TOPICS
TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science (2016) 1–17

ISSN:1756-8757 print / 1756-8765 online
DOI: 10.1111/tops.12183

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

Florian Sense,[a,b,c] Friederike Behrens,[d] Rob R. Meijer,[c] Hedderik van Rijn[a,b,c]

[a]*Department of Experimental Psychology, University of Groningen*
[b]*Behavioral and Cognitive Neuroscience, University of Groningen*
[c]*Department of Psychometrics and Statistics, University of Groningen*
[d]*Research School of Behavioral and Social Sciences, University of Groningen*

Received 30 June 2015; received in revised form 23 October 2015; accepted 23 October 2015

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 effect. 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. We 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 system allowed different rates of forgetting for different materials, it could adapt to individual learners more readily.

Keywords: Learning; Spacing; Testing; Tutoring; Parameter stability

1. Introduction

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

Correspondence should be sent to Florian Sense, Department of Psychology, Grote Kruisstraat 2/1, 9721 TS Groningen, The Netherlands. E-mail: f.sense@rug.nl (or) Hedderik van Rijn, Department of Psychology, Grote Kruisstraat 2/1, 9721 TS Groningen, The Netherlands. E-mail: hedderik@van-rijn.org

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, Maanen, & 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 made by retaining the estimated characteristics over sessions. However, this requires that these characteristics do not fluctuate too much between 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). It has also been shown convincingly, based on both behavioral (e.g., Lindsey, Shroyer, Pashler, & Mozer, 2014; Nijboer, 2011; Van Rijn et al., 2009) and psychophysiological data (e.g., 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, Vul, Rohrer, Wixted, & Pashler, 2008). However, the vast majority of students do not space their studying at all: They mass-study just before an exam (Taraban, Maki, & Rynearson, 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, Tabbers, Pecher, & Zeelenberg, 2012; Kornell & Bjork, 2008b; 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, Wixted, Pecher, Zeelenberg, & Huber, 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: (a) maximizing time between repetitions of an item to get the biggest spacing effect and

(b) 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 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 (e.g., Pavlik & Anderson, 2008) is an excellent control condition because many students report using similar procedures to study for exams (Hartwig & Dunlosky, 2012; Kornell & Bjork, 2008a; Kornell, Son, College, & York, 2009; Wissman, Rawson, & Pyc, 2012).

As a starting point for the development of such adaptive 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 his or her memory and therefore assumed to be a trait-like, stable property, regardless of whether he or she studies 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 3 weeks as well as across four different types of declarative fact material.

2. Methods

Participants were tested in three separate sessions, each spaced 1 week apart. Each session consisted of two blocks. 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 Fig. 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 dif-

ferent types of declarative fact material, we can investigate their stability across knowledge domains.

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.

2.1 The model

The model used in this experiment is based on ACT-R's declarative memory equations (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:

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

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,

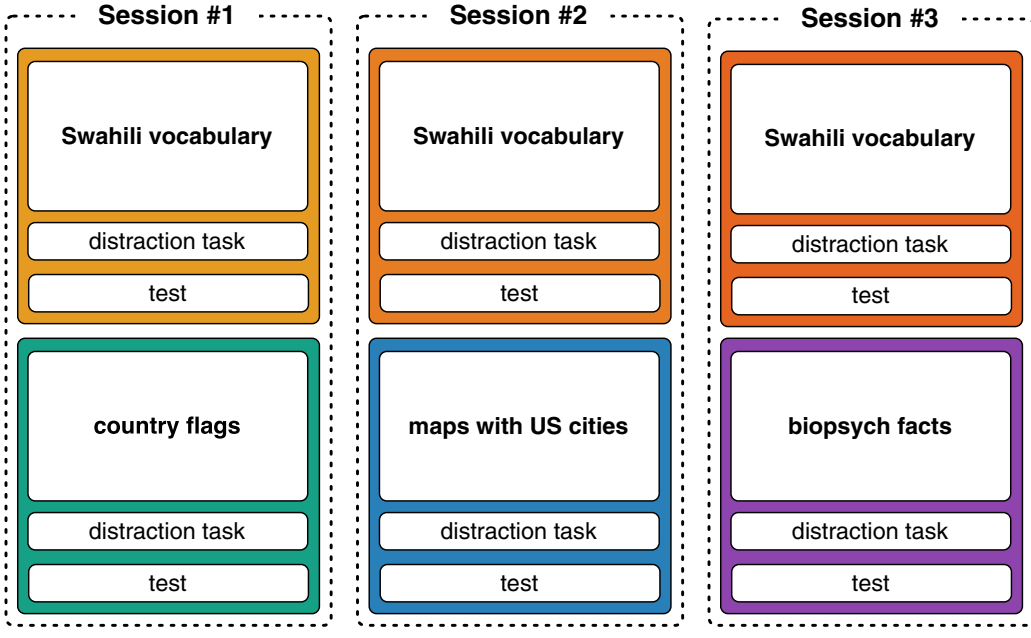

Fig. 1. Overview of the experimental design of the study. The sessions are spaced 1 week apart. A different type of declarative fact material is studied in each (color-coded) block. Six unique sets of items were used, one per block.

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:

$$d_j = ce^{Ai(t)} + \alpha$$

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. 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:

$$RT_i(t) = \mathrm{Fe}^{-Ai(t)} + \text{fixed time}$$

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 given 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 et al. (2009) and Nijboer (2011) have proposed an algorithm, based on the original Pavlik and Anderson work, which 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 with 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 on a trial-by-trial basis. This decay parameter is an operationalization of the *rate of forgetting*, which is the term we will use in the remainder of this article.

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 Fig. 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.

2.2 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

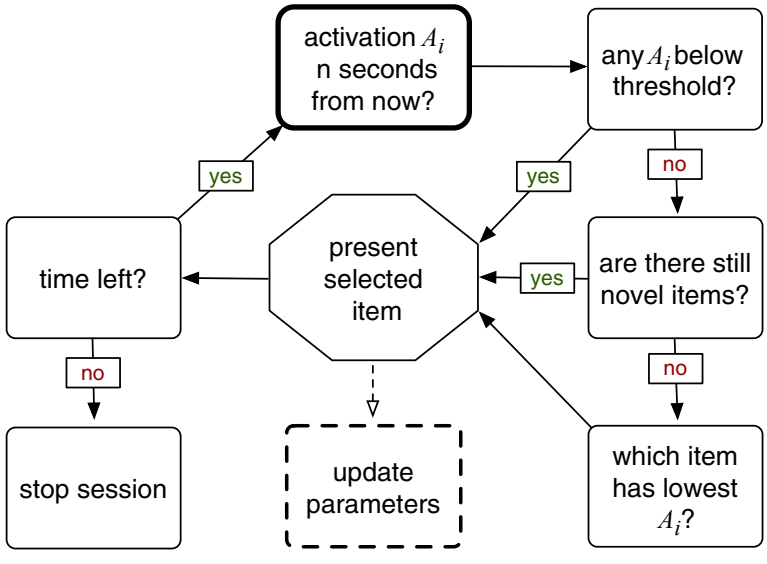

Fig. 2. A graphical representation of how the model determines the order in which to repeat old and present new items.

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.

2.3 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 <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).

2.4. 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:

2.4.1. Vocabulary

Seventy-five Swahili-English word pairs were selected from the list compiled by Van den Broek, Segers, Takashima, and Verhoeven (2014) and are listed in the Supplement. 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.

2.4.2. 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 in the Supplement. The

country's flag was the cue to which the participants had to respond by typing in the country's name.

2.4.3. 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 in the Supplement. Participants always saw a map of the United States 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.

2.4.4. Biopsychology facts

A list of 25 biopsychology facts was compiled from the Glossary in Kalat (2012), a textbook 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 in the Supplement. The description of the term was the cue to which the participants had to respond by typing in the described term.

2.5. Procedure

Each person participated in the study for three sessions on 3 days, each session spaced 1 week apart. Within each session, there were two blocks. Each block consisted of a 20-min study session, a 5-min distraction task, followed by a test of the studied declarative fact material that took about 5 more minutes (see Fig. 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 5-min distraction was a simple variation of the puzzle game Tetris, which participants played until they were automatically redirected 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 s 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 et al. (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 et al. (2012) showed that the optimal duration of a *study trial* is 4 s.

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.

2.6. 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. Finally, 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 https://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. Fig. 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.

As described in the subsection *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 (a common default; see 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.

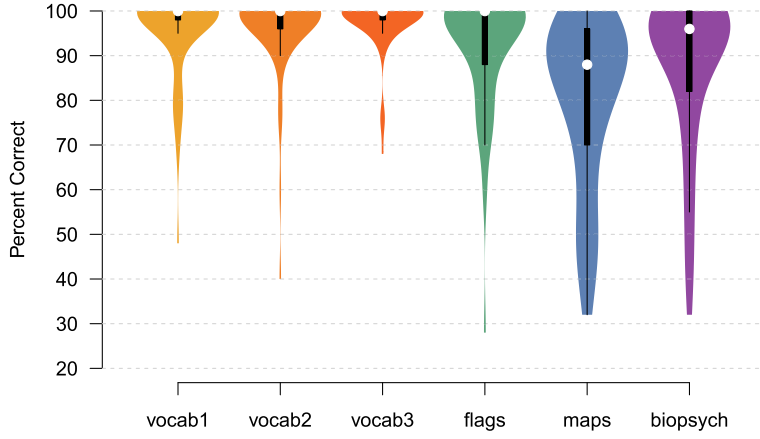

Fig. 3. Performance on the final test that was taken at the end of each block.

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 Fig. 4. For this plot, each participant contributed one mean *rate of forgetting* per block and his or her 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.

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 Fig. 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 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).

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

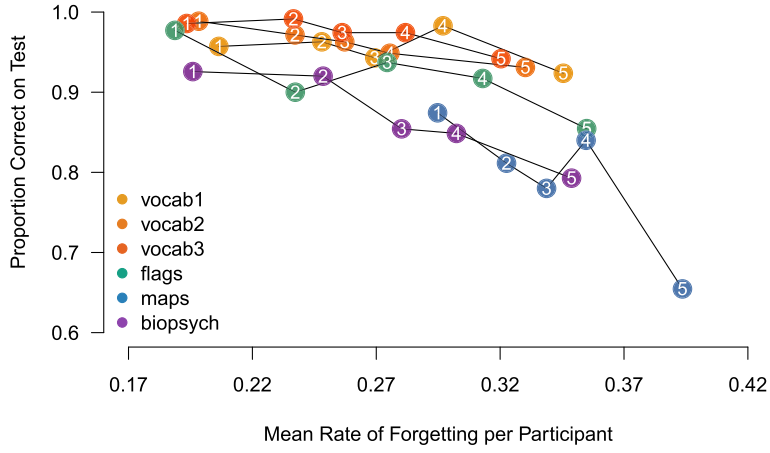

Fig. 4. Participants' mean rate of forgetting is plotted against their performance on the final test for each block. The data are binned to make the plot more readable. Bins were created by ordering the 76 individual data points in each condition on the *x*-axis, dividing them into five bins, and then computing the mean values for each bin.

Table 1
Correlations of mean rates of forgetting between all blocks

	vocab1	vocab2	vocab3	flags	maps
vocab1	–				
vocab2	0.758	–			
vocab3	0.771	0.863	–		
flags	0.603	0.630	0.549	–	
maps	0.499	0.556	0.495	0.604	–
biopsych	0.630	0.572	0.534	0.511	0.384

Note. All correlations differ significantly from 0 with $p < 0.001$.

violations of homoscedasticity and normality. Table 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.

The *rates of forgetting* do not significantly differ between sessions ($t(73)=1.3$, $p = .198$). However, the contrast between the Swahili and non-Swahili blocks significantly influences the *rates of forgetting* ($t(9506) = 16.5$, $p < .001$), 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 versus non-Swahili is not significant ($t(82) = 1.81$, $p = .074$). This means the difference in the *rate of forgetting* when performing a non-Swahili task compared with the Swahili task does not differ as a function of the session in which the task is performed in. Fig. 5 gives a visual overview of the mean *rate of forgetting* in each block and suggests that the significant difference of the Swahili versus non-Swahili comparison is driven by the higher *rate of forgetting* in

Table 2
Results from the linear mixed-effects regression with dummy coding

	$beta_{ln(ROF)}$	*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

Note. 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.

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 ($t(66) = 0.27$, $p = .785$.[1]

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 sessions were entered as independent predictors in a backward 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 (Fig. 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

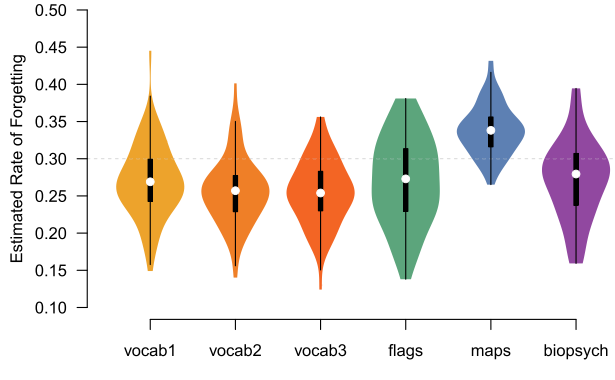

Fig. 5. The mean rate of forgetting per participant for each block as a violin plot to show the density estimates as well as classic box plots (in black) with superimposed medians (in white).

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 ($\beta = 0.64$, $t(64) = 7.3$, $p < .001$) than the performance during the first (and therefore earlier) session ($\beta = 0.24$, $t(64) = 3.0$, $p = .004$). 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 0.77 ($F(2, 64) = 111.4$, $p < .001$).

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 Fig. 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 (i.e., 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 Fig. 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 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 *biopsychology* conditions were similar to each other while the two conditions that used visual information (*flags* and *maps*) were as different as the *maps* and *biopsychology* con-

ditions, 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* to account for these individual differences.[2]

Acknowledgments

An earlier and shorter version of this article was published in the conference proceedings of the 13th International Conference on Cognitive Modeling.

Notes

1. Bayes factor *t*-tests (Rouder, Speckman, Sun, Morey, & Iverson, 2009) were performed with the default settings of the BayesFactor package (Morey & Rouder, 2015) 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 provide support for the null model when compared with the alternative model when comparing the *flags* and *biopsych* conditions (Bayes factor = 7.2).

2. The supplementary material, all analyses, and the raw data can be found at https://github.com/fsense/parameter-stability-paper.

References

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.

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* New York: Oxford University Press.

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

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

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. doi:10.1080/09541440701326097.

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. doi:10.3758/MC.36.2.438.

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition*, *20*(6), 633–642.

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), 354–380. doi: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), 1095–1102. doi:10.1111/j.1467-9280.2008.02209.x.

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. doi:10.1037/a0025897.

Delaney, F. 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 peter. In B. H. Ross (Ed.), *The psychology of learning and motivation* (1st ed., Vol. 53, pp. 63–147). Burlington: Elsevier. doi:10.1016/S0079-7421(10)53003-2

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.

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. doi:10.1037//0021-9010.84.5.795.

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. doi:10.1080/09658211.2013.798416.

Goossens, N. A. M. C., Camp, G., Verkoeijen, P. P. J. L., Tabbers, H. K., & Zwaan, R. A. (2014). 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. doi: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. doi:10.3758/s13423-011-0181-y.

Hintze, J. L., & Nelson, R. D. (1998). Violin plots: A box plot-density trace synergism. *The American Statistician*, *52*(2), 181. doi:10.2307/2685478.

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. *Quarterly Journal of Experimental Psychology*, *65*(5), 962–975. doi:10.1080/17470218.2011.638079.

Jastrzembski, T. S., Gluck, K. A., & Gunzelmann, G. (2006). Knowledge tracing and prediction of Future Trainee Performance. *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). Cengage Learning: Wadsworth.

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. doi:10.1016/j.jarmac.2014.05.006.

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science*, *319* (5865), 966–968. doi:10.1126/science.1152408.

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

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

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

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 – ICCM2009*. UK: Manchester.

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

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, *19*(4–5), 494–513.

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

Nijboer, M. (2011). *Optimal fact learning: Applying presentation scheduling to realistic conditions*. Groningen, The Netherlands: Unpublished master's thesis, University of Groningen.

Pavlik, P. I., Jr., & Anderson, J. R. (2003). An ACT-R model of the spacing effect. F. Detje, D. Doerner, & H. Schaub (Eds.), In *Proceedings of the Fifth International Conference on 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), 559–586. doi: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. doi:10.1037/1076-898X.14.2.101.

Pavlik, P. I., Bolster, T., Wu, S., Koedinger, K. R., & MacWhinney, B. (2008). Using optimally selected drill practice to train basic facts. In B. Woolf, E. Aimer, & R. Nkambou (Eds.), *Proceedings of the 9th International Conference on Intelligent Tutoring Systems* (pp. 593–602). Canada: Montreal.

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science*, *330*, 335.

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.

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.

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. doi: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. doi:10.1037/0033-295X.103.4.734.

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. doi:10.3758/BF03207718.

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), 803–812.

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), doi: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. In A. Howes, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 9th International Conference on Cognitive Modeling* (pp. 110–115). UK: Manchester.

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

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