Predicting University Students' Exam Performance Using a Model-Based Adaptive Fact-Learning System

Sense, Florian; van der Velde, Maarten; van Rijn, Hedderik

Published in:
Journal of Learning Analytics

DOI:
10.18608/jla.2021.6590

JOURNAL OF LEARNING ANALYTICS
SOLAR SOCIETY for LEARNING ANALYTICS RESEARCH
Volume 8(3), 155—169. https://doi.org/10.18608/jla.2021.6590

Predicting University Students' Exam Performance Using a Model-Based Adaptive Fact-Learning System

Florian Sense[*,1], Maarten van der Velde[*,2], Hedderik van Rijn[3]

Abstract
Modern educational technology has the potential to support students to use their study time more effectively. Learning analytics can indicate relevant individual differences between learners, which adaptive learning systems can use to tailor the learning experience to individual learners. For fact learning, cognitive models of human memory are well suited to tracing learners' acquisition and forgetting of knowledge over time. Such models have shown great promise in controlled laboratory studies. To work in realistic educational settings, however, they need to be easy to deploy and their adaptive components should be based on individual differences relevant to the educational context and outcomes. Here, we focus on predicting university students' exam performance using a model-based adaptive fact-learning system. The data presented here indicate that the system provides tangible benefits to students in naturalistic settings. The model's estimate of a learner's rate of forgetting predicts overall grades and performance on individual exam questions. This encouraging case study highlights the value of model-based adaptive fact-learning systems in classrooms.

Notes for Practice

- The spacing and testing effect are well-known and robust findings in cognitive psychology, but they are difficult to translate to educational practice. In this study, we deployed an adaptive fact-learning system to exploit these effects in an undergraduate cognitive psychology course. Study behaviour and exam performance were recorded for two consecutive cohorts of students using the system.
- The estimated model parameter of the adaptive fact-learning system predicted exam grades up to two weeks in advance (Section 4.2.3). In practice, this information could be used as an early warning signal to students and their instructors that additional study is likely required for a passing grade.
- Performance on individual items during practice was predictive of their eventual recall on the exam (Section 4.2.4). Thus, the adaptive system could provide tailored feedback to students and help them assess their mastery of the material. Furthermore, instructors may use this information to select appropriate materials for the exam, or even forgo exams altogether by relying exclusively on the assessments provided by the adaptive system.
- In terms of the expected gains in exam performance, students who have an additional hour to spend studying are better off spending it on the adaptive learning system than on other self-reported study activities (Section 4.2.6).
- One strength of the system deployed here is that it is agnostic to the material studied, which makes it applicable in a wide range of educational settings. In general, adaptive learning systems are a viable way of translating spacing and testing effects to educational practice, and they also provide useful insights to students and instructors.

Keywords
Adaptive learning, technology-enhanced learning, cognitive tutor, cognitive model, rate of forgetting, academic achievement

Submitted: 06/05/2019 — **Accepted:** 26/04/2021 — **Published:** 05/07/2021

[*] *Shared first authorship*

[1] *Email: f.sense@rug.nl Address: Department of Experimental Psychology, and Behavioral and Cognitive Neurosciences, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, Netherlands ORCID ID: https://orcid.org/0000-0001-9982-4701*
[2] *Email: m.a.van.der.velde@rug.nl Address: Department of Experimental Psychology, and Behavioral and Cognitive Neurosciences, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, Netherlands ORCID ID: https://orcid.org/0000-0003-4849-2676*
[3] *Email: d.h.van.rijn@rug.nl Address: Department of Experimental Psychology, and Behavioral and Cognitive Neurosciences, University of Groningen, Groningen, Netherlands, Grote Kruisstraat 2/1, 9712 TS Groningen, Netherlands ORCID ID: https://orcid.org/0000-0002-0461-9850*

1. Introduction

Living in an ever-changing environment compels us to be lifelong learners and means that learning extends far beyond structured classroom settings. Learners have to decide when and what to study, whether they have learned the material well enough to stop rehearsing it, and how to distribute their time across various materials. These behaviours and decisions are collectively known as self-regulated learning (see Bjork, Dunlosky, & Kornell, 2013, for a review). New educational technology could alleviate some of the problems that learners face when self-regulating their learning. Such digital tools can make learning more accessible, and learning analytics can be used to monitor progress and guide study decisions. Many modern educational tools have adaptive components that effectively make some of the self-regulation decisions for learners. The degree and nature of adaptiveness determine which decisions are made by learners and which are made by technological aids—thereby reshaping the self-regulated learning landscape.

1.1 The Current Study

We present empirical findings from a model-based adaptive fact-learning system that was deployed in the naturalistic context of a cognitive psychology undergraduate course. The system—called SlimStampen and described in Section 3.1—was designed to trace learners' memory strength while they study material related to the course. This is achieved by continuously adjusting a free parameter in the system that schedules within-session repetitions such that learning gains are maximized. The main objective of the current study was to answer the research question "Does the parameter that controls the adaptive nature of our fact-learning model capture individual differences related to exam performance?" Use of SlimStampen was entirely optional, and a subset of the materials that could be rehearsed with the system (which were also available to study independent of the system) appeared verbatim on the exam. This enabled us to investigate when and how students used the tool and—more important—whether the learning analytics recorded during study were predictive of exam performance. The ability to predict exam performance would suggest that the model parameter traces a relevant facet of knowledge acquisition. Given the observational nature of this study, we focused on the associative relations between the learning data and exam scores.

2. Literature Review

2.1 Sub-optimal Study Decisions in Self-Regulated Learning

Bjork and colleagues (2013) subtitled their review of the self-regulated learning literature "beliefs, techniques, and illusions" because learners' beliefs about the effectiveness of study techniques influence which techniques they use, and because of learners' susceptibility to meta-cognitive illusions that distort their beliefs. For example, the majority of learners in the studies reported by Kornell (2009) believed that massed study, in which items are repeated immediately, was more effective than spaced study, in which repetitions are spread out, even though their test scores showed the opposite to be true. Learners' propensity to mass rather than space study sessions is well documented, through both surveys (e.g., McAndrew, Morrow, Atiyeh, & Pierre, 2016) and experiments (e.g., Kornell, 2009), despite the fact that temporal spacing of repetitions is known to result in more durable memories—a benefit known as the *spacing effect* (see Dempster, 1988, for a review).

Similarly, many learners report favouring passive studying techniques, such as repeatedly reading over the materials (Karpicke, Butler, & Roediger, 2009), that are less effective than more active studying techniques, such as self-testing. If active self-testing is used, it is often as self-assessment, not as a learning tool (Kornell & Bjork, 2007; Wissman, Rawson, & Pyc, 2012). In fact, Kornell and Son (2009) report that when given both options, learners believed passive re-study to be more effective. Nevertheless, experiments have shown convincingly that active testing results in superior long-term retention relative to passive re-study—a benefit known as the *testing effect* (e.g., Karpicke & Roediger, 2008; van den Broek et al., 2016).

These benefits extend beyond the laboratory: learners that engage in more self-testing and spaced practice tend to obtain higher grades (Hartwig & Dunlosky, 2012; McAndrew et al., 2016). However, Blasiman, Dunlosky, and Rawson (2017) showed that even learners who had good intentions at the beginning of the semester—start studying early and use effective strategies—struggled to follow through, mainly studying just before the exam using ineffective methods. Actual study behaviour is often driven by impending deadlines (Hartwig & Dunlosky, 2012; Kornell & Bjork, 2007), not by optimal planning. Consequently, literature-based recommendations typically aim to make both teachers and learners aware of the meta-cognitive pitfalls and

ISSN 1929-7750 (online).

erroneous beliefs that lead to sub-optimal study decisions (Dunlosky & Rawson, 2015; Putnam, Sungkhasettee, & Roediger, 2016; Bjork et al., 2013).

2.2 Learning Analytics and Adaptive Systems

Increasingly, there are technological solutions for guiding learners toward more effective learning strategies. Digital learning systems allow learners to track their progress and make study decisions informed by data[1]. For example, Duolingo, a language-learning tool, shows learners an overview of their mastery of each lesson in a dashboard (Figure 1(a) in Settles & Meeder, 2016). Rosetta Stone, another language-learning tool, has a similar dashboard and includes a suggested next study activity (Ridgeway, Mozer, & Bowles, 2017).

Adaptive learning systems take this a step further by assuming control over some study choices that might otherwise be made by learners. Using an internal model of the learner that is informed by the learner's performance, such systems can adapt the learning experience in real time (VanLehn, 2006). The adaptation can include changing the difficulty of the problems presented to the learner, changing the amount of feedback that the learner receives, and changing the scheduling of repetitions within and between learning sessions. For example, the mathematics tutors from Carnegie Learning (Ritter, Anderson, Koedinger, & Corbett, 2007) and Math Garden (Klinkenberg, Straatemeier, & Van der Maas, 2011) use an internal computational model to track learners' skill development and adapt the choice of problems accordingly. Similarly, the ARTS system for memorizing facts (Mettler, Massey, & Kellman, 2016) uses learners' performance to change the scheduling of item repetitions during a learning session. What type and degree of adaptivity are most beneficial is an empirical question and depends on whether the adaptive system accurately traces the acquisition and forgetting of knowledge over time. If implemented well, adaptive learning systems can help students achieve more effective study behaviour by facilitating spaced repetition, active study, and other effective techniques.

Here, we report on a cognitive psychology course in which the topic of sub-optimal study decisions was also covered. Armed with this knowledge, students were explicitly advised to space their study over the duration of the course, though we did not enforce any particular schedule or technique. Rather, we focused our efforts on optimizing the learning process within a learning session once it was initiated by a student. This optimization was realized through an adaptive learning system that let students study course material using a scheduling algorithm that balances spacing and testing effects and allows for precise tracing of students' learning trajectories (described in more detail in Section 3.1). Focusing on within-session optimization made our approach an achievable application of learning theory in a typical educational setting.

3. Methods

3.1 SlimStampen: A Model-Based Adaptive Fact-Learning System

3.1.1 App

Learners could access the learning system on their computer via the course website (the university's content management system Blackboard; https://www.blackboard.com/about-us) or on their phone via an app. Learners started a session by picking a chapter to study, choosing a question type for the session (multiple-choice or open response questions), and selecting the desired duration of the session using a slider (see Figure 1a). By default, a session lasted eight minutes, but sessions could be set to last anywhere from one minute to twenty minutes. Additionally, sessions could be extended by five minutes every time the selected duration was completed.

During a learning session, two types of trials were presented. On *study trials*, both the cue (e.g., "The process through which…" in Figure 1c) and the correct answer were shown on the screen and learners progressed by pressing the "Next" button. Study trials were only used the first time a learner encountered an item—all subsequent repetitions were test trials. On *test trials*, only the cue was shown and learners had to provide the correct answer by typing (for open response questions; see Figure 1c) or by selecting from four alternatives (for multiple-choice questions). Typed responses were considered correct if they matched the expected answer, ignoring capitalization. We overlooked small typing errors: responses were still considered correct if they differed from the correct answer by a Levenshtein distance of no more than 1, 2, or 4 for answers shorter than 5 characters, between 5 and 10 characters, and longer than 10 characters, respectively. Multiple-choice questions always had a single correct response. A response was always followed by feedback (Figure 1b depicts a correct multiple-choice response and Figure 1d an incorrect open response question response). Regardless of the accuracy of the response, the correct answer was

[1] The resulting analytics can also offer instructors and educational researchers new insight into the learning process. Analytics gathered by such systems provide rich, naturalistic data on a scale previously unattainable for psychologists (Griffiths, 2015). Such data can be used to verify findings established in controlled laboratory settings (Kim, Wiseheart, & Rosenbaum, 2019) or mine for new psychological principles (Goldstone & Lupyan, 2016). For example, early work established that engagement with course materials is a good predictor of academic performance (Taraban, Maki, & Rynearson, 1999; Taraban, Rynearson, & Stalcup, 2001). The positive relationship between learning outcomes and participation in online learning environments has since been replicated in a range of studies (see Section 2.1 in Boulton, Kent, & Williams, 2018, for a summary).

ISSN 1929-7750 (online).

always shown in green. Following correct answers, the next trial commenced after one second. For incorrect answers, feedback remained on the screen until the learner pressed the "Next" button at the bottom of the screen (see Figures 1b and 1d), making the feedback similar to the study trials. The exact sequence of study and test trials was determined by the scheduling algorithm outlined below. Response time, defined as the time between the onset of the cue and the first response, and accuracy (ignoring capitalization) were recorded on every trial.

3.1.2 Scheduling Algorithm

The scheduling of items was determined by a computational cognitive model that is an extension of the adaptive item-learning model by Pavlik and Anderson (2005; 2008) and has been tested in laboratory settings (van Rijn, van Maanen, & van Woudenberg, 2009; Sense, Behrens, Meijer, & van Rijn, 2016; Sense, Meijer, & van Rijn, 2018) but has not been deployed in a university course before. This model capitalizes on the spacing effect (see Dempster, 1988, for a review) within a single session by scheduling repetitions as far apart as possible, while also optimizing for the testing effect (see van den Broek et al., 2016, for a review) by repeating items soon enough that most responses are correct.

The model represents every encountered item by a unique *memory chunk*, based on the ACT-R theory of declarative memory (Anderson, 2007). Each chunk has an activation—a representation of the ease with which that item could be retrieved—that receives a boost whenever an item is re-encoded and that decays over time. The activation A of a chunk i at time t, given n previous encounters at $t_1, \ldots, t_n$ seconds ago, is

$$A_i(t) = \ln\left(\sum_{j=1}^{n} t_j^{-d_i(t)}\right). \tag{1}$$

When a new trial commences, the model determines the activation of all items 15 seconds in the future, and if the item with the lowest activation has an activation value below a *retrieval threshold*, that item will be scheduled for presentation. If all predicted activations are above the retrieval threshold, the model will introduce a new item. By selecting items on the basis of their activation, items will be repeated with as much spacing as possible, while ensuring that, theoretically, a correct response can still be given.

The decay of the activation (parameter d in Equation (1)) varies between items to account for differences in difficulty. The higher this decay, the faster a chunk's activation will decrease, causing it to be repeated sooner than an item with a lower decay. The decay d of a chunk i at time t depends on the activation of the chunk at the time of its previous encounter, as well as an offset that we label the *rate of forgetting*, α_i:

$$d_i(t) = c * e^{A_i(t_{n-1})} + \alpha_i. \tag{2}$$

The model assumes that each item has a standard initial rate of forgetting when it is first presented. However, this value is updated during learning. At each presentation, the model calculates an expected response time, $\mathbb{E}(RT)$, based on the activation at the time of the presentation (e^{-A_i}, based on Equation (5) in Anderson, Bothell, Lebiere, & Matessa, 1998) and an estimated reading time of the prompt (based on the number of characters in the prompt; see Section 2.2.1 in Nijboer, 2011, for details).

The accuracy of the response and the mismatch between expected and observed response time are used to update the value of the rate-of-forgetting parameter. Using both accuracy and response time to update the model allows for adjustment of the parameter estimate after *any* response, not just after an incorrect response. A correct but slower-than-expected response signals that the memory trace has decayed further than assumed, meaning that the item's true rate of forgetting is higher than the current estimate. That is, when a learner arrives at the right answer but takes longer than anticipated, they likely struggled to recall the information. Conversely, an incorrect or missing response suggests that the activation of the item's memory trace actually dropped below the retrieval threshold, which means that the true rate of forgetting should be higher because this item's activation was expected to be above the threshold (which was fixed at ACT-R's default value). An unexpectedly fast correct response, on the other hand, indicates a stronger-than-expected memory trace and implies that the estimated rate of forgetting should be adjusted downward.

Since interruption or distraction can cause disproportionately large response times, observed response times are capped before their mismatch with the expected response time is calculated. The capped response time (RT') is limited to 1.5 times the estimated time it takes to read the prompt and retrieve an item with an activation at threshold value, representing the slowest possible correct retrieval. Incorrect responses are similarly coded as having this maximum response time.

To update the rate of forgetting after each trial, the model uses a binary search in a small window around the previous value to identify the rate of forgetting that minimizes the mismatch between $\mathbb{E}(RT)$ and RT'. Both the windowed binary search and the response time cap restrict the impact a single extreme response time can have on the estimated rate of forgetting.

ISSN 1929-7750 (online).

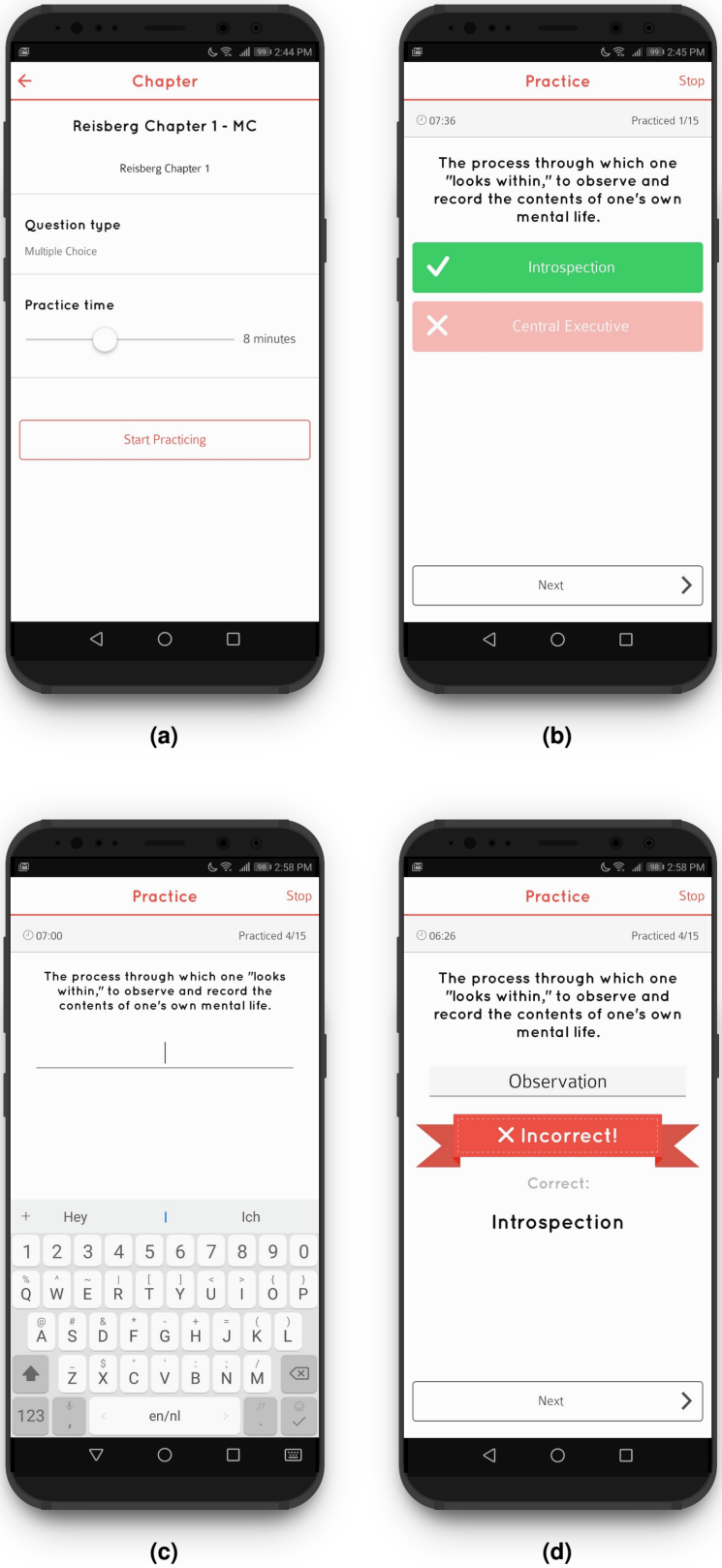

Figure 1. Screenshots of the smartphone app, showing examples of **(a)** the session configuration screen, **(b)** corrective feedback on a multiple-choice question, **(c)** an open response question, and **(d)** corrective feedback on an open response question. The interface was similar when the system was used on a computer instead of on a phone.

ISSN 1929-7750 (online).

Table 1. Summary statistics about the SlimStampen data from both cohorts. Q25 = first quartile; Q75 = third quartile.

Cohort	# Students at exam	# Students that used system	Total # trials	Average # sessions median (Q25, Q75)	Average session length (minutes) median (Q25, Q75)	Average response accuracy (%) median (Q25, Q75)
2017	147	138	262,978	29 (15, 49)	7.81 (4.08, 13.14)	91.4 (87.6, 94.0)
2018	138	129	205,639	24 (11, 46)	7.80 (2.98, 10.94)	88.9 (84.8, 93.5)
Combined	285	267	468,617	27 (13, 48)	7.80 (3.71, 12.36)	89.4 (86.2, 92.9)

3.2 Sample

SlimStampen was made available as an optional study method for the cognitive psychology course—a second-year elective undergraduate course at the University of Groningen, Netherlands—in two consecutive years. For each chapter of the textbook (*Cognitive Psychology* by Reisberg, 6th edition) covered in the course, students could learn the associated glossary items (see Figure 1). The course instructors, textbook, and glossary items were the same in both cohorts. Written consent for analyzing learning data was requested during the exam. Across both cohorts, 54% of students that took the exam gave consent. Ethical approval for using these educational data for research purposes was obtained from the Ethics Committee Psychology (ID: 18072-O). No additional demographic information was collected.

3.3 Available Data

For both cohorts, the course started in early September, and the exam was scheduled for mid-November. In both years, there were two lectures per week for seven weeks (without mandatory attendance), and students were made aware of the tool in the third and second (in 2017 and 2018, respectively) week of the course by means of a 45-minute lecture on the underlying principles of SlimStampen and related systems. Individual items were grouped into the same chapters as used in the textbook. In 2017, material was made available for rehearsal through the system after the associated chapter had been discussed in class. In 2018, all material was made available at once in the second week of the course. Table 1 summarizes the available data.

3.3.1 Study Data

Learners could install a mobile app to rehearse material using their phones (see Figure 1) or access the same tool through the Blackboard environment in a web browser. We had no control over, nor any information about, the context in which students interacted with the system. After a student chose a chapter and set the desired study duration, a study session commenced. Learning data were stored on a secure server, including a code identifying the user, chapter, and item ID (including the prompt); a timestamp; the question type (open response/multiple choice); the response time and accuracy; and three components of SlimStampen associated with the current item—the estimated activation, the estimated response time, and the current value of the rate of forgetting.

3.3.2 Exam Data

Grades are available for 285 students, including 21 students that did not use the tool and 3 students that did not take the exam but used the system. Grades range from 0 to 10 (highest)—with a 5.5 as the minimum passing grade. The exam data also contain the points awarded for individual items on the exam, including 10 items that are identical to the glossary items available through SlimStampen.

4. Results

All analyses were conducted separately on the two cohorts and on the combined data. The results reported here are based on the combined data, unless conclusions did not hold in either cohort separately. All data and analyses, including analyses by cohort, are provided in the online supplement (https://doi.org/10.17605/OSF.IO/E28NW).

4.1 Usage of the System

Table 1 provides an overview of students' usage of the system by cohort. A more comprehensive description of system usage can be found in the appendix (https://osf.io/y7efq/). Most students exhibited strong "cramming" behaviour, with much higher SlimStampen usage in the days leading up to the exam: in both cohorts, we observed a sharp increase in activity starting around 10 days before the exam and peaking on the last day. As the exam neared, usage intensified throughout the day and extended into the night.

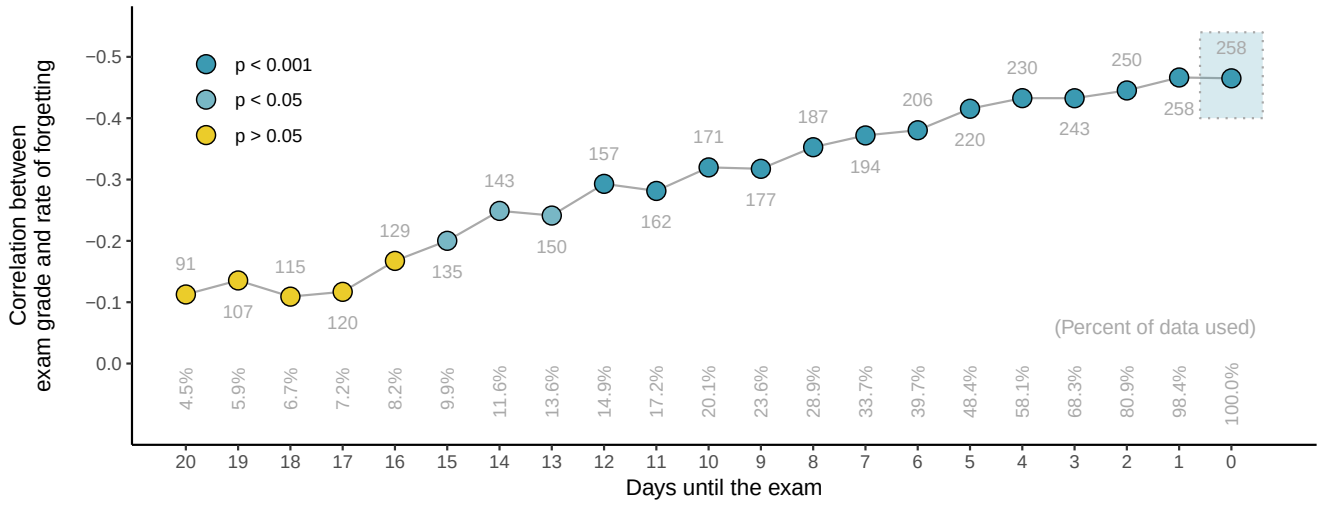

Figure 2. The correlation between exam grade and learner-specific rates of forgetting based on data available up to 20 days before the exam. The colour of the data points indicates the significance level of the correlation; the numbers alternating above/below each point indicate the number of participants for whom rates of forgetting could be computed; the percentage of data available at each moment is shown above each day. The correlation based on the complete data set is highlighted by the shaded blue square.

4.2 Exam Performance

In both cohorts, students that used SlimStampen (92.6% of students) obtained higher grades than those that did not—averaging 6.91 compared to 5.86, respectively—but a direct comparison of these groups is problematic due to selection effects and the imbalanced distribution. For this reason, the remaining analyses focus on students who used the system.

4.2.1 Amount of Practice

Among users of the system, the number of study trials completed was positively correlated with the final grade[2] ($r = 0.20$, $t(283) = 3.09$, $p = 0.002$); on average, completing more trials was associated with higher grades on the exam. The number of unique days on which a learner engaged with the tool—an index of spaced practice—was also positively correlated with exam grades ($r = 0.27$, $t(283) = 4.81$, $p < 0.001$). As one would expect, the two measures of engagement were strongly and positively correlated ($r = 0.75$, $t(283) = 18.78$, $p < 0.001$).

4.2.2 Studied versus Non-studied Items

Beyond overall exam performance, which was based on students' score on items studied with SlimStampen, and items not covered in SlimStampen, we also investigated performance on individual items. We observed a large difference between exam questions that learners had used the system to study and questions that they had not[3]: students' accuracy was 83.7% on studied items but only 53.6% on unstudied items. A mixed-effects logistic regression (with random intercepts for learners and items) confirmed that encountering an item during SlimStampen rehearsal considerably increased the chances of a correct answer on the exam ($b_{\text{studied/not studied}} = 1.70$, SE $= 0.18$, $z = 9.06$, $p < 0.001$).

4.2.3 Rates of Forgetting and Grades

Zooming in on the studied items, we used the rate of forgetting estimated during study as a predictor of exam performance. The rate of forgetting, which was initially estimated for each learner–item combination, was converted into a learner-specific rate of forgetting by averaging over all studied items. There was a strong negative correlation between a learner's rate of forgetting and their grade ($r = -0.47$, $t(256) = -8.41$, $p < 0.001$; shaded square in Figure 2), which explained $R^2 = -0.47^2 = 22.1\%$ of the variance in exam performance. The negative correlation shows that a learner who was estimated to forget material more slowly also tended to obtain higher grades.

[2]The reported numbers are the result of considering only students that completed at least 250 trials to ensure that the correlation is not artificially inflated by a cluster of students with few observations and low grades. If all data are considered (and assuming zero trials for students that did not use the system), the correlation is slightly higher ($r = 0.33$, $t(283) = 5.93$, $p < 0.001$).

[3]We consider an item studied if a learner has encountered any item—whether open response or multiple choice—with the cue used on the exam during SlimStampen rehearsal.

In practice, a possible relationship between someone's rate of forgetting and eventual exam performance would be most useful if it could be detected ahead of time rather than on the day of the exam—when it is too late to potentially help struggling students, for example. Figure 2 shows the development of the correlation as increasingly more days of observations were included: from left to right, the correlation became stronger over the 20 days leading up to the exam. This suggests that the correlation emerges well before the end of the course even though its computation necessarily relies on less data relative to the full dataset available just before the exam. Twenty days before the exam, only 4.5% of the data were available, yielding rates of forgetting for 91 learners across the two cohorts. Two weeks before the exam, when 11.6% of data were available, the 143 rates of forgetting were already significantly correlated with exam grades[4] ($r = -0.25$, $t(141) = -3.05$, $p = 0.003$).

This pattern could be driven by additional learners that start at the last minute and demonstrate poor learning performance and poor grades. However, correlations computed for the subset of 150 learners for whom rates of forgetting could be estimated 13 days before the exam increased at the same rate as in Figure 2 (see the last section at https://osf.io/7cjxn/). Therefore, more observations resulted in better estimates that were progressively more strongly related to exam performance. Taken together, these data suggest that learners' estimated rates of forgetting captured differences in exam performance—up to two weeks ahead of the exam.

4.2.4 Predicting Performance on Individual Exam Questions

The results reported so far confirm that the expected patterns emerged in the aggregate: a learner's average rate of forgetting was strongly related to their average performance on the exam (i.e., grade). Next, we tested whether item-specific rates of forgetting had explanatory power beyond two other, more commonly available measures of performance during study: the number of repetitions and the proportion of correct responses. Like the rate of forgetting, these measures can be computed for each studied item. We used a mixed-effects logistic regression model with accuracy on each exam question as the dependent variable and the three measures of interest—rate of forgetting, number of repetitions, and proportion of correct responses—along with all their interactions, as the independent variables. The model included random effects for learners and items. A step-wise backward elimination procedure was used to find the best model: starting with the full model, the term with the lowest absolute z-value was removed until the simpler model was no longer preferred on the basis of BIC and AIC (Gelman & Hill, 2006).

The best-fitting model retained only two effects[5]: the log-transformed number of repetitions ($b = 0.43$, $z = 3.56$, $p < 0.001$) and the normalized rate of forgetting ($b = -0.35$, $z = -3.99$, $p < 0.001$). Since the raw data are difficult to visualize, Figure 3 shows the model fits instead. These were generated for a range of item repetition counts (x-axis) and rates of forgetting (colour-coded). The figure makes apparent the positive association between additional repetitions and the probability of answering an exam question correctly. Additionally, the estimated rate of forgetting modulated the effect such that learner–item combinations with very low rates of forgetting have a higher chance of yielding a correct answer. The differences in rates of forgetting are especially pronounced at a low number of repetitions due to the non-linear mapping between the predictors and the predicted probability introduced by the logit function. Note that the apparent interaction pattern in Figure 3 is merely an artifact of this non-linear mapping.

4.2.5 Predicting Performance on the Exam

Finally, we present an analysis in which a range of features were used to predict performance on the exam. These features were derived from SlimStampen and captured both usage and performance statistics. We used lasso regression (Tibshirani, 1996) to predict grades using nine predictors: a student's accuracy during study, their cohort, their cumulative usage time, the number of days on which they used the system, the number of items they studied, the number of sessions they recorded, the number of trials they completed, their estimated rate of forgetting, and their median response time. Some of these predictors were correlated, especially the four predictors labelled "number of …," with correlations[6] ranging from 0.60 to 0.86. Lasso (least absolute shrinkage and selection operator) regression is a machine-learning approach to performing subset selection and regularization to increase prediction accuracy. The advantage of lasso regression is that the shrinkage term handles multicollinearity between the predictors by shrinking their coefficients (see Section 3.4 on Shrinkage Methods in Hastie, Tibshirani, & Friedman, 2009). The shrinkage is achieved by imposing a cost function on the magnitude of the coefficients themselves: the best fit is achieved by the model that minimizes the OLS with the smallest coefficients. In fact, coefficients are shrunk entirely if they do not explain sufficient variance to justify inclusion in the model. In lasso regression, predictors must be normalized to ensure that the shrinkage term affects all predictors equally. A convenient consequence of normalized predictors is that their post-shrinkage

[4] The increasingly stronger correlation between rate of forgetting and exam grade was present in both cohorts, though we found that the correlation first became significant much earlier in the 2018 cohort, at 36 days before the exam, compared to 8 days in the 2017 cohort; see https://osf.io/7cjxn/.

[5] See https://osf.io/7cjxn/ for details on the model-fitting and selection procedure, as well as plots that show the distributions of the predictors (before and after transformation). To meet model assumptions, the right-skewed numbers of repetitions were log-transformed; normalizing the rate of forgetting brings both predictors to similar scales and results in a coefficient that is more easily interpreted.

[6] See https://osf.io/vbnk3/ or the appendix at https://osf.io/y7efq/ for a correlation matrix and more details.

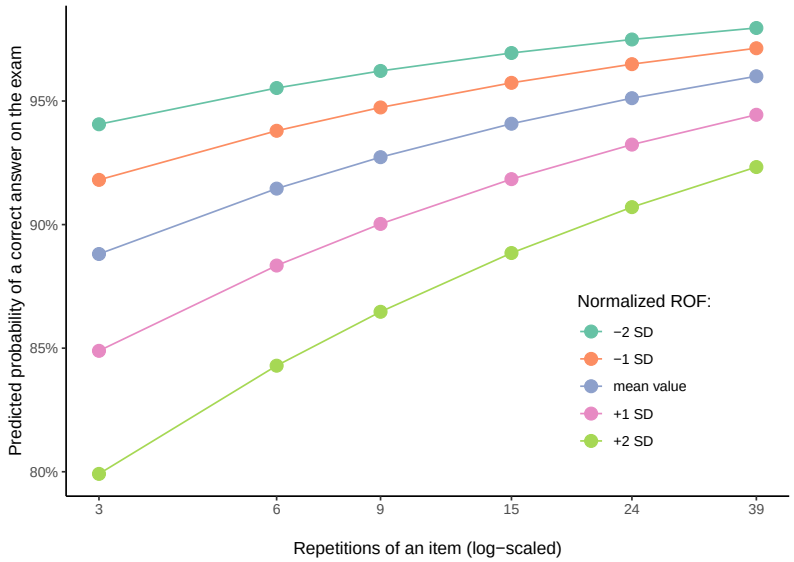

Figure 3. Predictions generated from the best-fitting mixed-effects logistic regression model fit to accuracy on the subset of exam questions that students could encounter during study—demonstrating the effects that the (log-transformed) number of repetitions and the (normalized) estimated rates of forgetting (ROF) had on answering an exam question correctly.

coefficients directly indicate their importance: since all predictors are on the same scale, the most important predictor retains the largest (absolute) coefficient.

To obtain reliable estimates for the coefficients *and* evaluate the predictive accuracy simultaneously, we implemented a 250-fold cross-validation procedure (Shmueli, 2010). On each fold, the available data were randomly split into a training set (80% of data) and a test set (20% of data). A lasso regression was fit to the training data to estimate coefficients. Subsequently, the model was used to predict the grades in the test data. Repeating this procedure 250 times yielded three relevant outcomes: (1) the inclusion rate, that is, the proportion of random data samples in which a predictor was not shrunk from the model completely; (2) a distribution of each predictor's estimated coefficients across the 250 training sets; and (3) the predictions made on each random subset. Figure 4 shows (2) ordered by (1), listed on the right along with the mean and standard deviation used for normalization. The most reliably important predictors will have high inclusion rates of relatively large (absolute) coefficients across the cross-validation folds.

Performance during study (response time and accuracy) was most important for predicting grades, with all model fits finding that better study performance increased the predicted exam grade. Most models also indicated that greater spacing of practice over time (number of unique days and sessions) was predictive of higher exam grades. Furthermore, students' estimated rate of forgetting was found to be predictive of their exam performance in 81.6% of model fits, even when accounting for the other usage and performance measures, with higher rates of forgetting being associated with lower grades.

On each of the 250 cross-validation runs, predictions were made for the 20% of data that were withheld. Figure 5 plots the actual grades against all generated predictions (see (3) above). Besides the distributions, the figure highlights the root-mean-square error, the mean absolute error, and the linear regression line that corresponds to a correlation coefficient of $r = 0.47$. Using all available information, the out-of-sample predictions from the cross-validated lasso regression model could explain about 22% (i.e., 0.47^2) of the variance in exam grades.

4.2.6 Comparing Self-Reported and Recorded Study Times

To disentangle the effects of general studiousness from potential benefits of SlimStampen as a study tool, we asked learners in the 2018 cohort to report the time they spent studying for the exam[7]. These self-reports, which were made right after the exam, focused on study behaviour during two periods: the week leading up the the exam and an average week before the final week[8]. Alongside self-reported study times, time spent studying with SlimStampen could be derived directly by summing the durations of each learner's sessions.

[7] It only occurred to us that this information would be worth collecting after the 2017 cohort's data had been collected.

[8] The exact questions were "How many hours did you spend studying (irrespective of method) for this course in the last week before the exam? (please enter with just a number, i.e., 12 if you studied 12 hours)" and "How many hours did you spend studying (irrespective of method) for this course in other weeks (on average per week)?"

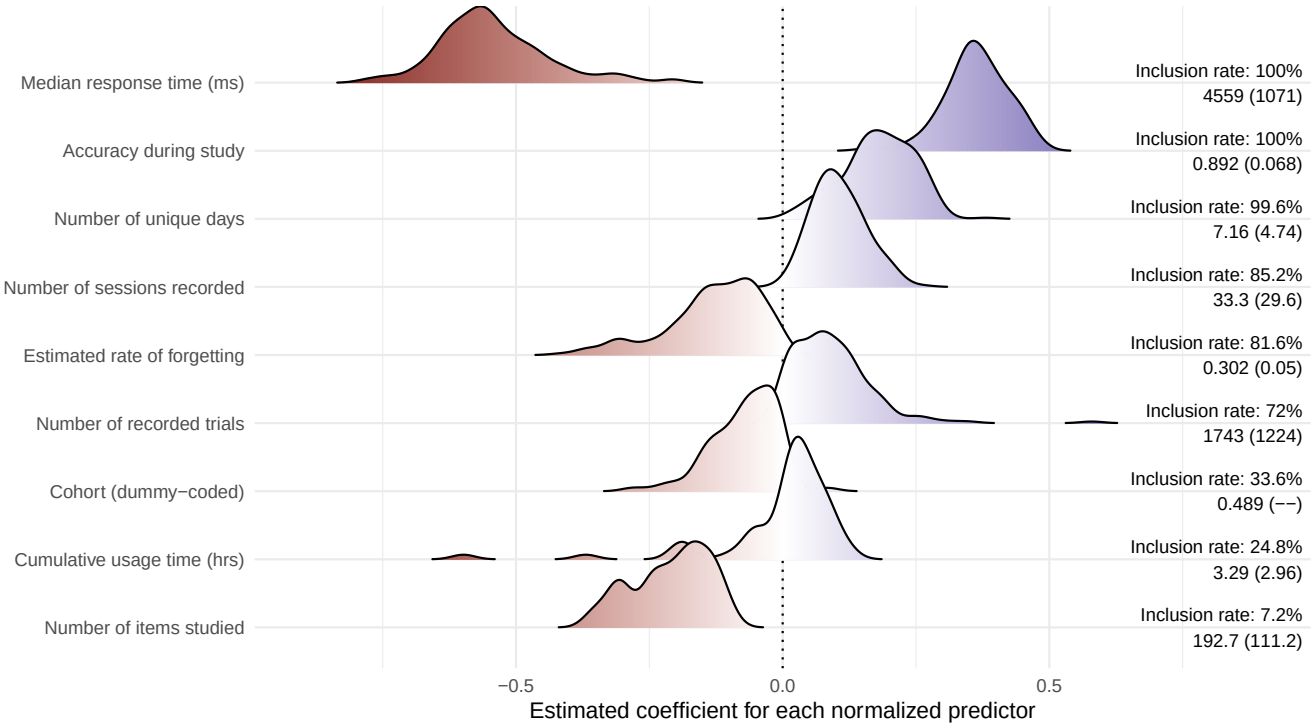

Figure 4. Kernel densities of the estimated coefficients for each normalized predictor of the cross-validated lasso model. Colour indicates the distance from zero. The right-hand side lists each predictor's inclusion rate across the cross-validation folds (see the main text for details) and its mean and standard deviation.

Here, we only report the study times from the last week before the exam because those from the preceding week showed no significant relationship with grades obtained on the exam[9]. The left panel in Figure 6 plots the two study time measures against each other. The superimposed linear regression line is very flat, and the corresponding correlation between the two measures is not significantly different from zero ($r = -0.02, t(118) = -0.18, p = 0.858$). This means that students who used SlimStampen more did not necessarily self-report studying more overall. Thus, the positive association between more SlimStampen usage and higher grades was unlikely to be a consequence of higher motivation alone.

The right panel of Figure 6 shows the relationship between study times and obtained grades. Plotting them on the same x-axis emphasizes the differences in the range of the values ($M_{\text{self-reported}} = 26.9$, $SD_{\text{self-reported}} = 14.2$; $M_{\text{recorded}} = 3.5$, $SD_{\text{recorded}} = 3.4$). Both study times showed the expected positive relationship with grades as indicated by the superimposed linear regression lines. The estimated linear regression slopes for self-reported and recorded study times are $b = 0.03$, $SE = 0.01$, $t(117) = 3.17$, $p = 0.002$, and $b = 0.11$, $SE = 0.05$, $t(117) = 2.54$, $p = 0.012$, respectively[10]. This suggests, unsurprisingly, that general studiousness led to higher exam performance. More interestingly, time spent studying with SlimStampen was time well spent, as the expected gain in grades associated with additional hours of study was 0.11 points, compared to only 0.03 points gained by an hour of unspecified study time.

5. Discussion

Digital learning environments—and adaptive learning systems in particular—can foster more effective study strategies by giving learners insight into their progress and implementing methods known to be effective, such as spaced retrieval practice. In this study we explored when and how learners used an adaptive fact-learning system that was freely available in the context of a university course. Our primary interest was whether and how the learning analytics extracted from the system were related to exam performance.

The rate-of-forgetting parameter, which controls the adaptive nature of the fact-learning system, did indeed capture

[9] See https://osf.io/kwxcm/ for two analyses that use all four variables in a multiple regression analysis. Estimated slope coefficients differ slightly, but the overall conclusion was unaffected. Online materials also include analyses with normalized predictors.

[10] Again, see https://osf.io/kwxcm/.

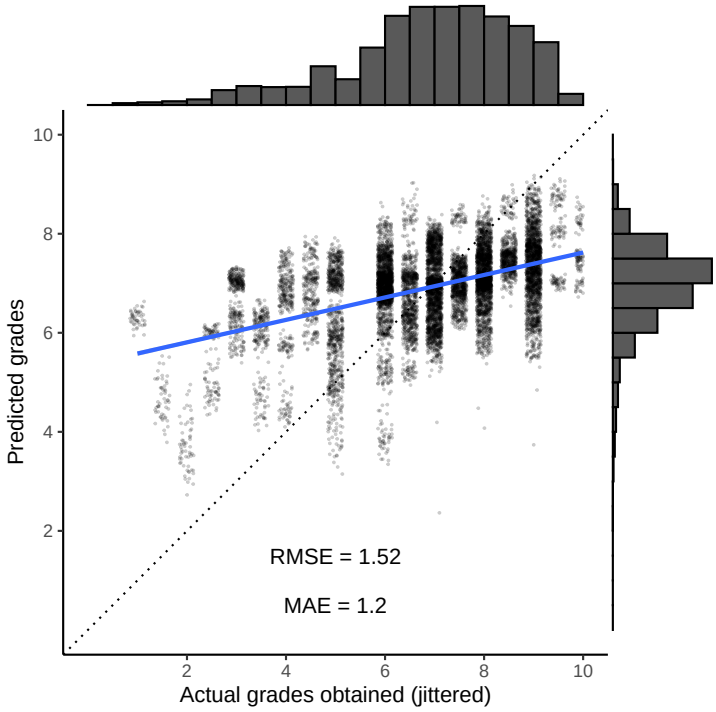

Figure 5. Predictions from the 250-fold cross-validation procedure plotted against the actual grades that were obtained (jittered slightly). Marginal histograms show each variable's distribution on the opposite axis. RMSE = root-mean-square error; MAE = mean absolute error.

individual differences related to exam performance. Students' rates of forgetting, estimated by the system during use, were correlated with exam performance up to two weeks before the exam (Figure 2), even though $< 5\%$ of the data were available at that point. Furthermore, rate-of-forgetting estimates for individual facts were predictive of learners' performance on the associated exam questions, along with the number of times these facts were repeated during study (Figure 3). A cross-validated lasso regression further corroborated the importance of the rate of forgetting as a predictor of exam performance (Figure 4). These findings can be seen as a high-face-validity extension of earlier work reporting a very high correlation ($r \approx 0.80$) between the estimated model parameters and a delayed-recall test of studied material (Sense et al., 2018).

Despite being informed about the benefits of spaced practice, most students exhibited cramming behaviour, primarily using the system in the final days before the exam. These findings are in line with prior observations (Taraban et al., 1999, 2001; Kornell, 2009; Blasiman et al., 2017) and directly oppose recommendations to start studying early (e.g., Putnam et al., 2016; Dunlosky & Rawson, 2015). Even if students intended to space their practice, this did not manifest in their usage of the system (similar to Blasiman et al., 2017). Students who spaced their practice over more days tended to obtain higher grades on the exam, suggesting that the last-minute learning displayed by most students was indeed not optimal. Consequently, there might be merit in policies that enforce spaced learning schedules through course requirements. Such schedules could be tailored to individual needs by an adaptive system to further amplify the benefits of spaced engagement with course materials.

Spending more time with the adaptive system in the week leading up to the exam was associated with a higher exam grade, as was spending time on other self-reported study activities, though to a lesser extent (Figure 6). This is consistent with other studies that looked at engagement and academic performance (Bloom, 1974; Gurung, Weidert, & Jeske, 2012; Meehan & McCallig, 2019; Boulton et al., 2018). One limitation of the sample was that we did not know what other study methods students may have used alongside the system. It is possible that the spike in activity in the days preceding the exam was caused by students verifying that they had retained the knowledge obtained through other study activities. Indeed, Wissman and colleagues (2012) report that many learners use self-testing for monitoring, rather than as a way to acquire new knowledge, which would fit this explanation. Nevertheless, our results suggest that both self-reported and recorded study times in the week before the exam were significant predictors of a learner's grade (Figure 6). Therefore, even if learners merely used the system for self-assessment, the self-testing itself might have provided tangible benefits that learners should exploit. This implication is corroborated and expanded on by the lasso regression, which suggests that the cumulative study time is less important than the

ISSN 1929-7750 (online).

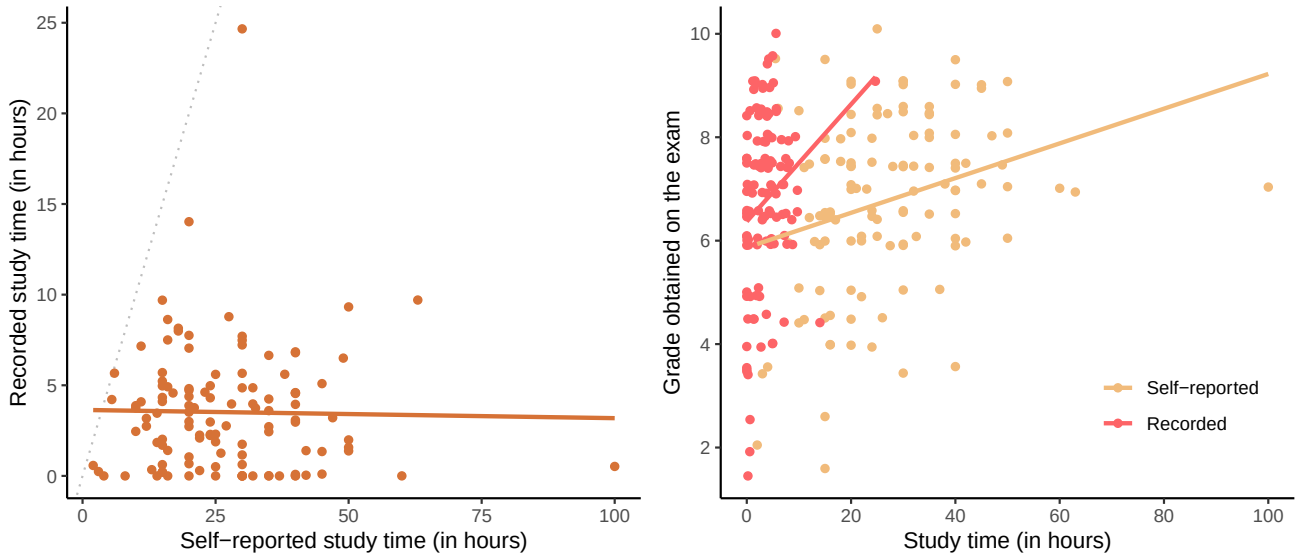

Figure 6. *Left*: Scatter plot contrasting self-reported with recorded study times with grey dotted equality line and superimposed linear regression fit. *Right*: Both study times plotted against the grades obtained on the exam (slightly jittered to avoid overlapping points); separate linear regression lines are superimposed for self-reported and recorded study times.

number of sessions and the number of days on which students used the system (Figure 4). One caveat to note is that causality cannot be inferred from the correlational findings reported by the current and cited studies.

Since using the system was optional, overall engagement with the tool could be considered a behavioural measure of motivation. Here, engagement was positively correlated with academic performance, which has also been observed for self-reported motivation (e.g., Fortier, Vallerand, & Guay, 1995). Relating self-reported motivation to a behavioural measure of motivation would be an interesting extension of the current work; the behavioural measure could be as simple as overall engagement with an optional rehearsal tool (as done here and in Taraban et al., 2001).

The adaptive learning system was made available to two consecutive cohorts of students. Given that the findings reported here held across both cohorts, these results seem highly reliable. The current situation provided an excellent real-world case of a model-based system that was first tested in the laboratory (van Rijn et al., 2009; Sense et al., 2016, 2018) and then released "into the wild." Any adaptive system that was developed on the basis of theoretical memory models would ultimately have to be deployed and tested in a comparable context.

5.1 Implications

While students in the cognitive psychology course were made aware of the psychological principles on which the adaptive learning system is based, it is difficult to imagine how this knowledge would have affected their speed or accuracy of responding. Therefore, regardless of whether students are aware of the underlying principles, the system is well suited for deployment in a wide range of (online) courses that involve a fact-learning component. Because the underlying model automatically identifies the difficulty of each fact for each learner based on the learner's responses, instructors and course designers, and even students themselves, can easily create new lessons in the system without needing to specify any difficulty information.

The adaptive fact-learning system may be extended with an analytics dashboard—similar to Duolingo and Rosetta Stone (Settles & Meeder, 2016; Ridgeway et al., 2017)—indicating a learner's progress on each lesson. Such a dashboard could also show students the items that are most difficult for them, along with their current estimated memory strength. The insight provided by such a dashboard would allow learners to make more informed decisions about which lessons to study at a given time. The system's analytics can also be useful for instructors, helping them identify students that are struggling with the material, as well as the individual items that their students find most difficult. Dynamic analytics dashboards that present progress in terms of estimated learning and forgetting rates would be especially useful as a theory-backed approach to making learning analytics digestible to students/instructors (Wise & Shaffer, 2015).

The current work focused on controlling within-session study decisions through the adaptive fact-learning system, leaving other study decisions—when to study, which chapter to study, how long to study, and whether to study with open response or multiple-choice questions—to the learner. This setup made the system easy to implement in the course, although we observed

ISSN 1929-7750 (online).

that, given the choice, students still made sub-optimal decisions about when to repeat a lesson that they had studied previously. A possible extension to the system would therefore be to use the system's internal model of a student's memory to notify the student of the optimal time to rehearse a particular lesson, thereby achieving better spacing between sessions. Alternatively, the system could suggest the lesson that would yield the largest learning gain at the moment a student decides to start a session. The feasibility of these approaches and their effect on the learning experience is a matter for future research, but we expect that further offloading study decisions to the adaptive learning system is likely to help learners study even more effectively.

6. Conclusion

We deployed a model-based adaptive fact-learning system in a university course to understand how and when students would use such a system and to test whether information gathered by the system could predict exam performance. We found that the learner-specific rate of forgetting measured by the system was predictive of learners' exam performance, as were other measures of engagement with the system. This study shows a real-world application of a model-based system that is based on psychological principles of learning. The current work suggests a range of possible extensions and modifications that might make the system more useful to instructors and more effective for students.

Acknowledgements

We would like to thank Tom Doesburg and Jasper Smit for their technical assistance and the anonymous reviewers for their constructive feedback, which has greatly improved the paper.

Data from the first cohort were presented as a poster at ICCM 2018 (see https://github.com/VanRijnLab/cogpsych-poster).

Declaration of Conflicting Interest

The authors declare no conflict of interest.

Funding

An earlier version of the adaptive learning system discussed in this manuscript is licensed to Noordhoff Publishers by the University of Groningen. This project was partially funded by these licence fees. However, the publishing house had no involvement in this study. FS was supported by InfiniteTactics LLC through the Air Force Research Laboratory at Wright-Patterson Air Force Base.

References

Anderson, J. R. (2007). *How Can the Human Mind Occur in the Physical Universe?* New York: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195324259.001.0001

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. https://doi.org/10.1006/jmla.1997.2553

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology*, *64*, 417–444. https://doi.org/10.1146/annurev-psych-113011-143823

Blasiman, R. N., Dunlosky, J., & Rawson, K. A. (2017). The what, how much, and when of study strategies: Comparing intended versus actual study behaviour. *Memory*, *25*(6), 784–792. https://doi.org/10.1080/09658211.2016.1221974

Bloom, B. S. (1974). Time and learning. *American Psychologist*, *29*(9), 682–688. https://doi.org/10.1037/h0037632

Boulton, C. A., Kent, C., & Williams, H. T. P. (2018). Virtual learning environment engagement and learning outcomes at a "bricks-and-mortar" university. *Computers & Education*, *126*, 129–142. https://doi.org/10.1016/j.compedu.2018.06.031

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. https://doi.org/10.1037/0003-066X.43.8.627

Dunlosky, J., & Rawson, K. A. (2015). Practice tests, spaced practice, and successive relearning: Tips for classroom use and for guiding students' learning. *Scholarship of Teaching and Learning in Psychology*, *1*(1), 72–78. https://doi.org/10.1037/stl0000024

Fortier, M. S., Vallerand, R. J., & Guay, F. (1995). Academic motivation and school performance: Toward a structural model. *Contemporary Educational Psychology*, *20*(3), 257–274. https://doi.org/10.1006/ceps.1995.1017

Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511790942

Goldstone, R. L., & Lupyan, G. (2016). Discovering psychological principles by mining naturally occurring data sets. *Topics in Cognitive Science*, *8*(3), 548–568. https://doi.org/10.1111/tops.12212

Griffiths, T. L. (2015). Manifesto for a new (computational) cognitive revolution. *Cognition*, *135*, 21–23. https://doi.org/10.1016/j.cognition.2014.11.026

Gurung, R. A. R., Weidert, J., & Jeske, A. (2012). Focusing on how students study. *Journal of the Scholarship of Teaching and Learning*, *10*(1), 28–35. Retrieved from `https://scholarworks.iu.edu/journals/index.php/josotl/article/view/1734`

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. https://doi.org/10.3758/s13423-011-0181-y

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer Science & Business Media. https://doi.org/10.1007/978-0-387-84858-7

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. https://doi.org/10.1080/09658210802647009

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

Kim, A. S. N., Wiseheart, M., & Rosenbaum, R. S. (2019). The spacing effect stands up to big data. *Behavior Research Methods*, *51*, 1485–1497. https://doi.org/10.3758/s13428-018-1184-7

Klinkenberg, S., Straatemeier, M., & Van der Maas, H. L. J. (2011). Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers & Education*, *57*(2), 1813–1824. https://doi.org/10.1016/j.compedu.2011.02.003

Kornell, N. (2009). Optimizing learning using flashcards: Spacing is more effective than cramming. *Applied Cognitive Psychology*, *23*(9), 1297–1317. https://doi.org/10.1002/acp.1537

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review*, *14*, 219–224. https://doi.org/10.3758/BF03194055

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

McAndrew, M., Morrow, C. S., Atiyeh, L., & Pierre, G. C. (2016). Dental student study strategies: Are self-testing and scheduling related to academic performance? *Journal of Dental Education*, *80*(5), 542–552. https://doi.org/10.1002/j.0022-0337.2016.80.5.tb06114.x

Meehan, M., & McCallig, J. (2019). Effects on learning of time spent by university students attending lectures and/or watching online videos. *Journal of Computer Assisted Learning*, *35*(2), 283–293. https://doi.org/10.1111/jcal.12329

Mettler, E., Massey, C. M., & Kellman, P. J. (2016). A comparison of adaptive and fixed schedules of practice. *Journal of Experimental Psychology: General*, *145*(7), 897–917. https://doi.org/10.1037/xge0000170

Nijboer, M. (2011). *Optimal Fact Learning: Applying Presentation Scheduling to Realistic Conditions.* University of Groningen, Groningen, Netherlands. (Unpublished master's thesis.)

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. https://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. https://doi.org/10.1037/1076-898X.14.2.101

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. https://doi.org/10.1177/1745691616645770

Ridgeway, K., Mozer, M. C., & Bowles, A. R. (2017). Forgetting of foreign-language skills: A corpus-based analysis of online tutoring software. *Cognitive Science*, *41*(4), 924–949. https://doi.org/10.1111/cogs.12385

Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. (2007). Cognitive Tutor: Applied research in mathematics education. *Psychonomic Bulletin & Review*, *14*(2), 249–255. https://doi.org/10.3758/BF03194060

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. https://doi.org/10.1111/tops.12183

Sense, F., Meijer, R. R., & van Rijn, H. (2018). Exploration of the rate of forgetting as a domain-specific individual differences measure. *Frontiers in Education*, *3*(112). https://doi.org/10.3389/feduc.2018.00112

Settles, B., & Meeder, B. (2016). A trainable spaced repetition model for language learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* 7–12 August 2016, Berlin, Germany (pp. 1848–1858). Association for Computational Linguistics. https://doi.org/10.18653/v1/p16-1174

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310. https://doi.org/10.2139/ssrn.1351252

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. https://doi.org/10.3758/BF03207718

Taraban, R., Rynearson, K., & Stalcup, K. A. (2001). Time as a variable in learning on the World-Wide Web. *Behavior Research Methods*, *33*(2), 217–225. https://doi.org/10.3758/bf03195368

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.1467-9868.2011.00771.x

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*(2), 52–66. https://doi.org/10.1016/j.tine.2016.05.001

VanLehn, K. (2006). The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education*, *16*, 227–265. Retrieved from `https://dl.acm.org/doi/10.5555/1435351.1435353`

van Rijn, H., van Maanen, L., & van Woudenberg, M. (2009). Passing the test: Improving learning gains by balancing spacing and testing effects. In A. Hoses, D. Peebles, & R. Cooper (Eds.), *Proceedings of the Ninth International Conference on Cognitive Modeling,* 24–26 July 2009, Manchester, UK (pp. 110–115). Retrieved from `https://iccm-conference.neocities.org/2009/proceedings/cd/papers/200/paper200.pdf`

Wise, A. F., & Shaffer, D. W. (2015). Why theory matters more than ever in the age of big data. *Journal of Learning Analytics*, *2*(2), 5–13. https://doi.org/10.18608/jla.2015.22.2

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