Cognitive Science 41 (2017) 924–949

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12385

# Forgetting of Foreign-Language Skills: A Corpus-Based Analysis of Online Tutoring Software

Karl Ridgeway,[a] Michael C. Mozer,[b] Anita R. Bowles[c]

[a]Department of Computer Science, University of Colorado [b]Department of Computer Science and Institute of Cognitive Science, University of Colorado [c]Rosetta Stone

Received 24 November 2015; received in revised form 28 February 2016; accepted 9 March 2016

## Abstract

We explore the nature of forgetting in a corpus of 125,000 students learning Spanish using the Rosetta Stone foreign-language instruction software across 48 lessons. Students are tested on a lesson after its initial study and are then retested after a variable time lag. We observe forgetting consistent with power function decay at a rate that varies across lessons but not across students. We find that lessons which are better learned initially are forgotten more slowly, a correlation which likely reflects a latent cause such as the quality or difficulty of the lesson. We obtain improved predictive accuracy of the forgetting model by augmenting it with features that encode characteristics of a student's initial study of the lesson and the activities the student engaged in between the initial and delayed tests. The augmented model can predict 23.9% of the variance in an individual's score on the delayed test. We analyze which features best explain individual performance.

Keywords: Forgetting; Big data; Corpus analysis; Computational modeling; Second language learning

## 1. Introduction

Psychologists have studied forgetting—the durability of memory over time—for over 130 years. Beginning with experiments that Ebbinghaus (1885/1964) conducted on himself, traditional controlled studies have involved learning some material—typically paired associates or facts—to criterion, and then probing residual memory strength after varying lags. Memory strength might be measured by recognition or recall tests, or—as Ebbinghaus did—the time saved when relearning the material. Observations of memory strength over time are used to fit a retention or forgetting function. This function shows a rapid,

Correspondence should be sent to Karl Ridgeway, Department of Computer Science, University of Colorado, 3535 19th St., Boulder, CO 80304. E-mail: karl.ridgeway@colorado.edu

monotonic decline shortly after initial study, but the curve's slope decreases over time and the curve levels off.

From Ebbinghaus forward, there has been intense interest in determining the quantitative relationship between the retention interval and memory strength. In a tour de force, Rubin and Wenzel (1996) reanalyzed 210 published data sets by fitting each to 105 different two-parameter functions. Across the data sets, four functions appeared to fit about equally well, all of which are based on a logarithmic (or logarithm-like) scale of time. One of these is a power function, which is both elegant and has a long theoretical tradition (Wixted & Carpenter, 2007; Wixted & Ebbesen, 1991, 1997):

$$\hat{y} = at^{b}; \quad (1)$$

in which $\hat{y}$ is a prediction of memory strength (e.g., recall accuracy), t is the lag between study and test, $a$ is a constant that represents strength of initial learning, and $b$ is a decay constant where larger $b$ corresponds to faster memory decay. Wickelgren (1974) proposed a three-parameter version of the power function,

$$\hat{y} = a(1 + ct)^{b}; \quad (2)$$

which has the advantage over Eq. 1 that $\hat{y}$ is defined at $t = 0$ and that $\hat{y}$ is constrained to lie in [0,1], allowing $\hat{y}$ to be interpreted as a probability correct or accuracy measure.

Forgetting has most often been studied in the lab in highly controlled conditions. Subjects are presented with novel materials, for example, nonsense syllables or arbitrary word-number associations or independent facts. Initial training is either designed to achieve an initial performance criterion or is undergone for a fixed number of trials or duration. Retention intervals are typically on the order of minutes to days. Re-exposure to the materials during the retention interval is either prevented or is highly unlikely due to the obscurity of the materials. In contrast, Harry Bahrick and colleagues have made heroic strides toward studying memory in naturalistic settings over retention intervals of up to 50 years, including such domains as: names and faces from high school (Bahrick, Bahrick, & Wittinger, 1975), the spatial layout of a city (Bahrick, 1983), various facets of Spanish as a foreign language (including grammar, idioms) (Bahrick, 1984), and algebra and geometry content (Bahrick & Hall, 1991). These studies show a rapid decline in memory and skill strength over the first 5 years, but are often characterized by a long period (35 years) of relative stabilization in a "permastore" (Bahrick, 1984) followed by further decline. However, given noise in behavioral observations, it is difficult to rule out continuous decay with an increasingly shallow slope, for example, power function (Wixted, 2004). Even highly motivated learners seem to show significant forgetting over long periods of time: Medical students forget roughly 25%–35% of basic science knowledge after 1 year, more than 50% by the next year (Custers, 2010), and 80%–85% after 25 years (Custers & ten Cate, 2011).

The emphasis of almost all studies of forgetting is on how memory strength varies with time and with characteristics of initial learning. Only occasionally are other

covariates considered, and then, only one or two at a time. For example, Bahrick and Hall (1991) examined retention of algebra knowledge contingent on a student's top level of math achievement.

The advent of modern electronic methods of education has created opportunities to analyze memory at scale. Large online educational programs such as Rosetta Stone Software, Khan Academy, and massively open online courses like Coursera and edX are capable of recording every interaction with a student at the level of keystrokes and mouse clicks. With such data, is it possible to examine memory in naturalistic learning settings with genuinely interested learners, and to explore the effects of confounds and interac-tions that psychologists have traditionally avoided in laboratory studies? Beyond using electronic-education tools to better understand memory, it should also be possible to apply our best memory models to enhance the course experience. For example, the tools could recommend study of the material predicted to be most fragile or the material whose study will obtain the greatest predicted learning gains.

Our research leverages data from a massively scaled on-line language learning application with 125,000 users studying subsets of 48 lessons. In this corpus, there is no clear notion of a retention interval between initial learning and delayed testing because the traditional definition of a retention interval is that students avoid all contact with the mate-rial during the interval. In our corpus, students continue to be exposed to related material. Instead of learning isolated facts that are easily distinguished from intervening activities, students are studying a series of interrelated and interdependent lessons. The lessons incorporate many varieties of knowledge, including vocabulary, syntax, morphology, inflections and derivations, phonetics, and phonology. Despite these confounds, we also have indicators to provide some information about the students' intervening activities, and can use these indicators—and other student-specific information—to model perfor-mance on a delayed test. With many lessons and many students, we can ask questions about how lessons differ from one another and how students differ from one another.

## 2. Background on the software

Rosetta Stone Ltd. develops technology-driven language and literacy training programs for use by schools, businesses, and government organizations. Its interactive software for foreign-language learning covers over 30 languages, from Arabic to Vietnamese. Each course is composed of up to five language levels, which are designed to be taken in ser-ies. Each successive level builds on material learned in the previous level. Each level is divided into 16 lessons. Lessons also have cumulative content and are typically studied in series. A lesson is composed of a set of primitive activities. The essential content of the lesson is introduced in an activity labeled as the core. Depending on student preferences, students may engage in various specialty activities that cover similar content to that intro-duced in the core activity, but focus on particular skills such as vocabulary, pronuncia-tion, grammar, and reading.

Between activities, students are taken to a home screen which displays a dashboard indicating the completion status of various activities within the current lesson. The home screen includes a recommendation for what to do next, for example, to begin the next les-son in the curriculum, to review an old lesson, or to schedule a live coaching session. From the home screen, students may navigate to any lesson and any activity in the curriculum.

Each lesson includes an activity that serves as a review test. No new material is presented, but students are evaluated on content from the lesson's core activity. Students taking the test receive a score indicating their mastery of the lesson. Because these scores are the basis for our investigation, we provide some details concerning the review test.

## 2.1. Review test

The review test requires students to respond to a set of challenges which vary by their prompting media—text, audio, or an image—and by the mode of interaction for responses—clicking an image or text, speaking a phrase out loud, or typing a free-response answer. For example, students may be prompted with an audio clip of a sen-tence spoken in the foreign language—for example, "The woman is running"—and may be required to select from among four images depicting various scenes. Fig. 1 shows three different combinations of challenge responses. In the left example, students are prompted to select a picture. In the center example, students are prompted to select from a number of text/audio options. In the right example, students must select from a set of text phrases to fill in the blanks in a sentence.

After a response is selected, the software provides feedback indicating whether the response was correct or incorrect. Fig. 2 shows an example interaction with a challenge on a review test. In this case, the student first makes an incorrect initial choice, and then selects the correct response. The score for the review test is based on students' initial responses, but because the activity is designed not just to evaluate but also to provide additional learning opportunities, each challenge is repeated until students respond cor-rectly. Students are not allowed to skip challenges in review tests.


Fig. 1. Three examples of different kinds of challenges, drawn from the Japanese product. In the first frame, the student listens to a phrase or sentence spoken in Japanese and responds by selecting the corresponding image. In the second frame, the student reads and/or listens to three possible descriptions of the image and responds by selecting the correct description. In the final frame, the student must complete a sentence in Japanese by selecting from a sequence of multiple-choice text options.

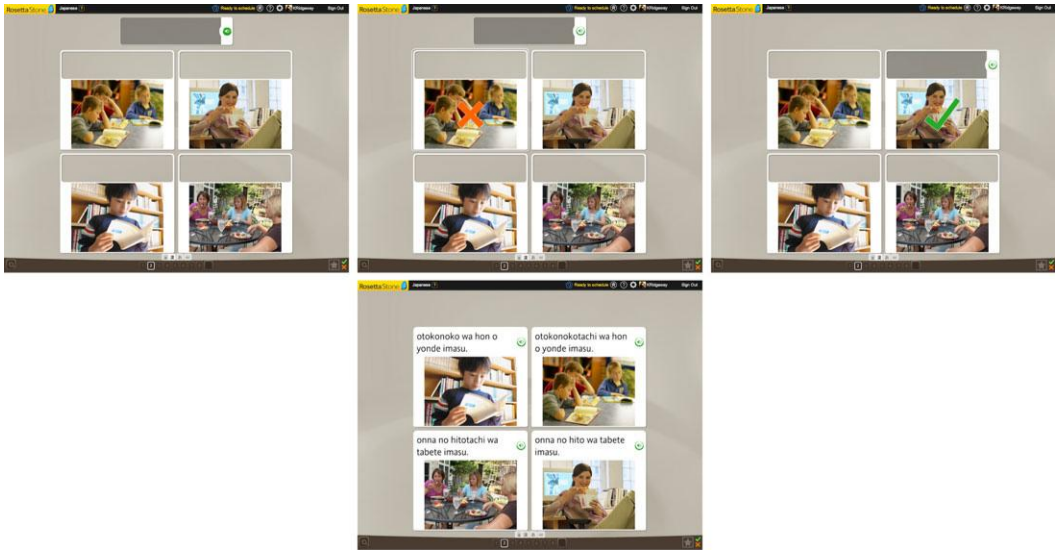


Fig. 2. An example sequence of interactions a student can have on the review test.

As Fig. 1 illustrates, each display in the review test has multiple panels. This design allows for multiple challenges to be created using the same display. Because each challenge can eliminate a response alternative, responses to later challenges in a sequence become increasingly constrained. Typically, panels in a display are rearranged following each challenge, and a display will have between 2 and 8 challenges associated with it. A typical review test consists of 8–10 displays, leading to on the order of 25–50 responses per activity. Since the challenges typically involve a prompt and selection from a set of responses, we consider the challenges to be recognition (as opposed to recall) tests.

Students may quit part way through a review test, but—in contrast to other activity types—students may not skip challenges or go back to previous challenges to amend responses and improve scores. Like the other types of activities, the only aspect of a review test that changes on consecutive attempts is the randomization of the order of pan-els within displays. These properties, along with the fact that only the initial response to a challenge can be scored as correct, make review test scores a sensible assessment of language skills.

Review tests are recommended to students at predetermined points in a lesson's curriculum. Additionally, review tests from previous lessons are periodically recommended in order to mitigate forgetting. The algorithm for selecting old lessons to revisit, called the Adaptive Recall function (Keim, 2009), suggests review of a lesson 14 days after the initial review. Following the second review, subsequent suggestions are temporally distributed according to an expanding spacing schedule (Kang, Lindsey, Mozer, & Pash-ler, 2014; Landauer & Bjork, 1978), contingent on the student's performance. Students can ignore the recommendation, and also they can choose to repeat a review test at any point in the curriculum.

Because students often perform a review test multiple times—either due to prompting from the software or on their own initiative—the difference in scores between successive attempts can be used as a measure of a lesson's retention. We use this measure to investigate factors influencing forgetting. However, we note that the measure is not a pure measure of forgetting for three reasons. First, review tests are designed to instruct as well as evaluate. Even material designed primarily as an assessment can support knowledge retention (Bjork, 1994). Thus, knowledge may be strengthened as a result of the activity. Second, the specific challenges used in the review activity are repeated each time the activity is performed (possibly in a different randomized order). Thus, the review score reflects both mastery of the core lesson skills as well as episodic memory from previous attempts on the specific challenge examples. Third, students typically have significant contact with course material between the two review tests—practicing new lessons and other activities in the tested lesson—in contrast to typical memory studies that control for exposure between tests.

## 3. Data set

The software can be run either as a web-client online or as a stand-alone application. Activity logs are available only for students using the online software; these logs are condensed and stored on servers. Our investigation utilized anonymized activity logs from insti-tutional usage. Some institutions mandate the use of the software; others make the use optional. We have no means of determining the usage policy governing individual students.

Our data set is drawn from the online Latin American Spanish course, levels 1–3, with the TOTALe™ software suite. This software suite, when it was launched in 2007, originally included only a self-study application, which is the core pedagogical activity in the suite. It was later expanded to include access to videoconferencing with a language coach to reinforce content from the Course. The data used in this study were collected between January 2008 and March 2014, and therefore correspond to various versions of the software depending on the date collected. All data points are anonymized, where each student is identified only by a unique integer value.

Our data set consists of 46.3 million observations of anonymized students performing activities, of which 6.1 million were review activities. These activities were distributed over a total of 48 lessons—16 in each of 3 levels. In the database, each review activity is associated with a total score, representative of the aggregate performance over the whole activity.

Fig. 3 illustrates a typical student's path through a lesson. The lesson is associated with a set of activities that are performed at various times. Students begin a lesson by completing the core activity before moving on to other activities (e.g., grammar) and eventually taking their initial review test. Following some lag, they may take the review test for a second time. Additional activities may be performed between the two tests. A student may choose to take the review test again after the delayed review. However, we removed these attempts from our data set and only consider the initial and first delayed reviews.

We selected from the complete database all students and all lessons for which the student completed two or more review activities. This subset consists of 545,629 student-lessons (i.e., instances where a student completed at least two review tests for a lesson) from 125,112 unique students. Fig. 4 shows the count of students by lesson, arranged in the order in which they appear in the curriculum. The vertical axis is log-scaled to better represent the dynamic range. The most populated lesson has over 86,000 students; the least populated lesson has fewer than 1,000. The sawtooth pattern is due to the fact that students tend to drop out within a level of a course, and new students join at the begin-ning of each of the three levels.

Fig. 5 shows the distribution of lags—the time between the initial and delayed review tests. This bimodal distribution can be attributed to two features of the software. First, the course allows students the freedom to repeat activities at will. So, after students com-plete a review, they are free to simply repeat it immediately. Many do so to raise their scores. Second, the mode of the distribution at roughly 14 days is due to the software design, which automatically schedules a repeat of the review activity 2 weeks after the initial review attempt. Although students have the ability to opt out of the scheduled review, this default suggestion is typically followed.

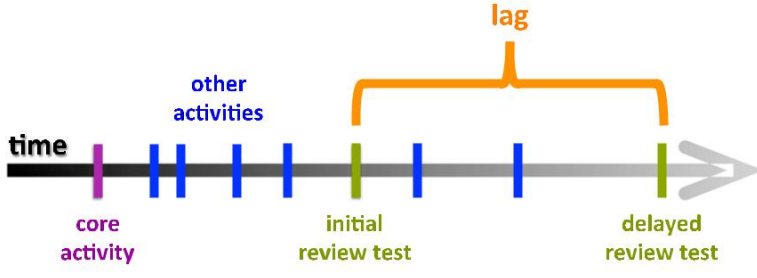


Fig. 3. A typical sequence of student activities within a lesson.

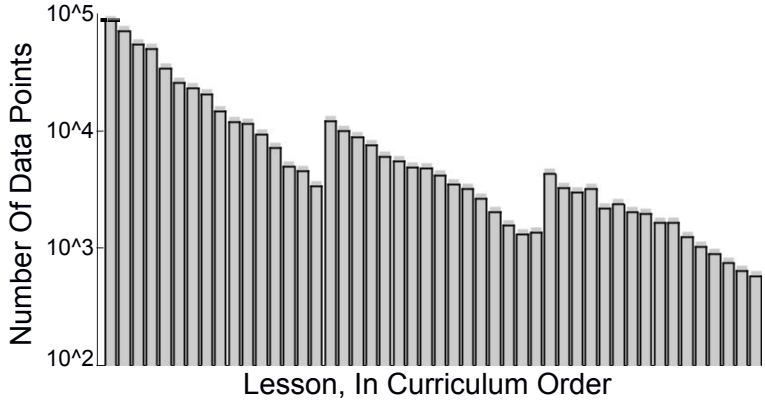


Fig. 4. Number of students per lesson for the three levels of our data set and all 16 lessons within a level. The ordinate is scaled logarithmically to represent the full dynamic range.

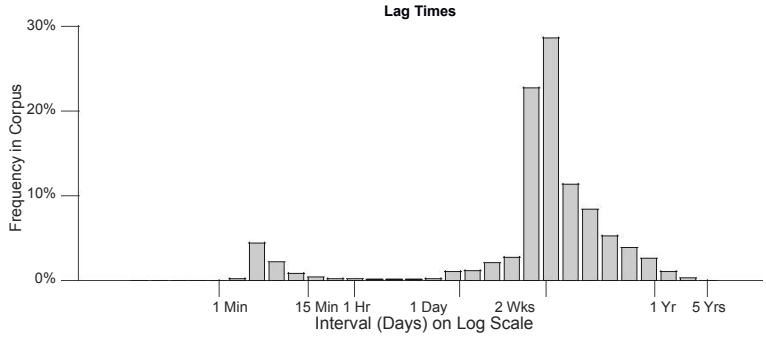


Fig. 5. The distribution of lags between the initial and delayed review test in the data set. The abscissa is displayed on a log scale to capture the dynamic range from minutes to years.

## 3.1. Review test scores

Each test is summarized in our database with a score ranging from 0 to 1. Only the aggregate score is available; responses to specific challenges are not. We use $y_0$ and y to denote the initial and delayed aggregate scores, respectively.

The values of $y_0$ and y for the review activity are the guessing-corrected proportion correct across the review activity. The guessing correction is made by subtracting from the raw proportion correct the baseline proportion correct that would be obtained by ran-dom selection of responses and then renormalizing to the range [0,1]. The baseline pro-portion is estimated for each display, and it depends on the type of response demanded from students. For displays that have N alternative responses, and M\N questions are asked in series, each requiring the selection of a distinct response, the expected propor-tion correct by guessing is $\sum_{i=1}^{M} \delta N \; i \; \beta \; 1 \Phi^{1}$, assuming that students use feedback from the first m questions to constrain their response choice for the m + 1'th question. The remainder of questions required spoken answers, prompted either by images, written phrases, or spoken phrases. For these questions, we set the baseline proportion to be the false alarm rate of the automatic speech recognition software.

Fig. 6 shows the distribution of guessing corrected scores on the initial and delayed tests, $y_0$ and y, respectively. Note that the vertical axis of the graph is log scaled, and that most of the guessing-corrected scores are above 0.7.

## 3.2. Forgetting in the wild

To begin our investigation of the data set, we ask whether forgetting is observed, and if so, whether it has the same qualitative properties as forgetting as assessed in controlled laboratory studies. We treat the lag between initial and delayed tests as a retention inter-val. However, unlike laboratory experiments, this lag is not an independent variable: Stu-dents determined when they wished to re-test themselves on a lesson. Our data set further deviates from laboratory experiments in that during the lag, students often used the lan-guage learning software and were thus engaged with the same or similar materials as that

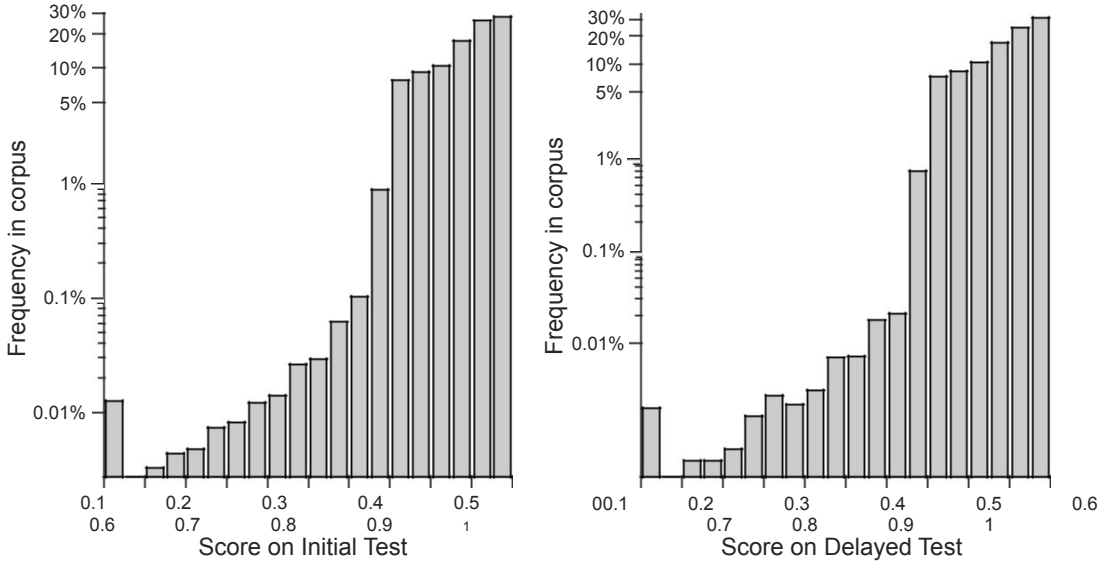


Fig. 6. Distributions of scores on the initial and delayed review tests.

covered by the test. Although we have information about the student's activities with the software during the lag, our initial exploration simply focuses on the effect of lag. Subsequent investigations in this article will consider the other activities a student performs during the lag.

We analyzed data from each of the 48 lessons separately. For each lesson, we created a scatter plot of scores on the delayed test versus lag. Due to the large number of stu-dents, it is difficult to visualize the relationship when a separate data point is plotted for each student. Consequently, we aggregated the data by forming groups of 50 students who took the delayed test at roughly the same lag. Fig 7 shows scatter plots for three les-sons, plotted on a log–log scale, where each point represents the mean score of a group of 50 students at their mean lag.

The strong linear relationship on a log–log scale indicates power function decay of knowledge. The solid line in each graph shows the best fit to a two-parameter power-law model (Eq. 1), $\hat{y} = at^{b}$, where $\hat{y}$ is the predicted score, t is the lag between initial and delayed review tests, and a and b are free parameters of the model. The power-law model is fit to the data of the individual students—not the group averages—in order to minimize the sum squared deviations of the log scores.[1]

The three lessons depicted in Fig. 7 are those whose data are best fit by the power function. The three fits explain, from left to right, 90.2%, 94.6%, and 90.7% of the variance in the aggregated data (on the log–log scale). The three fits explain 24.0%, 26.7%, and 28.1% of the variance in the individual student scores. Note that because the scores lie in a narrow range, 0.85–1.00, the log transform of the score does not induce a strong nonlinearity, and the fits are quite comparable for the untransformed scores.[1]

The three lessons depicted in Fig. 8 are those whose data are most poorly fit by the power function. The three fits explain, from left to right, 21.7%, 33.0%, and 39.1% of the variance in the aggregated data (on the log–log scale). Although these fits are not bad, they explain only 1.4% of the variance in the individual student scores in each of the three lessons.

The best fitting lessons tend to be those late within a level of a course and which have the fewest students enrolled. The worst fitting lessons tend to be those early within a level of the course, with the greatest number of students enrolled. This pattern makes sense given that early within a level, students of varying ability and degrees of interest participate, but those with the least interest tend to drop out over the lessons within a level. As a result, there is greater heterogeneity for the earlier, more populated lessons than for the later, less populated lessons. Additionally, the lessons often include, and serve as review of, material from earlier lessons. Therefore, later lessons are less likely to be reviewed and are thus retention can be better predicted by pure models of forgetting.

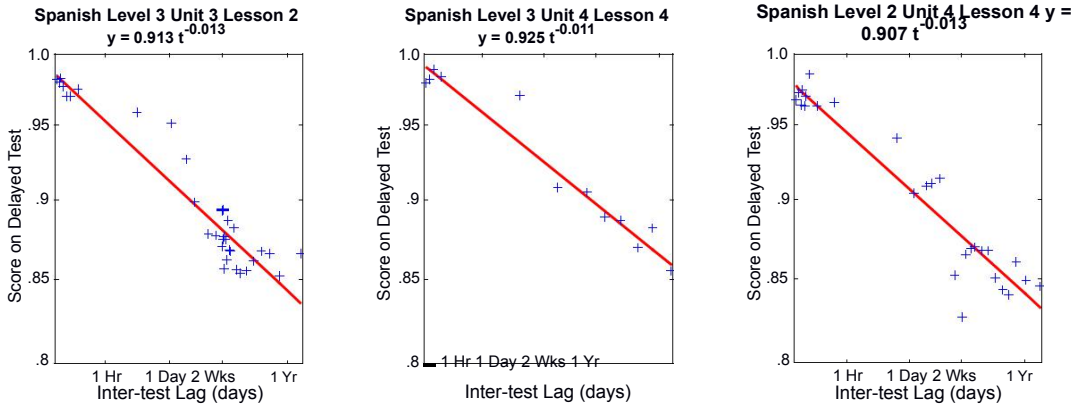


Fig. 7. Power-law model fits of forgetting for the three lessons whose data are well explained by the power function. The higher density of data around the 2-week lag is due to the default review scheduling policy.

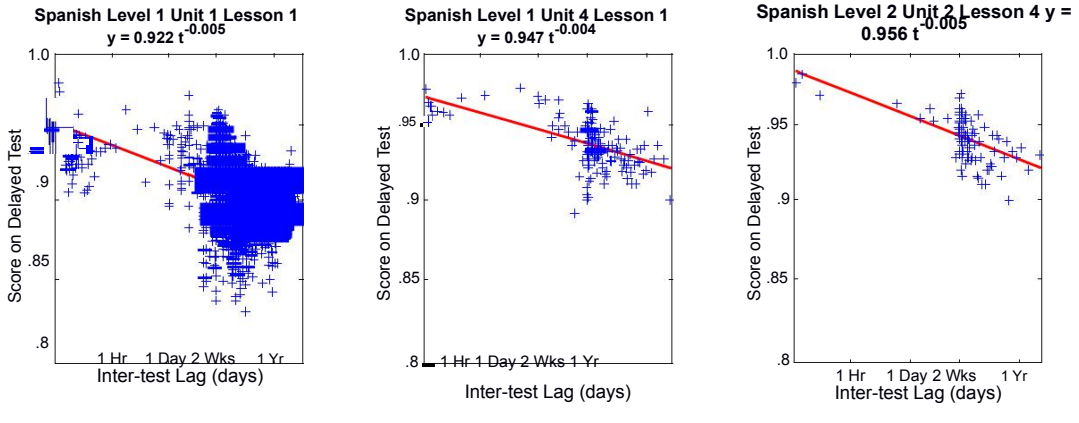


Fig. 8. Power-law model fits of forgetting for the three lessons whose data are poorly explained by the power function.

For each lesson, we estimate the model parameters a and b which represent initial learning and memory decay, respectively. Fig. 9 shows the resulting power-law forgetting curves for the 48 lessons. The marginal distributions of a and b, as well as a scatter plot of the pairwise values, are depicted in Fig. 10A. The strong negative correlation in the scatter plot suggests that lessons which are better learned initially are forgotten more slowly. The correlation does not inform us as to the cause of the relationship. The corre-lation could in principle imply that when students learn a lesson well initially, they forget it more slowly. But the correlation may instead be attributed to some underlying cause responsible for both effects. For example, more effective lessons might produce both bet-ter initial learning and slower forgetting. Another possibility is that noise in the data leads to the observed trade-off between a and b. To rule out this possibility, we performed a simulation under the null hypothesis that the true a and b parameters do not vary mean-
ingfully across data sets. We generated 48 synthetic data sets, analogous to the 48 lessons in the actual data, each consisting of 5,000 samples from $y = 0.95t^{-0.012} + g$, represent-
ing a typical lesson forgetting curve (Fig. 9) with additive noise g sampled from a mean-zero Gaussian density with standard deviation 0.15. Fitting the synthetic data in the same manner as the actual data, we obtain a and b estimates and generate a scatter plot for the synthetic data (Fig. 10B). Here, we observe a strong positive correlation—the opposite of what we observe with the actual data—suggesting that the observed a–b correlation is due to meaningful variation in a and b across data sets, and not to noise artifacts under the null hypothesis. The simulation with synthetic data is also interesting in that it pro-duces a roughly comparable distribution of b to what is found in the actual data, but the distribution of a is much tighter. This comparison of distributions provides evidence that initial learning (a) does vary across lessons, but it does not offer strong support for inter-lesson variability in decay rates (b).

In all fits, the forgetting rate, b, is relatively low compared to laboratory studies of free recall. Typical values in laboratory studies we reviewed range from 0.15 to 0.30. The

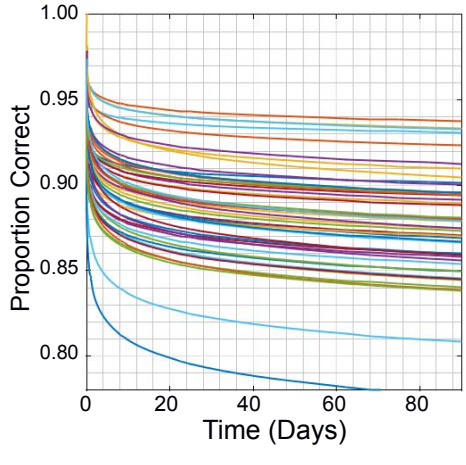


Fig. 9. Forgetting curves fit to the data from the 48 lessons using the two-parameter power-law model.

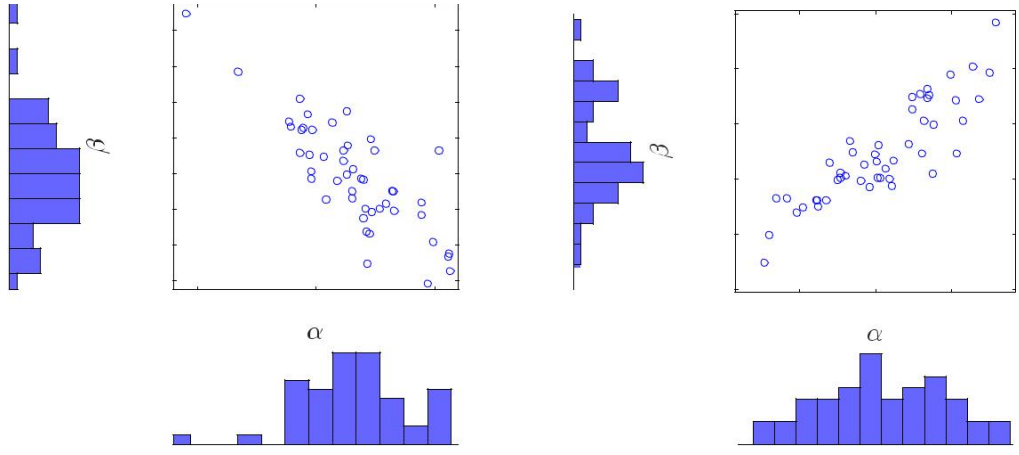


Fig. 10. (A) Scatter plot of power law parameters for the 48 lessons. Each point in the scatter plot represents a single lesson's {a, b} values. The histograms along the horizontal and vertical axes depict the marginal distributions of a and b, respectively. (B) Scatter plot of recovered parameters from an experiment with syn-thetic data in which 48 sets of observations were generated via noisy samples from a power function with fixed parameters a = 0.95 and b = 0.012.

decay rates in our fits are closer to 0.01. We suspect this lower decay rate is due to the fact that the material being tested is exercised in other activities the students perform. Further, even though we have corrected the scores for guessing, typical memory studies involve free recall, and the review tests in the software mostly involve recognition. Finally, the review tests include feedback, and the initial and delayed tests are nearly identical; consequently, there may be a learning benefit of performing the initial test.

### 3.3. Predictors of knowledge retention

Forgetting in the laboratory is typically characterized by a relatively small number of factors, such as the retention interval and the nature of initial study (e.g., reading vs. retrieval practice). In our data set, we have the potential for considering a much larger collection of variables that might contribute to the durability of knowledge. For each student in each lesson, we extracted features that seemed potentially useful as predictors of knowledge strength and retention. In the remainder of the article, we explore models that incorporate these features to predict performance. The following list enumerates 25 dis-tinct features that are included in our models.

1. Information about a student's performance on prior review tests.

   - The student's score on the initial review test, a fraction in [0, 1] representing the number of challenges in the review activity the student

answered correctly on the first attempt divided by the total number of challenges attempted. This score is guessing corrected, as described earlier.

- The lag, in days, between the initial and the delayed review test.
- The total time spent on the initial review test. This time, and all others that follow, is specified in seconds.
- A count of the number of times that a student began but did not complete the review test between the initial and the delayed tests.
- The total time that a student spent on incomplete review tests prior to the delayed test.

2. A binary variable indicating whether the delayed test was taken at the prompting of the software (i.e., whether it was scheduled with Adaptive Recall ).

3. The student's score on the core lesson and amount of time spent in the core lesson. As discussed in earlier, core lessons introduce the material being tested.

4. Information about other (non-review) activities performed within the same lesson.

- The total time spent between initial and delayed tests on all other activities. For
- each of eight distinct activity types (writing, grammar, listening, reading, listening and reading, speaking, pronunciation, and vocabulary), the count of the number of attempts on that activity type prior to the initial review test. For each
- distinct activity type, the count of the number of attempts on that activity type between the initial and the delayed review tests.

## 4. Methodology

One goal of this work is to compare the quality of alternative models. A model's quality is determined by how well the model predicts performance of a particular student on a second attempt at taking a lesson's review test, given the lag between the first and second attempts. In this section, we describe how we use our data set to evaluate each model.

Each lesson is modeled independently. We divide the overall data set by lesson; the lesson-specific data set includes all students who have performed the associated review activity at least twice. We use five-fold cross validation for model evaluation. This procedure involves randomly partitioning the data set into five roughly equal sized groups of students. For each partition, we form a training set consisting of data from all students except those in the partition. We fit a model using the training set, and then evaluate model predictive performance using data from students in the held-out partition, which we refer to as the validation set. Aggregating validation-set predictions across the five partitions, we obtain a prediction for each student using a model that was not fit to that student. In order to perform the most meaningful comparison across alternative models, we use the same cross validation partitions for all models.

### 4.1. Performance metric

We evaluate a model m by comparing the score on the second review test for each student in the validation set, which we denote as $y_{lpi}$ for student i in partition p of lesson l,

to the model's prediction, denoted $\hat{y}^{m}_{lpi}$. For a given partition p of lesson l, we compute the normalized mean squared error (NMSE) for model m as:

$$\mathrm{NMSE}_{lp}^{m} \quad \text{———————————————}$$

where $N_{lp}$ is the number of students held out in partition p of lesson l, $\bar{y}_{lp}$ is the mean score of all students in the training set for partition p of lesson l, and the model's predic-tions are restricted to the [0,1] range. The NMSE will typically range from 0 to 1, where 0 indicates that the model's predictions are perfect, and 1 indicates that the model does no better than predicting the mean score of students the training set. The NMSE can also be interpreted as the proportion of variance in the data that the model fails to explain.

To compute the mean performance of model m on a lesson l, we average over partitions:

$$\mathrm{NMSE}_{l}^{m}$$

where $N_p = 5$ is the total number of partitions. The overall mean for model m, $\mathrm{NMSE}^{m}$ is simply the average over lessons:

$$1$$

$$\mathrm{NMSE}^{m} ¼$$

where $N_l = 48$ is the total number of lessons. This error metric weighs all lessons equally. Despite the wide disparity in the number of students who complete a lesson in our data set, we chose to weight each lesson equally, rather than each student or each student-les-son equally. This choice allows us to interpret our NMSE measure as a prediction of how well a model will generalize to new lessons. If our goal was to focus on students who were the heaviest users in our data set or lessons that were more popular in our data set, it might be more appropriate to weight the NMSE by student-lessons or students, respec-tively. Although the results we report in this article are based on a lesson-weighted NMSE, we have run all of our simulations with an NMSE weighted by student-lessons, and the two weightings yield essentially the same conclusions.

To compare the performance of models $m_1$ and $m_2$, we perform two-tailed t tests with lesson as the random variable and $\mathrm{NMSE}^{m_1}_{l}$ and $\mathrm{NMSE}^{m_2}_{l}$ as paired samples.

## 5. Baseline models

In this section, we compare alternative models for predicting a student's delayed review test score, y. We begin with the two variants of the power model described earlier, one with two parameters (Eq. 1) and one with three parameters (Eq. 2). Both models

assume forgetting as a power function of lag. The normalized cross-validation error is almost identical for the two models (0.8858 and 0.8852 for the two- and three-parameter models, respectively, $t(47) = 1.67$, $p = .10$), indicating that the additional flexibility of the three-parameter model does not lead to superior predictions on the held-out data. Consequently, in subsequent comparisons involving the power model, we utilize the two-parameter model, $\hat{y} = at^{b}$ (Eq. 1).

Many alternatives to the power model have been proposed and explored by Rubin and Wenzel (1996). Rubin and Wenzel found several models to obtain roughly equivalent fits to the power model, including a model termed the exponential-power model, as defined by:

$$\hat{y}$$

Although Rubin and Wenzel were unable to distinguish power and exponential-power models based on goodness of fits to the data they had available, we find that the normalized cross-validation error is reliably worse for the exponential-power model than for the power model (0.9450 and 0.8858, respectively, $(47) = 11.36$, $p < .01$). Both models have two free parameters. The power function appears to be better suited for describing forgetting in our data set.

Beyond models traditionally used to characterize forgetting, we explored two addi-tional models based on a generic regression approach in which y is predicted from a vec-tor of features, $\mathbf{x}$, that characterize a student's specific study history for the given lesson. With linear regression, we have

$$\hat{y} = \sum_{j} w_j x_j; \qquad (4)$$

where y is a linear function of the feature vector and the model has coefficients w. (Although scores are bounded to lie in the range [0,1], linear regression predictions are not. Nonetheless, the scoring function, Eq. 5, does not penalize scores outside the range.) The model includes a constant feature, $x_0 = 1$, to provide a bias term on the prediction. With logistic regression, we have

$$\hat{y} =$$

1 þ

which restricts predictions to the [0,1] range.

Fig. 11 shows the normalized cross-validation error for the three models. All models have predictive value, explaining between 10% and 20% of the variance in the scores. With lesson as the random factor and normalized prediction error on the model's test set as a measure of model accuracy, the linear model performs significantly better than the power model ($t(47) = 8.08$, $p < .01$), indicating that prediction is enhanced by

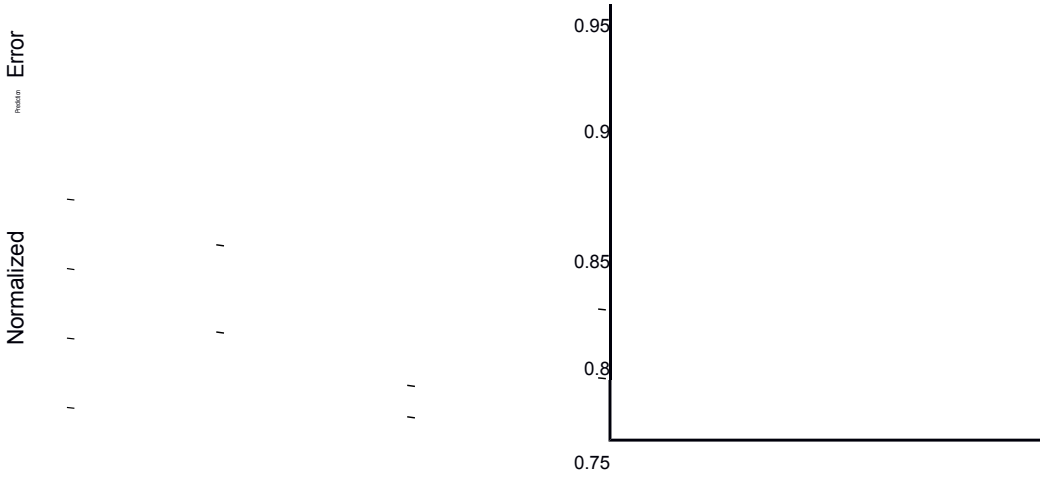


Fig. 11. Mean normalized prediction error for the power, linear, and logistic models. The error bars indicate 1 SE of the mean across the 48 lessons, and they have been repeated-measures corrected to remove between-lesson variance as described in Masson and Loftus (2003).

student-lesson specific features other than the lag between the first and second tests. The linear model also outperforms the logistic model ($t(47) = 7.47$, $p < .01$). This finding may be due to the fact that the y scores tend to be high and lie within a narrow range, even after guessing-rate correction: The mean y score is 0.90 with a standard deviation of 0.09. With scores this high, the logistic model operates in the nonlinear regime of the sig-moid response function, where the contribution of a feature to the prediction depends on the contribution of other features. The better performance of the linear model indicates that such interactions are not helpful in this data set.

## 6. Individualized models of forgetting

Student retention can be partly explained by a power model that takes into account the passage of time and also by a regression model that takes into account student-lesson specific features. Because the regression models lack the ability to represent power-law forgetting, and the power-law models have no notion of student-lesson specific features, it is natural to conjecture that integrating these two classes of models might yield even better predictions. In this section, we describe three variations on such a hybrid model.

In the power-law model of Eq. 1, $y\hat{} ¼ at^{b}$, the constants $a$ and $b$ are estimated for the entire population of students in a lesson. If we replace one or both of these constants with a function of the student-lesson specific features used in the linear model, then we can make power-law predictions that are individualized to the students' study history.

We define $a(x)$ to be a linear function of the student-lesson specific features $x$:

As in the linear model, $\mathbf{x}$ includes a constant feature, $x_0 = 1$, to provide a bias term. This function will predict an individualized scaling factor that takes into account

student-lesson specific features. By replacing $a$ with $a(x)$ in Eq. 1, we define our first combined model, called $\text{Hybrid}_{a(x)}$:

$$\hat{y} = a(x)(t+1)^{-b} \quad (7)$$

We can also individualize the forgetting rate $b$ via a function $b(x)$:

This function allows us to define two model variations that predict an individualized forgetting rate for a set of student-lesson specific features. In model $\text{Hybrid}_{b(x)}$, $b$ is individualized:

$$\hat{y} = a(t+1)^{-b(x)} \quad (8)$$

In model $\text{Hybrid}_{a(x),b(x)}$, both $a$ and $b$ are individualized:

$$\hat{y} = a(x)(t+1)^{-b(x)}: \quad (9)$$

The nonlinear procedure used for fitting parameters of the hybrid models is described in the Appendix. Fig. 12 shows the normalized error for these three models. In addition, the error for the power and linear models is copied from Fig. 11 for reference. Of the three hybrid models, only the models individualizing the scale factor $a$ showed any improvement over the reference linear and power models. $\text{Hybrid}_{a(x)}$ performs significantly better than the previous best—the linear model ($t(47) = 10.8$, $p < .01$), and explains 24% of the variance on scores, about 4% more than the linear model. The model with only an individualized forgetting rate, $\text{Hybrid}_{b(x)}$, performs significantly worse than the linear model ($t(47) = 15.8$, $p < .01$). The more complex $\text{Hybrid}_{a(x),b(x)}$ model, a superset of $\text{Hybrid}_{a(x)}$, does not provide a reliable performance advantage ($t(47) = 0.37$, $p = .72$).

Why do individualized forgetting rates fail to improve predictions? One hypothesis centers on overfitting: $\text{Hybrid}_{a(x),b(x)}$has many more free parameters than $\text{Hybrid}_{a(x)}$. To rule out this hypothesis, we conducted simulations with both L1 and L2 regularization, but did not improve on the results reported in Fig. 12. We are aware of few previous efforts that have explored individual differences in forgetting rates, other than a recent study by Van Vuuren and Cherney (unpublished data) in which aphasic patients were trained to learn scripts and were then tested on their retention of the scripts. Van Vuuren and Cherney also found that allowing the power law forgetting rate to vary across indi-vidual did not improve model predictions. Because of the nature of this corpus-based study, we are among the first to be able to explore inter-individual variability in forget-ting rate. It is an intriguing but tentative conclusion that inter-individual variability in for-getting rate does not appear to be large, at least at least relative to inter-individual variability in the strength of initial learning.

We reached a similar conclusion earlier for inter-lesson variability in forgetting rates versus initial-learning strength. Via the power model and a comparison of parameter val-ues obtained from human and synthetic data (Fig. 10), we concluded that inter-lesson variability in forgetting rates was consistent with sampling noise, whereas inter-lesson variability in initial learning strength was too large to be explained by sampling noise. The forgetting rate distribution obtained from the $\text{Hybrid}_{a(x)}$ model, shown in Fig. 13, is quite similar to that we obtained with the power model. One might be tempted to dismiss the variability as uninteresting, except the strong relationship with initial learning (Fig. 10A) suggests otherwise.

## 7. Incorporating student effects

The models we have described to this point predict a student's retention of a lesson based solely on information associated with that lesson. Might model predictions be

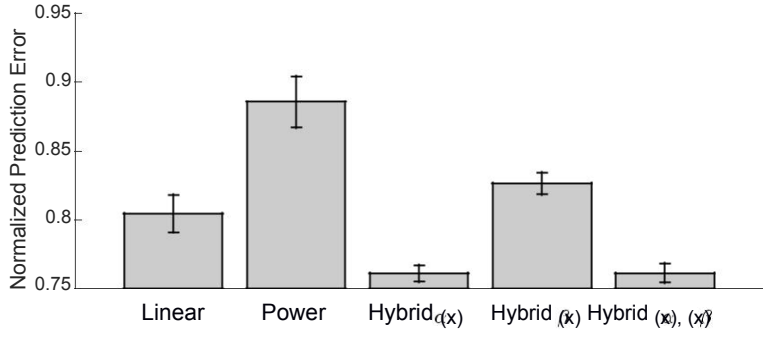


Fig. 12. Mean normalized prediction error for the three Hybrid models, as well as for the linear and power models (duplicated from Fig. 11). The error bars indicate 1 SE of the mean across the 48 lessons and have been corrected to remove between-lesson variance as described in Masson and Loftus (2003).

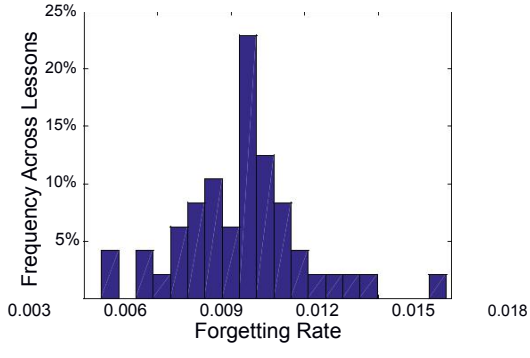


Fig. 13. Distribution of forgetting rates over the 48 lessons.

improved if we incorporated information from other lessons? After all, a student who has demonstrated good retention of lessons 1–10 seems likely to have good retention of les-son 11. A full investigation of incorporating prior-lesson data into our models is beyond the scope of this paper, but we conducted an initial exploration aimed at assessing an upper bound on the potential gains one could obtain by combining data across lessons.

To consider a student's performance across lessons, we construct a mixed-effects model in the style of item-response theory (De Boeck & Wilson, 2004). The model includes the lesson-specific features considered in our earlier model, as well as cross-les-son features that identify the specific student. Student i is identified by a one-hot vector $\mathbf{s}$ whose elements are all zero except for element i which has value 1. Extending our best model, $\text{Hybrid}_{a(\mathbf{x})}$, we obtain a new model $\text{Hybrid}_{a(\mathbf{x},\mathbf{s})}$, in which the coefficients associ-ated with $\mathbf{x}$ are lesson-specific, but the coefficient vector associated with $\mathbf{s}$ is shared across lessons. The coefficient associated with $s_i$ tells us about student i's ability. If stu-dent i consistently performs better across lessons than student j, this difference can be represented by a difference in the corresponding coefficients. Consequently, a student who is above average on a set of lessons will be predicted to be above average on other lessons as well.

For our initial investigation, we wished to select a subset of lessons and students such that each student had completed all lessons. We chose the 14 most popular lessons in the course and identified 1,755 students who had each completed all 14 lessons. We did not use the full course because there are very few students who completed all of the lessons. We evaluated the model on each lesson using the same cross validation procedure used for the other models. We found a non-significant increase in mean normalized vali-dation-set error with $\text{Hybrid}_{a(\mathbf{x},\mathbf{s})}$ over $\text{Hybrid}_{a(\mathbf{x})}$ (0.863 vs. 0.853; $t(13) = 0.37$, $p > .1$). $\text{Hybrid}_{a(\mathbf{x},\mathbf{s})}$ performs worse than $\text{Hybrid}_{a(\mathbf{x})}$ on only 5 of 14 lessons, although on one of those, it is 30% worse which washes out its gains.

Because the data set used in this experiment contains a lot of information about each student, our experimental conditions represent a best-case scenario for uncovering a bene-fit to the inclusion of student-specific factors in the model. With 14 lessons and five-fold cross validation, 11 or 12 lessons were part of a training set used to constrain the stu-dent-specific factor. In natural use, when predicting a student's performance on lesson n, we would have only the previous n 1 lessons for training.

Why aren't student-specific factors helpful for predicting performance? Although some students are surely better on average than others, it appears that a student's average per-formance level isn't pertinent given the other information available for prediction, most notably, the student's initial score on a lesson.

## 8. Interpreting model coefficients

In this section, we interpret the coefficients of our best model, $\text{Hybrid}_{a(\mathbf{x})}$, to obtain a better understanding of which features of a student's study history are critical for predict-ing the student's retention of a lesson. To remind the reader, the model incorporates 26

features which are linearly combined to determine the base performance level, a(x). Each feature j, $x_j$, is associated with a coefficient, $w_j$ (Eq. 6). We would like to interpret the magnitude of a coefficient as the importance of the corresponding feature in determining the base performance level, a(x). To facilitate this interpretation, all features in the train-ing set were renormalized to standard scores, that is, such that E[$x_j$] = 0 and E½$x^2_j$& ¼ 1. The mean and standard deviation used to renormalize the training set were also used to renormalize the test set. This procedure does not affect model predictions, but it does decouple the mean and variance of a feature from the magnitude of its corresponding coefficient, and thereby enables us to interpret the coefficient magnitude as the feature's predictive utility.

Fig. 14 shows the coefficient magnitudes of 15 features of the $\text{Hybrid}_{a(x)}$ model, sorted by importance. The values depicted are the mean across the 48 lessons. Shown are all coefficients that are robustly nonzero across lessons, as determined by a t test at the p = .05 level. Black and white coloring of the bars indicate negative and positive correla-tions with the score, respectively. Unsurprisingly, the student's score on the initial review test (first row), and to a lesser extend their score on the core activity (fourth row), is a

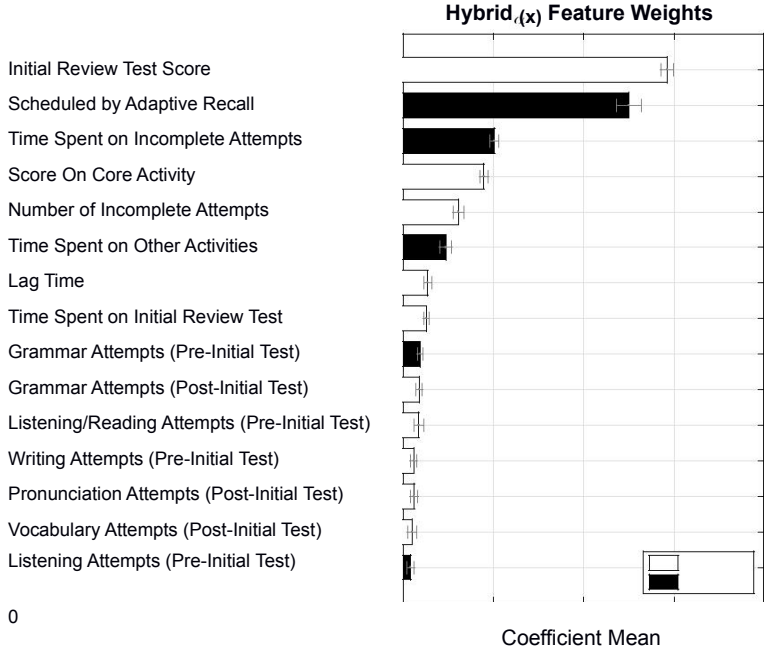


Fig. 14. The features that reliably contribute to predictions in the $\text{Hybrid}_{a(x)}$ model, as determined by the magnitude of the feature's corresponding coefficient (all feature values first converted to standard scores). The length of a bar indicates the coefficient magnitude and the sign is indicated by coloring, with black and white for negative and positive, respectively. Each bar indicates the mean coefficient across the 48 lessons and the 5 validation splits within each lesson. Error bars indicate 1 SE of the lesson means. Shown are all features with nonzero means by a t test at the p = .05 level.

positive predictor of their score on the delayed review test. Students who take the delayed review test at the point when recommended by the software ("scheduled by adaptive recall," second row) perform worse than students who have initiated the exam on their own. When students start but do not complete earlier attempts at the review examination, the time they spent on these incomplete attempts (third row) has a negative correlation with score on the examination they eventually complete but total number of attempts (fifth row) has a positive correlation. Perhaps the reason for these opposing correlations is that the more time spent relative to the number of review attempts is an indication that students are trying but quitting due to difficulty with the material. The time spent on the initial review test (eighth row) is slightly positively predictive, which might indicate that students who take the test more carefully tend to perform better. Oddly, the linear lag between initial and delayed test ("lag time," seventh row) is a positive predictor of score, conflicting with the power-function effect of lag embodied in the forgetting term. The rel-atively small compensatory effect of lag in $a(x)$ may suggest that forgetting might be bet-ter modeled by a function other than the power function. There is a similar opposing effect between the time spent on other than the core activity (sixth row) and the number of various specific activities conducted before and after the initial test (rows 9–15): attempting a large number of other activities relative to the amount of time spent on these activities predicts better delayed-test scores, but engaging in fewer activities or spending a lot of time per activity predicts worse scores.

In explaining student retention, how important is the temporal dimension to forgetting, $t^{b}$, versus the effect of student-specific features, embodied in $a(x)$? One answer is obtained by examining Fig. 12, which shows that power-function forgetting explains 11.4% of the variance in the data, whereas $\text{Hybrid}_{a(x)}$, which incorporates the student-spe-cific features, explains 23.9% of the variance. Thus, the contribution of temporal forget-ting is about as great as the contribution from all the student-specific features in aggregate. As an independent means of examining the contributions of different factors to model prediction, we constructed a modification of $\text{Hybrid}_{a(x)}$ in which $a(x)$ is defined as a product of terms instead of as a sum, i.e., $a(x) = \prod_j x_j^{w_j}$. This model, $\text{Hybrid}_{a^P}$, turns out to have the same explanatory power as $\text{Hybrid}_{a(x)}$. The advantage of $\text{Hybrid}_{a^P}$ is that it expresses the log score as a linear function of log features:

$$\ln \hat{y} = \sum_j w_j \ln x_j - b \ln t.$$

Because this formulation places forgetting, represented by coefficient $b$, on the same footing as the individual features, represented by coefficients $\{w_j\}$, we can directly compare the magnitudes of all coefficients—the $\{w_j\}$ as well as $b$. With all variables expressed as standard scores, $b$ has a mean (across lessons) of 0.32, whereas the largest feature coefficient, associated with the initial review test score, has a mean of 0.29, indicating that forgetting is at least as important as initial test performance. The ranking of coefficients and their magnitudes are almost identical to those we obtained from $\text{Hybrid}_{a(x)}$ (Fig. 14).

## 9. Conclusions

In this article, we examined retention of foreign language skills using a large naturalis-tic data set of self-directed students. We attempted to understand the factors contributing to retention by constructing models that predict performance on a delayed test. We con-sidered many factors influencing retention, including information about a student's perfor-mance on prior tests, the lag between initial and delayed tests, and detailed information about the student's participation and performance in specific instructional activities cover-ing material related to the test. The key take-home messages we have gleaned from our analysis are as follows.

1. Power-law forgetting is observed in the naturalistic corpus. As in controlled labo-ratory studies, power-law decay appears to be a reasonable characterization of for-getting over time. This parallel seems nontrivial, given that our data contrast with those from laboratory studies in three key respects. First, students in our corpus were likely exposed to the test material between tests, either by engaging in addi-tional activities related to the corresponding lesson or by advancing to new lessons with overlapping material. Language courses tend to build on earlier material; con-sequently, material covered in lesson 1 is likely to be encountered again in subse-quent lessons. A delayed test on lesson 1 will not assess pure forgetting because a student's knowledge state has been contaminated by exposure to the same material in other lessons. Second, the review tests are designed as instructional activities; students receive feedback that likely alters their knowledge state. Third, many of the test questions were of a multiple-choice format, serving as more of a recogni-tion memory test than a cued-recall test. As a result of these three factors, the for-getting rate we observe in our naturalistic data set is low compared to that in controlled laboratory studies. Indeed, we are not modeling pure forgetting as observed in laboratory studies, but rather the interaction between memory mecha-nisms and the specific curriculum. The presence of these interactions suggestion that modeling forgetting in situ requires a big-data approach: If the curriculum is altered such that a lesson is placed in a different context, previously built models will no longer fit well. Nonetheless, it is intriguing that the basic form of forgetting parallels that observed in laboratory-based research.
2. Our corpus included 48 lessons which yield different levels of initial learning by students and different degrees of retention. We find that lessons which are better learned initially are forgotten more slowly. To be clear, we have merely observed a correlation. This correlation is strong (Fig. 10A) and we ruled out an artifactual explanation for the correlation in terms of data set noise (Fig. 10B). The correlation could imply one of two possibilities. First, when a particular student learns a lesson better, that student will forget more slowly. Second, an underlying unobserved cause influences both initial learning and forgetting rates. The cause might be vari-ation in lesson difficulty or quality, or different amounts of overlap between lessons

such that material which is better integrated into the course is learned better ini-tially and appears to be forgotten more slowly due to indirect exposure via other lessons. To distinguish these two possibilities, randomized controlled experiments would be required in which students are taught to various criteria and then forget-ting rates are assessed (e.g., Rohrer, Taylor, Pashler, Wixted, & Cepeda, 2005). Even if the correlation does not reflect an underlying memory process, it is of prac-tical use in predicting student performance: Lessons that are learned more slowly should be reviewed sooner.

3. We find no evidence that forgetting rates vary from student to student. In contrast, individuating baseline scores via student-specific features improves prediction accuracy. We wish to be cautious in generalizing these results beyond our data set for two reasons. First, in related work using laboratory-based cued-recall measures (Mozer & Lindsey, 2016), we do find evidence for meaningful inter-student varia-tion in forgetting rates. Second, although forgetting rates vary meaningfully by les-son (perhaps reflecting inter-lesson interactions as described above), the narrow range of scores in our data sets may reduce the opportunity to observe individual differences in forgetting rates.

4. Typically in cognitive modeling, researchers are concerned with fitting population data. In memory studies, the data are means across a population of subjects and a population of items. Moreover, the data are cross-sectional because each individual can be tested at only one point in time due to the observer effect, that is, memory retrieval affects subsequent memory strength. Consequently, the cross-sectional pop-

ulation data may not reflect the longitudinal trajectory of an individual. However, in the present work, our models make highly specific predictions—for a particular student on a particular test for a particular lesson. The predictive methodology we use for evaluation overcomes limitations of cross-sectional studies and allows us to draw conclusions concerning the longitudinal trajectory of individual memory traces. The methodology also appears promising to help discriminate among competing theories that were heretofore difficult to distinguish. For example, Rubin and Wenzel (1996) were unable to discriminate between power and exponential-power models of forgetting, yet our data reveal the superiority of the power model.

5. Performance following a retention interval has traditionally been modeled by power-law decay of knowledge. Power-law decay could explain 11.4% of the variation in delayed-test scores of individual students. In contrast to this approach based on psychological theory, we also investigated a black-box modeling approach in which scores were predicted from a set of features describing the study history of a student. The most successful of these models, simple linear regression, explained

19.5% of score variance. Our key modeling insight was to combine the two approaches—leveraging insights from both psychological theory and data-driven modeling—to obtain predictions that explain 23.9% of score variance.

Should we be satisfied with models that explain only one quarter of the variance in the scores? Although we hope that others could improve on our results, we suspect that

predictive accuracy is limited by the information missing from our data set. For example, we know only the overall score on the initial review test; it may be useful to have this score broken down by question or activity. We also have no information about the con-tent of each lesson and the interrelationships among lessons, which might increase the value of cross-lesson data for prediction. Nonetheless, there are certainly avenues that can be investigated even with the current data. For example, we may be able to leverage the data set itself to draw inferences about the set of knowledge components—skills and facts—required for each lesson (Lindsey, Khajah, & Mozer, 2014) and use these inferred knowledge components to better represent the activities a student engages in between the initial and delayed tests.

Our research has practical implications, not only for language learning software, but for computer-assisted tutoring in general. Because all knowledge and skills are forgotten if not practiced, review is critical. The software's review tests are designed to serve this function. However, as students progress through a course—regardless of the subject—the body of knowledge and skills they are tasked to master continues to grow. For example, even a single level of the Rosetta Stone software has as many as 16 lessons that could potentially be reviewed. Students typically are not excited about review activities that interfere with the ongoing demand to master new material. Even if willing to review, stu-dents are not particularly adept at metacognitive judgments about when to review (e.g., Cohen, Yan, Halamish, & Bjork, 2013; Nelson & Dunlosky, 1991). Ill-timed review—re-view that occurs too soon or too late—has less benefit than review at the point of desir-able difficulty (Bjork, 1994).

Review must therefore be efficient and well-timed. Predictive models offer the poten-tial of prioritizing review in a manner that is optimal to a particular student. For example, review might be recommended at the point when the predicted knowledge strength drops below a certain threshold. This heuristic has been successful in improving long-term retention (Khajah, Lindsey, & Mozer, 2014; Lindsey et al., 2014; Pavlik & Anderson, 2008).

Many electronic tutoring systems, including the Rosetta Stone software and Khan Academy, provide students with a dashboard showing students the state of master of each lesson or skill, and possibly identifying which are due for review. We envision that this dashboard might provide more nuanced predictions concerning the student's knowledge state. Such individualized dashboards offer the metacognitive insight that students lack and should serve to guide students in a more directed manner than qualitative guidance typically offered by psychological theory.

## Acknowledgments

This research was supported by NSF grants SES-1461535, SBE-0542013, and SMA-1041755. We thank Maryellen MacDonald, Andrew Butler, and an anonymous reviewer for constructive feedback on the manuscript.

## Note

1. We also fit data with the three-parameter power-law model but found, using a cross-validation measure of model performance to be described in a later section, that the three-parameter model explained no more of the variance in the data than the two-parameter model. On the grounds of parsimony and interpretability, all results we report are for the two-parameter model.

## References

Bahrick, H. P. (1983). The cognitive map of a city: Fifty years of learning and memory In G. Bower (Ed.), The psychology of learning and motivation: Advances in research and theory, (Vol. 17) (pp. 125–163). New York: Academic Press.

Bahrick, H. P. (1984). Semantic memory content in permastore: Fifty years of memory for Spanish learned in school. Journal of Experimental Psychology: General, 113, 1–29.

Bahrick, H. P., Bahrick, P. O., & Wittinger, R. P. (1975). Fifty years of memory for names and faces: A cross-sectional approach. Journal of Experimental Psychology: General, 104, 54–75.

Bahrick, H. P., & Hall, L. K. (1991). Lifetime maintenance of high school mathematics content. Journal of Experimental Psychology: General, 120(1), 20.

Bjork, R. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe et al. (Eds.), Metacognition: Knowing about knowing (pp. 195–205). Cambridge, MA: MIT Press.

Cohen, M. S., Yan, V. X., Halamish, V., & Bjork, R. A. (2013). Do students think that difficult or valuable materials should be restudied sooner rather than later? Journal of Experimental Psychology: Learning, Memory, and Cognition, 39(6), 1682. doi:10.1037/a0032425

Custers, E. (2010). Long-term retention of basic science knowledge: A review study. Advances in Health Science Education: Theory & Practice, 15(1), 109–128.

Custers, E., & ten Cate, O. (2011). Very long-term retention of basic science knowledge in doctors after graduation. Medical Education, 45(4), 422–430.

De Boeck, P. & Wilson, M. (Eds.) (2004). Explanatory item response models: A generalized linear and nonlinear approach. New York: Springer.

Ebbinghaus, H. (1885/1964). Memory: A contribution to experimental psychology. Mineola, NY: Dover.

Kang, S. H. K., Lindsey, R. V., Mozer, M. C., & Pashler, H. (2014). Retrieval practice over the long term: Expanding or equal-interval spacing? Psychological Bulletin & Review, 21, 1544–1550.

Keim, G. (2009). Adaptive recall. Google Patents. US Patent App. 12/052,435. Available at http://www.google.com/patents/US20090061407. Accessed November 24, 2015.

Khajah, M. M., Lindsey, R. V., & Mozer, M. C. (2014). Maximizing students' retention via spaced review: Practical guidance from computational models of memory. Topics in Cognitive Science, 6, 157–169.

Landauer, T. K., & Bjork, R. A. (1978). Optimum rehearsal patterns and name learning. Practical Aspects of Memory, 1, 625–632.

Lindsey, R. V., Khajah, M., & Mozer, M. C. (2014). Automatic discovery of cognitive skills to improve the prediction of student learning. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Weinberger (Eds.), Advances in neural information processing systems 27 (pp. 1386–1394). Red Hook, NY: Curran Associates, Inc.

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

Masson, M. E., & Loftus, G. R. (2003). Using confidence intervals for graphically based data interpretation. Canadian Journal of Experimental Psychology, 57(3), 203.

Mozer, M. C., & Lindsey, R. V. (2016). Predicting and improving memory retention: Psychological theory matters in the big data era. In M. Jones (Ed.), Big data in cognitive science. Sussex, UK: Psychology Press.

Nelson, T., & Dunlosky, J. (1991). When people's judgments of learning (JOL) are extremely accurate at predicting subsequent recall: The delayed-JOL effect. Psychological Science, 2, 267–270.

Pavlik, P., & Anderson, J. (2008). Using a model to compute the optimal schedule of practice. Journal of Experimental Psychology: Applied, 14, 101–117.

Rohrer, D., Taylor, K., Pashler, H., Wixted, J. T., & Cepeda, N. J. (2005). The effect of overlearning on long-term retention. Applied Cognitive Psychology, 19(3), 361–374. doi: 10.1002/acp.1083

Rubin, D. C., & Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. Psychological Review, 103, 734–760.

Wickelgren, W. A. (1974). Single-trace fragility theory of memory dynamics. Memory & Cognition, 2, 775–780. Wixted, J. T. (2004). On common ground: Jost's (1897) law of forgetting and Ribot's (1881) law of retrograde amnesia. Psychological Review, 111(4), 864–879.

Wixted, J. T., & Carpenter, S. K. (2007). The Wickelgren power law and the Ebbinghaus savings function. Psychological Science, 18, 133–134.

Wixted, J. T., & Ebbesen, E. (1991). On the form of forgetting. Psychological Science, 2, 409–415.

Wixted, J. T., & Ebbesen, E. B. (1997). Genuine power curves in forgetting: A quantitative analysis of individual subject forgetting functions. Memory & Cognition, 25(5), 731–739.

## Appendix: Nonlinear Optimization Procedure

Fitting the power law models to the data requires nonlinear optimization. We used MATLAB's fminsearch function, which performs black-box optimization on arbi-trary functions. Specifically, we used a wrapper function called fminsearchbnd, which adds bound constraints to the parameters of the optimized function. For the two-para-meter power law model, $a$ was bound to lie between $[0,+\infty)$ and $b$ between $[-1,0]$. Initial values for $a$ were drawn from a uniform random distribution between $[0.9,1]$, and initial val-ues for $b$ were drawn from a random uniform distribution between $[-0.5,0]$, which represent a common range for forgetting rates. For the three-parameter power law model, the same bounds and initial values were used for $a$ and $b$. The $c$ parameter was bound to lie between $[0,+\infty)$ and was drawn from a uniform random distribution between $[0.9,1]$.

This procedure can be extended to fit the two-parameter models that replace $a$ and/or $b$ with functions $a(x)$ and $b(x)$.

First, we fix the coefficients of $a$ so that the value of $a(x)$ is fixed for each data point. These fixed $a(x)$ values are then used to estimate the coefficients of $b(x)$, using the fminsearch procedure. Likewise, the new $b(x)$ coefficients are then fixed, and the $a(x)$ coefficients are estimated using least-squares regression. This procedure is repeated until the percent change in root-mean-square error of the prediction on the training set, compared to the last iteration, falls below a threshold (in our case, 0.0001%).