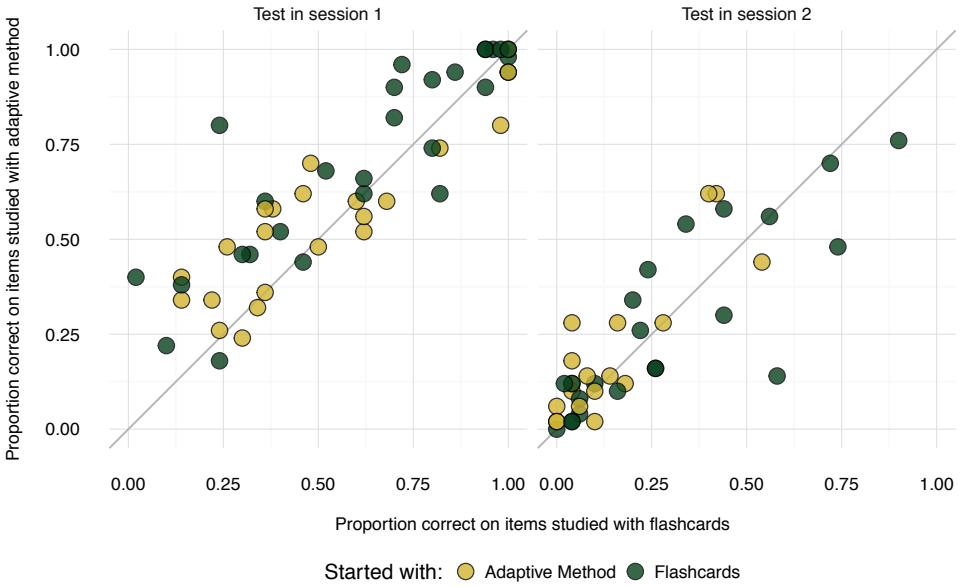


Figure 2.1. Test performance as a function of session, study method, and order. Each semi-transparent point shows a participant's proportion of correct responses on the test across all items in each condition and is color-coded based on the method that each participant started with (i.e., order). For each participant, performance on items studied with the two study methods is plotted against each other to show the relative difference performance.Consequently, points falling above the diagonal line indicate that a participant's score on the test was higher after studying with the adaptive method than after studying with the flashcard method.

The best-fitting model contained the all four main effects and the two second-order interactions listed in Table 2.1 and the approximated Bayes factor indicates that it is 35,738 times more likely than the full model given the data. The best-fitting model is only 1.8 times better than the next best model (that does not contain the interaction between study method and session) according to the approximated Bayes factor but both the differences in AICs (9131.6 vs. 9123.3) and the Chi-squared test ($\chi^2(1) = 10.4$ with $p = 0.001$) favor the more complex model summarized in Table 2.1.

Table 2.1. Estimated coefficients of best-fitting logistic regression model to predict accuracy of responses on the two tests.

	Estimate	Std. Error	z-value	p-value
Intercept	-0.747	0.218	-3.4	<0.001
Study method	0.255	0.055	4.6	<0.001
Order[a]	0.829	0.420	2.0	0.048
Session[a]	-2.262	0.084	-26.9	<0.001
Retention interval[b]	-0.115	0.029	-4.0	<0.001
Study method * session	-0.354	0.110	-3.2	0.001
Retention interval * order[a]	-0.356	0.058	-6.1	<0.001

Note: [a] = variable was contrast-coded; [b] = variable was centered.

The effect we were primarily interested in is that of *study method*. The results listed in Table 2.1 show that studying with the adaptive method has a significant, positive effect on test performance independent of *session* and *order* for a median *retention interval*. Overall, a participant's probability of answering an item correctly is 0.32 when studying with flashcards and 0.38 when studying with the adaptive method[4]. This effect should be interpreted considering the significant interaction between *study method* and *session*, though: the difference in test performance is much larger in the first session than in the second. More specifically, assuming a retention interval of 7 days, the predicted probability of answering an item correctly on the first test is 0.605 for someone studying with flashcards and 0.693 for someone studying with the adaptive method – a difference of about 10% in test scores. This difference decreases to less than 1% in the second session (i.e., to 0.133 and 0.141, respectively).

The main effect of *retention interval* indicates that longer intervals between the first and second *session* affect recall negatively. Keeping all other predictors constant, the expected decrease in the probability of answering an item correctly in the second session is 0.013 for

[4] Since this is a logistic regression, these numbers were computed by taking the inverse logit of the sum of the relevant coefficients. For example, $\text{logit}^{-1}(-0.747) = 0.32$ and $\text{logit}^{-1}(-0.747 + 0.255) = 0.38$.

every additional day between the study phase and the test. This effect interacts significantly with *order* such that someone studying with flashcards can expect a decrease of 0.040 for each additional day since studying if they started studying with flashcards. If they started studying with the adaptive method, on the other hand, they can expect their performance to decrease by 0.079 for each additional day since studying.

In summary, the analysis of the test performance data shows that there is a substantial effect of *study method*. The overall benefit of studying with the adaptive method as well as the interaction with *session* are also apparent in Figure 2.1: Most participants perform better on the first test when studying with the adaptive method than with flashcards (i.e., there are more points above the diagonal line in the left panel) but this benefit is diminished on the test in the second session. The analysis also suggests that an extension of the *retention interval* generally decreases the probability of answering an item correctly but that participants starting with the adaptive method are affected about twice as much as those starting with flashcards.

Performance during studying

Not all items were studied by all participants, especially when using the adaptive system. For items that were studied, the number of repetitions per item and the proportion of correct responses during learning were recorded. An overview of the data is provided in Figure 2.2. To test whether the proportion of correct responses during learning differs between the two study methods, the best-fitting linear mixed-effects regression was determined as described above. The included predictors were *study method*, *order*, and *repetitions* and random intercepts were added for participants and items again. As in the first analysis, the predictor *order* was contrast-coded as {-0.5 = started with flashcards; 0.5 = started with adaptive method} and *study method* was coded as {0 = flashcards; 1 = adaptive method}. The additional predictor *repetition* indicates how many test trials a participant completed for each item. The mean number of *repetitions* across all participants and conditions was five and the predictor was centered by subtracting this average.

The best-fitting model includes all three main effects and the three second-order interactions. Its fixed effects are summarized in Table 2.2. According to the approximated Bayes factors, the best-fitting model was 62.9 times more likely than the full model given the data but only 1.2 times more likely than the model that did not include the interaction between *order* and *repetitions*. The more complex model was considered the best-fitting because both the Chi-squared test ($\chi^2(1) = 7.94$ with $p = 0.005$) and the AICs (-4652.4 vs. -4658.3) favored the more complex model listed in Table 2.2.

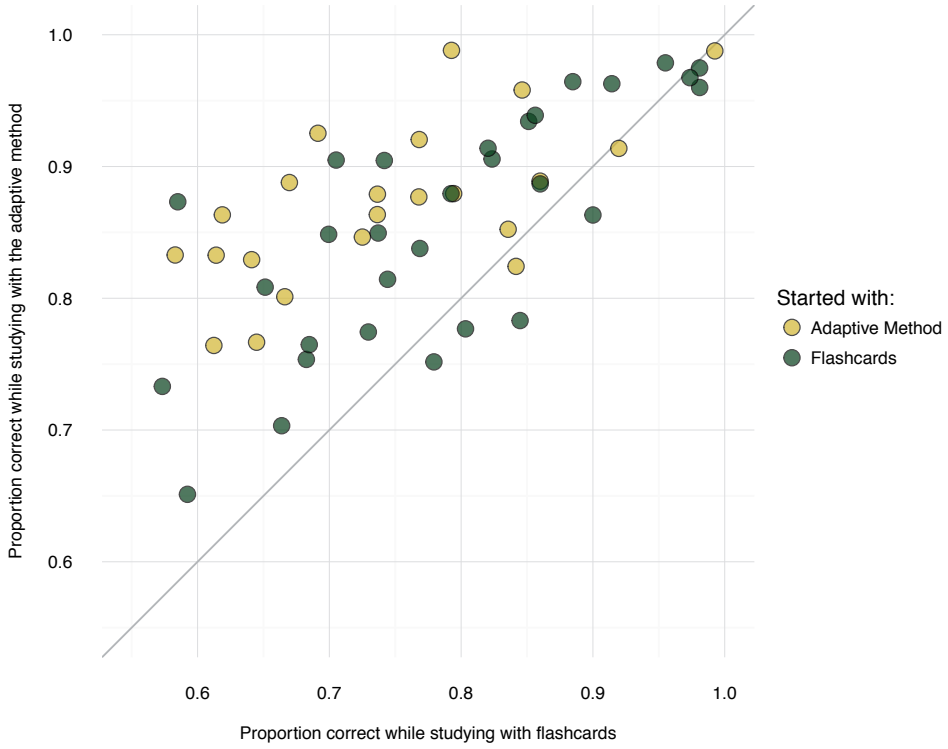


Figure 2.2. Performance during learning as a function of study method and order. Each point indicates a participant's proportion of correct responses across all items when using the two study methods and are color-coded based on which condition that particular participant started with (i.e., order).

The coding of the predictors ensured that the estimated coefficient for *study method* expressed the change in proportion correct during learning for an item with an average number of *repetitions* independent of *order* for a learner studying with the adaptive method rather than flashcards. Overall, the expected proportion of correct responses during learning is .672 when studying with flashcards (i.e., the model's intercept) and the proportion increases by .197 (i.e., the main effect of *study method*) to .869 when studying with the adaptive method. One interesting pattern emerges from the interaction between *study method* and *repetitions*: The coefficient for the main effect of *repetition* and the one for the interaction between *study method* and *repetition* almost cancel each other out (i.e., -0.09 vs. 0.08, respectively). This means the proportion of correct responses for a participant studying with flashcards is expected to decrease as the number of repetitions increases (because *study method* = 0) while the expected proportion of correct responses for a participant studying with the adaptive system will hardly be affected by the number of repetitions. For example, ignoring the effect of *order*, the predicted proportion decreases from .67 to .22 when the repetitions increase from 5 (the

average) to 10 for a flashcard learner. For someone studying with the adaptive system, the same change is much less extreme: from .87 to .82.

Table 2.2. Estimated coefficients of best-fitting model to predict the proportion of correct responses during learning.

	Estimate	Std. Error	t-value	p-value
Intercept	0.672	0.016	42.9	<0.001
Study method	0.197	0.005	40.3	<0.001
Order[a]	0.041	0.031	1.3	0.191
Repetitions[b]	-0.090	0.002	-46.0	<0.001
Study method * order	-0.051	0.010	-5.4	<0.001
Study method * repetitions	0.080	0.002	37.8	<0.001
Order * repetitions	-0.004	0.002	-2.8	0.005

Note: [a] = variable was contrast-coded; [b] = variable was centered.

The predictor *study method* also interacts significantly with *order* such that the proportion of correct responses is generally lower when studying with flashcards but the exact difference between *study methods* (i.e., the benefit of the adaptive method) depends on *order*. An average participant starting with flashcards can expect their proportion of correct responses to increase from 0.69 to 0.86 when switching to the adaptive method (a difference of 0.17; for an average number of *repetitions*). Conversely, if the same hypothetical participant had started with the adaptive method, they could expect their performance to drop from 0.87 to 0.65 – a difference of 0.22 – when switching to flashcards (for an average number of *repetitions* again).

While the main effect of *order* is not significant by itself, it does interact significantly with *repetitions*. The negative effect that more *repetitions* have on the proportion of correct responses during learning is slightly larger (0.004, see Table 2.2) for a learner that starts with flashcards. For example, for someone studying with flashcards who started with flashcards, the expected proportion of correct responses drops by 0.092 for every additional *repetition* while the same decrease is 0.088 (i.e., a difference of 0.004) had they started with the adaptive method.

In summary, the linear mixed-effects regression highlights the benefit of the adaptive method relative to flashcards on the proportion of correct responses while studying, an effect also evident in Figure 2.2. The effect of the number of *repetitions* interacts significantly with *study method* such that additional repetitions have a large negative effect on expected performance during learning when studying with flashcards but the decrease is markedly smaller when using the adaptive method. This difference is slightly amplified for participants that started with flashcards, as indicated by the significant interaction between *order* and *repetitions*.

Descriptive analysis of performance differences between the study methods

The overall advantage of the adaptive method revealed by the previous two analyses is closely linked to four measures that capture performance during learning: the time spent receiving feedback, the total number of trials completed in the 20-minute study phase, the average response times on test trials, and the mean number of repetitions per item (see Figure 2.3). All four measures are closely interconnected and any difference in the measures between the study methods is a result of how the introduction of new items and the repetition of old items is implemented in the two methods. The current experiment was not designed to disentangle which of these factors has a unique or dominant influence on the improved performance (both during learning and on subsequent tests) but it is nevertheless interesting to consider how they interact to appreciate how better outcomes emerge in the adaptive method for many of the participants.

Figure 2.3 provides a graphical summary of the four measures that are of interest when comparing performance during the study phase between the two methods. In all four panels of Figure 2.3, each point represents a participant's value for each study method. Two participants have been highlighted so they can be compared across the four panels: The red triangle indicates a relatively high-performing participant and the red square indicates a relatively low-performing participant (see the Discussion for an in-depth comparison). To further aid interpretation, the diagonal line in each panel indicates equal performance with both study methods and data points are color-coded such that green points signal participants that benefit from studying with the adaptive method and gray points those that perform better with the flashcard method (or equal in both conditions).

Figure 2.3A depicts the time participants spent receiving corrective feedback in both study conditions. Since feedback after an incorrect response always lasted four seconds, the durations are a linear transformation of the total number of errors made during study. The average time receiving feedback is 213 seconds when studying with flashcards and 140 seconds when studying with the adaptive method. A Bayesian paired t-test[5] confirms that participants' durations are shorter when studying with the adaptive method (BF_{H1} = 667,441).

The total number of trials completed in the two conditions are contrasted in Figure 2.3B. The average number of trials is higher when participants study with the adaptive method (225 vs. 195) and the data overwhelmingly support the alternative hypothesis of unequal means in a Bayesian paired t-test (BF_{H1} = 1,143,511).

[5] All Bayesian paired t-tests reported here were conducted using the default settings in the BayesFactor package (R. D. Morey & Rouder, 2015b).

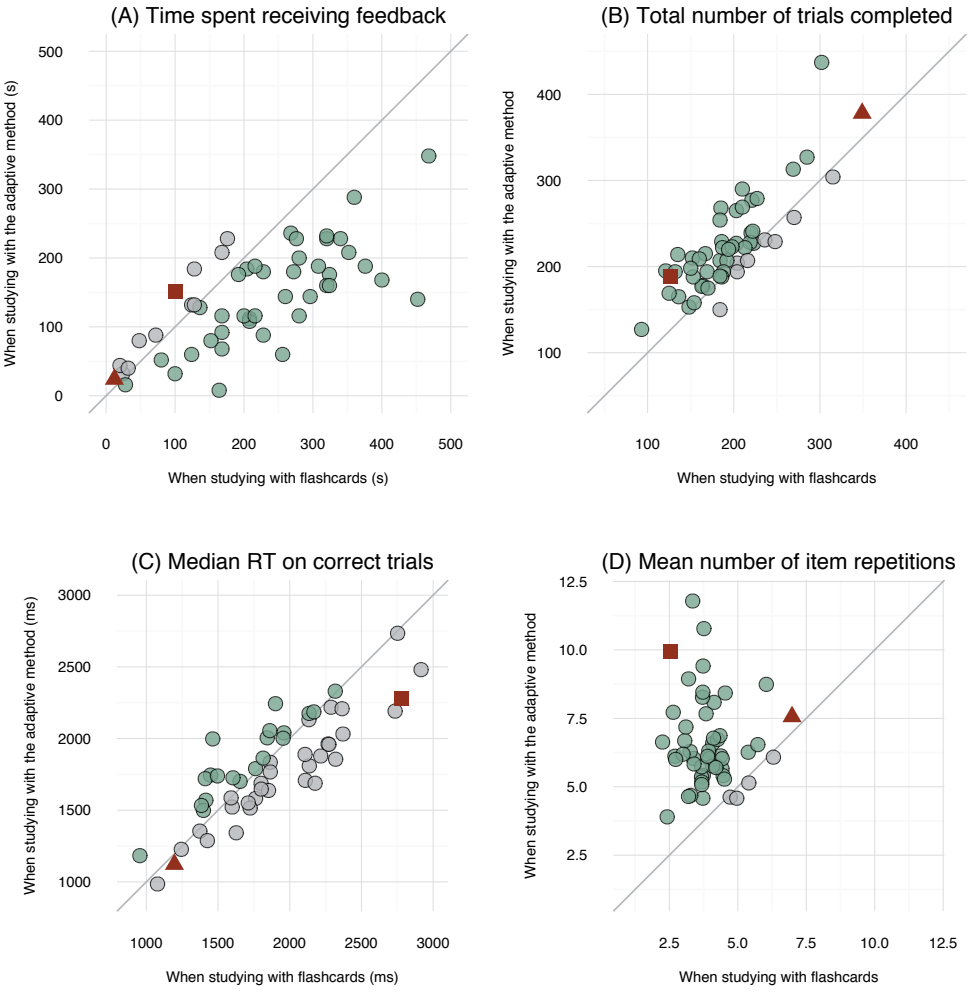


Figure 2.3. Four performance measures contrasted for the two study methods. In all four panels, green points indicate participants that have a relative advantage when studying with the adaptive method and gray points those with an advantage using the flashcard method. Two participants are highlighted in red symbols so the relationship between the four performance measures can be compared across the four panels. The performance measures for each study method are: (A) the number of seconds a participant spent receiving corrective feedback (i.e., the number of errors multiplied by four seconds); (B) the total number of trials completed, including both study and test trials; (C) the median response time on correct test trials; and (D) the average number of repetitions for each item studied.

The only measure in Figure 2.3 that does not show a substantial difference between study methods in the median response time on correct trials per participant and is shown in panel C. The Bayesian paired t-test is inconclusive (BF_{H1} = 1.89) and the mean of the median response times is only 70 ms lower for the adaptive method than for the flashcards (1,803 vs. 1,873 ms). Participants whose median response time is lower for the adaptive method were coded in green indicating an advantage for that condition because only response times from correct trials were included and a longer response time that leads to a correct response is a proxy of more effortful retrieval, which is beneficial for learning[6]. Considering only incorrect trials yields median response times that are, on average, much longer (4,598 ms in the adaptive method vs. 4,929 ms with flashcards) and a Bayesian paired t-test suggest that there is positive evidence for the equality of the response times between the two study methods (BF_{H0} = 5.0).

Figure 2.3D reveals the most extreme difference between the two conditions: When comparing the average number of repetitions per item that each participant did in both conditions, only four participants show near-equal numbers in both conditions – all other participants have a higher average number of repetitions when using the adaptive method. On the group level, this results in an enormous Bayes factor for the paired t-test ($BF_{H1} = 2.4 \times 10^{10}$) and a mean number of repetitions that is 166% higher for the adaptive method than for flashcards (6.57 vs. 3.96, respectively).

Another factor that is directly related to the four measures shown in Figure 2.3 is the number of Swahili-English word pairs that participants were introduced to in the two methods. This number differs drastically: when studying with the adaptive method, participants saw, on average, 36.1 items (median = 35.5; SD = 11.2) but the average when studying with flashcards was 48.6 (median = 50; SD = 2.9). In the adaptive condition, only 32.7% participants saw 45 or more items (of 50) while 92.3% of the participants saw 45 or more when studying with flashcards. In the light of this stark difference, the finding that the adaptive method produces higher absolute scores on the tests may seem surprising: Participants tend to perform better despite having seen fewer items. This effect becomes especially apparent if we create a variation of Figure 2.1 – which is based on absolute test scores – in which the proportion of correct responses on the test is shown only for words that participants were exposed to (see Figure 2.4).

One striking pattern that becomes apparent in the left panel of Figure 2.4 is that none of the participants remember fewer than 50% of the items studied with the adaptive method

[6] However, one could also argue that faster response times are better because they leave more time for additional trials and more retrieval practice. Since maximizing the number of trials is not the goal, however, we chose the retrieval effort-inspired color-coding.

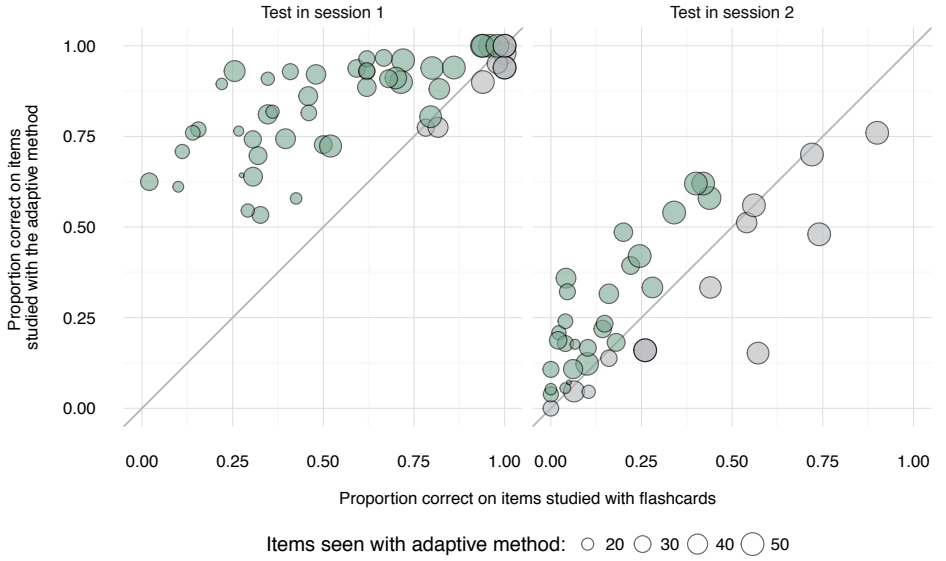


Figure 2.4. Proportion correct on the tests considering only items actually studied. The size of each data point indicates the number of items studied with the adaptive method. Points are displayed in green if someone's performance was higher after studying with the adaptive method and gray otherwise.

(after a 15 minute retention interval), whereas 23 out of 52 participants score below 50% when studying with flashcards. The size of the points in Figure 2.4 indicates the number of items that participants were exposed to when studying with the adaptive method[7] and it is clear that (1) most participants with near-perfect performance have seen most items and (2) participants that saw fewer items have the largest relative benefit (i.e., are furthest from the diagonal line). The right panel of Figure 2.4 suggests that learners that see more items tend to show superior long-term retention.

Taken together, these results imply that the adaptive method identifies participants capable of learning a larger number of word pairs and protects poorer learners by withholding some of the word pairs. Preventing relatively poor learners from being exposed to too many items ensures that fewer errors are made during learning (i.e., less time is spent on corrective feedback), leaving more time for additional repetitions of word pairs and more trials overall (see Figure 2.3).

[7] Since almost everyone was exposed to all 50 items when studying with flashcards, it does not make much sense to discriminate along that dimension.

DISCUSSION

The current study was designed to contrast two study methods, addressing two specific questions: First, do the methods produce different learning outcomes? And second, do they result in a different number of errors during learning? The two contrasted methods were a digital flashcard system and an adaptive, computerize fact-learning system. The former was chosen as a benchmark because many students report using flashcards and the latter is a theory-based model that seeks to alleviate some of the short-comings of flashcards as typically used by students. With regards to the two questions, the analyses show a clear advantage of the adaptive method for both the proportion of correct responses during learning and the proportion of correct responses on subsequent tests at varying retention intervals.

With regards to the test performance, there is an overall advantage of the adaptive method that is especially pronounced on the test at the end of the first session. Due to the mechanics of the adaptive method, not all participants have studied all items. On average, a participant has seen 12.5 fewer items when studying with the adaptive method. To keep the within-subject comparison fair, the logistic mixed-effects regression was conducted on *all* items a participant could have encountered and revealed that the estimated probability of answering any of the 50 items correctly on the test is higher if someone studied with the adaptive method (see Table 2.1). If only the subset of items is considered that participants studied – as illustrated in Figure 2.4 –, the contrast between the two conditions is amplified substantially.

Furthermore, Figure 2.3 and Figure 2.4 suggest that a contributing factor to the difference is the way the adaptive method introduces new items: By withholding new items until old items are estimated to be learned well, relatively low-performing learners can spend a larger proportion of the 20-minute study phase repeating a subset of items to ensure they are still remembered when the test is administered. Conversely, there is no indication that relatively high-performing learners are held back by the adaptive method. Someone scoring at or near ceiling when using flashcards is likely to score similarly when studying with the adaptive method. However, 50 Swahili-English word pairs for a 20-minute learning session might not be enough to reveal a possible disadvantage for very high-performing learners studying with the adaptive method and the number of items would have to be increased to test for it. Exploring potential differences between the two study methods in the high-performance range would be an interesting extension of the current work.

With regards to the proportion of correct responses during learning, the linear mixed-effects regression summarized in Table 2.2 corroborates the large overall benefit of studying with the adaptive method that is also visible in Figure 2.2. The significant interaction between *study method* and *repetition* is primarily an artifact of how the flashcard method was implemented. Incor-

rect items were repeated until all in a stack of five were answered correctly once. This means that more repetitions of an item either indicate that more errors were made or that the participant made few errors and completed many trials. This contrast is illustrated by the two participants highlighted in Figure 2.3 and whose exact values are summarized in Table 2.3. Both participants make relatively few errors and subsequently spent little time receiving corrective feedback. They differ dramatically with regards to the rest of their performance profile, though.

Table 2.3. Comparing the values of the two participants highlighted in Figure 2.3. Note that the retention intervals for the two participants were 7 (triangle) and 6 (square) days. The columns labeled Feedback, Trials, RT, and Reps correspond to panels A through D in Figure 2.3 and display the exact values that pinpoint the two participants' position in the four plots. Additionally, the number of items each participant studied with each method are listed (column Items) as well as their proportion of correct responses on the tests (across all 50 possible items, not just those studied; columns Test 1 and Test 2).

		Feedback	Trials	RT	Reps	Items	Test 1	Test 2
▲	Adaptive method	24s	378	1.12s	7.6	50	1.00	0.62
	Flashcards	12s	349	1.20s	7.0	50	1.00	0.42
■	Adaptive method	152s	189	2.28s	10.0	19	0.34	0.02
	Flashcards	100s	127	2.78s	2.5	50	0.22	0.00

As can be seen in Figure 2.3, the high-performing participant (red triangle) has the greatest number of trials in the flashcard condition, has a very low median response time and their mean number of repetitions hardly differ between the two study methods. Furthermore, for both study methods, they have been exposed to all 50 word pairs and score perfectly on both tests in the first session (see Table 2.3). A week later, they still remember 62% and 42% of the Swahili words they studied with the adaptive method and flashcards, respectively.

The low-performing participant (red square), on the other hand, was exposed to only 19 of 50 items when studying with the adaptive method but correctly recalled 17 of them on the test in the first session (34% correct out of 50). When studying with flashcards, they saw all 50 items – they did not make many errors but completed relatively few trials yielding, on average, only 2.5 repetitions of each of the 50 items – compared with an average of 10 repetitions of the 19 items studied with the adaptive method. On the test in the first session (i.e., 15 minutes after study), they only remembered 11 items (22% correct out of 50). The fact that they had forgotten virtually all studied items by the time the second test was taken six days later further indicates below-average memory performance for this participant.

In summary, the advantage of the adaptive method is twofold: Errors are avoided by repeating

items *before* they are forgotten and retrieval errors are made, resulting in less time spent receiving feedback (see Figure 2.3A) and the optimal time for additional repetitions can be determined for each individual participant. The consequence is that fewer items are studied by relatively low-performing participants but the number of repetitions for each item is higher. The significant interaction between *study method* and *repetitions* in the linear regression analysis confirms that the higher number of repetitions hardly increases the number of errors during learning. The results of the logistic regression indicate that participants generally have higher test scores when studying with the adaptive method and the descriptive analyses summarized in Figure 2.3 and Figure 2.4 suggests that the benefit largely stems from protecting low-performing participants. The comparisons between the two highlighted participants explicitly exemplify effects that hold on the group level and illustrate the large individual differences between participants.

The results and discussion so far have demonstrated and focused on the advantage of the adaptive method. However, the benefit hinges on the comparison with the flashcard method that was implemented to serve as a control. As mentioned in the introduction, flashcards were used as a control condition because many students report using flashcards to study (especially when learning vocabulary; Wissman et al., 2012), which provides a natural benchmark for an adaptive system to be pitched against. However, there are many possible ways to employ flashcards when studying (some more effective than other, e.g., Adragna, 2016; van Houten & Rolider, 1989). The choice of splitting the 50 Swahili-English word pairs into stacks of five was inspired by the fact that many students report creating relatively small stacks (e.g., Kornell, 2009). While it might be possible to create a more effective way to utilize flashcards, the data presented here suggest that we did not create a "strawman control": the spread on the x- and y-axes in Figure 2.1 are comparable, suggesting that both methods permitted individual differences to emerge. In fact, learners' test scores in the two conditions are highly correlated ($r = 0.88$ in the first session and $r = 0.83$ in the second session) but the statistical analyses indicate that most participants perform better on the majority of the outcome measures explored here.

Outside the laboratory, students have the option of dropping flashcards and they do (often too early, e.g., Kornell & Bjork, 2008b)– something that neither method tested here permitted. A possible extension of the current work would be a comparison between the adaptive method and a flashcard system that allowed participants more freedom. For example, at the end of each stack, participants could get the choice to either proceed to a new stack or pick a previous stack to return to. Also included could be the option to drop individual items or entire stacks by marking them as learned. Such a comparison would provide the ultimate test of whether an adaptive method can make more accurate and appropriate judgments of learning and strategic choices regarding the repetition of old and introduction of new material.

3

An Individual's Rate of Forgetting is Stable Over Time, but Differs Across Materials

Acknowledgements

This chapter has been published as Sense, F., Behrens, F., Meijer, R. R., & van Rijn, H. (2016). An Individual's Rate of Forgetting Is Stable Over Time but Differs Across Materials. *Topics in Cognitive Science, 8*(1), 305–321. www.doi.org/10.1111/tops.12183

Supplementary materials for this chapter are available at www.osf.io/yt9qs

Abstract

One of the goals of computerized tutoring systems is to optimize the learning of facts. Over a hundred years of declarative memory research have identified two robust effects that can improve such systems: the spacing and the testing. By making optimal use of both and adjusting the system to the individual learner using cognitive models based on declarative memory theories, such systems consistently outperform traditional methods (Van Rijn, Van Maanen & Van Woudenberg, 2009). This adjustment process is driven by a continuously updated estimate of the rate of forgetting for each item and learner on the basis of the learner's accuracy and response time. In this study, we investigated to what extent these estimates of individual rates of forgetting are stable over time and across different materials, and demonstrate that they are stable over time but not across materials. Even though most theories of human declarative memory assume a single underlying rate of forgetting, we show that, in practice, it makes sense to assume different materials are forgotten at different rates. If a computerized, adaptive fact-learning systems would allow for different rates of forgetting for different materials, it could adapt to individual learners more readily.

In many school curricula, students are partly evaluated based on how well they learn sets of facts. With the advance of computers into classrooms and workplaces, tutoring systems have been developed to help learners master the required declarative fact material. Over a hundred years of declarative memory research have singled out two robust effects that developers of such systems can use to enhance them: the spacing effect and the testing effect (Delaney, Verkoeijen, & Spirgel, 2010). By making optimal use of both of them *and* adjusting the system to the individual learner, such tutoring systems can make learning a lot more efficient. As of now, however, each learning session is treated in isolation in the system that we have been developing (van Rijn, van Maanen, & van Woudenberg, 2009): user-specific characteristics are estimated during a session to optimize learning in that session but are not preserved *between* learning sessions, something which this system shares with most other adaptive learning systems. A potential improvement for these adaptive systems could be found in retaining the estimated characteristics over sessions. However, this requires that these characteristics do not fluctuate too much between different learning sessions. In this study, we investigated to which extent user-specific characteristics relevant to such a tutoring system are stable over time and across different materials.

The optimization of fact learning is often based on balancing the benefits of the spacing and the testing effect. The spacing effect describes the finding that performance on tests of recall is improved when study time is distributed over multiple occasions (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Dempster, 1988; Donovan & Radosevich, 1999; Jastrzembski, Gluck, & Gunzelmann, 2006), a benefit that persists even after a delay (Godbole, Delaney, & Verkoeijen, 2014). However, the same spacing principle can also be used within a single learning session as it has been shown convincingly, based on both behavioral (Lindsey, Shroyer, Pashler, & Mozer, 2014; Nijboer, 2011; van Rijn et al., 2009) and psychophysiological data (van Rijn, Dalenberg, Borst, & Sprenger, 2012), that retention can be increased by spacing items within a single learning session. The optimal spacing schedule ultimately depends on how much time is available and when the material is tested (Cepeda et al., 2008). However, the vast majority of students do not space their studying at all: they mass-study just before an exam (Taraban et al., 1999).

The testing effect describes the finding that active memory retrieval during practice is more beneficial for long-term retention than passive study (Karpicke & Roediger, 2008; Roediger & Butler, 2011). That is, being forced to retrieve the answer from declarative memory leads to better learning than simple re-studying (i.e., looking at) the cue-answer pair (Carrier & Pashler, 1992). This effect has been studied extensively in the laboratory (de Jonge et al., 2012; Kornell & Bjork, 2008a; Verkoeijen & Bouwmeester, 2014) but also holds in more realistic classroom settings (Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008; Butler & Roediger,

2007; Goossens, Camp, Verkoeijen, Tabbers, & Zwaan, 2014; McDaniel, Anderson, Derbish, & Morrisette, 2007). Importantly, the testing effect is strongest if the item can be successfully retrieved (Carrier & Pashler, 1992) and practically disappears for non-retrievable items (Jang et al., 2012).

The goal of learning systems is to devise a learning schedule that makes optimal use of each effect's benefits. This requires balancing two seemingly opposing goals: (1) maximizing time between repetitions of an item to get the biggest spacing effect, and (2) minimizing time between repetitions of an item to make sure it can still be retrieved from declarative memory (to take advantage of the testing effect). Such computer adaptive practice models have been developed and implemented with great success (Lindsey, Mozer, Cepeda, & Pashler, 2009) and have been shown to outperform flashcard control conditions (Atkinson, 1972; Nijboer, 2011; van Rijn et al., 2009). A simple flashcard procedure (Pavlik & Anderson, 2008) is an excellent control condition because many students report using similar procedures to study for exams (Hartwig & Dunlosky, 2012; Kornell & Bjork, 2008b; Kornell & Son, 2009; Wissman et al., 2012).

As a starting point for the development of such models, Anderson and Schooler (1991) showed that data on declarative memory performance (i.e., practice and retention) across time courses ranging from seconds to years can be fit by power functions (also see Rubin & Wenzel, 1996). Pavlik and Anderson (2003, 2005) argued that the practice and retention of facts can be approximated using the same equations that can be used to describe the behavioral effects in the data. They developed a model that formalizes this process and showed how it can be used to compute the optimal schedule of practice, taking into account the effects of practice, retention, and spacing (Pavlik & Anderson, 2008; Pavlik, Bolster, Wu, Koedinger, & MacWhinney, 2008). Like most other models of human declarative memory (e.g., Raaijmakers & Shiffrin, 1980), their model assumes that there is some stable effect based on each individual's *rate of forgetting* and additional effects based on *item difficulty*. Someone's *rate of forgetting* is (implicitly) assumed to be a property of their memory and therefore assumed to be a trait-like, stable property, regardless of whether they study vocabulary, topographical information, or glossary definitions.

The learning outcomes after studying declarative fact materials using the models based on Pavlik and Anderson's approach (Nijboer, 2011; Pavlik & Anderson, 2008; van Rijn et al., 2009) are promising. However, the models have only been tested within a single session in one domain at a time. The stability of participants' *rates of forgetting* across time and knowledge domains is assumed but has not been demonstrated empirically. The goal of the present study was to investigate to which extent participants' *rates of forgetting* vary over the course of three weeks as well as across four different types of declarative fact material.

METHODS

Participants were tested in three separate sessions, each spaced one week apart. In each session, participants learned Swahili-English vocabulary during the first block, whereas in the second block one of three other types of declarative fact material was presented (see Figure 3.1 for an overview). Participants' *rates of forgetting* were estimated for each block. This way, participants' estimated *rate of forgetting* for each of the three Swahili-English blocks can be determined to assess the stability of the *rate of forgetting* over time. By comparing the *rates of forgetting* between the different types of declarative fact material, we can investigate their stability across knowledge domains.

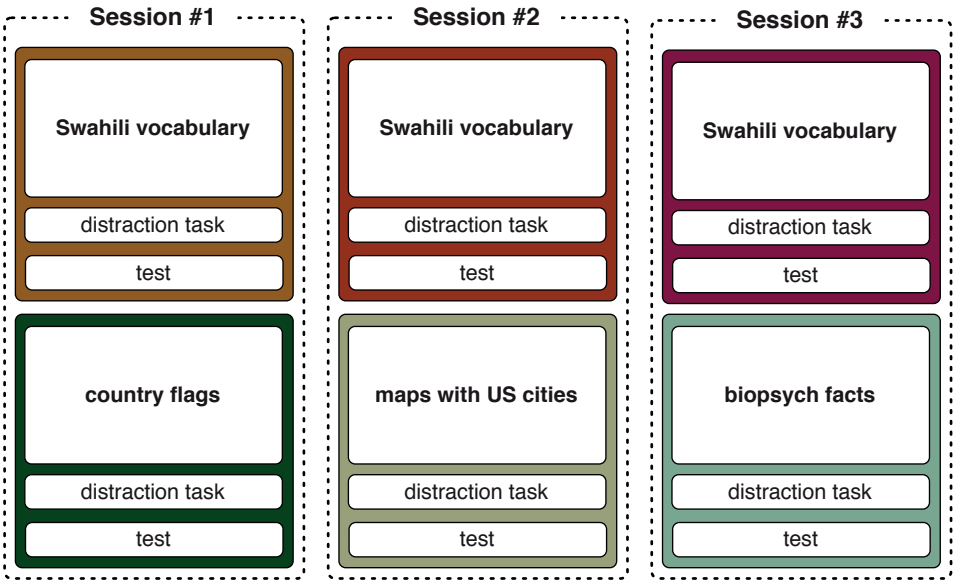


Figure 3.1. Overview of the experimental design of the study. The sessions are spaced one week apart. A different type of declarative fact material is studied in each (color-coded) block. Six unique sets of items were used, one per block.

In the following, we describe in more detail how the *rates of forgetting* were estimated for each participant, what types of material were used, and which analyses were conducted to address the research question.

The model

The model used in this experiment is based on ACT-R's declarative memory equations (J. R. Anderson, 2007). In the ACT-R framework, each item that is learned is assigned an *activation* value. Activation is highest at the moment an item is encountered and then decays as a function of time. The activation of an item at any point in time can be computed using the following equation:

$$A_i(t,n) = \left(\sum_{j=1}^{n} (t - t_j)^{-d_j} \right)$$

According to this equation, the activation of item *i* at time point *t* depends on all *n* previous time points at which item *i* has been encountered. The activation of each previous encounter *j* decays over time with d_j, which, as *d* values are always positive, translates to a smaller contribution to the current activation if encounter *j* has occurred long before time point *t*. The rate with which the activation of an encounter decays is determined at the presentation of that encounter:

$$d_j = ce^{A_i t} + \alpha_i$$

In this equation, which is evaluated at the time of the more recent encounter, but without taking that encounter into account for calculating the activation, *c* is the decay scale parameter that determines the relative contribution of the activation component ($A_i(t)$). The activation is calculated using the earlier discussed equation over all previous encounters, excluding the current encounter for which the decay is calculated. Alpha *(α)* represents the decay intercept that is used as the decay value for the first encounter and is subsequently adjusted for each item I individually. This α parameter will be the main focus of the adaptive algorithm discussed later.

An activation value can be converted to an estimated response time by scaling the activation and adding a *fixed time* that accounts for non-memory related processes. The following equation is used to convert the scaled (*F*) activation of item *i* at time point *t* ($A_i(t)$) to an estimated reaction time:

$$RT_i(t) = Fe^{-A_i(t)} + fixed\ time$$

Pavlik and Anderson (2003, 2005, 2008) have shown that the three equations outlined here can be used to fit a wide range of data from learning-related experiments and can account for additional benefits gained through the spacing effect (however, see Lindsey et al., 2009 for some limitations).

The system has not only been used to describe collected data but also to devise a system that predicts, in real-time, the order in which items should be repeated to yield optimal retention. To account for individual variability, the system designed by Pavlik and Anderson adjusted the decay parameter based on accuracy scores. When an incorrect response was giv-

en to an item of which the model assumed that it had an activation that should have resulted in a successful retrieval, the decay rate for that item was increased, and vice versa when an unexpected correct response was given. However, more information can be gathered from the answers, as the deviation with the predicted reaction time can also be used for updating of decay parameters.

More recently, Van Rijn and colleagues (van Rijn et al., 2009) and Nijboer (2011) have proposed an algorithm, based on the original Pavlik and Anderson work, that also takes into account the response times. In a series of laboratory and classroom studies, they showed that this refined algorithm leads to improved performance compared to both the Pavlik and Anderson (2008) model and to flashcard procedures. This updating algorithm adjusts an item's alpha (α) parameter by comparing the estimated reaction time with the observed reaction time. If the estimated reaction time was longer than the observed reaction time, the activation estimated by the model was too low, and thus a better fit to the empirical data would have been observed if the decay had been higher. To compensate for this discrepancy, the α parameter for the given item is adjusted using a binary search algorithm to improve the model's estimate on the following trial (see Nijboer (2011) for details). Using this adaptation procedure, the α parameter is modified per item per learner to best capture the behavioral variation observed during learning. This decay parameter is an operationalization of the *rate of forgetting*, which is the term we will use in the remainder of this manuscript.

The sequence in which the items are presented is based on the activation values calculated using these *rate of forgetting* parameters using a procedure graphically depicted in Figure 3.2.

The procedure is relatively straightforward: When an item needs to be presented, the model calculates the estimated activation *n* seconds from the current time for all items that have been encountered earlier. If, *n* seconds from now, the activation of any item has dropped below the retrieval threshold, that item is presented next. If no item has dropped below the threshold, a new, not-yet-presented item is scheduled for presentation as long as novel items are still available. Otherwise, the item with the lowest activation *n* seconds from now is presented. Based on the answer on this presentation, the *rate of forgetting* of this item is updated, and the model checks whether a next repetition needs to be scheduled.

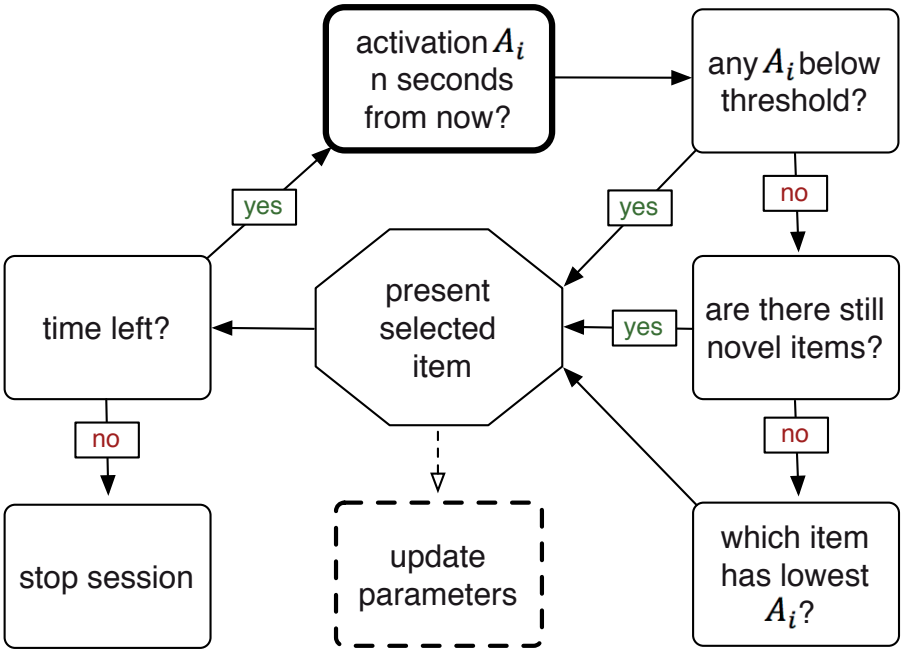


Figure 3.2. A graphical representation of how the model determines the order in which to repeat old and present new items.

Exclusion criteria

Before starting data collection, we defined three exclusion criteria meant to ensure that participants contributed complete data sets and actively engaged with the task. First, participants must have completed all three sessions. Second, participants must have performed well enough to encounter at least 10 unique items in each block. As the model only introduces new items when the activation of the previous items is high enough (see above for details), we used this as a proxy to determine which participants actively engaged in learning. Third, participants must have answered at least 25% of the items they studied correctly on the delayed recall test. Participants who did not meet these criteria were unlikely to have been actively engaged in the task, potentially distorting the data.

Participants

Participants were 76 first-year psychology students from the University of Groningen, 71 completed all three sessions and 70 fulfilled the minimum requirement of having seen at least 10 unique items in each study block. Another three participants were removed because they performed at less than 25% on the final test. Of the remaining 67 participants (88%), 50 were female (75%) and the median age was 20 (SD_{age} = 1.73; $range_{age}$ = [17; 26]). No participant

indicated familiarity with Swahili, 35.8% were Dutch, and 52.2% were German. The remaining students were from the English-language Bachelor program with other nationalities. All participants indicated to be fluent in English and gave informed consent as approved by the Ethical Committee Psychology (ID: 14017-NE).

Materials

For each block, a list of 25 items was compiled. The lists of items were identical for all participants but during each study block, the model randomized the order in which items were presented based on participants' identification numbers. There were four types of declarative fact material that were studied by each participant:

Vocabulary. 75 Swahili-English word pairs were selected from the list compiled by Van den Broek, Segers, Takashima, and Verhoeven (2014) and are available on the Open Science Framework[1]. Swahili-English word pairs are common stimuli in vocabulary learning (e.g. Carpenter, Pashler, Wixted, & Vul, 2008; Kang & Pashler, 2014; Pyc & Rawson, 2010; van den Broek et al., 2014) because participants from countries in which a Germanic language is the majority language are typically not familiar with it but it uses familiar letters and phonemes. The written Swahili word was the cue to which the participants had to respond by typing the correct English word.

Flags. A list of 25 countries and their flags was compiled from Wikipedia's list of sovereign states. We strived to pick flag/country combinations that were not likely to be known by the participants, using the experimenters' familiarity with the countries' flags and a pilot study as a benchmark. A full list of all countries can be found online[1]. The country's flag was the cue to which the participants had to respond by typing in the country's name.

Maps with U.S. cities. A list of 25 items was compiled by searching for small cities on Google Maps, making sure the cities were more or less evenly spaced across the United States of America. Cities were picked so that their names were unique, not too difficult to spell, and did not contain information about their geographical location. A full list can be found online[9]. Participants always saw a map of the U.S. with state borders on which all cities from the set were marked with gray dots. The cue for a city was the same map with the city in question highlighted in bright right. The participants had to respond by typing in the city's name.

Bio-Psychology Facts. A list of 25 bio-psychology facts was compiled from the Glossary in Kalat (2012), a text book used in a mandatory Biopsychology course scheduled for the following semester for all students participating in this experiment. The facts were chosen so that the answer would always be a single word and that there was some variation in how difficult the words are to spell. A full list can be found online[1]. The description of the term was the cue to which the participants had to respond by typing in the described term.

[1] All materials and the data are available at: www.osf.io/hm3ma/

Procedure

Each person participated in the study for three sessions on three days, each session spaced one week apart. Within each session, there were two blocks. Each block consisted of a 20-minute study session, a five-minute distraction task, followed by a test of the studied declarative fact material that took about five more minutes (see Figure 3.1). At the beginning of the first session, each participant also completed a short questionnaire regarding demographic information (age, gender, nationality, and language skills). The five-minute distraction was a simple variation of the puzzle game Tetris which participants played until they were automatically re-directed to the test that concluded the block.

While learning the material of each block, novel items were presented on *study trials* and subsequent repetitions were presented on *test trials*. On a *study trial*, participants saw both the cue and the correct response and had to type in the correct response to proceed. On a *test trial*, participants only saw the cue and had to type in the correct response. Feedback ("correct"/"incorrect") was provided in both trial types and lasted 0.6 and 4 seconds for correct and incorrect responses, respectively. The feedback on incorrect trials always resembled a *study trial* and displayed both the cue and the correct response. Jang, Wixted, Pecher, Zeelenberg, & Huber (2012) have shown that for non-retrievable items, an additional *study trial* is very effective because participants do not benefit from the testing effect and De Jonge, Tabbers, Pecher, and Zeelenberg (de Jonge et al., 2012) showed that the optimal duration of a *study trial* is four seconds.

During the test at the end of each block, participants were provided with a list of all cues and were asked to provide their responses (in any order they preferred). There was no explicit time limit for completing the test.

Analysis

The main analysis is based on two regression analyses. The first analysis addressed the main research question and focused on whether the estimated *rates of forgetting* were stable over time and materials. An additional backwards regression was conducted to corroborate the findings and to investigate which *rates of forgetting* from the first two sessions (i.e., days) best predicted the *rates of forgetting* in the last session (i.e., on the last day).

RESULTS

First, we present participants' performance on the final test to verify that they actively learned the declarative fact material. Then correlations between the operationalized *rates of forgetting* are presented to show the coherence between blocks. Lastly, two regression analyses are presented to show in more detail how the *rates of forgetting* differ between types of material but not over time. Note that all data and associated analysis scripts can be downloaded from: www.github.com/fsense/parameter-stability-paper.

Performance on the final test for the six different blocks was plotted to verify that participants were exposed to and learned the material sufficiently. Figure 3.3 depicts violin plots (Hintze & Nelson, 1998) that show local density estimates added to each side of a traditional boxplot. Specifically, the white dots represent medians and the black portions correspond to the traditional "box and whiskers" of a box. The plot demonstrates that overall performance was very high, suggesting that the participants were exposed to the material sufficiently to learn it well. The material in the *maps* condition was more difficult to learn than the material from the other conditions. Furthermore, many people perform at ceiling in the *Swahili vocabulary* and *flags* conditions.

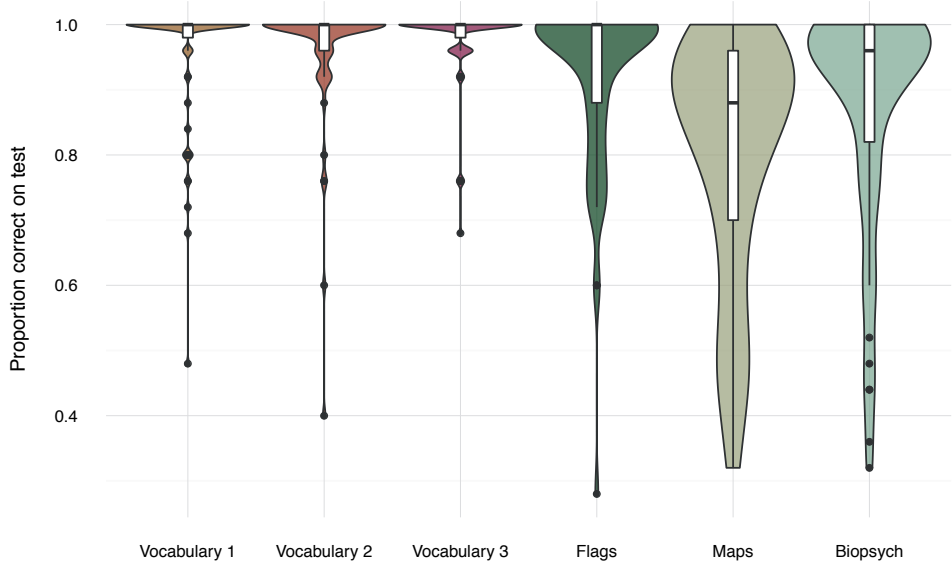


Figure 3.3. Performance on the final test that was taken at the end of each block as a violin plot to show the density estimates as well as classic box plots (in white). . Note that each violin is scaled so that its maximum density is one.

As described the in the sub-section *The Model* above, the *rate of forgetting* was operationalized as the model's α parameter. Each item started with a *rate of forgetting* of 0.3 (Nijboer, 2011; van Rijn et al., 2009) and then the value was adjusted on each repetition of the item. The adjustment depended on how the participant responded to the item (correct/incorrect) and how well the observed response latency corresponded with the model's prediction. For the following analyses, the final *rates of forgetting* of items that were presented at least three times were included. That is, each participant contributed multiple *rates of forgetting* and the exact number depended on how many items each participant encountered at least three times within each block. By averaging the *rates of forgetting* observed in one block, a mean *rate of forgetting* for a participant was computed that indicates how quickly information learned in that block was forgotten.

It is expected that participants with a low *rate of forgetting* have a higher chance to perform well on the final test - and vice versa. This assumed relationship is depicted graphically in Figure 3.4. For this plot, each participant contributed one mean *rate of forgetting* per block and their performance in that block. For each block, participants were binned with respect to their mean *rate of forgetting*. The mean *rate of forgetting* and mean performance were calculated in each bin and plotted against each other.

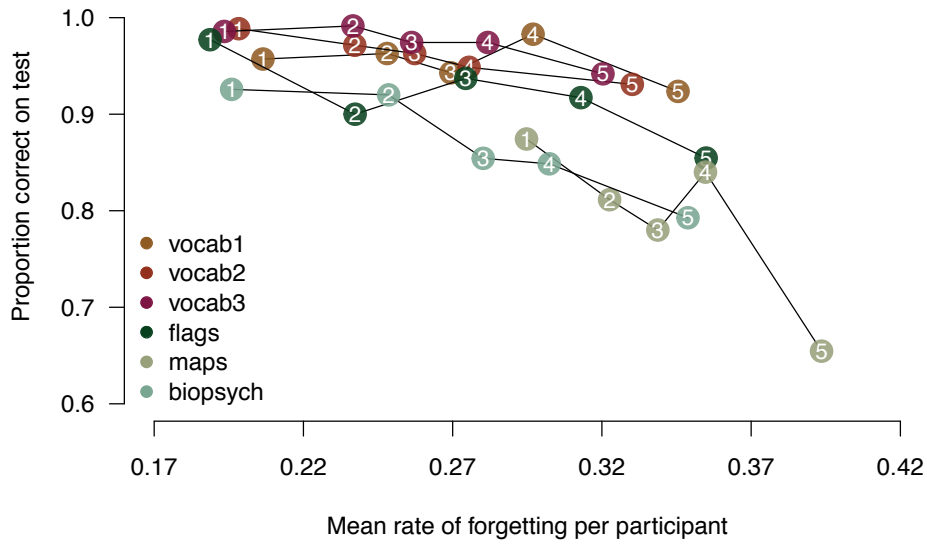


Figure 3.4. Participants' mean rate of forgetting is plotted against their performance on the final test for each block. The data is binned to make the plot more readable. Bins were created by ordering the 67 individual data points in each condition on the x-axis, dividing them into five bins and then computing the mean values for each bin.

The three vocabulary blocks (in shades of orange) form one cluster and the performance on the test is near or at ceiling regardless of the rate of forgetting. For the other three conditions, however, a higher *rate of forgetting* implies slightly lower performance on the final test. Additionally, one can see that for the most difficult condition (*maps*, also see Figure 3.3), the *rates of forgetting* are generally higher (i.e., shifted to the right on the x-axis) than in the other conditions. This indicates the material was forgotten more quickly. Overall, the relationship between the *rate of forgetting* and performance on the test suggests that the model's α parameter is a useful and informative operationalization.

Correlations between the *rates of forgetting* in each block are shown in Table 3.1. The relatively high correlations indicate coherence between the conditions. They do not, however, tell us whether the variation in mean *rate of forgetting* is greater between domains than it is within a domain or how stable the values are over time (even though the high correlations between the vocabulary blocks are promising).

Table 3.1. Correlations of mean rates of forgetting between all blocks. All correlations differ significantly from 0 with $p < .001$.

	vocab1	**vocab2**	**vocab3**	**flags**	**maps**
vocab1	-				
vocab2	.758	-			
vocab3	.771	.863	-		
flags	.603	.630	.549	-	
maps	.499	.556	.495	.604	-
biopsych	.630	.572	.534	.511	.384

To directly address the research question, we looked at the variation in *rates of forgetting* across time and materials using linear mixed-effects model regression. Two dummy-coded variables were included in the model: The first variable coded the *session* in which a block was completed. This tests whether there is any significant variation over time across all blocks. The second variable was coded 0 for blocks in which participants studied Swahili vocabulary and 1 for those in which non-Swahili material was studied. This directly compares the differences between multiple blocks of learning Swahili to non-Swahili blocks. For this analysis the *rates of forgetting* were log-transformed to prevent violations of homoscedasticity and normality. Table 3.2 summarizes the estimated regression coefficients of the log-transformed *rates of forgetting*, the standard error of the estimate, the degrees of freedom, the absolute t-value, and the p-value.

Table 3.2. Results from the linear mixed-effects regression with dummy coding. The columns show the estimated regression coefficients of the log-transformed rates of forgetting values, the standard error of the estimate, the degrees of freedom, the absolute t-value, and the p-value, respectively.

	$\mathbf{beta}_{\mathbf{ln(ROF)}}$	SE	df	| t |	p
intercept	-1.394	0.040	112	34.38	<0.001
session	-0.036	0.028	73	1.30	0.198
SW vs. non-SW	0.182	0.011	9506	16.50	<0.001
session * (SW vs. non-SW)	-0.051	0.028	82	1.81	0.074

The *rates of forgetting* do not significantly differ between sessions (t(73)=1.3, p=0.198). However, the contrast between the Swahili and non-Swahili blocks significantly influences the *rates of forgetting* (t(9506)=16.5, p<0.001) indicating that the *rates of forgetting* in Swahili blocks are significantly different from non-Swahili blocks. The interaction between the *session* and the contrast Swahili vs. non-Swahili is not significant (t(82)=1.81, p=0.074). This means the difference in the *rate of forgetting* when performing a non-Swahili task compared to the Swahili task does not differ as a function of the session in which the task is performed in. Figure 3.5 gives a visual overview of the mean *rate of forgetting* in each block and suggests that the significant difference of the Swahili vs. non-Swahili comparison is driven by the higher *rate of forgetting* in the *maps* condition. This was confirmed by Bonferroni-corrected post-hoc paired t-tests, which revealed differences between the *maps* and the *flags* conditions ($t(66) = 12$, $p < .001$) and the *maps* and the *biopsych* conditions ($t(66) = 10.7$, $p < .001$) but no significant differences between the *flags* and the *biopsych* conditions ($t(66) = .27$, $p = .785$.[2]

[2] Bayes factor t-tests (Rouder, Speckman, Sun, Morey, & Iverson, 2009) were performed with the default settings of the BayesFactor package (R. D. Morey & Rouder, 2015a) and corroborate the findings: When tested against the default null model, the alternative model indicating inequality between both groups is favored when comparing the *maps* and *flags* conditions (Bayes factor = $1.8*10^{15}$) and the *maps* and *biopsych* conditions (Bayes factor = $1.2*10^{13}$). However, the data provides support for the null model when compared with the alternative model when comparing the *flags* and *biopsych* conditions (Bayes factor = 7.2).

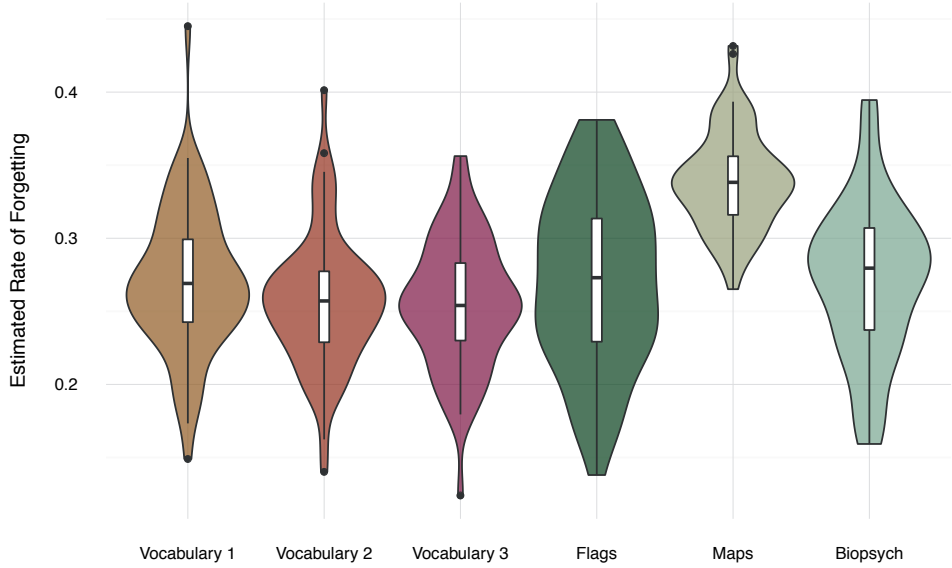


Figure 3.5. The mean rate of forgetting per participant for each block as a violin plot to show the density estimates as well as classic box plots (in white). Note that each violin is scaled so that its maximum density is one.

An additional analysis investigated to which extent the *rate of forgetting* in one session can predict the *rate of forgetting* in a subsequent session. Specifically, the *rate of forgetting* in the third session was predicted based on the values from previous sessions. The Swahili block in the third session was the dependent variable and data from the four blocks that were completed in the two previous session were entered as independent predictors in a backwards regression analysis. The *rates of forgetting* from the *biopsych* block were not included in the analysis because these are possibly confounded due to being completed in the same session as the predicted values (see Figure 3.1).

We expected that (1) the same and (2) more recent tasks would have more predictive power than another and earlier tasks. Both expectations are confirmed by the backward regression analysis. The first expectation was supported by the fact that the best-fitting model includes only the rates of forgetting of the second and first Swahili blocks (in that order). The blocks with declarative fact material from other domains (*flags* and *maps*) do not significantly improve goodness of fit when added to the model. The second expectation is supported by the finding that the *rate of forgetting* of the more recent Swahili task is a stronger predictor (beta = .64, $t(64) = 7.3$, $p < .001$) than the performance during the first (and therefore earlier) session (beta = .24, $t(64) = 3.0$, $p = .004$). Given that the estimate of the more recent task is almost three times higher than the earlier task, the prediction of the third session is closer to the *rate of forgetting* of the second session. The final model, including the rates of forgetting from the second and the first session as predictors, has an adjusted R^2 of .77 ($F(2, 64) = 111.4$, $p < .001$).

DISCUSSION

In this study, we investigated the stability of individual *rate of forgetting* parameters in a model of optimal fact learning. The emphasis was on scrutinizing the stability of the parameter values across time and across different materials. Knowing more about the circumstances under which a learner's estimated *rate of forgetting* is stable in time and across declarative fact materials enables us to further develop the model by carrying over what we learned about the participant in one learning session to the next.

The results of the analyses show that the estimated *rates of forgetting* do not differ significantly over time. There is a difference in estimated *rates of forgetting*, however, when different types of declarative fact material are studied. When looking at the data of the performance on the final test depicted in Figure 3.3, one can see that there was a clear ceiling effect. The effect is especially pronounced in the three Swahili blocks and the block in which participants learned flags. This might be considered to be an issue because it would facilitate the stability of results within those Swahili-learning blocks. It should be noted, however, that by using the parameter values that were estimated throughout the learning session instead of the *results* of the learning session (test performance), one gets a much more fine-grained view of the differences between conditions. There is much more variation in estimated rates of forgetting than the corresponding results on the test suggest (i.e., more variation on the x-axis of Figure 3.4 than on the y-axis). This conclusion is further supported by the fact that there was no significant difference across the three sessions (see both main effect of *session* and the interaction between *session* and *type* of materials shown in Table 3.2), even though the comparison *did* include the blocks for which final performance was not at ceiling. Therefore, analyses based on the estimated *rates of forgetting* are more informative than those based on the performance on the final test.

The data presented here suggest that participants' *rates of forgetting* are stable over time within one type of material (*Swahili vocabulary*), but less stable between materials. We deliberately picked the types of declarative fact material to be different from each other (i.e., vocabulary, visual, topographical, and factual) so perhaps it is not surprising that a difference was found. What is surprising to us, however, is that the *flags* and the *bio-psychology* conditions were similar to each other while the two conditions that used visual information (*flags* and *maps*) were as different as the *maps* and *bio-psychology* conditions, indicating that simple surface features are unlikely to provide good predictors for the similarity of the estimates of the *rate of forgetting*.

According to the framework outlined by Pavlik and Anderson (2008), based on an idea that is also present in many other theories of human declarative memory (Raaijmakers &

Shiffrin, 1992), participants' *rate of forgetting* is considered to be a stable property of their memory system. If this was the case, an adaptive learning system could estimate a learner's *rate of forgetting* while studying one type of material and then re-use that parameter when the learner starts learning a different type of material. However, the data presented here do not support the idea that someone's *rate of forgetting* is stable regardless of the type of material and thus preserving estimated *rates of forgetting* between study sessions of different types of material is more complicated. The data presented here do, however, support the idea that someone's *rate of forgetting* is stable over time for the same type of declarative fact material, which means that we can preserve estimated parameter values and re-use them when the learner returns to the same type of material. How similar materials have to be to yield sufficiently similar *rates of forgetting* is an empirical question that has not yet been answered.

Besides such practical implications, there is also something to be said about models of human declarative memory in general. Most models that assume memories are encoded as traces assume that all traces are treated equally when it comes to learning and forgetting. The models propose certain rules under which traces are encoded, retrieved, and maintained but those rules usually apply regardless of the type of declarative fact material or the context. One exception might be the Search of Associative Memory model (SAM; Raaijmakers & Shiffrin, 1980), in which context information is encoded along with the trace itself. The experiment presented here is neither designed nor intended to falsify any of those theories. It is self-evident, however, that not all types of material are equally difficult for a single person. As long as it is not clear how a single underlying process that treats all types of material equally results in varying learning and forgetting in different tasks/materials, it makes sense - from a practical point of view - to use varying *rates of forgetting*.

4

The Rate of Forgetting as a Useful Individual Differences Measure

Acknowledgements

An earlier, shorter version of this chapter has been published as Sense, F., Meijer, R. R., & Van Rijn, H. (2016). On the Link between Fact Learning and General Cognitive Ability. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Meeting of the Cognitive Science Society.* Cognitive Science Society.

We thank Susan Niessen for her invaluable help in planning the experiment, collecting the data, and all aspects related to the Q1000 test. We also thank Atser Damsma, Anna Leonte, and Rob Nijenkamp for their help with the translations and Lukas Preis and Ron Woytaszek for their help with data collection.

Supplementary materials for this chapter are available at www.osf.io/cqypb

Abstract

Adaptive, personalized fact-learning systems can help learners make the most of time spent studying. Ideally, such systems should not only tailor towards each learner but should also be able to identify strengths and weaknesses of learners. One benchmark would be trying to predict delayed recall of studied material. One important predictor is probably the rate at which encoded material is forgotten, but additional cognitive processes related to attentional control might influence delayed recall. Here we show that the *rate of forgetting* estimated during learning alone is the best predictor of delayed recall. Two additional constructs – working memory capacity and general cognitive ability – were assessed and did not explain a significant amount of variance in delayed recall and were not correlated with the *rate of forgetting*. These findings suggest that, at least in the current small and relatively homogeneous sample, executive functioning and attentional control did not play important roles in predicting delayed recall of items studied using the adaptive fact-learning system developed in our lab. More generally, this might indicate that computerized learning environments should focus primarily on exploiting memory-related individual differences to optimize learning.

Few findings in psychology are as reliably reproduced as the spacing effect (Donovan & Radosevich, 1999), the finding that learning yields better long-term results if repetitions are spaced over time rather than crammed together. The spacing effect holds over various time scales (e.g., Cepeda et al., 2008), indicating that it reflects a fundamental property of human memory. Therefore, the spacing effect should be exploited when we want to optimize fact learning (J. R. Anderson & Schooler, 1991; Dempster, 1988; Rohrer, 2015). Pavlik and Anderson (2003) extended ACT-R's declarative memory module – a computational model of human memory – to account for spacing effects and subsequently made first steps in using the decay-based memory model to enhance the learning of facts (Pavlik & Anderson, 2005, 2008)[1]. Their model combines spacing with retrieval practice (Roediger & Butler, 2011; van den Broek et al., 2016). Making most trials a test event has two advantages. Firstly, long-term retention is improved by exploiting the testing effect and, secondly, trial-by-trial information about the state of the learner's memory can be obtained during learning. Van Rijn and colleagues (2009) showed that the estimation can be further improved by taking both accuracy *and* response latencies into account. This allows for continuous estimation of the strength of a memory trace while minimizing the number of errors during study. This way, items that are likely to be forgotten soon can be identified and scheduled for repetition accordingly, allowing personalized learning schedules to be devised on the fly.

As explained in detail in Chapter 2, the adaptive learning system developed in our lab estimates the speed with which each encoded item becomes less available in memory. The model has one parameter per item per learner that is continuously adjusted based on the learner's response accuracy and latency to each test event. By averaging across the item-specific parameters obtained at the end of the study session, we can compute a value that indicates how quickly, on average, each learner forgets the items in a particular set. We will refer to this value as the *rate of forgetting* (Sense, Behrens, Meijer, & van Rijn, 2016; van Rijn et al., 2009).

[1] Note that this is not the first and only adaptive learning model of this kind. It is merely the first version of the current model discussed here.

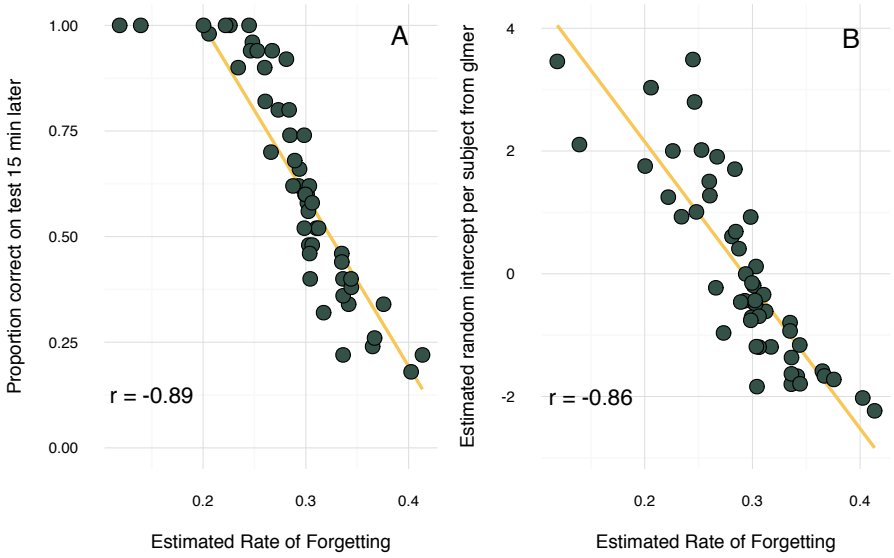


Figure 4.1. A: Each participant's estimated rate of forgetting (x-axis) plotted against their proportion of correct responses on a test (y-axis) of the studied material taken 15 minutes later. B: Each participant's estimated rate of forgetting (x-axis) plotted against their random intercept extracted from the best-fitting logistic mixed-effects regression reported in Chapter 2. The data for both plots were taken from the experiment reported in Chapter 2. In both plots, the yellow line shows linear regression model fit and the corresponding correlation coefficients are displayed in the lower left corner.

Recently, we have shown that estimated *rates of forgetting* are stable over time within a participant (Sense, Behrens, et al., 2016). Interestingly, the same data revealed that even though many participants performed at ceiling on the delayed recall test (see Figure 3.3 in this thesis), there was considerable variation in participants' *rates of forgetting*, even in a relatively homogenous group of learners. In datasets that were not plagued by ceiling effects in delayed recall performance, we have observed that the estimated *rate of forgetting* is very strongly correlated with subsequent test performance. Using the data from the experiment reported in Chapter 2, for example, we can compute the *rate of forgetting* for each of the 52 participant by averaging the final α value for each item they studied. Plotting the estimated *rate of forgetting* against the proportion of correct responses on the test taken 15 minutes later (Figure 4.1A) reveals a strong negative relationship (r = -0.89). This pattern is exactly what we would expect intuitively: A higher *rate of forgetting* (i.e., faster forgetting) results in lower scores on the test. In this case, the *rate of forgetting* is the aggregate end product of updating a model parameter during a 20-minute learning session and it might not be too surprising that this measure is strongly related to performance on the same items after only 15 minutes.

In the experiment this data was taken from (see Chapter 2 for details), participants studied both with the adaptive method and with digital flashcards and were tested twice on all items they studied (up to 100 in total). A logistic mixed-effects regression model was fit to predict whether individual items (not just those studied with the adaptive method) were answered correctly on two subsequent tests with varying retention intervals. Interestingly, we found a strong negative relationship of similar magnitude between the estimated *rate of forgetting* and the random intercepts from this mixed-effects regression model as well (r = -0.86; Figure 4.1B).

The random intercepts were added to the logistic mixed-effects model specifically to capture between-subject variance in ability[2]. The strong relationship apparent in Figure 4.1B suggests that the estimated *rate of forgetting* captures many of the individual differences encapsulated by the participant-specific random effects term in the logistic mixed-effects regression reported in Chapter 2. What is not clear, however, is which individual differences are reflected in the measures shown in Figure 4.1. Are they purely memory related processes or do they reflect more general aspects of the learners such as cognitive functioning and resources? The goal of this chapter is to explore the relationship between the *rate of forgetting* estimated during learning and established measures of individual cognitive differences. Specifically, we want to know whether the *rate of forgetting* simply captures well-known individual differences in general cognitive ability (GCA; i.e., fluid intelligence) and working memory capacity (WMC) or whether it could be a useful indicator of individual differences in the context of rote learning in and of itself.

The reason we chose to test a possible link between the *rate of forgetting* and GCA and WMC is that both have been used extensively as predictors of individual differences in the ability to learn. For example, Gathercole and Baddeley report that children with lower WMC were slower at learning unfamiliar names of toys (Gathercole & Baddeley, 1990) and demonstrated that WMC can predict the development of vocabulary in young children beyond what chronological age and GCA can predict (Gathercole & Baddeley, 1989). Unsworth, Brewer, and Spillers (2009) suggest that both individual differences in WMC and the ability to retrieve information from long-term memory account for the link between GCA and WMC (Unsworth, 2016; Unsworth & Engle, 2007). Both of these examples indicate that GCA and WMC are related but not identical constructs (Ackerman, Beier, & Boyle, 2005; Kane, Hambrick, & Conway, 2005) and that they can explain individual differences in learners' ability. Therefore, the *rate of forgetting* estimated by our adaptive learning system might simply reflect individual differences in GCA and/or WMC.

[2] The random intercepts are very highly correlated with the test performance displayed on the y-axis of Figure 4.1A as well: r = 0.92.

If between-subject variance in executive functioning and attentional control played an important role in fact-learning using computerized learning environments, we would expect those processes to exert their influence on two outcome measures: First, the estimated *rate of forgetting* might be correlated with GCA and/or WMC and second, delayed recall performance could be predicted by both the estimated *rate of forgetting* and WMC and GCA (or an interaction between these measures). For example, a higher GCA might lead to better recall independent of the *rate of forgetting* but WMC only modulates recall when the *rate of forgetting* is low. If, on the other hand, the *rate of forgetting* did not correlate with either GCA or WMC and is the only significant predictor of delayed recall performance in the experiment reported here, we would conclude that the model's parameter captures useful individual differences that are not expressed by individual differences in GCA and WMC.

Furthermore, given the work cited above, we have directional hypotheses about possible relationships that should be tested as such (Cho & Abe, 2013). GCA and WMC should be positively correlated with each other while their correlation with the estimated *rate of forgetting* should be negative: Higher cognitive functioning is expected to be an indicator of a lower *rate of forgetting* in a vocabulary task. With regards to the scores on a delayed recall test, the correlation coefficients are expected to be negative again with the estimated *rate of forgetting* (cf. Figure 4.1A) but positive for GCA and WMC.

METHODS

Procedure

All participants were invited for two sessions that were spaced three days apart.

Session 1. In the first session participants spent 20 minutes learning 35 Swahili-Dutch word-pairs. Participants were randomly assigned to study with one of two methods: Either they used digital flashcards or the adaptive learning method used and explained in detail in Chapters 2 and 3. As we will focus here on the results of the adaptive learning method, used by 66 participants, we will refrain from further discussion of the digital flashcard method. During learning, words were introduced on *study trials,* which showed both the cue (Swahili word) and the correct response (Dutch word) alongside an input field. *Study trials* were self-paced and participants proceeded by typing in the Dutch word. All subsequent repetitions of an item were *test trials,* which only displayed the cue and the input field. *Test trials* were followed by feedback: either a 600 ms display saying "correct" or a four second display of a *study trial* without the input field (as recommended by Zeelenberg et al., 2015).

After the study session, participants completed a personality questionnaire which took an average of 11.3 minutes to complete (range = [7; 32]). The results from the questionnaire will not be discussed here as they were part of an unrelated study.

Next, participants completed the three complex span tasks used by Foster and colleagues (2015). In these tasks, participants were shown items that needed to be recalled in the correct order at the end of each trial. Each to-be-remembered item is followed by a distractor, which requires the participant to engage executive attentional processes. This is to reduce the ability to rehearse to-be-remembered items. In the Operation Span task, for example, to-be-remembered items are letters and distractors are simple true/false equations (e.g., (2 x 2) - 1 = 3). The order of the tasks was identical across participants (Foster et al., 2015): first Operation Span, followed by Rotation Span, and then Symmetry Span. The computation of a participant's WMC based on performance in these tasks is explained in detail in the Materials sub-section below.

Finally, a test of the word-pairs that were studied at the beginning of the session was administered. All 35 Swahili cues were shown on screen as a list and the participant had to provide the correct Dutch translation. The test was self-paced and because all words were visible at the same time, participants were able to provide answers in any order they preferred. No feedback was provided.

The time it took to complete the personality questionnaire and the complex span tasks varied between participants. To ensure that the retention interval between the word-learning task and the test was identical across participants, a simple lexical decision task was administered as a filler task before the test. The task was setup in a way that it would terminate as soon as the retention interval was 80 minutes, irrespective of the number of trials completed. For the task, five-letter strings were presented on screen and participants had to press one of two buttons to indicate whether the string was a Dutch word or not. By using high frequency words, the task was made relatively easy to avoid fatigue. Although accuracy levels were below 75% for three participants, visual inspection of the response time distributions did not indicate failure to respond to the instructions[3]. Thus, no participant was excluded based on their performance in the lexical decision filler task. The data from this filler task will not be analyzed or discussed further.

Session 2. Three days later, participants came back for the second session which started with a second test of the Swahili-Dutch word-pairs learned at the beginning of the first session. The test was identical to the one completed at the end of Session 1.

Subsequently, we assessed the participant's general cognitive abilities (GCA) by adminis-

[3] One participant started responding randomly after about 500 trials. This participant performed very well in all other tasks and finished them very quickly, which meant the participant had to complete a large number of lexical decision task trials.

tering the Q1000 High Capacity test (Van Bebber, Lem, & Van Zoelen, 2010), which took participants between 32 and 87 minutes to complete. Mean completion time was 56 minutes. In contrast to more traditional tests of GCA, this test can be administered online at multiple computers simultaneously and upon completion, the website provided participants with feedback indicating how their performance compared to that of a norm group. For our analyses, we utilized raw scores on the test rather than the normed scores communicated to the participants.

Materials

Swahili-Dutch Word-Pairs. The 35 items were randomly sampled from the list of 100 Swahili-English word-pairs provided by Nelson and Dunlosky (1994). The English responses were translated to Dutch and all participants studied the same subset of 35 word-pairs. The order in which words were introduced was randomized.

Complex Span Tasks. The code for the three complex span task was obtained from the Engle lab's website[4] and used with permission. It is the same code used by Foster and colleagues (Foster et al., 2015) but all instructions were translated to Dutch. Scores reported in Table 4.1 are partial-credit unit scores (Conway et al., 2005). To express a single measure of working memory capacity (WMC), the scores on the three complex span tasks are summarized into a single composite score. This is done by calculating a participant's z-score for each task and then computing a z-score average for each participant (Foster et al., 2015). For brevity's sake, the composite score will be referred to as a participant's WMC (Conway et al., 2005).

General Cognitive Ability. As a measure of general cognitive ability, we used *Q1000 Capaciteiten Hoog* ("High Capacity"; normed for university-educated individuals) developed by Meurs HRM[5]. The test has been developed as a selection tool to determine whether a candidate has the necessary intellectual ability to perform well in cognitively demanding jobs. There are seven sub-scales that are ordered hierarchically with the goal of measuring general intelligence. The seven sub-scales can be reduced to reflect three scores for verbal, numerical, and figural capacity, respectively. Those three scores are then averaged to yield a participant's general cognitive ability (GCA). The Committee on Test Affairs Netherlands (Dutch abbreviation: COTAN) has evaluated the test and concluded that it is a valid and reliable measure of GCA (see also Van Bebber, Lem, & Van Zoelen, 2010). The ability score reported here was the mean across the proportions of correct items on each sub-scale. The resulting scores were subsequently multiplied by ten to create an easy to interpret 0-to-10 scale.

[4] www.englelab.gatech.edu/tasks.html

[5] www.meurshrm.nl/

Participants

A total of 126 participants were recruited from the Dutch first-year participant pool at the University of Groningen and participated for course credit. Of those, 89 were female (71%) and the median age was 20 (SD_{age} = 1.55; $range_{age}$ = [18, 26]). All participants spoke Dutch and no one indicated any familiarity with Swahili. All participants gave informed consent and the Ethics Committee Psychology approved the study (ID: 15006-N).

Due to technical issues, data in the Rotation Span task was lost for one participant and in the Symmetry Span task for another. The composite scores (i.e., WMC) for these two individuals are based on the z-score average of the two remaining tasks. A total of 15 participants did not show up for the second session[6], which meant that data from the second session (that is, scores on the second vocabulary test and the GCA scores) were missing. Another three participants entered invalid participant IDs on the GCA test so their data could not be recovered and five participants finished the test very quickly (in less than 30 minutes) with poor scores so their data were disregarded. Additionally, technical difficulties resulted in the loss of scores on the first test for three participants. Data on all measures were available for 103 participants (82%) but even if a participant's data were partially missing, the available subset of data were included in the analyses reported below. The number of data points contributing to each measure are summarized on the diagonal of Figure 4.2.

Analysis

Our first question – whether the *rate of forgetting* is related to GCA and/or WMC – was addressed by computing correlations among the measured constructs. More specifically, we computed Pearson's product-moment correlations and employed one-sided tests because we have clear, directional expectations about the relationships among all obtained measures (Cho & Abe, 2013). For the one-sided tests, we reported both traditional *p*-values corresponding to the one-sided tests as well as Bayes factors (Wagenmakers, Verhagen, & Ly, 2016). The Bayes factors' ability to express the likelihood of observing the data under the null hypothesis has several theoretical and practical advantages over a "null finding" in the traditional null hypothesis testing framework (Gallistel, 2009; Mulder & Wagenmakers, 2016; Wagenmakers, Morey, et al., 2016). The Bayes factors quantify the evidence the data provide for the null hypothesis relative to the alternative hypothesis such that a Bayes factor of, say, 7 can be interpreted as the data being 7 times more likely under the null model than the alternative model (Kass & Raftery, 1995). We will use the subscript "H0" for Bayes factors expressing evidence in favor of the null model and the subscript "H1" for Bayes factors expressing evidence in favor of the alternative model. A graphical summary of the results is provided in Figure 4.2.

[6] Four of which could not attend the second session because the university building was closed due to extreme weather conditions.

The second question – whether GCA and/or WMC can make significant contributions to predicting delayed recall – was addressed by fitting multiple linear regression models and comparing them using Bayesian model comparison (Rouder & Morey, 2012) to determine which (combination of) variables can best predict variance in delayed recall performance. This was done using the BayesFactor package (R. D. Morey & Rouder, 2015b) in R (R Development Core Team, 2016). Again, Bayes factors were used to express the evidence the data provided for each model relative to a reference model. As a reference model, we chose the *full model* that contains as predictors all three measures – the *rate of forgetting*, WMC, and GCA – and all interactions. A graphical summary of the results is provided in Figure 4.3.

RESULTS

A participant's working memory capacity (WMC) and their general cognitive ability (GCA) were computed as outlined in Materials above. Table 4.1 and Table 4.2 provide descriptive statistics and summarize the correlations among the components that make up someone's WMC and their GCA, respectively. For WMC, the correlations among scores on the three complex span tasks and their composite scores are positive. The two correlation coefficients between the scores on the Operation Span task and the other two complex span task are somewhat lower (.16 and .19), most likely due to the participants' relatively high scores which imposes a bit of a limit on observable correlations. For GCA, the correlations among scores on the three components and their combined score are also all positive and the data provide strong evidence that all coefficients differ from 0: the Bayes factors for the lowest coefficient is 205 in favor of the alternative model and the Bayes factors for all other coefficients are greater than 100,000.

Table 4.1. Descriptive statistics for the complex span tasks and their composite score (WMC) as well as the correlations between all measures.

	Mean	**S.D.**	**Range**	**OSpan**	**RotSpan**	**SymSpan**
Operation Span	58.9	9.6	[21, 75]			
Rotation Span	28.8	5.9	[10, 40]	.19[a]		
Symmetry Span	31.4	6.1	[14, 42]	.16[b]	.37	
WMC	0.0	0.7	[-1.8, 1.4]	.64	.74	.72

Note – [a] The one-sided *p*-value is < .05 and the corresponding BF_{H0} is 1.8; [b] The one-sided *p*-value is < .05 and the corresponding BF_{H0} is 0.9; The one-sided *p*-values for all other coefficients are < .001 and the corresponding BFs in favor of the null hypothesis are well over 1,000.

The relationships among all measures are summarized in Figure 4.2. The plot depicts the distribution of each measure – along with the number of available observations – on the diagonal. On the off-diagonal, scatterplots with fitted linear regression lines are shown (Anscombe, 1973). The corresponding Pearson correlations are shown on the other off-diagonal, along with one-sided *p*-values expressing the probability of observing the data assuming the coefficient is 0. The one-sided alternative hypothesis is that the coefficient is greater than 0 (except for the correlations with the *rate of forgetting* for which we would assume a negative correlation, if any). Also shown are the Bayes factor equivalents of the directional null-hypothesis significance tests (Wagenmakers, Verhagen, et al., 2016).

Table 4.2. Descriptive statistics for the Q1000 Capacity Test for the three components and the combined general cognitive ability (GCA) score. Also shown are the correlations between all measures.

	Mean	**S.D.**	**Range**	**Numeric**	**Figural**	**Verbal**
Numeric	6.1	1.6	[2.0, 9.7]			
Figural	6.7	1.5	[2.5, 9.6]	.50		
Verbal	6.5	1.1	[3.3, 8.4]	.48	.36	
GCA	6.4	1.1	[3.6, 9.0]	.83	.77	.78

Note – All correlation coefficients differ significantly from 0 with p < .001 (one-sided). The corresponding BF_{H0} is 205 for the lowest coefficient and well over one hundred thousand for all others.

For each of the 66 participants that studied with the adaptive method, the *rate of forgetting* was computed by averaging the final α values of all items with at least three test trials. The mean *rate of forgetting* is .295 with a standard deviation of 0.045 (range = [0.186; 0.442]). Since a higher *rate of forgetting* indicates faster forgetting, we would expect all correlations in the leftmost column of Figure 4.2 to be negative. This is confirmed for both the first and second test scores reported in the second and third row/column. The correlations are very high suggesting that the estimated *rate of forgetting* is a good predictor of test performance both 80 minutes and three days later.

This replicates earlier, unpublished findings from our lab such as those shown in Figure 4.1A. The correlation between someone's *rate of forgetting* and their WMC is also negative, but the data does not provide strong evidence for either the null or alternative hypothesis (BF_{H1} = 1.4) even though the traditional one-sided frequentist test suggests a significant difference from 0 ($t(64) = -1.81$; $p = 0.038$). With regards to the correlations between *rates of forgetting* and GCA, the Bayes factors provide positive evidence that the null model is more likely given the data (BF_{H0} = 3.5) and the *p*-values corroborate that the coefficient does not differ significantly

from zero ($t(51) = -.59$; $p = 0.281$). The scores from the two Swahili-Dutch tests are highly and positively correlated with each other, which is not surprising. We would expect test scores to correlate positively with WMC and GCA and all coefficients are indeed positive (see Figure 4.2). The data provide weak evidence that the score on the first test might be related to WMC ($r = 0.21$; $BF_{H1} = 3.1$) but this relationship breaks down for the test score obtained three days later ($r = 0.11$; $BF_{H0} = 2.5$).

For the relationship between GCA and the two test scores, the null hypothesis that there is no correlation cannot be rejected and the Bayes factors favor the null model as well but evidence is weak ($r = 0.10$ with $BF_{H0} = 3.1$ and $r = 0.13$ with $BF_{H0} = 2.0$ for the first and second test, respectively). Visual inspection of the corresponding scatter plots in Figure 4.2 confirms the absence of a clear linear (or any) relationship with any other measure. The weak positive correlation between the first test score and WMC might be partially driven by at ceiling performance on the test by some participants, which might explain why the correlation is not significant anymore when the performance decreases from the first to the second test.

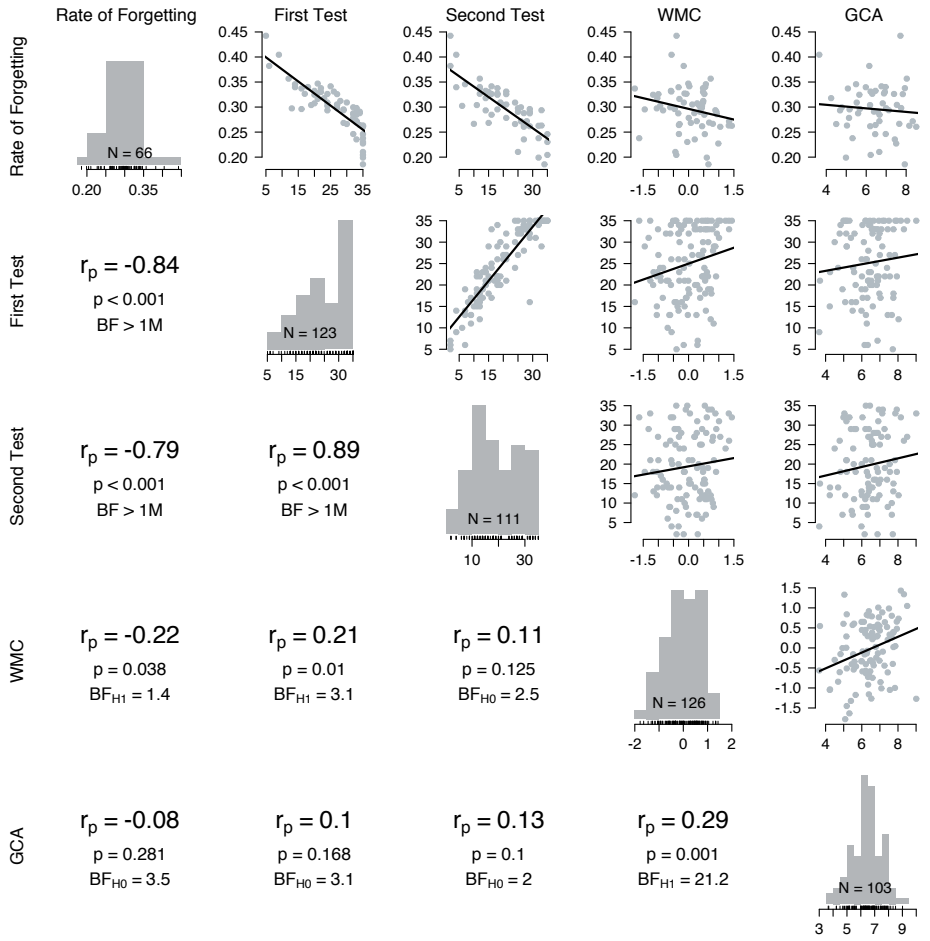


Figure 4.2. Depicted are the measures of interest: the estimated rate of forgetting, the scores on the first and second test, working memory capacity (WMC), and general cognitive ability (GCA). Note that p-values and Bayes factors correspond to the directional hypothesis that correlation coefficients are positive (except for the correlations with the rate of forgetting in the leftmost column: those are expected to be negative).

Given the large body of previous research on the relationship between WMC and GCA (Conway, Kane, & Engle, 2003; Kane et al., 2005), we would expect WMC and GCA to be positively correlated. This directional hypothesis is confirmed in our data: The *p*-value indicates a significant, positive deviation from 0 and the Bayes factor confirms that there is very strong evidence for a positive correlation (i.e., BF_{H1} = 21.2). The magnitude of the correlation coefficient (i.e., .29) is also roughly in the range we would expect (Conway et al., 2005).

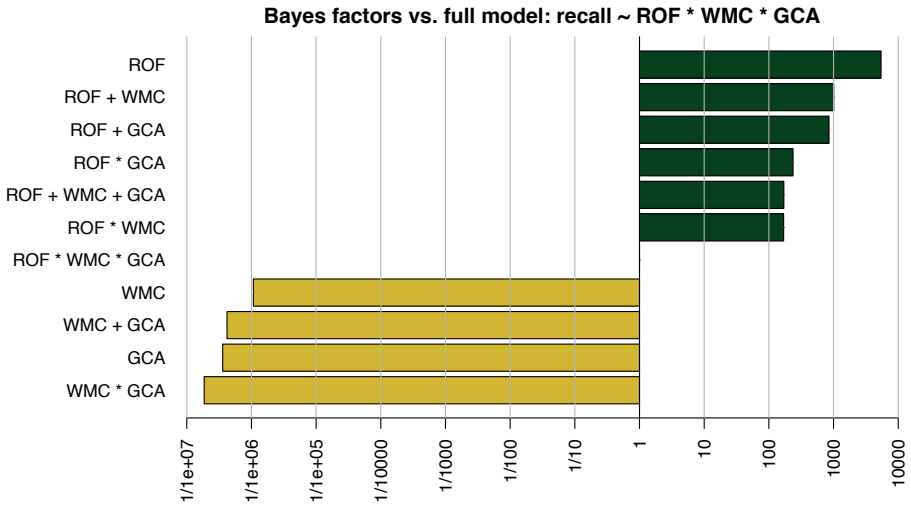


Figure 4.3. The Bayes factor (x-axis) for each model (y-axis) relative to the full model including the three constructs and all their interactions. In the model names on the y-axis, a "+" denotes that only main effects for the listed variables are included and an "*" denotes that both main and interaction effects are included. The variables abbreviations are ROF = rate of forgetting, WMC = working memory capacity, and GCA = general cognitive ability. The dependent variable in all models is the delayed recall performance three days later.

To test to which extent the three measures of interest – *rate of forgetting*, GCA, and WMC – can predict delayed recall, we fit a series of (multiple) linear regression models and compared them using Bayes factor model selection procedures (Rouder & Morey, 2012). The outcomes are summarized in Figure 4.3 and show that the best-fitting model includes only the estimated *rate of forgetting* as a main effect. The worst model, on the other hand, includes both main and interaction effects of WMC and GCA but not the *rate of forgetting*. Since all models are compared with the *full model*, that particular model has a Bayes factor of 1. A striking pattern in Figure 4.3 is that the *full model* demarcates two clusters of models: The first cluster has Bayes factors much greater than 1 (i.e., provide better fits relative to the *full model*) and all models in that cluster contain the *rate of forgetting* as a predictor. Furthermore, the models in that cluster decrease in complexity as the Bayes factors increase. The simplest model includes *only* the *rate of forgetting* as a predictor and fits the data 5.4 and 6.4 times better than the second and third models, respectively. The second cluster of models has Bayes factors much smaller than 1 and does not include the estimated *rate of forgetting*. The same pattern is apparent: the more complex the model, the worse it does.

To summarize, the results of comparing the models provide clear evidence that the estimated *rate of forgetting* alone is the best predictor of delayed recall. Adding either WMC or

GCA (and any interaction) increases the model complexity without explaining more variance in recall performance. This finding is substantiated by the correlations reported in Figure 4.3. The correlation coefficients can be squared to obtain the amount of explained variance in delayed recall for the three models that only contain a single predictor. This, again, clearly demonstrates that the *rate of forgetting* can explain a lot more variance in recall three days later than WMC or GCA: 62.4% compared to 1.2% and 1.7%, respectively.

DISCUSSION

Here, we explored whether two commonly used measures of individual cognitive differences – working memory capacity (WMC) and general cognitive ability (GCA) – are related to the *rate of forgetting* estimated while studying with the adaptive fact-learning system developed in our lab. We also tested whether those measures can predict delayed recall performance independent of or in conjunction with the *rate of forgetting*. The data presented here suggest that the *rate of forgetting* is neither correlated with GCA nor WMC and that only the *rate of forgetting* can explain significant amounts of variance in delayed recall performance.

The very high correlations between the *rate of forgetting* and subsequent test performance on the studied material reported here (see Figure 4.3) directly replicate earlier, unpublished findings (see Figure 4.2 but also section 3.2.2.4 in Nijboer's (2011) Master thesis). Estimating how well a learner has mastered the studied material is crucial to adaptive learning environments because this estimate will directly influence which items are selected for repetition (see Settles & Meeder, 2016 for a good example). Therefore, the ability to capture individual differences in the mastery of studied material is a crucial benchmark of any adaptive learning system. Based on the data presented here, we can conclude that the model developed in our lab passes this test.

In the setup used here, the learner has considerable control over certain aspects of the task. The adaptive system determines the order in which items are repeated but the learners can respond as they please and employ any additional rehearsal strategies or mnemonic devices they deem useful. This, presumably, leaves room for variation in executive functioning and attentional control between learners to exert their influence on the testable end product of the learning session: delayed recall performance. Thus, we would expect that the setup of the current experiment provides ample opportunity for individual differences in cognitive functioning to exercise their influence; both during learning and on subsequent test performance. Surprisingly, we find no relationship between the two measured constructs of executive functioning and attentional control – GCA and WMC – and delayed recall performance

in the homogenous sample tested here. The estimated *rate of forgetting*, on the other hand, does make substantial contributions to explaining variance in delayed recall performance. In fact, Figure 4.3 makes clear that any model that includes the estimated *rate of forgetting* as a predictor vastly outperforms any model that does not. The preferred model is the one including *only* the estimated *rate of forgetting*, which indicates that adding either WMC or GCA (and any interaction between the two) increases the complexity of the model without explaining a sufficient amount of additional variance in delayed recall performance.

These findings suggest that – at least in the current sample – executive functioning and attentional control do not play important roles in predicting delayed recall of items studied using the adaptive fact-learning system developed in our lab. The results of the analysis summarized in Figure 4.3 demonstrate that the *rate of forgetting* is the single best predictor of delayed recall performance. This means that the adaptive method is able to utilize a learner's responses during learning to dynamically adjust its internal parameters in a way that reflects the learner's current ability. Figure 4.3 also shows that the estimated *rate of forgetting* is not only the best predictor but any model that does not include it is at least a million times less likely given the data than even the most complex model that does include it.

It should be noted that GCA and WMC are both constructs that are validated as indicators of cognitive performance that are domain-general (e.g., WMC: Conway et al., 2005; GCA: Deary, 2012). The same cannot be said about the parameter extracted from the ACT-R-based model employed here. The model is based on a theoretical framework specifically designed to trace the temporal dynamics of declarative memory processes (see, e.g., J. R. Anderson, Bothell, Lebiere, & Matessa, 1998). As such, it is reassuring that the model's parameter can capture relevant memory processes that allow the design of effective, personalized learning schedules (see Chapter 2) and allows predicting learners' delayed recall of the studied material. As outlined in the Introduction, both WMC and GCA have been reported to be implicated in memory-related processes. However, the findings reported here suggest that such domain-general measures do not capture individual differences relevant to constructing adaptive learning environments of the kind studied here. The domain-specific *rate of forgetting*, on the other hand, does. More generally, this suggests that it is not necessary for computerized learning environments to gather information regarding learners' GCA and/or WMC to predict fact-learning ability.

There are many other adaptive models specialized in fact-learning (e.g., Lindsey et al., 2014; Papoušek, Pelánek, & Stanislav, 2014; Settles & Meeder, 2016; Woźniak & Gorzelańczyk, 1994). The exact mechanism employed by these models differ but the approach is the same: A new learner starts using the system and the underlying model assumes the learner's behavior can be described by a set of equations. These equations have free parameters to accommo-

date between-learner variation. However, the optimal parameters for that particular learner are unknown and models differ with regards to the techniques they use to approximate a learner's optimal parameters based on information gathered through responses. Before such responses can be collected, however, nothing is known about the new learner and the models use default parameters as starting values. Here, we explored the possibility of using additional information about participants obtained *outside* the learning session. Specifically, we chose to test two commonly used measures of cognitive functioning (WMC and GCA), hypothesizing that they might exert an influence alongside purely memory-based processes. We explored whether such measures are related to relevant aspects of fact-learning – in this case, the parameter estimated by our model and the delayed recall performance on the studied material. Because if they were, one could maybe use such additional information to pick personalized starting parameters. Unfortunately, we could neither find evidence that WMC or GCA were correlated significantly with the estimated *rate of forgetting* nor that they contributed to predicting delayed recall.

A certain degree of caution is appropriate when interpreting our results and generalizing them to other samples and contexts. Our sample was relatively small and very homogeneous. We only recruited participants who spoke Dutch because the test of GCA was only available in Dutch. This means that prior academic achievement played a big role in the pre-selection of participants, who will have to have completed their pre-university education to be admitted to the program. The participants in our sample are also around the same age with 80% of the participants between the age of 18 and 21. We think it is reasonable to assume that variations in WMC and GCA would have been related to delayed recall performance if a broader sample of the population had been tested.

For the given sample, it is interesting to note that neither WMC nor GCA can distinguish between participants on the outcome measure. Delayed recall on a set of studied word pairs is an outcome measure with high face-validity – for example, it is often used in school to assess students' progress. Therefore, the strong relationship between estimated *rates of forgetting* and delayed recall performance reported here is both of theoretical interest and practical relevance and an attempt to replicate it in a more heterogeneous sample would be an interesting extension of the current work.

To summarize, we present data from a correlational study in which we measured someone's *rate of forgetting* during a fact-learning session alongside their working memory capacity and general cognitive ability. Neither of the two common individual differences measures are related to someone's *rate of forgetting*. Moreover, we replicate previous findings showing a very high negative correlation between the estimated *rate of forgetting* and subsequent recall. Additionally, we show that neither working memory capacity nor general cognitive

ability are related to delayed recall. Keeping the limitations of the sample in mind, this implies that someone's *rate of forgetting* is the only individual differences measure tested here that can predict delayed recall. Furthermore, these findings suggest that even in the relatively homogenous sample tested here, between-subject variation in *rates of forgetting* capture useful individual differences that are independent of both working memory capacity and general cognitive ability. This implies that the model's parameter is not just an artifact of differences in higher cognitive functioning but encapsulates information about a learner that is useful in the context of a personalized fact-learning system.

5

Opportunity for Verbalization Does not Improve Visual Change Detection Performance: A State-Trace Analysis

Acknowledgements

This chapter has been published as: Sense, F., Morey, C. C., Prince, M., Heathcote, A., & Morey, R. D. (2016). Opportunity for verbalization does not improve visual change detection performance: A state-trace analysis. *Behavior Research Methods,* 1-10.

We would like to thank Yongqui Cong, Christian Hummeluhr, and Mareike Kirsch for valuable assistance with data collection.

Supplementary materials for this chapter are available at www.osf.io/wzdvq

Abstract

Evidence suggests that there is a tendency to verbally recode visually-presented information, and that in some cases verbal recoding can boost memory performance. According to multi-component models of working memory, memory performance is increased because task-relevant information is simultaneously maintained in two codes. The possibility of dual encoding is problematic if the goal is to measure capacity for visual information exclusively. To counteract this possibility, articulatory suppression is frequently used with visual change detection tasks specifically to prevent verbalization of visual stimuli. But is this precaution always necessary? There is little reason to believe that concurrent articulation affects performance in typical visual change detection tasks, suggesting that verbal recoding might not be likely to occur in this paradigm, and if not, precautionary articulatory suppression would not always be necessary. We present evidence confirming that articulatory suppression has no discernible effect on performance in a typical visual change-detection task in which abstract patterns are briefly presented. A comprehensive analysis using both descriptive statistics and Bayesian state-trace analysis revealed no evidence for any complex relationship between articulatory suppression and performance that would be consistent with a verbal recoding explanation. Instead, the evidence favors the simpler explanation that verbal strategies were either not deployed in the task or, if they were, were not effective in improving performance, and thus have no influence on visual working memory as measured during visual change detection. We conclude that in visual change detection experiments in which abstract visual stimuli are briefly presented, pre-cautionary articulatory suppression is unnecessary.

During his seminal experiments on human memory, Sperling noticed that many of his participants verbalized and repeated to-be-remembered material during retention, even if the studied material was not aurally presented. Sperling (1967) pointed out that visual information can be verbalized and many people reported doing so. This reflection confirmed intuitions that regardless of presentation modality, information may be encoded with some flexibility of representation: visual materials might be maintained in a verbal code, and imagery corresponding to verbal input may likewise become active.

However, demonstrating that recoding can occur does not imply that it always occurs, nor that it is beneficial. Murray (1965) showed that saying visually-presented verbal stimuli out loud improves recall performance relative to mouthing them silently. However, this relationship only seems to persist if the visually-presented material can be verbalized effectively (e.g., verbal stimuli, nameable visual images). The idea that it is the opportunity to rehearse these verbal codes that improves performance also remains a matter for debate, even for serially-ordered verbal stimuli (Lewandowsky & Oberauer, 2015). Attempts to verbalize stimuli that are difficult to describe succinctly and accurately (e.g., faces) might actually harm performance (Schooler & Engstler-Schooler, 1990). Brandimonte et al. (1992) showed that verbal recoding can be detrimental to a subsequent mental rotation task when the remembered verbal label is not relevant or helpful. What such experiments suggest is that there is a strong tendency to verbally recode visually-presented information, and that in some cases verbal recoding may boost memory performance. This logic is consistent with multi-component models of working memory, which propose that separate short-term memory stores for phonological and visual information can be applied to a short-term memory task (Baddeley, 1986). Naturally, if task-relevant information can be maintained simultaneously in two useful codes, one would expect memory performance to improve.

The possibility of dual encoding is problematic though if the goal is to measure capacity for visual information exclusively. Levy (1971) suggested a method of preventing such recoding via meaningless concurrent articulation. By repeating irrelevant syllables out loud during presentation and retention of visual information, participants' ability to verbally recode visually-presented stimuli is restricted. This procedure is known as *articulatory suppression* and is commonly used alongside visual change detection tasks with the specifically-stated intention that it is meant to prevent verbalization of visual stimuli (Allen, Baddeley, & Hitch, 2006; Brockmole, Parra, Sala, & Logie, 2008; Delvenne & Bruyer, 2004; Hollingworth & Rasmussen, 2010; Logie, Brockmole, & Vandenbroucke, 2009; Makovski & Jiang, 2008; Makovski, Sussman, & Jiang, 2008; Matsukura & Hollingworth, 2011; Treisman & Zhang, 2006; van Lamsweerde & Beck, 2012; Woodman & Vogel, 2005, 2008). This precaution is undertaken to ensure that task performance reflects visual memory, rather than some combination of memory for visual

images and verbal codes.

The use of precautionary articulatory suppression is common practice despite evidence that articulatory suppression has not been shown to have a measurable effect on some visual change detection tasks (Luria, Sessa, Gotler, Jolicœur, & Dell'Acqua, 2010; Mate, Allen, & Baqués, 2012; C. C. Morey & Cowan, 2004, 2005), nor have small verbal memory loads (Vogel, Woodman, & Luck, 2001). These studies imply that the precaution of employing articulatory suppression may be unnecessary: participants performed no better without articulatory suppression than with it, suggesting that verbal recoding is not the default strategy for visual change detection tasks as typically administered. However, these findings simply report null effects of meaningless articulatory suppression on visual memory tasks, and therefore cannot be taken as strong evidence of the absence of some effect, given sufficient power to detect it. Until a stronger case against verbal recoding during visual change detection can be made, enforcing articulatory suppression to prevent verbalization of visual images is a reasonable way for researchers to better ensure that their measure of visual memory performance is pure. However, enforcing articulation adds a substantial burden to an experiment from both the participant's and the experimenter's point of view. If a strong case could be made that possible verbal recoding of visual memoranda does not affect visual memory performance, researchers would be free to forgo including articulatory suppression from some designs.

We report evidence suggesting that articulatory suppression has no discernible effect on performance in a typical visual change-detection task. The experiment was designed so that some change-detection conditions encouraged verbalization by presenting memoranda one at a time. In all cases, the stimuli were arrays of distinctly-colored squares, and the object was to remember the location of each color. We manipulated the number of items in each array, whether the squares were presented simultaneously or sequentially, and whether participants performed articulatory suppression or not. If participants tend to verbally label the stimuli, and if verbal labeling assists the recognition decision, we would expect to observe at least a small benefit of silence over articulation in all conditions. It may also be the case that participants strategically choose when to verbally recode stimuli. If so, we would expect to see selective impairments with articulation for sequentially-presented items, perhaps most strongly for small set sizes where naming all the items might have occurred. In order to discern between small effects of articulation and the null hypothesis of no effect at all, we employ two modes of analysis: first, we provide a straightforward analysis based on descriptive statistics that shows that the effects tend to go in the reverse direction to what is predicted, ruling out evidence for the predicted effect; and second, we employed Bayesian state-trace analysis to show that participants show data patterns more consistent with a single-parameter explanation (visual short term memory) than a more complicated explanation (visual short term memory plus verbal short term memory).

METHODS

Participants performed a visual array change detection task under four conditions formed by the cross of two presentation conditions (sequential and simultaneous) and two articulation conditions (silent and articulatory suppression). The simultaneous presentation condition was the same as a standard visual array change detection task; in the sequential condition, the stimuli were presented one after another. We assumed that presenting visual stimuli sequentially would afford a better opportunity to engage in verbalization, if such verbalization occurs (Woodman, Vogel, & Luck, 2012). Articulatory suppression is supposed to prevent participants from employing subvocal verbalization. The combination of simultaneous/sequential and silent/articulate conditions creates combinations of conditions that discourage participants from recruiting verbal resources (i.e. articulate, simultaneous trials) as well as those that make it more likely they could benefit from verbalization (i.e. silent, sequential trials).

Participants

Fifteen participants (8 female) between the age of 21 and 31 (M = 25.4, SD = 2.67) were recruited from the population of Groningen. Participants were paid 10€ per 90-minute session and recruited through a local online social media group.

All participants were pre-screened for colorblindness and medication use that might affect their cognitive abilities, and all participants reported normal or corrected-to-normal vision and normal hearing. Furthermore, participants were only invited for subsequent sessions if they scored at least 85% correct on set-size-two trials (across all conditions) in the first session. This cut-off value was chosen based on an unpublished pilot study in which 14 out of the 15 pilot participants performed above 85%, and the remaining low-performing participant scored near chance (50%) and was assumed to have ignored the instructions. All fifteen participants in our final sample met this criterion. One of these participants completed only four sessions due to scheduling difficulties, while the remainder of the final sample completed five sessions.

Apparatus and Stimuli

The experiment was conducted using MATLAB (2011) using the Psychophysics Toolbox extensions (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007; Pelli, 1997). The stimuli were colored squares approximately 0.65° x 0.65° presented within a 7.3°x 9.8° area around the screen's center. On each trial, the colors were randomly sampled without replacement from a set of nine easily-discriminable colors and presented on a gray background. The set of possible colors was identical to the one used by Rouder et al. (2008) with the exception that black was

excluded, since Morey (2011) showed that black exhibited markedly different effects in a similar change detection task. Stimuli were shown against a neutral gray background. The items within a single array were always arranged with a minimum distance of 2° from one another and participants sat approximately 50 *cm* from the monitors. This setup allowed them to see the entire display without moving their heads.

Feedback was given via one of three clearly-discriminable sounds signaling a correct, incorrect, or invalid response (i.e., a key that was not assigned to either of the two valid responses). The sounds were played through headphones worn throughout the entire experiment.

Procedure

Within each session, participants completed one block of trials in which subvocal articulation was suppressed by requiring them to repeat aloud the syllables "ta" and "da" (*articulation* block) and one block in which no such articulatory suppression was enforced (*silent* block). Both the articulation and the silent blocks were further sub-divided in two blocks: one in which stimuli were presented simultaneously and one in which they were presented sequentially. The order in which blocks were completed was determined based on the participants' IDs and identical in each session. There were 504 trials in each session, yielding a total of 2,520 trials per participant (except for participant 10 who came in for four sessions, contributing 2,016 instead of 2,520 trials).

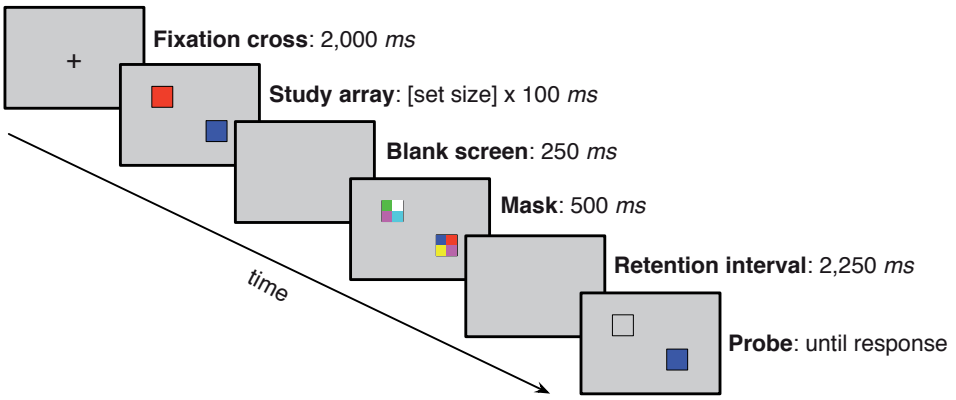


Figure 5.1. A schematic representation of a set size two trial in the simultaneous presentation condition. Note that the image is not to scale.

The overall structure of the task is depicted in Figure 5.1. The trial started with a fixation cross that was on screen for 2,000 *ms*. The study time in the simultaneous block was a linear function of the set size (study time = set size x 100 *ms*) and the set sizes were 2, 4, and 8.

We adopted this deviation from the more typical visual change detection task in which the timing of the stimulus display is constant in order to ensure that exposure time to the objects was constant across the simultaneous/sequential manipulation. In the sequential block, the stimuli appeared one after another. Each stimulus was shown with a thin, black outline and remained on the screen for 100 *ms*. The stimulus color was then replaced with the background gray color and the black outline remained. After an inter-stimulus interval of 200 *ms* the following stimulus appeared on screen. The outlines of all stimuli remained on screen until a mask appeared. There was a 250 *ms* blank screen between the study array (or the final stimulus color in the sequential presentation) and the mask. The mask was displayed for 500 *ms*. Each individual stimulus mask was made up of a 4-by-4 grid of colored rectangles and the colors were randomly chosen from the same color set as the whole array. After the mask disappeared, a 2,250 *ms* retention interval (blank screen) delayed the onset of a single probe. The probe remained on screen until the participant made a response. Alongside the probe were thin, black outlines of the other stimuli from the study array, which were displayed to prevent the participant from being unsure about which of the studied stimuli was probed.

RESULTS

Prior to data analysis, all trials containing invalid responses (0.1% of trials) were removed, and trials with unusually long or short response times (<200 *ms* or >3 *s*; 2% of trials) were excluded. The overwhelming majority of these were too slow, possibly because participants took unscheduled breaks by deliberately delaying their response. Overall, 36,495 trials across the 15 participants remained for analysis. Descriptive statistics for task performance across conditions are summarized in Figure 5.2. Overall accuracy is high in the set size 2 condition, as expected, and decreases as set size increases. In addition, Table 5.1 shows the mean hit and false alarm rates across all participants.

Table 5.1. Mean hit and false alarm rates for all conditions across all participants. Numbers in parentheses are standard deviations of the corresponding means.

	Hits				
	Simultaneous			**Sequential**	
Set size	**Articulate**	**Silent**		**Articulate**	**silent**
2	0.95 (0.022)	0.95 (0.036)		0.94 (0.026)	0.94 (0.032)
4	0.84 (0.065)	0.88 (0.063)		0.82 (0.076)	0.82 (0.097)
8	0.72 (0.079)	0.70 (0.100)		0.70 (0.101)	0.70 (0.174)
	False alarms				
2	0.08 (0.040)	0.06 (0.035)		0.11 (0.066)	0.07 (0.039)
4	0.25 (0.131)	0.22 (0.132)		0.31 (0.166)	0.26 (0.154)
8	0.41 (0.120)	0.39 (0.138)		0.41 (0.142)	0.42 (0.159)

In order to assess the performance while controlling for response bias, for each condition-participant-set size combination we subtracted the false alarm rate from the hit rate to form an overall performance measure *d* (Cowan et al., 2005; Rouder, Morey, Morey, & Cowan, 2011). Of particular interest is how the performance advantage for the silent condition is affected by the type of presentation. If participants verbalize when the presentation is sequential, we would predict that articulation would hurt performance more with sequential presentation, and thus the advantage for the silent condition would be larger with sequential presentation.

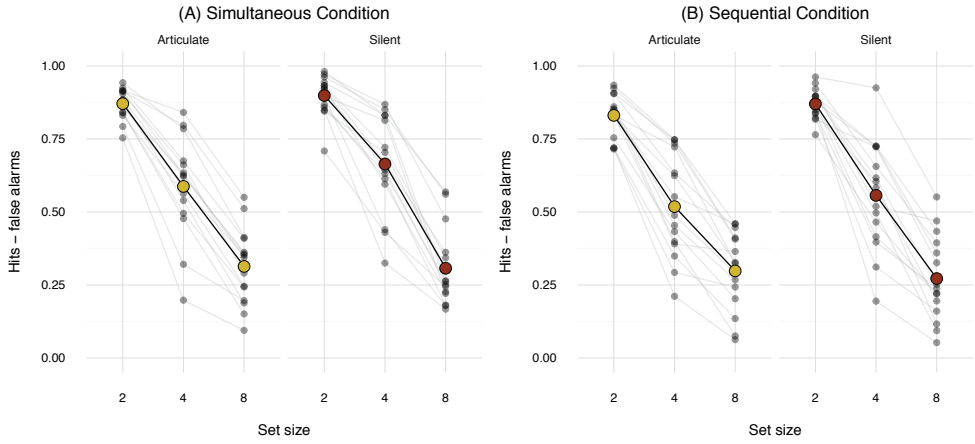


Figure 5.2. Descriptive statistics for the relevant performance measure d across the different conditions of the experiment. Semi-transparent black circles show the mean performance in each condition per participant and lines connect individual participants' means. Larger, colored symbols are group means for each condition, connected by thicker, black lines.

Figure 5.3A plots the silent advantage in the simultaneous condition as a function of the same for the sequential condition for all participant by set size combinations. If participants were verbalizing, then being silent should aid performance. Moreover, being silent should aid performance *more* in the sequential-presentation condition than in the simultaneous-presentation condition. This prediction would appear in Figure 5.3A as points falling below the diagonal. However, 28 out of the 45 points actually fall *above* the diagonal, inconsistent with the verbalization hypothesis.

It is plausible to suppose that participants only sometimes engage in verbal recoding, perhaps when it is most natural, or when they believe it will be most helpful (Larsen & Baddeley, 2003). Larsen and Baddeley surmised that participants abandon articulatory rehearsal with long or otherwise difficult-to-rehearse verbal lists. Building on this assumption, one might imagine that participants engage in strategic verbal recoding for small set sizes where helpful, distinct labels may be generated for each item, but abandon this strategy for larger set sizes. However, for all set sizes, the number of points above the diagonal in Figure 5.3A is greater than one-half: 8/15, 10/15, and 10/15 points lie above the diagonal for set sizes 2, 4, and 8, respectively. There is no evidence of the predicted effect in these data; instead, the effect appears to go in the wrong direction.

We also examined whether the apparent lack of an effect may be due to differences in strategy over the experimental sessions; however, a similar picture emerges when the effect is examined across time, as in Figure 5.3B. The verbalization hypothesis would predict that

points would fall above the horizontal line at 0 on average; however, if anything, the points tend to fall *below* the line.

Given the descriptive analysis above, we eschew typical ANOVA analyses in favor of reliance on a state-trace analysis[1]. We have the luxury of avoiding the assumption-laden ANOVA because we have directional predictions that are violated in the data. Thus, there cannot be evidence for the prediction of interest. Furthermore, we are interested in the dimensionality of the latent system that has produced the observed data – a question that an ANOVA, unlike state-trace analysis (Prince, Brown, & Heathcote, 2012), cannot provide a reliable answer to. The state-trace analysis complements the descriptive analysis by showing that the data are highly consistent with a simple explanation: that performance is governed by a single latent variable (interpreted as visual short term memory capacity) and no more complicated explanation involving verbalization is needed.

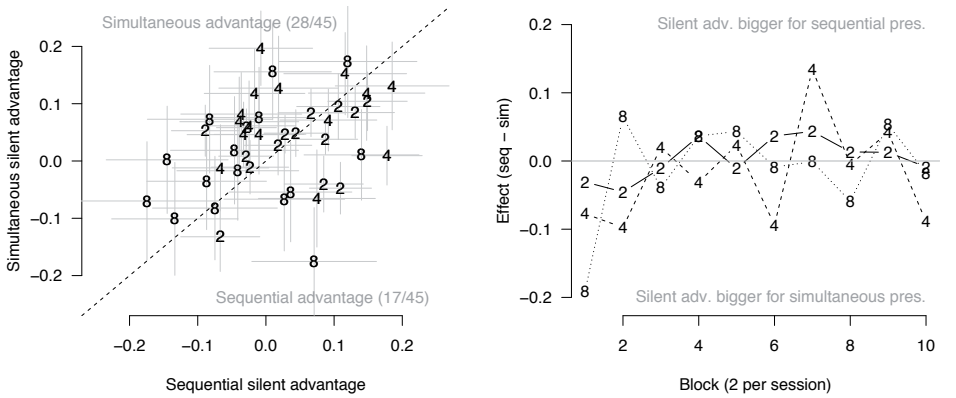


Figure 5.3. A: Advantage for silent condition (i.e., d in silent condition minus d in articulate condition) with simultaneous presentation as a function of the same for sequential presentation. Each point represents a single participant and set size. Error bars are approximate standard errors. B: The difference between the advantage for the silent condition in the sequential and simultaneous presentation conditions as a function of experimental block. In both plots, the number for each point represents the set size.

State-trace analysis

Another way to examine whether there is any evidence for verbalization is a *state-trace analysis*. State-trace analysis, outlined in its original form by Bamber (1979), is a data analysis technique intended to reveal how many latent dimensions a system requires to produce observed empirical results (see Prince et al., 2012 for an overview and the application of Bayes-

[1] For those interested, a traditional repeated measures ANOVA has been included in the online supplement available at www.osf.io/wzdvq/.

ian analysis). A simple system may have only one latent dimension (e.g., working memory capacity in general, or visual working memory capacity specifically), and all experimental manipulations affect performance along that latent dimension. More complex systems may show relationships that are impossible to explain by a single dimension, and therefore require positing more latent constructs (see section *Diagnosing Dimensionality* in Prince et al., 2012 for a detailed explanation based on hypothetical examples).

Considering visual change detection performance, one might imagine that only one latent memory dimension contributes to recognition accuracy or alternatively that separate visual and verbal memory systems jointly contribute to recognition accuracy. The multi-component model of working memory (Baddeley, 1986) proposes sub-systems for verbal and visual short-term memory, and would be consistent with the suggestion that both verbal and visual codes are stored during visual array memory, with both codes contributing to recognition accuracy. This assumption is the reason why precautionary articulatory suppression is so often employed during visual memory tasks. One reasonable prediction of the multi-component model is thus that at least two latent factors, verbal and visual memory, contribute to visual change recognition accuracy. Another reasonable expectation is that whether or not verbal encoding occurs, it is insufficient to affect recognition accuracy in this task, and in that case, a single dimension would better explain recognition accuracy in visual change detection. If visual change detection performance in our study, which was explicitly designed to allow verbalization to exert effects in specific conditions, can be explained by a single latent dimension then we would conclude that articulatory suppression is not needed to prevent verbalization in tasks with similar designs.

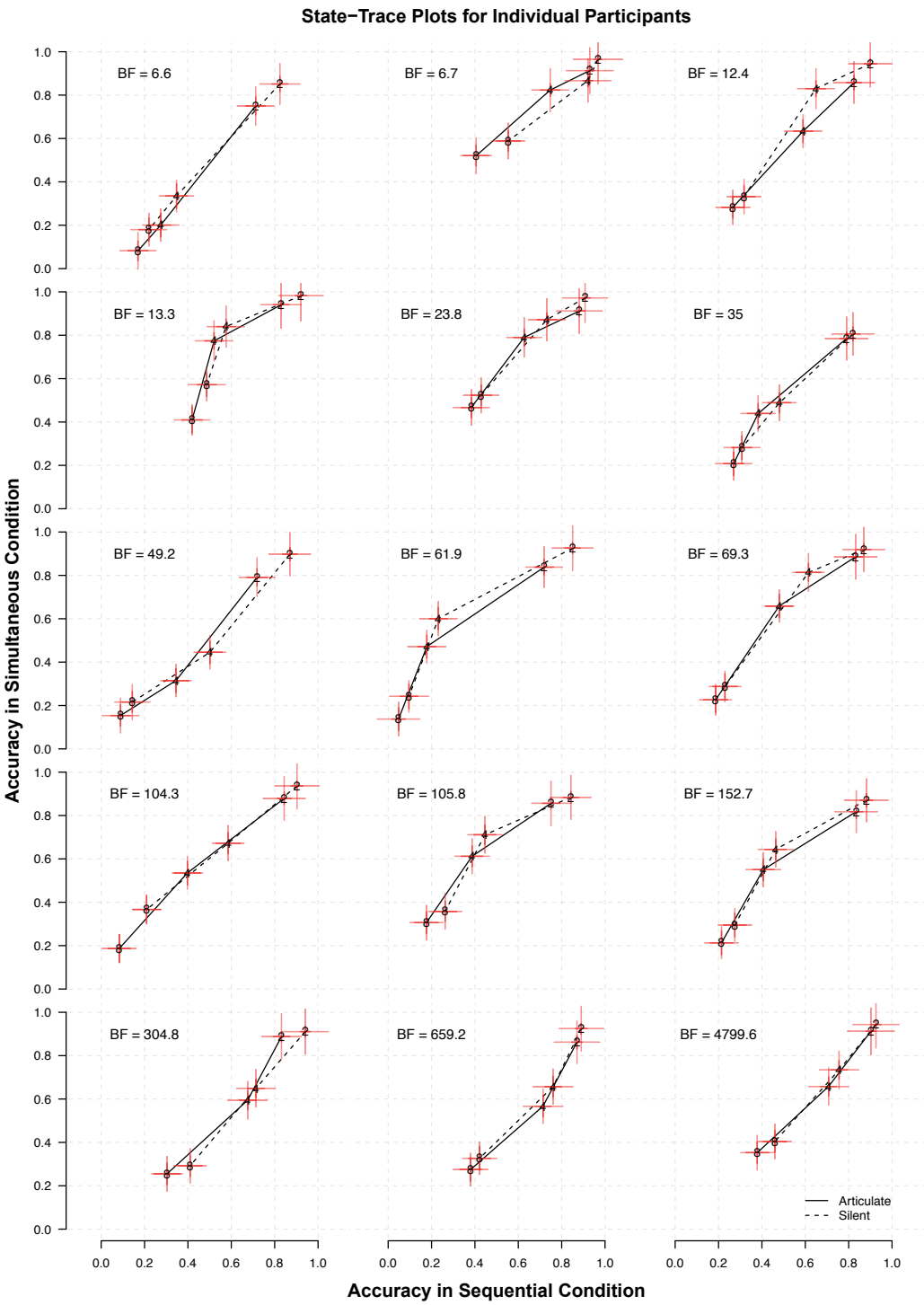
In the logic of state-trace analysis, performance in the sequential and simultaneous presentation conditions arise from either one or more latent constructs. If they both arise from a single latent variable, such as (visual) working memory capacity –- and if performance in both is a monotone function of the latent variable -- then performance in the sequential presentation must be a monotone function of performance in the simultaneous condition. To the extent that no monotone function can describe the relationship between simultaneous and sequential task performance, two latent constructs – perhaps distinct visual and verbal working memory capacities – are assumed to be needed to describe the performance.

For the state-trace analysis, we again used *d*, the hit rate minus the false alarm rate, as a measure of performance in our simulations. To reduce possibly spurious deviations in our simulations, we computed Bayesian estimates of *d* applying three reasonable constraints: first, we assumed that the true hit rate was greater than the true false alarm rate, and thus performance was truly above chance. Second, for both the sequential and the simultaneous condition, *d* must decrease with increasing array set size; for instance, true *d* to a set size of

8 cannot be better than performance to set size 4, all other things being equal. Third, it was assumed that suppression cannot benefit performance; for each set size and presentation condition, the true *d* in the articulate condition must be less than in the silent condition. This restriction was applied because a small dual-task cost appearing in all conditions would be consistent within any working memory theory, and with our distinctly-colored stimuli and meaningless articulation instructions, no benefit of articulation was reasonably expected. When a simulation produced one of these patterns, we excluded it and replaced it. Estimating the true discrimination under these restrictions yields a less error-prone measure of performance due to the exclusion of simulations with implausible data patterns.

Figure 5.4 shows the state-trace plots for each participant, formed by plotting estimated performance in the simultaneous presentation condition against the performance in the sequential condition. State-trace logic says that more than one latent construct is needed to explain the data when these points cannot be joined by a single, monotone curve; however, as can be seen from the state-trace plots for all participants, the state-trace plots are strikingly monotone. There does not appear to be any evidence that more than a single latent construct – (visual) working memory capacity – is needed, and thus no evidence that verbalization plays a role in performance in this task.

Figure 5.4. Individual state-trace plots for all 15 participants. The dependent variables are hit rate minus false alarm rate d for the three set sizes (2, 4, and 8) and are plotted with standard errors. In the top left corner, each plot also features the Bayes factor in favor of a monotone ordering of the points over a non-monotone ordering.


State−Trace Plots for Individual Participants
BF = 6.6
BF = 6.7
BF = 12.4
BF = 13.3
BF = 23.8
BF = 35
BF = 49.2
BF = 61.9
BF = 69.3
BF = 104.3
BF = 105.8
BF = 152.7
BF = 304.8
BF = 659.2
BF = 4799.6
Articulate
Silent
Accuracy in Simultaneous Condition
Accuracy in Sequential Condition
0.0
0.2
0.4
0.6
0.8
1.0

To quantify the support for monotonicity in the state-trace plots, we computed Bayes factors comparing the evidence for two hypotheses: first, that the true performance underlying the state-trace plots are ordered the same on both axes (that is, they can be described by a monotone curve), and second, that they are not ordered the same on both axes (Prince et al., 2012). We refer the reader to Prince et al. (2012) for technical details, and to the supplement to this article for details of how these Bayes factors were computed (Davis-Stober, Morey, Gretton, & Heathcote, 2016).

In addition to the state-trace plots for each participant, Figure 5.4 also contains the Bayes factor favoring a monotone ordering of the points over a non-monotone one. The Bayes factors uniformly favored the monotone ordering of the points. The Bayes factors ranged from about 7 to almost 5000. These data do not appear to provide any evidence for a deviation from monotonicity. Because our manipulations were designed to introduce effects of articulation consistent with the notion that verbal labeling can occur during visual memory tasks and can sometimes aid performance, this persistent monotonicity suggests that, at least for paradigms like this one, verbal labeling does not contribute to visual change detection performance.

DISCUSSION

The main question motivating the experiment we report was whether verbalization assists with other processes to influence visual memory performance. In that case, the application of articulatory suppression would be required to disengage a verbal memory dimension so that a pure measure of visual memory performance could be obtained. Neither a straightforward descriptive analysis nor a state-trace analysis revealed evidence that participants engaged in verbalization or that verbalization helped visual recognition memory, despite the fact that the experimental design favored the use of verbalization even more than the typical design of visual change detection tasks. The absence of a complex relationship between suppression, presentation type, and performance provides evidence that verbal recoding was not a strategy adopted by the participants in this task. Unlike previous studies which did not show effects of articulation on visual change detection performance, we were able to quantify evidence in favor of the null hypothesis for each individual participant using Bayesian state-trace analysis, providing novel positive evidence for the absence of this effect.

These results do not rule out any particular model of working memory. One interpretation of the multi-component working memory model (Baddeley, 1986), namely that both verbal and visual codes would be generated and maintained during visual change detection tasks, was unsupported by our analysis. The assumption that verbal codes could be generat-

ed during visual change detection is not a proposal of the model, but merely an assumption made by researchers that is consistent with the model. Verbal encoding of visual materials is not necessarily obligatory. However, our results do have important practical implications for researchers interested in measuring visual working memory capacity. Our analyses confirm that for briefly-presented, abstract visual materials whose to-be-remembered elements are not readily encompassed by a verbal label, verbal labeling either does not occur at all, or if it does occur, does not contribute to recognition accuracy. These results are not inconsistent with the multi-component working memory model, but suggest that it is not reasonable to invoke this influential model to support arguments that verbal encoding of visual materials necessarily contaminates estimates of visual working memory capacity.

Another possible interpretation of the multi-component model of working memory is that the central executive component, which directs attention within the system, may only be applied to a single sub-system at once. This supposition might lead to predictions that individuals strategically choose to encode visual materials in verbal code or alternatively in visual code. Though it would be difficult to eliminate such a flexible account of the encoding of visual materials entirely, we think that our data tend to rule out this idea as applied to visual change detection. If this strategic choice of coding occurred, then it might reasonably have occurred only in the sequential condition, or only for small set sizes, or might have been especially prevalent in the sequential conditions for small set sizes. Evidence against the interactions that would support these predictions are provided by descriptive analysis: for all set sizes, more participants showed a greater silent advantage in the simultaneous condition, contrary to predictions. Note that the multi-component working memory model generates no explicit prediction that participants must strategically switch between encoding materials in verbal or visual code; indeed, it has been shown that encoding verbal and visual-spatial stimuli can proceed with little if any dual-task cost (Cowan & Morey, 2007; C. C. Morey, Morey, van der Reijden, & Holweg, 2013), which rather suggests that adopting a switching strategy would be unnecessary if one assumes that separate verbal and visual short-term memory sub-systems are available.

One caveat for the interpretation of these results is that state-trace analysis, like all methods, is limited by the resolution of the data. Detecting deviations from monotonicity in a curve depends on how finely points on the curve are measured. It is possible that with finer gradations of set size, we might be able to detect non-monotonicities that are not apparent in these data. However, visual inspection of the state-trace plots in Figure 5.4 suggests that any effect of articulatory suppression is small; detecting such a small deviation from monotonicity would require finer gradations of sets size and more trials per set size. Our design already included thousands of trials per participant, and detected no positive effect of articulation condition

while providing robust positive evidence for monotonicity. Even if a small deviation from monotonicity existed, it would be unlikely to have any substantial effect on measurements of visual working memory capacity.

In some instances, verbalization clearly effects visual memory performance (Brandimonte et al., 1992), but features of the stimuli and the task likely limit the potential effects of verbalization. Stimulus presentation duration is likely a crucial factor determining whether verbalization strategies are employed in visual memory tasks. The abstractness of the stimuli employed likely also influences the extent to which verbalization occurs. In instances in which verbalization appeared to assist visual memory, abstract visual patterns were shown for 3 seconds, with retention intervals of 10 seconds, allowing plenty of time for both the generation and rehearsal of verbal labels (Brown, Forbes, & McConnell, 2006; Brown & Wesley, 2013). Moreover, in each of these studies demonstrating effects of verbalization, stimuli that were amenable to verbalization (determined by pilot testing) were chosen. In an investigation of effects of articulation on color-shape memory in which only articulation of visually imaginable phrases harmed visual recognition, participants were given 4 seconds, one second per visual object, to study the objects for a later memory test (Mate et al., 2012). In contrast, the stimulus presentation timings we employed (100-300 *ms* per item, depending on whether inter-stimulus intervals are considered) were substantially faster than those used in paradigms meant to encourage verbalization, and our stimuli were random patterns of colors. Recognition of the color and its spatial location was required to respond correctly. These design features are representative of visual change detection paradigms generally. The timings we chose are within the range of the visual change detection papers cited in our Introduction, which range from 8 *ms* per item (Woodman & Vogel, 2005) to as much as 500 *ms* per item (Brockmole et al., 2008). We conclude that for presentations as fast or faster than the 100 *ms* per item rate that we measured, it appears safe to assume that verbalization does not augment visual change detection performance.

Researchers employing nameable visual stimuli at paces enabling verbalization should still consider employing precautionary articulatory suppression if their goal is to isolate visual memory specifically. However, based on our data, we conclude that for many typical visual memory paradigms, such as those using brief presentations of randomly-generated abstract images, this precaution is unnecessary. Enforcing precautionary articulatory suppression does not seem to be necessary to get interpretable data from visual change detection tasks.

6

Making the Most of Human Memory

Acknowledgements

I would like to thank Michael LeKander for many inspiring discussions that have greatly contributed to the *Future directions* section.

Supplementary materials for this chapter are available at www.osf.io/q64bp

The work in the first part of this thesis has focused on a computerized fact-learning model[1]. The model incorporating the accuracy and latency of responses during retrieval practice is outlined in detail in Chapter 2, in which we show that it outperforms a traditional flashcard system on a range of performance measures. The work in Chapter 3 scrutinizes the stability of the model's parameters over time and materials and the data show that the model's estimate of a participant's rate of forgetting is stable within the same material. This indicates two important features of the underlying process captured by the parameter: it is stable over time and it can be estimated reliably by tracking a learner's performance during study. The combination of these two features suggests that the estimated rate of forgetting could be a useful indicator of individual differences between participants. In Chapter 4, we explored whether the parameter estimated by the model is merely an artifact of high-level, domain-general variation between participants captured by two widely used individual differences measures: working memory capacity and general cognitive ability. The analyses revealed no relationship between the estimated rate of forgetting and either of the two domain-general constructs in the tested sample, indicating that the individual differences captured by the model's parameter are likely not confounded by variation in cognitive functioning and attentional control.

In the remainder of this chapter, I will explore a number of aspects related to the estimated rate of forgetting in more detail. First, I will discuss the usefulness of the estimated rate of forgetting as an individual differences measure. Next, I will review how the model's parameter compares to other performance measures recorded during study in their ability to predict delayed recall. Finally, I will look beyond the current implementation of the model and consider conceivable future developments for the adaptive fact-learning system. This chapter will end with an overview of the discussed issues and how they relate to other relevant models.

The estimated rate of forgetting as an individual differences measure

Individual differences measures are useful to the extent that they capture differences between individuals that are predictive of (or related to) other relevant outcome measures. The experiment presented in Chapter 4 was designed to reveal a possible relationship between two established, widely used measures of individual differences in cognitive functioning – working memory capacity (WMC) and general cognitive ability (GCA) – and the estimated rate of forgetting extracted from the adaptive fact-learning system at the center of this thesis. No such relationship could be established in the tested sample, which suggests that estimated rate of forgetting captures differences between individuals that are not equivalent to differences in cognitive functioning.

[1] The work presented in Chapter 5 is self-contained and thematically different from the other chapters and will not be discussed further here.

Here, I will discuss in more detail how individual differences are captured by the model's parameters. Through a series of visualizations and descriptive analyses, I will show that there are substantial individual differences between participants in the studies reported in this thesis. It should be noted that the sample is relatively small and reported findings might not generalize entirely to the general population. However, being able to quantify substantial differences even within a rather homogeneous sample suggests that (presumably even larger) differences in the general population should be quantifiable in a similar fashion.

In the current implementation of the model, each item that is introduced to a learner is assigned an initial α value of 0.3. The accuracy and response times on subsequent repetitions of the item are then used to adjust the α value to best capture the model's estimation of how quickly an item is forgotten (see the Methods section of Chapter 2 for details). This gradual development of a participant's item-level parameters over the course of a 20-minute study session is presented graphically in Figure 6.1.

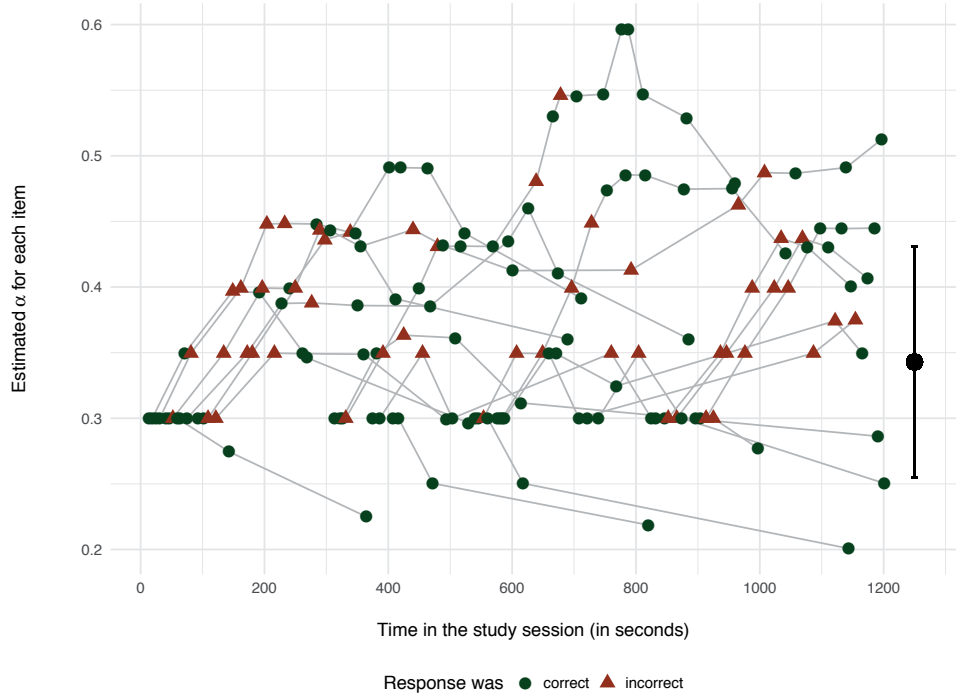


Figure 6.1. Development of each item's α parameter over time for one participant. Each data point indicates a trial; green circles signal correct responses and red triangles signal incorrect responses. Gray lines are used to visually group repetitions of one item together, they do not indicate a gradual change in the α values. The mean and standard deviation of the final α value across all items are summarized in the error bar on the right-hand side of the figure.

In Figure 6.1, time is plotted on the x-axis and the estimated α on the y-axis. Each response of the participant is shown as a symbol (157 in total) such that green circles indicate correct responses (106) and red triangles incorrect responses (51)[2]. The gray lines connecting symbols are used to group responses to the same item together. It is important to note that the lines only signal grouping and do not indicate a gradual change in α: Adjustment of the α value happens step-wise at the moment the response is collected. The participant in Figure 6.1 studied a total of 20 items (out of 35) and the development of each item as a function of the accuracy of responses can be traced in the plot. As discussed in earlier chapters, the response time is also used to adjust α on each trial but omitted from the graphical representation for the sake of clarity.

Figure 6.1 illustrates that all items have an α value of 0.3 when they are introduced but that subsequent responses shift the α values up or down. The items that get adjusted downwards are exclusively associated with correct responses and, as visible in the plot, lower α values result in wider spacing for subsequent repetitions of these items. The opposite pattern emerges for items that get adjusted upwards: a series of incorrect (or relatively slow) responses will yield higher α values and result in shorter spacing for repetitions.

The estimated rate of forgetting that is used as an outcome measure in Chapter 2 through Chapter 4 is the mean across the α values associated with the final repetition of each of the items in the study session. For the participant in Figure 6.1, this value is is 0.34, based on 20 α values (because they studied 20 of the 35 items). The rate of forgetting and the associated standard deviation are presented graphically on the right-hand side of Figure 6.1.

Visualizing the data from a single participant's study session is a good way to gain insight into how the model adjusts its internal parameters based on the incoming trial-by-trial data provided by their responses. It is also clear that differences between items affect the scheduling of repetitions and that these differences emerge very soon after an item has been introduced. Here, however, we are not interested in the differences between items within a single participant but rather in difference between participants independent of the studied material.

To make an analogous comparison between individual learners, we can compute the estimated rate of forgetting not only at the end of the 20-minute learning session but at any moment during the session (by considering only data collected up until that point in time). That way, the data shown in Figure 6.1 can be condensed into a single *trace* that summarize the development of a participant's estimated rate of forgetting throughout the learning session. An overview of all participants included in the experiments presented this thesis is shown in

[2] Please note that I have picked a participant that made a relatively large number of errors to better illustrate the influence of errors on a trial-by-trial basis.

Figure 6.2, in which each line represents a participant's rate of forgetting *trace*.

In Figure 6.2, each panel corresponds to an experiment. In all experiments, participants completed 20-minute learning sessions (x-axis) with different materials. Each line in the plot represents one participant's rate of forgetting *trace* (N = 183 across all experiments). The participants in Chapter 3 completed six sessions each (see Figure 3.1) and the data have been pooled across those sessions to produce a single trace for each participant.

The average trace across the participants in each panel is drawn as a black line and is strikingly similar across the experiments using different participants and materials. The estimated rate of forgetting for a participant that was used as an outcome measure in many of the analyses throughout this thesis is the last value in each trace at 1,200 seconds (i.e., 20 minutes). The distribution of these final values are shown in the violin plots (Hintze & Nelson, 1998) in the right margin of each panel. The black dot in each violin indicates the mean across the final values and – as suggested by the mean traces – is almost identical across the four panels: 0.318, 0.315, and 0.321 for the three panels, respectively (left to right).

Figure 6.2 suggests that individual differences emerge early in all three experiments. This suggests the study session could be shortened without losing accuracy in determining individual differences. To investigate whether and when a stable pattern in the individual differences emerges, we can estimate the rate of forgetting at earlier time points in the study session and correlate those earlier estimates with the final estimate. In the most extreme case, participants' earlier rates of forgetting are identical to their final rates of forgetting and the correlation would be 1 before the end of the study session. If they differed, the correlation would be less. The larger the difference between the rate of forgetting estimated at an earlier time and the final rate of forgetting, the smaller the correlation coefficient. The results from this analysis are presented in Figure 6.3.

Each participant's estimated rate of forgetting at that time
0.25
0.30
0.35
0.40
0
400
800
1200
Chapter 2
Chapter 3
Chapter 4
Time in the study session (in seconds)

Figure 6.2. Development of each participant's rate of forgetting over time. Each "trace" in the plot represents one participant and they are split up into panels based on experiments. In all experiments, the total study duration was 20 minutes (i.e., 1,200 seconds). The rate of forgetting at each point in time is based on the data collected up until that time. For participants in the experiments from Chapter 3, the data have been pooled across all sessions they completed. The black line in each panel represents the mean across the individual traces and the violin plots on the right-hand side of each panel show the distribution of the rates of forgetting at the end of the 20-minute study session (with the black dot inside each violin representing the mean).

As in the previous two figures, the time in the study session is shown on the x-axis. On the y-axis, the correlation coefficients between the rate of forgetting estimated after every minute and the final rate of forgetting are shown. Consequently, the correlation in all three experiments at time point 20 is 1: the final values are correlated perfectly with themselves. The colored lines indicating the three experiments are surrounded by areas demarcating the 95% confidence interval around the estimated correlation coefficient. From this information, we can gather that all correlation coefficient differ significantly from zero by the second minute into the study session. However, to determine a reasonable level of stability, the correlation should be very high, rather than merely non-zero. Figure 6.3 suggests that all coefficients are greater than 0.75 after 9 minutes of studying and close to 0.9 after 15 or 16 minutes.

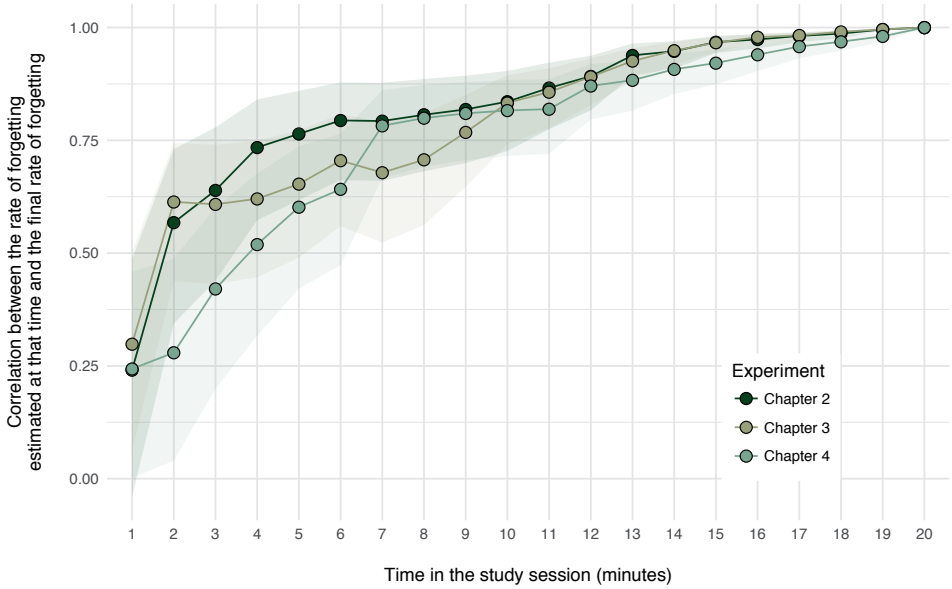


Figure 6.3. Correlations with final rates of forgetting as a function of time and experiment. For each participant in each experiment, the rate of forgetting at each minute during the study session were correlated with their rates of forgetting after 20 minutes. The shaded area around each line indicates the 95% confidence interval for the correlation coefficients.

Taken together, this exploration of participants' estimated rates of forgetting suggests that not much new information about a person's rate of forgetting is acquired during the last third of the study session. This indicates that the captured individual differences stabilize before the end of the session, which implies that the duration of the study session could be shortened without sacrificing information that distinguishes individual learners from each other. Presumably, the duration could be made even shorter in more heterogeneous samples or if less precision is required. It should be noted, however, that the traces of participants with the very low rates of forgetting in Figure 6.2 indicate that shortening the duration of the study session has a possible disadvantage: Not enough time might be left for differences between high-performing participants to emerge.

In summary, this investigation of the individual differences captured by the α parameter has demonstrated that information about participants can be used on different levels[3]. For a

[3] Theoretically, all descriptive analyses performed in this section could also be conducted for items. That is, instead of aggregating the α values across items to get a performance measure for a participant, information could be aggregated across *participants* to indicate which items are forgotten more quickly than others and how differences between items emerge and stabilize over time.

single participant, we can track the trial-by-trial fluctuations in the parameter as a function of their responses (Figure 6.1). By compressing these high-resolution data into a single trace for each participant, we can see individual differences emerge and stabilize over the time course of a study session (Figure 6.2 and Figure 6.3). On the highest level, we have the option to compare the performance of groups of participants with each other. For example, the violin plots in Figure 6.2 summarize the distributions of three groups. In this specific situation, comparing the groups might not be particularly interesting because they all come from the same student population. However, this might be an interesting approach to explore differences between groups from distinct populations.

This exploration has shown that the estimated rate of forgetting captures differences between individuals. To constitute a *useful* individual differences measure, however, the rate of forgetting must be related to relevant outcome measures. In the following section, I will explore how the estimated rate of forgetting relates to performance on delayed recall tests of the studied material.

Predicting delayed recall

In Chapter 4, we tested which of the three obtained measures – the estimated rate of forgetting, working memory capacity (WMC), and general cognitive ability (GCA) – could best predict delayed recall. The Bayes factors summarized in Figure 4.3 demonstrate that the estimated rate of forgetting is the single best predictor, while neither WMC nor GCA explain additional variance in delayed recall scores. One reason for the lack of a relationship could be that both WMC and GCA are constructs intended to capture domain-general cognitive processing abilities of participants. The estimated rate of forgetting, on the other hand, is the aggregate of memory-specific model parameters. Furthermore, the rate of forgetting is estimated while studying the material that is later tested, whereas both WMC and GCA are combined scores from independent tasks. If the goal is solely to determine the best predictor of delayed recall performance (which it was not in Chapter 4), measures directly based on the studied material might fare better.

During a computerized learning session, it is possible to keep track of various performance-related measures. Such measures are potentially useful predictors of delayed recall performance. Here, I will explore which performance measure extracted from the data collected during study can best predict delayed recall. Specifically, I will focus on three straightforward measures: the proportion of correct responses, the average response times on test trials, and the number of trials completed. These can be extracted regardless of the method that participants studied with. For participants studying with the adaptive method, however, we have one additional performance measure: the rate of forgetting. It is estimated by incor-

porating both accuracy and response times information on a trial-by-trial basis. Moreover, it also directly influences these measures because subsequent trials are scheduled based on its value. Thus, comparing the predictive power of the three universal measures extracted from an adaptively scheduled study session against the estimated rate of forgetting from that session would be subject to various confounds. Luckily, performance measures from flashcard learners can serve as a convenient benchmark against which the rate of forgetting can be evaluated. The amount of adaptivity while studying with flashcards is considerably less, thus reducing such confounds. The data presented in Chapter 2 and Chapter 4 allow us to do exactly that.

Participants in the experiment reported in Chapter 2 studied with both the adaptive method and flashcards and took a delayed recall test 15 minutes after completing each study session. Therefore, we can extract both the rate of forgetting (from the adaptive method session) and the additional learning measures (from the flashcard session) and correlate them with the scores on the delayed recall performance. The scores on the second test will not be included because the time between the two tests was not identical for all participants. There are data on all measures from 52 participants. In the experiment reported in Chapter 4, participants were randomly assigned to study with either the adaptive method or flashcards. Each participant was tested twice: after approximately 80 minutes and again 3 days later. Data from the adaptive method is available from 64 participants on the first and 56 on the second test, respectively. From the flashcards session, performance measures were extracted for all participants and can be correlated with scores from 59 participants on the first test and 55 on the second.

Consequently, a single performance measure is extracted from the learning session in which participants studied with the adaptive method: the estimated rate of forgetting. As in the other chapters in this thesis, the rate of forgetting is the average across the item-level final α values for each participant. From the flashcard data, on the other hand, three performance measures were extracted. First, the total number of trials – including both the initial study trials and all subsequent test trials – was computed for each participant. Second, study trials were removed and the proportion of correct responses across all test trials was computed for each participant. Third, trials with response times longer than ten seconds were removed (3.45% and 3.33% in Chapter 2 and Chapter 4, respectively) and the median response time across the remaining trials was computed.

The extracted performance measures from the two sources were correlated with the available test scores and the resulting Pearson's product-moment correlation coefficients are reported in Table 6.1. The three coefficients associated with each measure are of similar magnitude and retain their signs consistently. The rate of forgetting and the median response

times are negatively correlated with subsequent test performance, indicating that smaller values – that is, slower estimated forgetting and faster responses during learning – are related to better delayed recall performance. For the other two measures, the relationship with test scores is positive. The right-most column in Table 6.1 indicates the proportion of variance in test scores that can be explained by each measure and was computed by averaging the three coefficients in each row and squaring the resulting mean coefficient.

Table 6.1. An overview of correlation coefficients between performance measures extracted from the two study methods (adaptive method and flashcard method). Each of the measures is an aggregate of information recorded during the learning session: the estimated rate of forgetting is the mean across the final α values for all items a participant has studied. The proportion of correct responses is based only on test trials. All response times longer than ten seconds were removed before calculating the median response time. The total number of trials includes study and test trials The average amount of explained variance reported in the right-most column is the square of the mean of the three coefficients in each row. All correlation coefficients differ significantly from 0 at the 0.05 level.

		Chapter 2	Chapter 4		
	Measure	**Test 1**	**Test 1**	**Test 2**	**Avg. R^2**
Adaptive method	Estimated rate of forgetting	-0.84	-0.79	-0.89	0.71
Flashcard method	Proportion correct responses	0.77	0.80	0.70	0.59
	Median response time	-0.32	-0.36	-0.50	0.10
	Total number of trials	0.42	0.39	0.59	0.18

On average, the estimated rate of forgetting based on the data from the adaptive method has the highest absolute coefficients and can thus explain the largest amount of variance in test scores (71%). Of the three measures extracted from the data based on the flashcard method, none can explain as much variance in test scores as the estimated rate of forgetting (see last column in Table 6.1). The proportion of correct responses on test trials during study would be the single best predictor of delayed recall for data from the flashcard conditions in the two experiments.

Alternatively, one could combine the information from the three measures extracted from the flashcard condition to explain more of the variance in test scores after studying with the flashcard method. Given that the estimated rate of forgetting integrates information regarding the accuracy and response times of individual responses, such a comparison should at least put the combined information from the flashcard condition on par with the estimated rate of forgetting. This was tested by fitting multiple linear regression models to the test data using the three measures extracted from the flashcard data as predictors. The adjusted R^2 statistics are 0.70, 0.72, and 0.74 for test scores from Chapter 2, Chapter 4 Test 1, and Chapter 4 Test 2, respectively. This

suggests that together, the three performance measures from the flashcard condition can explain about the same amount of variance in test scores as the estimated rate of forgetting.

Taken together, the estimated rate of forgetting emerged as the single best predictor of delayed recall after comparing a number of performance measures from the two study methods. If the three performance measures based on the flashcard condition are combined, however, they can explain similar amounts of variance in test scores. Nevertheless, I would argue that using the estimated rate of forgetting has a number of distinct advantages over the combination of performance measures from the flashcard method. First, estimating the rate of forgetting requires learners to study with the adaptive method. The data and analyses presented in Chapter 2 demonstrate that many – but especially low-performing – learners benefit from using the adaptive method rather than traditional flashcards (also see van Rijn et al., 2009). Second, the ability to summarize the relevant aspects of performance during learning in a single measure (rather than three) is parsimonious and convenient. Third, and most importantly, the rate of forgetting is conceptually easier to grasp and interpret than a collection of regression model coefficients. Furthermore, the functional role the parameter plays in the larger underlying theoretical framework means that it could potentially be used to make "out of sample" (or: material) personalized predictions, an idea discussed in more detail in the next section.

Future directions

The work presented in this thesis has largely focused on an adaptive fact learning system. The system is introduced in detail in Chapter 2 and Chapter 3, in which we show that learners generally benefit from using the system (relative to a traditional flashcard system) and that a learner's parameters extracted from the system are stable over time. The long-term goal of developing this adaptive fact-learning system is twofold: (1) to learn more about the properties of human declarative memory and how to use its regularities to make personalized predictions about when studied material will be forgotten, and (2) to deploy the system at large to help people use their study time more effectively. Thus far, we have focused on (1). In this section, I want to discuss possible future steps that pertain to (2). The long-term goal in this context is using (1) to achieve (2).

The biggest limitation of the system as it is currently implemented is that each study session is treated in complete isolation. We used this to our advantage in Chapter 3, where we estimated the rate of forgetting for each participant in six distinct learning sessions to test their stability over time and materials. In the current implementation, the α parameter was set to 0.3 for everyone at the beginning of each 20-minute learning session and the model used the learner's input to fine-tune the parameter for each learner during the session. Given that the average final α values at the end of the session were found to be highly stable (within

the same type of material). This suggests that it does not make sense to treat each session in isolation: the parameters extracted during the first study session clearly capture information that is relevant in subsequent sessions. One possible way to integrate what has been learned about a learner in a previous session would be to set all starting values of α for that learner to the mean of the final α values from the previous session (instead of 0.3). Theoretically, this should reduce the number of trials the system needs to adapt to the appropriate α values for that particular learner, especially for relatively low- and high-performing participants.

An analogous approach could be taken even in a learner's first session. Since the α values are estimated and adjusted on an item-level, they could also be aggregated across participants. If the mean is computed across the *participants*, the aggregate would indicate how quickly each item in the set is estimated to be forgotten. Thus, assuming there are data from other participants that have studied the same set of items, a new learner's α values could be assigned starting values based on the average α values extracted from other learners' data. In this case, instead of setting all starting values to 0.3 or setting them all to a different value based on past performance, each item would have a unique starting value based on the performance of other learners.

If there are data from a previous session for a learner *and* other learners that have studied the same material, the starting values could be set taking both sources of information into account. If the number of previous learners is large enough, those similar to the current learner could be identified and the starting value for the current learner could be based on their item-level α values.

These ideas are conceptually straightforward. However, they do require data from study sessions to be stored in a way that the program could easily access each learner's previous records and aggregate across the stored data of other learners. This is almost entirely a software development problem that is beyond the scope of this thesis. However, it is important to be mindful of the crucial role software development will play in all future developments.

Adjusting the model's α values is the most obvious avenue towards making the model adapt to the learner more quickly, mainly because that is the parameter that is currently adjusted based on the trial-by-trial information the participant provides. However, looking at the model specifications detailed in Chapter 2 and Chapter 3 reveals that the equations depend on a range of other parameters, all of which are currently fixed. For example, the equation used to convert the estimated activation of an item to an expected response time consists of two parts: the scaled activation of the item and the *fixed time costs*:

$$L_i(t) = Fe^{-A_i(t)} + \textit{fixed time cost}$$

The scaling factor (F) is currently set to 1, effectively eliminating its effect on the estimated activation, and the *fixed time costs* are set to 300 ms as long as the cue is a single word. In the context of its ACT-R roots, the *fixed time costs* represent the time it takes to execute basic per-

ceptual and motor functions (Byrne & Anderson, 1998). In ACT-R models, the *fixed time costs* are kept constant unless those functions are part of the behavior that is modelled.

However, in the context of our adaptive fact-learning model, the *fixed time costs* could conceivably be repurposed to capture individual differences in reading and typing speed. Both of which are factors that will influence the response time but are not memory-related processes. In the bio-psychology material participants studied in the third session of the experiment reported in Chapter 4, for example, the cue consisted of definitions (e.g., "A tiny area of the retina specialized in acute, detailed vision") to which participants had to respond by typing in the corresponding concept (e.g., "fovea"). We would assume that there are individual differences in typing speed but that, for a given participant, they are not item dependent[4]. For the reading speed, on the other hand, one would expect both participant-level effects as well as item-level effects; some item's definitions are longer than others (or linguistically more complex) *and* some people read faster than others. Currently, this idea is reflected in the model by adjusting the *fixed time costs* based on the number of characters for cues that have more than two words[5].

Similar arguments and considerations could be made for other parameters that are currently fixed (e.g., the retrieval threshold). However, the more general questions are how to pick parameters to vary and how to evaluate whether varying them improves the model.

Parameters could be chosen based on their implication in the theoretical framework. As indicated above, it could be argued that the *fixed time costs* should not be fixed given the parameter's functional role in the model. Alternatively, *all* parameters could be varied (sequentially or in any conceivable combination simultaneously) regardless of their theoretical function. This would either require a massive amount of recorded learning sessions that can be data-mined or an absurd number of experiments to test empirically. Since the latter is not feasible, we will focus on the practicality of using existing data to retrospectively determine candidate parameters.

If one had data of thousands of learners using the current system, parameter optimization procedures could be deployed across the trial-by-trial response data to test whether varying a candidate parameter results in better performance. Given that this would be "historical" data, however, participants' responses are fixed. Thus, the only tenable way of quantifying "error", is by considering the mismatch between the predicted response time and the observed response time. (Because the method for making predictions can be changed after the fact but

---

[4] Even for the same participant, however, typing speed will most likely be platform-dependent, assuming they type faster on a laptop/desktop computer than on their mobile devices. These platform-dependent differences are also relevant if the system is implemented on a mobile platform or accessed on other devices.

[5] Specifically, the adjustment is max(300, (-157.9 + 19.5x)), where x is the number of characters in the cue. For the example "A tiny area of the retina specialized in acute, detailed vision", the reading time would be 1,071 ms because the cue has 63 characters. The max() function ensures that the *fixed time costs* cannot be less than 300 ms.

the recorded responses cannot.) Consequently, the optimization procedure should minimize an error term based on this mismatch[6]. The added benefit could then be evaluated based on the difference between the cumulative errors of the current model and the model that allows the candidate parameter(s) to vary, or the distribution of the errors more generally.

Such a data-driven approach could be used to identify candidate parameters that could potentially make the model adapt to individual learners more efficiently. Whether this would have a discernable effect in practice, however, would have to be determined empirically. Experiments in controlled lab conditions could provide data to evaluate the "new" model against the current model by looking at various performance measures. Next to the mismatch in predicted and observed response times, other – potentially more relevant – measures can be extract: the proportion of retrieval errors, the number of trials completed, the number of items studied, and, of course, performance on delayed recall tests. In other words, this comparison could be comparable to the one between the current model and a traditional flashcard method presented in Chapter 2 but would also allow the comparison of estimated model parameters that were not available in the flashcard condition. Importantly, testing any alteration of the current model is also necessary to verify that a theoretical, data-mining-based benefit actually emerges in practice (and is not simply the result of overfitting, for example) and yields an improvement whose magnitude is of practical relevance[7].

For this approach to be feasible, however, large datasets are needed. Ideally, the data should come from a heterogeneous group of people and be based on studying a wide range of materials. Recently, the implementation of the current model has been adopted by Noordhoff[8], a Dutch publisher of text books used in schools across the country. As a result, approximately 250,000 high school students have access to the adaptive system that allows them to study English, French, and German vocabulary. Additionally, the development of a mobile application has been sponsored by an internal e-learning grant at the University of Groningen. This application is called Rugged

[6] This mismatch could be expressed in different ways, of course. Most straightforward would be the absolute difference between the predicted and recorded response times. Alternatively, the difference could be raised to the power of *n*, where *n* could be set depending on how harshly large deviations should be penalized.

[7] In the context of this discussion, it is also worth thinking about reasons for avoiding excessive adaptation and personalization. This might seem counter-intuitive but there are arguments for keeping the model in its current form – not assuming anything about the learner (or items). Currently, the model is entirely agnostic with regards to the source of variation captured by the estimated parameters. If the parameter value associated with an item gets adjusted downwards (i.e., is estimated to be forgotten more quickly than previously assumed), this could imply any number of things: the item is difficult, the item is easy but it is studied by a low-performing learner, the learner was distracted and the recorded response time is not a "pure" measure of retrieval time, or many other. Keeping the model agnostic with regards to the source of these differences could serve a protective function.

[8] www.noordhoffuitgevers.nl/wps/portal

Learning[9] and will allow any course instructor to specify a set of materials that students enrolled in the course can study on their smartphones at their own convenience. Rugged Learning is currently in the last testing phase and will go "live" in the next academic year. Furthermore, a collaboration has recently been initiated with *HoeGekIsNL?*[10], which is an online research platform offering the general (Dutch-speaking) population the opportunity to complete a variety of questionnaires and tests; both to contribute to scientific research and as a self-measurement tool (primarily for mental health). Both Rugged Learning and the collaboration with *HoeGekIsNL?* will provide useful data as soon as they go live and more data will be collected continuously as long as it is live.

The goals of these collaborations differ: for Noordhoff and Rugged Learning students, the goal is primarily to make the most optimal use of their study time and to increase long-term retention relative to the study method they would have used otherwise. For the *HoeGekIsNL?* users, on the other hand, the adaptive system is used primarily to have fun while learning something interesting (psychological disorders of famous people, in this case) and receiving feedback regarding their estimated rate of forgetting and how it compares to the rest of the population.

Either way, these collaborations are an excellent opportunity to obtain large amounts of data from different populations. These data can be used as a testbed for the current model as well as an opportunity to apply the "historical data-mining approach" outlined above. Ideally, the data-mining will uncover options to improve the model further, which can then be tested empirically. If verified in the lab, improved versions of the model can be deployed at large to start a new iteration of improvement.

Access to data on this scale would also enable our group to further explore an issue that was briefly mentioned in the discussion of Chapter 2: is there any benefit for high-performing participants in using the adaptive method? Or even: might there be a disadvantage? The adaptive model will not introduce new items unless old items are estimated to have been learned well enough. Achieving an α value low enough for an item to not be repeated for a while will take a couple of fast and accurate responses to that item. This is because the binary search used to determine the best-fitting α value across the last five trials has limits of α ± 0.05 (see Methods of Chapter 2 for details). As a result, the α value cannot change by more than 0.05 in either direction. This limitation has been put in place to avoid overfitting extreme response times, which are common in noisy data and are expected to be even more prevalent if people use the system outside controlled lab settings. For very high-performing learners, however, the system might be too slow to adapt. And in the current implementation, each item starts with an α of 0.3, which, for them, will be too low every time. This could make the learning ex-

[9] Note: The Dutch name of the University of Groningen is *Rijksuniversiteit Groningen*, abbreviated as RUG. Hence "Rugged".

[10] www.hoegekis.nl/; the English name of the project is "how nuts are the Dutch?".

perience frustrating and repetitive[11]. Given that we already see sizable individual differences in the first-year psychology population tested in the experiments presented here (see Figure 6.2 specifically), it will be interesting to see the individual differences that emerge in a much more heterogeneous population and how well the current model can adapt to them.

Another future development that is partially dependent on access to enough data is predicting future test performance (see previous section). It is important to realize, however, that these are retrospective predictions based on simple linear regression models (since a correlation is a linear regression with a single predictor). In terms of the user experience and the usefulness of the system, it would be interesting to fit a more complex model – regression or other – on a large dataset that includes test scores after a range of retention intervals. Such a model's fit could be used to make true predictors of a learner's future performance in real-time. That is, the user interface could display a message such as "Given your performance, you are expected to still remember 60-70% of the studied items 3 days from now".

Furthermore, the fit of a truly predictive model could be used to provide the learner with an alternative stopping rule. In the current implementation of Rugged Learning, for example, the learner picks the set of materials from a list and then picks the time they want to spend studying. Alternatively, they could be given the option to study until the model estimates their performance to be at a certain level for a given retention interval. This would be an excellent way to counteract prevalent biases in our own judgments of learning – assuming the model is sufficiently accurate.

In summary, the work in our group has demonstrated that the current implementation of the adaptive fact-learning model works well. The results are promising but largely limited to single session learning in a relatively homogeneous sample. The priority of future work should be to extend the model so it can be used across multiple sessions and with larger sets of items. Simultaneously, the current implementation should be tested "in the wild" under realistic conditions by a more heterogeneous group of users. Collaborations with Noordhoff and *HoeGekIsNL?*, as well as the Rugged Learning implementation for smartphones, offer a chance to deploy the model at scale. The data collected through these channels has the potential to identify possible extensions and improvements of the model. Specifically, data-mining approaches could be exploited to pinpoint possible enhancements that can then be tested in controlled experiments. Ideally, this approach will single out weaknesses of the current model and point towards ways to refine it. Ultimately, the goal of this line of research should be to bridge the gap between our theoretical understanding of human memory and apply this understanding in practice. That way, we can make the most of human memory.

---

[11] Note that the same holds for learners at the other end of the spectrum: very low-performing learners could find the system frustrating to use because they have already forgotten an item on the first couple of repetitions because the model needs too long to adjust the parameters accordingly.

References

Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2005). Working memory and intelligence: the same or different constructs? *Psychological Bulletin*, *131*(1), 30–60. http://doi.org/10.1037/0033-2909.131.1.30

Adragna, R. (2016). Be Your Own Teacher: How to Study with Flashcards. Retrieved October 21, 2016, from http://www.learningscientists.org/blog/2016/2/20-1

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger, H. L., & McDermott, K. B. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology*, *22*, 861–876.

Allen, R. J., Baddeley, A. D., & Hitch, G. J. (2006). Is the binding of visual features in working memory resource-demanding? *Journal of Experimental Psychology: General*, *135*, 298–313.

Anderson, J. R. (2007). *How can the human mind exist in the physical universe? Oxford Series on Cognitive Models and Architectures*. New York, NY: Oxford University Press.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An Integrated Theory of the Mind. *Psychological Review*, *111*(4), 1036–1060. http://doi.org/10.1037/0033-295X.111.4.1036

Anderson, J. R., Bothell, D., Lebiere, C., & Matessa, M. (1998). An integrated theory of list memory. *Journal of Memory And Language*, *38*(4), 341–380. http://doi.org/10.1006/jmla.1997.2553

Anderson, J. R., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review*, *96*(4), 703–719. http://doi.org/10.1037/0033-295X.96.4.703

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the Environment in Memory. *Psychological Science*, *2*(6), 396–408. http://doi.org/10.1111/j.1467-9280.1991.tb00174.x

Anderson, M. C., Bjork, E. L., & Bjork, R. A. (2000). Retrieval-induced forgetting: Evidence for a recall-specific mechanism. *Psychonomic Bulletin & Review*, *7*(3), 522–530. http://doi.org/10.3758/BF03214366

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering Can Cause Forgetting: Retrieval Dynamics in Long-Term-Memory. *Journal of Experimental Psychology-Learning Memory and Cognition*, *20*(5), 1063–1087. http://doi.org/10.1037/0278-7393.20.5.1063

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, *27*(1), 17–21. http://doi.org/10.1080/00031305.1973.10478966

Atkinson, R. C. (1972). Optimizing the Learning of a Second-Language Vocabulary. *Journal of Experimental Psychology*, *96*(1), 124–129.

Baddeley, A. D. (1986). *Working memory*. Oxford, England: Clarendon Press.

Bamber, D. (1979). State trace analysis: method of testing simple theories of causation. *Journal of Mathematical Psychology*, *19*, 137–181.

Bartol, T. M., Bromer, C., Kinney, J., Chirillo, M. A., Bourne, J. N., Harris, K. M., & Sejnowski, T. J. (2015). Nanoconnectomic upper bound on the variability of synaptic plasticity. *eLife*, *4*, 1–18. http://doi.org/10.7554/eLife.10778

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. http://doi.org/10.18637/jss.v067.i01

Benjamin, A. S., & Bjork, R. A. (1996). Retrieval fluency as a metacognitive index. *Implicit Memory and Metacognition*, 309–338.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. (J. Metcalfe & A. Shimamura, Eds.)*Metacognition: Knowing about Knowing*. Cambridge, MA: MIT Press.

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *From Learning Processes to Cognitive Processes: Essays in Honor of William K. Estes* (pp. 35–67). Hillsdale, NJ: Erlbaum.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-Regulated Learning: Beliefs, Techniques, and Illusions. *Annual Review of Psychology*, *64*, 417–44. http://doi.org/10.1146/annurev-psych-113011-143823

Brainard, D. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436.

Brandimonte, M. A., Hitch, G. J., & Bishop, D. V. M. (1992). Verbal recoding of visual stimuli impairs mental image transformations. *Memory & Cognition*, *20*(4), 449–455.

Brewin, C. R., & Andrews, B. (2016). Creating Memories for False Autobiographical Events in Childhood: A Systematic Review. *Applied Cognitive Psychology*. http://doi.org/10.1002/acp.3220

Brockmole, J. R., Parra, M. A., Sala, S. Della, & Logie, R. H. (2008). Do binding deficits account for age-related decline in visual working memory? *Psychonomic Bulletin & Review*, *15*(3), 543–547.

Brown, L. A., Forbes, D., & McConnell, J. (2006). Limiting the use of verbal coding in the Visual Patterns Test. *The Quarterly Journal of Experimental Psychology Section A*, *59*(7), 1169–1176.

Brown, L. A., & Wesley, R. W. (2013). Visual working memory is enhanced by mixed strategy use and semantic coding. *Journal of Cognitive Psychology*, *25*(3), 328–338.

Busey, T. A., Tunnicliff, J., Loftus, G. R., & Loftus, E. F. (2000). Accounts of the confidence-accuracy relation in recognition memory. *Psychonomic Bulletin & Review*, *7*(1), 26–48. http://doi.org/10.3758/BF03210724

Butler, A. C., & Roediger, H. L. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology*, *19*(4–5), 514–527. http://doi.org/10.1080/09541440701326097

Byrne, M. D., & Anderson, J. R. (1998). Perception and action. In J. R. Anderson & C. Lebiere (Eds.), *The atomic components of thought* (pp. 167–200). Mahwah, NJ: Lawrence Erlbaum.

Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory & Cognition*, *36*(2), 438–448. http://doi.org/10.3758/MC.36.2.438

Carrier, M. (2003). College students' choices of study strategies. *Perceptual and Motor Skills*, *96*, 54–56. http://doi.org/10.2466/pms.2003.96.1.54

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition*, *20*(6), 633–42. http://doi.org/10.3758/BF03202713

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, *132*(3), 354–80. http://doi.org/10.1037/0033-2909.132.3.354

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: a temporal ridgeline of optimal retention. *Psychological Science*, *19*(11), 1095–102. http://doi.org/10.1111/j.1467-9280.2008.02209.x

Cho, H. C., & Abe, S. (2013). Is two-tailed testing for directional research hypotheses tests legitimate? *Journal of Business Research*, *66*(9), 1261–1266. http://doi.org/10.1016/j.jbusres.2012.02.023

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user' s guide. *Psychonomic Bulletin & Review*, *12*(5), 769–786. http://doi.org/10.3758/BF03196772

Conway, A. R. A., Kane, M. J., & Engle, R. W. (2003). Working memory capacity and its relation to general intelligence. *Trends in Cognitive Sciences*, *7*(12), 547–552. http://doi.org/10.1016/j.tics.2003.10.005

Cowan, N., Elliott, E. M., Saults, J. S., Morey, C. C., Mattox, S., Hismjatullina, A., & Conway, A. R. A. (2005). On the capacity of attention: Its estimation and its role in working memory and cognitive aptitudes. *Cognitive Psychology*, *51*, 42–100.

Cowan, N., & Morey, C. C. (2007). How can dual-task working memory retention limits be investigated? *Psychological Science*, *18*, 686–688.

Davis-Stober, C., Morey, R. D., Gretton, M., & Heathcote, A. (2016). Bayes factors for state-trace analysis. *Journal of Mathematical Psychology*, *72*, 116–129.

de Boer, H., Donker, A. S., & van der Werf, M. P. C. (2014). Effects of the Attributes of Educational Interventions on Students' Academic Performance: A Meta-Analysis. *Review of Educational Research*, *84*(4), 509–545. http://doi.org/10.3102/0034654314540006

de Jonge, M., Tabbers, H. K., Pecher, D., & Zeelenberg, R. (2012). The effect of study time distribution on learning and retention: A Goldilocks principle for presentation rate. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(2), 405–412. http://doi.org/10.1037/a0025897

Deary, I. J. (2012). Intelligence. *Annual Review of Psychology*, *63*(1), 453–482. http://doi.org/10.1146/annurev-psych-120710-100353

Delaney, P. F., Verkoeijen, P. P. J. L., & Spirgel, A. (2010). Spacing and Testing Effects: A Deeply Critical, Lengthy, and at Times Discursive Review of the Literature. *Psychology of Learning and Motivation*, *53*, 63–147. http://doi.org/10.1016/S0079-7421(10)53003-2

Delvenne, J. F., & Bruyer, R. (2004). Does visual short-term memory store bound features? *Visual Cognition*, *11*(1), 1–27.

Dempster, F. N. (1988). The spacing effect: A case study in the failure to apply the results of psychological research. *American Psychologist*, *43*(8), 627–634.

Donker, A. S., de Boer, H., Kostons, D., Dignath van Ewijk, C. C., & van der Werf, M. P. C. (2014). Effectiveness of learning strategy instruction on academic performance: A meta-analysis. *Educational Research Review*, *11*, 1–26. http://doi.org/10.1016/j.edurev.2013.11.002

Donovan, J. J., & Radosevich, D. J. (1999). A meta-analytic review of the distribution of practice effect: Now you see it, now you don't. *Journal of Applied Psychology*, *84*(5), 795–805. http://doi.org/10.1037/0021-9010.84.5.795

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving Students' Learning With Effective Learning Techniques: Promising Directions From Cognitive and Educational Psychology. *Psychological Science in the Public Interest*, *14*(1), 4–58. http://doi.org/10.1177/1529100612453266

Foster, J. L., Shipstead, Z., Harrison, T. L., Hicks, K. L., Redick, T. S., & Engle, R. W. (2015). Shortened complex span tasks can reliably measure working memory capacity. *Memory & Cognition*, *43*(2), 226–36. http://doi.org/10.3758/s13421-014-0461-7

Gabrieli, J. D. E. (1998). Cognitive Neuroscience of Human Memory. *Annual Review of Psychology*, *49*, 87–115. http://doi.org/10.1146/annurev.psych.49.1.87

Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review*, *116*(2), 439–453. http://doi.org/10.1037/a0015251

Gardiner, F. M., Craik, F. I. M., & Bleasdale, F. A. (1973). Retrieval difficulty and subsequent recall. *Memory & Cognition*, *1*(3), 213–216. http://doi.org/10.3758/BF03198098

Gates, A. I. (1917). Recitation as a factor in memorizing. *Archives of Psychology*, *6*.

Gathercole, S. E., & Baddeley, A. D. (1989). Evaluation of the role of phonological STM in the development of vocabulary in children: A longitudinal study. *Journal of Memory and Language*, *28*(2), 200–213. http://doi.org/10.1016/0749-596X(89)90044-2

Gathercole, S. E., & Baddeley, A. D. (1990). The role of phonological memory in vocabulary acquisition: A study of young children learning new names. *British Journal of Psychology*, *81*(4), 439–454. http://doi.org/10.1111/j.2044-8295.1990.tb02371.x

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York, NY: Cambridge University Press.

Godbole, N. R., Delaney, P. F., & Verkoeijen, P. P. J. L. (2014). The spacing effect in immediate and delayed free recall. *Memory*, *22*(5), 462–469. http://doi.org/10.1080/09658211.2013.798416

Goossens, N. A. M. C., Camp, G., Verkoeijen, P. P. J. L., Tabbers, H. K., & Zwaan, R. A. (2012). The benefit of retrieval practice over elaborative restudy in primary school vocabulary learning. *Journal of Applied Research in Memory and Cognition*, *3*(3), 177–182. http://doi.org/10.1016/j.jarmac.2014.05.003

Hartwig, M. K., & Dunlosky, J. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement? *Psychonomic Bulletin & Review*, *19*(1), 126–134. http://doi.org/10.3758/s13423-011-0181-y

Hintze, J. L., & Nelson, R. D. (1998). Violin plots: A box plot-density trace synergism. *American Statistician*, *52*(2), 181–184. http://doi.org/10.1080/00031305.1998.10480559

Hollingworth, A., & Rasmussen, I. P. (2010). Binding objects to locations: The relationship between object files and visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(3), 543–564.

Jacoby, L. L., Kelley, C. M., & Dywan, J. (1989). Memory attributions. *Varieties of Memory and Consciousness: Essays in Honour of Endel Tulving*, 391–422.

Jang, Y., Wixted, J. T., Pecher, D., Zeelenberg, R., & Huber, D. E. (2012). Decomposing the interaction between retention interval and study/test practice: The role of retrievability. *The Quarterly Journal of Experimental Psychology*, *65*(5), 962–975. http://doi.org/10.1080/17470218.2011.638079

Jastrzembski, T. S., Gluck, K., & Gunzelmann, G. (2006). Knowledge tracing and prediction of future trainee performance. In *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference* (pp. 1498–1508). Orlando, FL: National Training Systems Association.

Kalat, J. W. (2012). *Biological psychology* (11th ed.). Wadsworth, Cengage Learning.

Kane, M. J., Hambrick, D. Z., & Conway, A. R. A. (2005). Working memory capacity and fluid intelligence are strongly related constructs: Comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin*, *131*(1), 66–71. http://doi.org/10.1037/0033-2909.131.1.66

Kang, S. H. K., & Pashler, H. (2014). Is the benefit of retrieval practice modulated by motivation? *Journal of Applied Research in Memory and Cognition*, *3*, 183–188. http://doi.org/10.1016/j.jarmac.2014.05.006

Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General*, *138*(4), 469–486.

http://doi.org/10.1037/a0017341

Karpicke, J. D., Butler, A. C., & Roediger, H. L. (2009). Metacognitive strategies in student learning: Do students practise retrieval when they study on their own? *Memory*, *17*(4), 471–479. http://doi.org/10.1080/09658210802647009

Karpicke, J. D., & Roediger, H. L. (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(4), 704–719. http://doi.org/10.1037/0278-7393.33.4.704

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science*, *319*(5865), 966–8. http://doi.org/10.1126/science.1152408

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Kelley, C. M., & Lindsay, D. S. (1993). Remembering Mistaken for Knowing: Ease of Retrieval as a Basis for Confidence in Answers to General Knowledge Questions. *Journal of Memory and Language*. http://doi.org/10.1006/jmla.1993.1001

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox-3? *Perception*, *36*.

Kornell, N. (2009). Optimizing learning using flashcards: spacing is more effective than cramming. *Applied Cognitive Psychology*, *23*, 1297–1317. http://doi.org/10.1002/acp.1537

Kornell, N., & Bjork, R. A. (2008a). Learning concepts and categories: Is spacing the "enemy of induction"? *Psychological Science*, *19*(6), 585–592. http://doi.org/10.1111/j.1467-9280.2008.02127.x

Kornell, N., & Bjork, R. A. (2008b). Optimising self-regulated study: the benefits - and costs - of dropping flashcards. *Memory*, *16*(2), 125–136. http://doi.org/10.1080/09658210701763899

Kornell, N., & Son, L. K. (2009). Learners' choices and beliefs about self-testing. *Memory*, *17*(5), 493–501. http://doi.org/10.1080/09658210902832915

Kraemer, P. J., & Golding, J. M. (1997). Adaptive forgetting in animals. *Psychonomic Bulletin & Review*, *4*(4), 480–491. http://doi.org/10.3758/BF03214337

LaBar, K. S., & Cabeza, R. (2006). Cognitive neuroscience of emotional memory. *Nature Reviews Neuroscience*, *7*(1), 54–64.

Larsen, J. D., & Baddeley, A. D. (2003). Disruption of verbal STM by irrelevant speech, articulatory suppression, and manual tapping: Do they have a common source? *The Quarterly Journal of Experimental Psychology Section A*, *56*(8), 1249–1268.

Levy, B. A. (1971). Role of Articulation in Auditory and Visual Short-Term Memory. *Journal of Verbal Learning and Verbal Behavior*, *10*, 123–132.

Lewandowsky, S., & Oberauer, K. (2015). Rehearsal in serial recall: An unworkable solution to the non-existent problem of decay. *Psychological Review*, *122*, 674–699.

Lewis, O., Lindsey, R., Pashler, H., & Mozer, M. C. (2010). Predicting Students' Retention of Facts from Feedback During Training. In S. Ohlsson (Ed.), *Proceedings of the 32nd annual conference of the cognitive science society* (pp. 2332–2337). Portland, OR: Cognitive Science Society.

Lindsey, R., Mozer, M., Cepeda, N. J., & Pashler, H. (2009). Optimizing memory retention with cognitive models. In A. Howes, D. Peebles, & R. Cooper (Eds.), *9th International Conference on Cognitive Modeling (ICCM)* (pp. 74–79). Manchester, UK: ICCM.

Lindsey, R., Shroyer, J. D., Pashler, H., & Mozer, M. C. (2014). Improving Students' Long-Term Knowledge Retention Through Personalized Review. *Psychological Science*, *25*(3), 639–647. http://doi.org/10.1177/0956797613504302

Logie, R. H., Brockmole, J. R., & Vandenbroucke, A. R. E. (2009). Bound feature combinations in visual short-term memory are fragile but influence long-term learning. *Visual Cognition*, *17*(1), 160–179.

Luria, R., Sessa, P., Gotler, A., Jolicœur, P., & Dell'Acqua, R. (2010). Visual Short-term Memory Capacity for Simple and Complex Objects. *Journal of Cognitive Neuroscience*, *22*(3), 496–512.

Madigan, S., Neuse, J., & Roeber, U. (2000). Retrieval latency and "at-risk" memories. *Memory & Cognition*, *28*(4), 523–8. http://doi.org/10.3758/BF03201242

Makovski, T., & Jiang, Y. V. (2008). Indirect assessment of visual working memory for simple and complex objects. *Memory & Cognition*, *36*(6), 1132–1143.

Makovski, T., Sussman, R., & Jiang, Y. V. (2008). Orienting attention in visual working memory reduces interference from memory probes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(2), 369–380.

Mate, J., Allen, R. J., & Baqués, J. (2012). What you say matters: Exploring visual-verbal interactions in visual working memory. *The Quarterly Journal of Experimental Psychology*, *65*(3), 395–400.

MATLAB. (2011). *Version 7.13.0 (R2011b)*. Natick, Massachusetts: The MathWorks Inc.

Matsukura, M., & Hollingworth, A. (2011). Does visual short-term memory have a high-capacity stage? *Psychonomic Bulletin & Review*, *18*(6), 1098–1104.

Matvey, G., Dunlosky, J., & Guttentag, R. (2001). Fluency of retrieval at study affects judgments of learning (JOLs): an analytic or nonanalytic basis for JOLs? *Memory & Cognition*, *29*(2), 222–233. http://doi.org/10.3758/BF03194916

McCabe, J. (2011). Metacognitive awareness of learning strategies in undergraduates. *Memory & Cognition*, *39*(3), 462–76. http://doi.org/10.3758/s13421-010-0035-2

Mcdaniel, M. A., Anderson, J. L., & Mary, H. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, *19*(4/5), 494–513. http://doi.org/10.1080/09541440701326154

Mettler, E., & Kellman, P. J. (2014). Adaptive response-time-based category sequencing in perceptual learning. *Vision Research*, *99*, 111–123. http://doi.org/10.1016/j.visres.2013.12.009

Mettler, E., Massey, C. M., & Kellman, P. J. (2011). Improving Adaptive Learning Technology through the Use of Response Times. In L. Carlson, C. Hoelscher, & T. F. Shipley (Eds.), *Proceedings of the 33nd Annual Conference of the Cognitive Science Society* (pp. 2532–2537). Boston, MA: Cognitive Science Society.

Morey, C. C., & Cowan, N. (2004). When visual and verbal memories compete: Evidence of cross-domain limits in working memory. *Psychonomic Bulletin & Review*, *11*, 296–301.

Morey, C. C., & Cowan, N. (2005). When Do Visual and Verbal Memories Conflict? {T}he Importance of Working-Memory Load and Retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 703–713.

Morey, C. C., Morey, R. D., van der Reijden, M., & Holweg, M. (2013). Asymmetric cross-domain interference between two working memory tasks: Implications for models of working memory. *Journal of Memory and Language*, *69*, 324–348.

Morey, R. D. (2011). A Hierarchical Bayesian model for the measurement of working memory capacity. *Journal of Mathematical Psychology*, *55*, 8–24.

Morey, R. D., & Rouder, J. N. (2015a). BayesFactor: 0.9.11-1 CRAN. http://doi.org/10.5281/zenodo.16238

Morey, R. D., & Rouder, J. N. (2015b). BayesFactor 0.9.12-2 CRAN.

Mozer, M. C., & Lindsey, R. (2016). Predicting and Improving Memory Retention: Psychological Theory Matters in the Big Data Era.

Mulder, J., & Wagenmakers, E.-J. (2016). Editors' introduction to the special issue "Bayes factors for testing hypotheses in psychological research: Practical relevance and new developments." *Journal of Mathematical Psychology*, *72*, 1–5. http://doi.org/10.1016/j.jmp.2016.01.002

Murdock, B. B. (1985). An analysis of the strength-latency relationship. *Memory & Cognition*, *13*(6), 511–521. http://doi.org/10.3758/BF03198322

Murray, D. J. (1965). Vocalization-at-presentation and immediate recall, with varying presentation-rates. *Quarterly Journal of Experimental Psychology*, *17*, 47–56.

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

Nijboer, M. (2011). *Optimal Fact Learning: Applying Presentation Scheduling to Realistic Conditions*. University of Groningen, Groningen, The Netherlands.

Norman, D. A., & Wickelgren, W. A. (1969). Strength theory of decision rules and latency in retrieval from short-term memory. *Journal of Mathematical Psychology*, *6*, 192–208.

Otgaar, H., Merckelbach, H., Jelicic, M., & Smeets, T. (2016). The Potential for False Memories is Bigger than What Brewin and Andrews Suggest. *Applied Cognitive Psychology*. http://doi.org/10.1002/acp.3262

Pachur, T., Schooler, L. J., & Stevens, J. R. (2014). We'll meet again: Revealing distributional and temporal patterns of social contact. *PLoS ONE*, *9*(1), 1–10. http://doi.org/10.1371/journal.pone.0086081

Papoušek, J., Pelánek, R., & Stanislav, V. (2014). Adaptive Practice of Facts in Domains with Varied Prior Knowledge. *Proceedings of the 7th International Conference on Educational Data Mining (EDM)*, 6–13.

Pavlik, P. I., & Anderson, J. R. (2003). An ACT-R model of the spacing effect. In F. Detje, D. Doerner, & H. Schaub (Eds.), *Proceedings of the Fifth International Conference of Cognitive Modeling* (pp. 177–182). Bamberg, Germany: Universitäts-Verlag Bamberg.

Pavlik, P. I., & Anderson, J. R. (2005). Practice and forgetting effects on vocabulary memory: an activation-based model of the spacing effect. *Cognitive Science*, *29*(4), 559–86. http://doi.org/10.1207/s15516709cog0000_14

Pavlik, P. I., & Anderson, J. R. (2008). Using a Model to Compute the Optimal Schedule of Practice. *Journal of Experimental Psychology. Applied*, *14*(2), 101–117. http://doi.org/10.1037/1076-898X.14.2.101

Pavlik, P. I., Bolster, T., Wu, S., Koedinger, K., & MacWhinney, B. (2008). Using optimally selected drill practice to train basic facts. In B. P. Woolf, E. Aïmeur, R. Nkambou, & S. Lajoie (Eds.), *Proceedings of the 9th International Conference on Intelligent Tutoring Systems* (pp. 593–602). Heidelberg, Germany: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-540-69132-7_62

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*, 437–442.

Prince, M., Brown, S. D., & Heathcote, A. (2012). The design and analysis of state-trace experiments. *Psychological Methods*, *17*(1), 78–99.

Putnam, A. L., Sungkhasettee, V. W., & Roediger, H. L. (2016). Optimizing Learning in College: Tips From Cognitive Psychology. *Perspectives on Psychological Science*, *11*(5), 652–660. http://doi.org/10.1177/1745691616645770

Pyc, M. A., & Dunlosky, J. (2010). Toward an understanding of students' allocation of study time: why do they decide to mass or space their practice? *Memory & Cognition*, *38*(4), 431–440. http://doi.org/10.3758/MC.38.4.431

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, *60*(4), 437–447. http://doi.org/10.1016/j.jml.2009.01.004

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: mediator effectiveness hypothesis. *Science (New York, N.Y.)*, *330*(October), 335. http://doi.org/10.1126/science.1191465

Raaijmakers, J. G. W., & Shiffrin, R. M. (1980). SAM: A theory of probabilistic search of associative memory. *Psychology of Learning and Motivation*, *14*, 207–262.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1992). Models for recall and recognition. *Annual Review of Psychology*, *43*, 205–234.

R Development Core Team. (2016). R: A Language and Environment for Statistical Computing. manual, Vienna, Austria.

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20–27. http://doi.org/10.1016/j.tics.2010.09.003

Rohrer, D. (2015). Student Instruction Should Be Distributed Over Long Time Periods. *Educational Psychology Review*, *27*(4), 635–643. http://doi.org/10.1007/s10648-015-9332-4

Rouder, J. N., & Morey, R. D. (2012). Default Bayes Factors for Model Selection in Regression. *Multivariate Behavioral Research*, *47*(6), 877–903. http://doi.org/10.1080/00273171.2012.734737

Rouder, J. N., Morey, R. D., Cowan, N., Zwilling, C. E., Morey, C. C., & Pratte, M. S. (2008). An Assessment of Fixed-Capacity Models of Visual Working Memory. *Proceedings of the National Academy of Sciences*, *105*, 5976–5979.

Rouder, J. N., Morey, R. D., Morey, C. C., & Cowan, N. (2011). How to Measure Working-Memory Capacity in the Change-Detection Paradigm. *Psychonomic Bulletin & Review*, *18*, 324–330.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. http://doi.org/10.3758/PBR.16.2.225

Rubin, D. C., & Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review*, *103*(4), 734–760. http://doi.org/10.1037/0033-295X.103.4.734

Schooler, J. W., & Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology*, *22*, 36–71.

Sense, F., Behrens, F., Meijer, R. R., & van Rijn, H. (2016). An Individual's Rate of Forgetting Is Stable Over Time but Differs Across Materials. *Topics in Cognitive Science*, *8*(1), 305–321. http://doi.org/10.1111/tops.12183

Sense, F., Meijer, R. R., & Van Rijn, H. (2016). On the Link between Fact Learning and General Cognitive Ability. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Meeting of the Cognitive Science Society*. Philadelphia, PA: Cognitive Science Society.

Settles, B., & Meeder, B. (2016). A Trainable Spaced Repetition Model for Language Learning. *Association for Computational Linguistic (ACL)*, 1848–1858.

Sperling, G. (1967). Successive approximations to a model for short-term memory. *Acta Psychologica*, *27*, 285–292.

Spitzer, H. F. (1939). Studies in retention. *Journal of Educational Psychology*, *30*(9), 641–656. http://doi.org/10.1037/h0063404

Stevens, J. R., Marewski, J. N., Schooler, L. J., & Gilby, I. C. (2016). Reflections of the social environment in chimpanzee memory: Applying rational analysis beyond humans. *Royal Society Open Science*, *3*. http://doi.org/10.1098/rsos.160293

Taraban, R., Maki, W. S., & Rynearson, K. (1999). Measuring study time distributions: implications for designing computer-based courses. *Behavior Research Methods, Instruments, & Computers*, *31*(2), 263–269. http://doi.org/10.3758/BF03207718

Treisman, A., & Zhang, W. (2006). Location and binding in visual working memory. *Memory & Cognition*, *34*(8), 1704–1719. article.

Tulving, E. (1995). Organization of Memory: Quo Vadis? *The Cognitive Neurosciences*. http://doi.org/10.1017/S0140525X00047257

Unsworth, N. (2016). Working memory capacity and recall from long-term memory: Examining the influences of encoding strategies, study time allocation, search efficiency, and monitoring abilities. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *42*(1), 50–61. http://doi.org/10.1037/xlm0000148

Unsworth, N., Brewer, G. A., & Spillers, G. J. (2009). There's more to the working memory capacity-fluid intelligence relationship than just secondary memory. *Psychonomic Bulletin & Review*, *16*(5), 931–937. http://doi.org/10.3758/PBR.16.5.931

Unsworth, N., Brewer, G. A., & Spillers, G. J. (2013). Focusing the search: proactive and retroactive interference and the dynamics of free recall. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *39*(6), 1742–56. http://doi.org/10.1037/a0033743

Unsworth, N., & Engle, R. W. (2007). The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. *Psychological Review*, *114*(1), 104–132. http://doi.org/10.1037/0033-295X.114.1.104

Van Bebber, J., Lem, J., & Van Zoelen, L. (2010). Q1000 Capaciteiten Hoog.

van den Broek, G. S. E., Segers, E., Takashima, A., & Verhoeven, L. (2014). Do testing effects change over time? Insights from immediate and delayed retrieval speed. *Memory*, *22*(7), 803–12. http://doi.org/10.1080/09658211.2013.831455

van den Broek, G. S. E., Takashima, A., Segers, E., Fernández, G., & Verhoeven, L. (2013). Neural correlates of testing effects in vocabulary learning. *NeuroImage*, *78*(April 2016), 94–102. http://doi.org/10.1016/j.neuroimage.2013.03.071

van den Broek, G. S. E., Takashima, A., Wiklund-Hörnqvist, C., Karlsson Wirebring, L., Segers, E., Verhoeven, L., … Nyberg, L. (2016). Neurocognitive mechanisms of the "testing effect": A review. *Trends in Neuroscience and Education*, *5*, 52–66. http://doi.org/10.1016/j.tine.2016.05.001

van Houten, R., & Rolider, A. (1989). An analysis of several variables influencing the efficacy of flash card instruction. *Journal of Applied Behavior Analysis*, *22*(1), 111–118. http://doi.org/10.1901/jaba.1989.22-111

van Lamsweerde, A. E., & Beck, M. R. (2012). Attention shifts or volatile representations: What causes binding deficits in visual working memory? *Visual Cognition*, *20*(7), 771–792.

van Rijn, H., Dalenberg, J. R., Borst, J. P., & Sprenger, S. A. (2012). Pupil Dilation Co-Varies with Memory Strength of Individual Traces in a Delayed Response Paired-Associate Task. *PLoS ONE*, *7*(12). http://doi.org/10.1371/journal.pone.0051134

van Rijn, H., van Maanen, L., & van Woudenberg, M. (2009). Passing the Test: Improving Learning Gains by Balancing Spacing and Testing Effects. *Proceedings of the 9th International Conference on Cognitive Modeling*, 110–115.

Verkoeijen, P. P. J. L., & Bouwmeester, S. (2014). Is spacing really the "friend of induction"? *Frontiers in Psychology*, *5*, 1–8. http://doi.org/10.3389/fpsyg.2014.00259

Vogel, E. K., Woodman, G. F., & Luck, S. J. (2001). Storage of features, conjunctions, and objects in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *27*, 92–114.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomical Bulletin & Review*, *14*(5), 779–804.

Wagenmakers, E.-J., Morey, R. D., & Lee, M. D. (2016). Bayesian Benefits for the Pragmatic Researcher. *Current Directions in Psychological Science*, *25*(3), 169–176. http://doi.org/10.1177/0963721416643289

Wagenmakers, E.-J., Verhagen, J., & Ly, A. (2016). How to quantify the evidence for the absence of a correlation. *Behavior Research Methods*, *48*, 413–426. http://doi.org/10.3758/s13428-015-0593-0

Wissman, K. T., Rawson, K. A., & Pyc, M. A. (2012). How and when do students use flashcards? *Memory*, *20*(6), 568–579. http://doi.org/10.1080/09658211.2012.687052

Wixted, J. T. (2004). The Psychology and Neuroscience of Forgetting. *Annual Review of Psychology*, *55*, 235–269. http://doi.org/10.1146/annurev.psych.55.090902.141555

Wixted, J. T., & Carpenter, S. K. (2007). The Wickelgren Power Law and the Ebbinghaus Savings Function, *18*(2), 133–134.

Woodman, G. F., & Vogel, E. K. (2005). Fractionating Working Memory: Consolidation and Maintenance Are Independent Processes. *Psychological Science*, *16*(2), 106–113.

Woodman, G. F., & Vogel, E. K. (2008). Selective storage and maintenance of an object's features in visual working memory. *Psychonomic Bulletin & Review*, *15*(1), 223–229.

Woodman, G. F., Vogel, E. K., & Luck, S. J. (2012). Flexibility in visual working memory: Accurate change detection in the face of irrelevant variations in position. *Visual Cognition*, *20*(1), 1–28.

Woźniak, P. A., & Gorzelańczyk, E. J. (1994). Optimization of repetition spacing in the practice of learning. *Acta Neurobiologiae Experimentalis*, *54*(1), 59–62.

Zeelenberg, R., de Jonge, M., Tabbers, H. K., & Pecher, D. (2015). The effect of presentation rate on foreign-language vocabulary learning. *Quarterly Journal of Experimental Psychology*, *68*(6), 1101–15. http://doi.org/10.1080/17470218.2014.975730

Summary
Samenvatting
Zusammenfassung

Acknowledgements

I would like to thank Aafke van Mourik Broekman and Niklas Sense for their help with the translations.

Thesis Summary "Making the most of human memory"

Memory is absolutely essential to the survival of almost all living organisms. Without memory, we would be helpless and confused because we could not access what we had previously learned and would not be able to make sense of the world around us. In the broadest sense, memory encompasses everything we have stored about our experiences throughout our lives. Many of these stored memories can be accessed and reported verbally and others are apparent in the effortless performance of skills we have learned: You can tell me the name of your best childhood friend while tying your shoelaces, and both would be highly dependent on memory.

Memory research focusses on how we encode, store, and retrieve memories in/from the brain. Contrary to what we might assume intuitively, most meaningful learning is highly dependent on forgetting. For most of us, however, forgetting is annoying at best and frustrating, embarrassing, or dangerous at worst. What we would therefore like to know is how *not* to forget.

The last 150 years of research in psychology have indicated a number of techniques and conditions that help us learn more quickly and forget more slowly (i.e., better *retention*). Such techniques – and the knowledge gathered in this field of research in general – are particularly useful if we make a deliberate effort to learn and want to remember what we have learned. The most common situation is probably the preparation for an exam in an educational setting. Two reliable and powerful techniques are known as *testing* and *spacing*. Testing is usually used as assess memory (like on an exam) but is actually one of the most powerful study strategies we can use: Forcing ourselves to make active memory retrievals to questions (rather than re-reading the answer, for example) results in much better performance on subsequent tests because we have practiced the action that we will need to perform. Consequently, this technique is also known as *retrieval practice*. Spacing refers to the way in which we arrange repetitions of study events over time: We can learn a lot by studying for 10 hours the day before an exam but will forget it very quickly after the exam. If we *space out* the same amount of study time over five days, on the other hand, we will retain the information much longer. And even longer if we spread those five days over multiple weeks.

What is important to understand is that forgetting over time is not random, even if it often seems like it to us. In fact, the time course of forgetting can be described by relatively simple mathematical functions (known as *the power law of forgetting*) and those allow us to make *predictions*. If we know when a student has repeated, say, a set of French vocabulary and how well they performed on each repetition, we can use the regularity of forgetting expressed through mathematical functions to predict whether they will still know the word when we

test them. The predictions of such a *model* are not perfect, of course. But they can be tested and corrected.

If a student is studying a set of French vocabulary for 20 minutes, we can present individual words one by one in a random order. On each repetition, we can let the model make a prediction. Specifically, the model will predict the probability that the answer is still known and how quickly a response will be given. This prediction will probably not be perfect but by testing the student on each repetition, we immediately get information about whether they still knew the answer and how long it took them to give it. With this new information, we can update our mathematical model and our predictions will get better and better the longer we can observe the student's behavior. Therefore, we can use continuous testing not only as an effective learning technique but as a source of information that tells us something about the student.

Given the information we gather during learning in this way, we can do a lot better than repeating words randomly. Since we can predict which words are learned well and which will be forgotten soon, we can decide at each moment during the study session which word should be repeated. Here, we rely on the benefits of spacing and wait as long as possible before we repeat a word such that it is only tested again just before it is forgotten. This way, time is used efficiently and both testing and spacing are incorporated in the study session. One consequence of this is that the actual events during a study session are completely different for different students because the model uses the information from their responses to construct the optimal order of repetitions. This is referred to as adaptive or personalized learning.

The work in this thesis focuses on one such adaptive learning model that focuses on personalizing the schedule of repetitions within a single study session. In Chapter 2, we explain the background and mechanics of the model in detail and discuss how trial-by-trial accuracy and response time data can be used to adapt to individual students. We report the results from an experiment in which we contrast the adaptive system with a simple flashcard system. The data suggest that most students learn more material with the adaptive method and the chapter ends on a detailed discussion of how this benefit emerges from the mechanics of the model.

During the study session, the model keeps track of its own predictions and updates its parameters to reflect what it learns about each student. In Chapter 3, we focus on these parameters to test how stable they are over time and different study materials. This is relevant because we want to make sure that the model learns something about the student and does not just pick up on random fluctuations that have no meaning outside each study session. The data from the experiment reported in Chapter 3 suggest that parameters are very stable over time and less stable over materials. That is, if a student is estimated to forget a set of French

words slowly this week, they are very likely to also forget another set of French words slowly two weeks later. How quickly they forget French words, however, will tell us less about how quickly they forget the flags of different countries around the world.

The work reported in Chapter 3 suggests that the model's parameters capture something about the student that is relevant to how quickly they learn and forget. It is not clear, however, what exactly these parameters reflect: Are they a true measure of memory or could they simply capture something more general such as differences in cognitive functioning? The work presented in Chapter 4 addresses this question by estimating two prominent measures of cognitive functioning (general cognitive ability (i.e., IQ) and working memory capacity) and testing whether they are related to the parameters estimated by the model. In the data presented in Chapter 4, we could not find any evidence for such a relationship. This suggests that the model's parameters are not simply an artifact of higher-order cognitive processes but could be a true measure of memory performance.

In the final chapter of the thesis, I discuss the usefulness of the model's parameters more generally. Specifically, I show that if we know someone's parameter values estimated during study, we can predict how well they will perform on a test of the studied material. Furthermore, I discuss in more detail how the parameters can be used to distinguish the ability of different learners from each other. This is interesting because it can be done based on the data that is collected while students are *learning* rather than, for example, based on a grade on an exam. This implies that a lot of useful information (both for the students themselves but also for educators) can be extracted from study sessions. Together with the ability to predict future test performance, this might eliminate the need to administer tests for assessment purposes. Finally, future directions for the development and extension of the model are discussed in the last section of the Discussion chapter.

We all have to live with our own version of a limited memory that is prone to forgetting. The work in this thesis illustrates that we can use our theoretical understanding of regularities in forgetting to build computerized learning systems that identify the strengths and weaknesses of each student. Understanding how forgetting happens over time allows us to construct computerized, adaptive learning systems that personalize the repetitions of to-be-learned material such that we can all make the most of human memory.

Samenvatting proefschrift "Making the most of human memory"

Geheugen is een essentieel onderdeel voor de overleving van bijna alle levende organismen. Zonder geheugen zouden we hulpeloos en verward zijn omdat we zonder geheugen niet kunnen leren. Als gevolg daarvan zouden we geen zin kunnen geven aan de wereld om ons heen. In brede zin omvat het geheugen alles dat we opslaan over onze ervaringen in het leven. Veel van deze opgeslagen herinneringen kunnen worden teruggehaald en mondeling worden gerapporteerd of zijn terug te zien in moeiteloze uitvoeringen van vaardigheden die we geleerd hebben: Men kan tegelijkertijd de naam van een jeugdvriend noemen en veters strikken, beide zijn sterk afhankelijk van geheugen.

Onderzoek over geheugen focust op hoe we herinneringen encoderen, opslaan, en terughalen uit het brein. In tegenstelling tot wat we intuïtief denken, is vergeten essentieel voor het correct functioneren van het geheugen. Voor de meeste van ons is vergeten echter hinderlijk, maar vaak ook frustrerend, beschamend, en in het ergste geval zelfs gevaarlijk. Daarom is het begrijpen van hoe *vergeten werkt* cruciaal om een volledig beeld te krijgen over hoe het *geheugen werkt*. En wat we vervolgens eigenlijk willen weten is; hoe zorgen we ervoor dat we *niet* vergeten.

Psychologisch onderzoek van de laatste 150 jaar wijst uit dat er een aantal technieken zijn waarmee we sneller leren en langzamer vergeten. Deze technieken zijn met name zinvol als we een bewuste inspanning maken om te leren en willen herinneren wat we leren. De meest voorkomende situatie waarin dit soort leren vereist is, is waarschijnlijk in het onderwijs (denk bijvoorbeeld aan het studeren voor een examen). Twee betrouwbare en krachtige technieken zijn *testing* en *spacing*. *Testing* is een methode waarbij men zichzelf test tijdens een studiesessie (zoals ook tijdens een examen wordt gedaan) en is een van de meest krachtige studiestrategieën die we kunnen gebruiken. Bij *testing* dwingen we onszelf om actief antwoorden op vragen te herinneren (in plaats van bijvoorbeeld het herlezen van antwoorden). Wanneer men bijvoorbeeld Franse vocabulaire leert, is het proberen te herinneren van de vertaling van een woord (d.w.z. onszelf testen) in plaat van het te herlezen, een goede strategie om te leren. Dit resulteert in een betere prestatie op een toets naderhand omdat we precies de actie hebben geoefend die noodzakelijk is voor het goed presteren tijdens de toets. Deze techniek wordt daarom ook wel *retrieval practice* genoemd. *Spacing* gaat over de manier waarop we herhalingen van studeermomenten over tijd indelen. We kunnen veel leren als we de dag vóór een examen tien uur studeren, maar zullen deze informatie snel vergeten na het examen. Echter, als we de studietijd verdelen over bijvoorbeeld vijf dagen zullen we de informatie langer vasthouden. De retentie van informatie wordt langer naarmate de studietijd over een langere periode wordt verspreid.

Belangrijk om te weten is dat vergeten gedurende de loop van de tijd niet willekeurig is,

zelfs al lijkt het soms wel zo. Sterker nog, het verloop van vergeten kan worden omschreven met een relatief simpele wiskundige functie (bekend als *the power law of forgetting*). Dit biedt ons de mogelijkheid om het proces van vergeten te voorspellen. Als we weten wanneer een student haar Franse vocabulaire heeft geoefend en herhaald, en we weten hoe deze student op elke herhaling heeft gepresteerd, kunnen we de wiskundige functie gebruiken om te voorspellen of de student een bepaald woord zal herinneren als we haar testen. Voorspellingen als deze zijn niet volledig perfect. Maar door de voorspellingen continu te testen, kunnen ze worden verbeterd.

Als een student 20 minuten Franse vocabulaire studeert, kunnen we individuele woorden een-voor-een aan deze student presenteren. Voor elke herhaling van een woord kunnen we het model een voorspelling laten maken. Dat wil zeggen, het model schat de kans dat de student het antwoord herinnert en hoe snel het antwoord zal worden gegeven. De voorspelling zal in eerste instantie niet erg correct zijn, maar hoe meer informatie we verzamelen over de retentie en responstijd van de student, hoe accurater de voorspellingen worden. Het model wordt continu bijgewerkt met de input van de student. Hoe langer we het studiegedrag van de student volgen, hoe beter het model zal kunnen voorspellen wat de kans is dat de student een woord herinnert en wat de responstijd zal zijn. Herhaaldelijk testen is dus niet alleen een goede techniek om te leren, maar ook een bron van informatie over de leercapaciteiten van de student.

Gegeven de informatie die we op deze manier verzamelen, kunnen we, in plaats van het willekeurig aanbieden van woorden, het leerprogramma aanpassen voor de individuele student. Omdat we weten welke woorden al bekend zijn en welke woorden makkelijk vergeten worden, kunnen we elke moment tijdens de studiesessie bepalen welk woord herhaald moet worden. Hier vertrouwen we op de voordelen van *spacing*; we wachten met het herhalen van een woord net tot het moment vóórdat een woord waarschijnlijk zal worden vergeten. Op deze manier wordt tijd efficiënt gebruikt en worden zowel *testing* als *spacing* opgenomen in de studiesessie. Omdat het model informatie van de individuele student gebruikt om de studiesessie vorm te geven, betekent dit dat iedere student een unieke en op maat afgestemde studiesessie krijgt. Dit wordt ook wel adaptief of gepersonaliseerd leren genoemd.

Het werk in dit proefschrift gaat over één van deze adaptieve leermodellen. Dit adaptieve leermodel focust zich op het personaliseren van herhalingsschema's binnen één enkele studiesessie. In Hoofdstuk 2 beschrijven we in detail de achtergrond en mechanismes van het model. Daarnaast bespreken we hoe data van de trail-by-trail accuraatheid en de responstijd worden gebruikt om de studiesessie op het individu aan te passen. In dit hoofdstuk rapporteren we de resultaten van een experiment waarin we het adaptieve systeem vergelijken met een simpel *flashcard* systeem. De uitkomsten suggereren dat de meeste studenten meer mate-

riaal kunnen leren met het adaptieve systeem dan met het *flashcard* systeem. In de discussie bespreken we uitgebreid hoe dit voordeel voorkomt uit de mechanismes van het model.

Gedurende de studiesessie houdt het model zijn eigen voorspellingen bij een werkt zijn parameters bij op basis van het leertraject van elke student. In Hoofdstuk 3 focussen we op de parameters om te testen hoe robuust deze zijn in de loop van de tijd en voor verschillend studiemateriaal. Dit is belangrijk omdat we er zeker van willen zijn dat het model iets unieks meet over de individuele student en niet willekeurige fluctuaties oppikt die geen betekenis hebben buiten de studiesessie. De data van het experiment in Hoofdstuk 3 laten zien dat de parameters erg stabiel zijn in de loop van de tijd maar minder stabiel voor verschillend studiemateriaal. Dat wil zeggen, als het model schat dat een student een aantal Franse woorden deze week erg snel zal vergeten, zal deze student een aantal andere Franse woorden over twee weken ook snel vergeten. Echter, hoe snel deze student Franse woorden vergeet vertelt ons minder over hoe snel deze student vlaggen van verschillende landen zal vergeten.

Het onderzoek dat in Hoofdstuk 3 wordt gerapporteerd suggereert dat de parameters van het model informatie bevatten over de student die relevant zijn voor hoe snel deze zal leren en vergeten. Het is echter onduidelijk wat deze parameters betekenen: Zijn het ware metingen van geheugen, of meten ze simpelweg iets algemeens zoals verschil in cognitief functioneren? Het werk in Hoofdstuk 4 besteed aandacht aan deze vraag door twee prominente maten van cognitief functioneren (algemene cognitieve vaardigheden (d.w.z. IQ) en de capaciteit van het werkgeheugen) te schatten en te testen of deze een relatie hebben met de parameters die door het model worden geschat. De uitkomsten van Hoofdstuk 4 boden geen bewijs voor de aanwezigheid van een relatie. Dit suggereert dat de parameters van het model niet slechts het resultaat zijn van algemene cognitieve processen, maar een ware meting van geheugenprestatie.

In het laatste hoofdstuk van dit proefschrift bespreek ik de bruikbaarheid van de parameters van het model in algemene zin. In het bijzonder laat ik zien dat als we iemands parameters schatten, we kunnen voorspellen hoe goed ze zullen presteren op een test van het studiemateriaal. Verder beargumenteer ik dat de parameters kunnen worden gebruikt om de vaardigheden van verschillende studenten te kunnen onderscheiden. Dit is interessant omdat men daarmee de vaardigheden van een student kan beoordelen terwijl men leert in plaats van slechts ná een toetsing. Dit betekent dat er veel bruikbare informatie (zowel voor studenten als onderwijzers) kan worden verzameld uit een studiesessie. Gegeven de rijke bron aan informatie en de mogelijkheid om toekomstig presteren te kunnen voorspellen, elimineert dit mogelijkerwijs de noodzaak voor toetsing ter beoordeling. Tot slot worden in dit hoofdstuk toekomstige richtingen voor ontwikkeling en uitbreiding van het model besproken.

We moeten allemaal leven met een beperkt geheugen dat vatbaar is voor vergeten. Het

werk in dit proefschrift illustreert dat we onze theoretische kennis over wetmatigheden van het vergeten kunnen gebruiken om geautomatiseerde leersystemen te ontwikkelen die de sterktes en zwaktes van elke student kunnen identificeren. Door deze adaptieve leersystemen die herhaling van de te leren stof personaliseren, kunnen we allemaal het beste halen uit ons geheugen.

Zusammenfassung der Dissertation “Die bestmögliche Nutzung des menschlichen Gedächtnisses“

Das Gedächtnis ist für das Überleben fast aller lebenden Organismen absolut notwendig. Ohne unser Gedächtnis wären wir hilflos und verwirrt. Wir wären nicht in der Lage auf Gelerntes zurückzugreifen und die Welt um uns herum zu verstehen. Im weitesten Sinne umfasst das Gedächtnis alle Erinnerungen an gemachte Erfahrungen. Auf viele dieser abgespeicherten Erinnerungen können wir jederzeit bewusst zurückgreifen und sie verbal kommunizieren. Andere finden ihren Ausdruck hingegen in der problemlosen Ausführung gelernter Fähigkeiten: es wäre kein Problem sich die Schuhe zu binden und dabei die Namen der Kindheitsfreunde aufzuzählen – beide Aktivitäten sind im hohen Maße abhängig von unserem Gedächtnis.

Die Gedächtnisforschung untersucht, wie wir Erinnerungen speichern und später wieder abrufen. Entgegen unserer Intuition ist die Grundlage für die meisten wichtigen Lernprozesse tatsächlich das Vergessen. Für die meisten von uns ist das Vergessen jedoch frustrierend, peinlich oder gar gefährlich. Was wir deshalb oft lieber wissen wollen, ist wie wir das Vergessen *vermeiden* können. Die letzten 150 Jahre psychologischer Forschung haben eine Vielzahl von Techniken und Bedingungen hervorgebracht, die uns dabei helfen können, schneller zu lernen und langsamer zu vergessen (Erhöhung der sogenannten „Retention“ (*retention*)). Diese Techniken – und generell das in diesem Feld der Forschung gesammelte Wissen – sind besonders hilfreich, wenn wir bewusst versuchen, etwas zu lernen und dieses später zu erinnern. Die wohl häufigste Situation dieser Art ist die Vorbereitung auf eine Prüfung im schulischen oder universitären Kontext. Zwei zuverlässige und effektive Techniken sind das „Testen“ (*testing*) und das „Strecken“ (*spacing*). Meistens wird das Testen genutzt um unsere Gedächtnisleistung zu messen – zum Beispiel bei einem Vokabeltest. Tatsächlich ist es jedoch auch eine der wirkungsvollsten Lernstrategien: die aktive Gedächtnisabfrage durch das wiederholte Beantworten von Fragen – im Gegensatz zum wiederholten Lesen der vorformulierten Antworten – führt zu wesentlich besseren Ergebnissen, da so die konkrete Tätigkeit trainiert wird, die in der Prüfungssituation ausgeübt werden muss. Folglich wird diese Technik auch als „Wiederabruf-Übung“ (*retrieval practice*) bezeichnet. Das Strecken hingegen beschreibt die Art und Weise, in der wir die Wiederholungen von Lerneinheiten über einen bestimmten Zeitraum gestalten: Zwar können wir eine Menge lernen, wenn wir am Tag vor einer Prüfung für 10 Stunden lernen, doch werden wir das Gelernte auch sehr schnell wieder vergessen. Wenn wir diese 10 Stunden jedoch über fünf Tage *strecken*, können wir das Gelernte wesentlich länger behalten; dieser Effekt ist noch größer, wenn wir diese fünf Tage auf mehrere Wochen verteilen.

Was in diesem Zusammenhang wichtig zu verstehen ist, ist die Tatsache, dass das Vergessen über einen bestimmten Zeitraum kein willkürlicher Prozess ist, auch wenn dem oft so scheint. Tatsächlich lässt sich dieser Vergessensprozess durch vergleichsweise einfache mathematische Formeln beschreiben (man spricht auch vom „Potenzgesetz des Vergessens" (*power law of forgetting*)). Diese Formeln erlauben es uns, *Vorhersagen* zu formulieren. Angenommen wir wissen, wann eine Person beispielsweise eine Reihe von Französischvokabeln gelernt hat und wie erfolgreich diese Person bei jeder Wiederholung war, so erlaubt uns die Regelmäßigkeit des Vergessens – ausgedrückt in Form von mathematischen Funktionen – vorherzusagen, ob sie die Vokabeln immer noch weiß wenn wir sie testen. Natürlich sind die Vorhersagen eines solchen Modells nicht perfekt, aber sie können bei jeder Wiederholung geprüft, angepasst und somit verbessert werden.

Wenn eine Person also 20 Minuten lang eine Reihe von Französischvokabeln lernt, können wir dieser die einzelnen Wörter in willkürlicher Reihenfolge einzeln vorlegen. Wir können dann das Modell nutzen, um für jede Wiederholung eine Vorhersage zu treffen. Genauer gesagt trifft das Modell eine Vorhersage nicht nur darüber, mit welcher Wahrscheinlichkeit die Vokabel noch abrufbar ist, sondern auch, wie schnell eine Antwort gegeben wird. Diese Vorhersage ist selbstverständlich nicht perfekt, aber da jede Wiederholung auch ein Test ist, erhalten wir sofort Informationen darüber, ob und wie schnell die Person die Antwort geben konnte. Anhand der neu gewonnenen Informationen kann das mathematische Modell stets verbessert werden. Somit werden unsere Vorhersagen umso besser je länger wir das Verhalten der Versuchsperson beobachten können. Das wiederholte Testen kann somit nicht nur als effektive Lernstrategie genutzt werden, sondern liefert darüber hinaus nützliche Informationen über das Verhalten der untersuchten Person.

Durch die Nutzung der Informationen, die wir durch diese Lernstrategie erhalten, sind die Lernergebnisse wesentlich besser als es bei willkürlicher Wiederholung der Wörter der Fall wäre. Da wir vorhersagen können, welche Wörter schnell gelernt und welche schnell vergessen werden, kann zu jedem Zeitpunkt der Lerneinheit entschieden werden, welches Wort wiederholt wird. Dabei können wir auf die Vorteile des Streckens zurückgreifen und ein Wort erst kurz vor dem geschätzten Zeitpunkt des Vergessens wiederholen. Somit wird der Zeitfaktor bestmöglich genutzt und beide Lerntechniken (Testen und Strecken) sinnvoll in einer Lerneinheit verknüpft. Da das Modell die individuellen Informationen der einzelnen Versuchsperson nutzt, um eine optimale Reihenfolge der Wiederholungen festzulegen, variieren die konkreten Schritte im Lernprozess von Person zu Person (adaptives oder personalisiertes Lernen).

Ein solches adaptives Lernmodell, basierend auf dem personalisierten Wiederholungsplan innerhalb einer einzelnen Lerneinheit, ist der Kern der in dieser Dissertation enthal-

tenen Arbeiten. Das zweite Kapitel gibt einen detaillierten Überblick über die theoretischen Grundlagen und praktischen Mechanismen dieses Modells. Darüber hinaus wird diskutiert, inwiefern die gewonnenen Informationen – sowohl über die Genauigkeit der Vorhersagen als auch über die Reaktionszeit – genutzt werden können, um das Modell individuellen Versuchspersonen anzupassen. Zudem wird unser Experiment präsentiert, in welchem das adaptive Modell mit einen einfachen Lernkarten-Modell verglichen wurde und dessen Ergebnis verdeutlicht, dass das adaptive Modell eine effektivere Lerntechnik darstellt. Abschließend wird ausführlich diskutiert, wie diese Vorteile auf die konkrete Arbeitsweise des Modells zurückzuführen sind.

Während der Lerneinheit verfolgt das Modell seine eigenen Vorhersagen und aktualisiert die Parameter anhand der neugewonnen Informationen kontinuierlich. Im dritten Kapitel prüfen wir, wie beständig diese Parameter im Laufe der Zeit und bezüglich unterschiedlicher Lernstoffe sind. Dieser Schritt ist wichtig, um sicherzustellen, dass das Modell tatsächlich etwas über die Versuchsperson lernt und nicht einfach willkürliche und außerhalb der Lerneinheit bedeutungslose Veränderungen misst. Die Daten des im dritten Kapitel vorgestellten Experiments zeigen, dass die Parameter zwar im Laufe der Zeit stabil bleiben, jedoch nicht bezüglich unterschiedlicher Lernstoffe. Wenn eine Versuchsperson laut dem Modell eine Reihe von Französischvokabeln in dieser Woche relativ langsam vergisst, ist die Wahrscheinlichkeit hoch, dass diese eine Reihe anderer Französischvokabeln in zwei Wochen ebenfalls langsam vergisst. Wie langsam die Versuchsperson Französischvokabeln vergisst, sagt jedoch weniger darüber aus, wie schnell die gleiche Versuchsperson beispielsweise die Flaggen von Staaten vergessen wird.

Die im dritten Kapitel vorgestellte Arbeit lässt den Schluss zu, dass die Parameter des Modells etwas über die Versuchsperson abbildet, das Einfluss darauf hat wie schnell diese etwas lernt und vergisst. Jedoch wird nicht deutlich, was genau diese Parameter abbilden: Messen sie tatsächlich konkrete Gedächtnisprozesse oder könnte es sein, dass sie eher etwas Allgemeineres abbilden, wie beispielsweise Unterschiede bezüglich der kognitiven Funktionen? Diese Frage wird im vierten Kapitel behandelt: zum einen werden zwei typische Maßeinheiten für die kognitive Funktion geschätzt (die generelle kognitive Fähigkeit (der IQ) und die Kapazität des Arbeitsgedächtnisses) und zum anderen wird geprüft, ob diese Maßeinheiten mit den geschätzten Parametern des Modells zusammenhängen. Die im vierten Kapitel präsentierten Daten suggerieren, dass eine solche Beziehung nicht besteht. Dies legt den Schluss nahe, dass die Parameter des Modells keine übergeordneten kognitiven Prozesse abbilden, sondern tatsächlich die Gedächtnisleistung messen.

Das abschließende Kapitel bietet eine allgemeinere Diskussion über die Nützlichkeit dieser Parameter. Zum einen wird aufgezeigt, dass das Wissen über die während einer Lernein-

heit geschätzten Parameterwerte einer Versuchsperson es ermöglichen, treffende Vorhersagen über die Leistung dieser Person bei einer Abfrage über das Gelernte zu machen. Zum anderen wird ausführlicher diskutiert, inwiefern die Parameter genutzt werden können, um zwischen den Fähigkeiten der individuellen Lernenden zu unterscheiden. Dies ist besonders interessant, weil somit Aussagen über die individuellen Fähigkeiten gemacht werden können, die während des *Lernens* ermittelt werden und nicht auf Grundlage von Prüfungsnoten. Aus den Lerneinheiten lassen sich also wichtige Information ziehen – sowohl für die Lernenden als auch für die Lehrenden. Dies, zusammen mit der Fähigkeit zukünftige Testergebnisse vorherzusagen, könnte dazu beitragen, dass Prüfungen zum Zweck der Bewertung nicht länger nötig sind. Der letzte Abschnitt der Diskussion widmet sich abschließend dem zukünftigen Entwicklungs- und Erweiterungspotential des Modells.

Wir alle müssen damit leben, dass unser Gedächtnis individuell eingeschränkt ist und stets zum Vergessen neigt. Die hier vorgestellte Arbeit zeigt jedoch, dass wir unser theoretisches Wissen über die Regelmäßigkeit des Vergessens sinnvoll nutzen können, um computerbasierte und adaptive Lernprogramme zu entwickeln, die den Wiederholungzirkel des zu lernenden Materials personalisieren. Diese Lernprogramme beziehen die individuellen Stärken und Schwächen der Lernenden mit ein und erlauben somit die bestmögliche Nutzung des menschlichen Gedächtnisses.

Acknowledgements

When I designed the cover for this thesis, I felt a bit strange putting only my name on it. I'm the one getting all the credit but there's no way I could have finished my PhD without the help of a large number of people. This section is meant to acknowledge some of them. I am immensely grateful to all of you. My life in the last four years was amazing because I always had something interesting to do and was always surrounded by loving, supportive, and incredibly fun people. Thank you all.

First and foremost, I would like to thank my family. You have always supported me and encouraged me to pursue my interests. You contributed nothing to the content of this thesis. Yet, your contribution was the most important because you did all the difficult groundwork that made me the kind of person that'd tackle a project like this. There was never a doubt in my mind that you'd have my back no matter what. I couldn't have done it without that peace of mind.

Most directly responsible for the completion of this thesis are my supervisors. I remember feeling overwhelmed by the complexity of both work and my life shortly after I started my PhD. I am grateful to Richard and Candice for your patience and guidance. I couldn't have hoped for better role models and mentors and was sad to see you leave Groningen. Rob was the perfect promotor, always having an open ear, giving me advice, and keeping me on track. Hedderik saw me most and had the biggest influence on me; thank you for always making time, counsel me through the endless obstacle course that is academia, and guiding me towards becoming a more independent researcher. A PhD project is a team effort and I can't thank you all enough for playing on my team.

The psychology faculty has always been a pleasant place to work. Thanks to my various office mates – Tanja, Tam, Mariska, Rasa, Edyta, Sabine, and the two Michaels – and colleagues – both at Psychometrics and Statistics and at Experimental Psychology –, spending time at work was only as stressful as I made it for myself. I appreciate that doors, ears, and minds were open and the flat hierarchy that made everyone easy to approach.

I also want to thank all the students. Lecturing, being an academic mentor, and grading reports always made me feel a bit awkward because I still see myself as a student. These interactions taught me a lot about myself. The best part was to see students turn into colleagues. Friederike, Sarah, and Charlotte: This thesis would have been at least a chapter shorter if it weren't for you. You set a very high bar for any future collaborators, thank you.

A big thank you should also go to the "Minnaars en Minnaressen". Aafke, Ana, Anne Marthe, Aytaç, Berry, Catia, Daniel, Darya, Elliot, Felicity, Jolien, Julia, Kim, Lowie, Luzia, Maja, Marloes, Nadine, Nico, Rob, Russell, Susie, Tassos, Tineke, Tomas, and Wisnu. Thanks for the drinks after work and the countless other social events over the last couple of years. A special thanks to the inner circle – Aafke, Darya, Kim, and Rob – and a very special thanks to Nico

and Berry: living with you was amazing. Also on this list should be George and Garrett, who I didn't see nearly enough. Having you in my life made me a better person. Thanks for calling me out on my shit and the never-ending support. I love you guys.

There were also some people outside of academia that I want to thank. Adam, Arun, Brenda, and Harmen (and everyone else at Squadraat) for countless hours of squash. I always had a blast and knew I'd get a good workout in if you guys were on court. Thanks for letting me win occasionally. I'd also like to thank my friends in Germany for making the effort to stay in touch and all the good times and trips that were a necessary break from my normal PhD life: Franze, Max, Tobi, Katchi, Lutz, Tobi, Caro, Lena, Catharina, and Basti and everyone else that was around occasionally.

The last four years were the best years of my life. I am sad to see them come to an end and incredibly excited for the future. A future in which I hope to cross paths with all of you again.