university of groningen

faculty of science and engineering

Cold Start Mitigation in an Adaptive Fact Learning System: Bayesian Prediction of Secondary School Student Performance.

Jelle Bosch

university of groningen

faculty of science and engineering

University of Groningen

Cold Start Mitigation in an Adaptive Fact Learning System: Bayesian Prediction of Secondary School Student Performance.

Master's Thesis

To fulfill the requirements for the degree of
Master of Science in Human-Machine Communication
at University of Groningen under internal supervision of
Dr. Jelmer Borst (Artificial Intelligence, University of Groningen)
and external supervision of
Maarten van der Velde, Msc. (Experimental Psychology, University of Groningen)

Jelle Bosch (s4210883)

December 12, 2021

Abstract

Digital learning aids have become increasingly popular to support classical education. In particular, adaptive fact learning systems (AFLSs) have been shown to lead to better learning outcomes as their underlying cognitive models allow them to train facts of appropriate difficulty at the appropriate time to optimise declarative memory reinforcement. With each given response, the cognitive model adapts to better reflect how well each fact is represented in the user's declarative memory, leading to a more personalised learning experience over time. When a new, unknown fact or user is encountered, the cognitive model will need to adapt a number of times before the model can provide an accurate representation of that fact in declarative memory, meanwhile the learning experience is not optimal. This is known as the 'cold start problem'. The current study sought to mitigate the cold start problem by predicting an AFLS parameter that reflected fact difficulty and student learning ability through five distinct Bayesian prediction methods. Predictions were carried out in a post-hoc simulation on a large and varied naturalistic dataset consisting of more than 117 million AFLS trials performed by over 135 thousand Dutch secondary school students. Out of the five prediction methods, a *fact-level* prediction method that made predictions per fact based on previous performances on each fact and a *hybrid* prediction method that combined *fact-level* and *student-level* predictions were found to consistently make more accurate model parameter and response time predictions than the *default* method which induced the cold start problem. It was concluded that these prediction methods are likely to mitigate the cold start problem in an applied setting. Further deliberation revealed that the *fact-level* prediction method was the best candidate to mitigate the cold start problem of an AFLS in practice.

Introduction

In the current digital age, computers have become an indispensable commodity in education. Multimedia presentations, serious games and online tutoring are commonplace in all levels and grades of students' academic careers Dahlstrom, Brooks, and Bichsel (2014); Ennouamani and Mahani (2017); Jeong and Kim (2017). Digital learning allows students to see learning material from a different perspective to the regular classroom setting, and has hence been found to enrich their learning experience FitzPatrick (2012). One increasingly popular instance of digital education is that of fact learning systems Ennouamani and Mahani (2017). Fact learning systems help students to expand their factual knowledge by repeatedly testing them on factual inquiries, such as the translations of words in case of language classes or multiplication tables when learning mathematics. As students typically need to memorise the meaning of thousands of words to achieve a moderate level of mastery in a foreign language Webb and Nation (2017), digital fact learning systems represent a welcome alternative to conventional learning techniques.

The simplest of digital fact learning systems increase learning outcomes just by repeatedly testing users' declarative knowledge. More advanced fact learning systems seek to further increase learning outcomes through the integration of adaptive models. As a type of intelligent tutoring system (ITS), such adaptive fact learning systems (AFLSs) may keep track of users' responses, overall knowledge level, cognitive and emotional state, and learning ability depending on the complexity of the AFLS's user model Nkambou, Mizoguchi, and Bourdeau (2010). An AFLS will use this information to provide learning material and feedback that best suits the user's current state in order to optimise the learning process VanLehn (2006). In this way, AFLSs create personalised learning environments tailored to their users. For instance, an AFLS may manipulate the schedule which dictates when each fact is presented. If a fact proves difficult to memorise, it may be repeated more often, while easier fact could be repeated only rarely. This may enhance the learning process because difficult facts require more repetition in order to be memorised than easy facts Pavlik Jr and Anderson (2005).

By providing students with a personalised learning environment that suits their knowledge level AFLSs have been found to lead to better learning outcomes Alshammari, Anane, and Hendley (2016). Meta-analyses by Ma, Adesope, Nesbit, and Liu (2014) and Steenbergen-Hu and Cooper (2014) have shown that ITSs can lead to better learning outcomes than many traditional education methods, such as large-group instruction, textbooks, passive computer-assisted learning and practical assignments. ITSs overall were found to lead to similar learning outcomes as much more labour-intensive methods such as small-group instruction and personal tutoring Steenbergen-Hu and Cooper (2014), promoting their use in current learning environments where there is often little capacity for personalised teaching Travers (2017).

For an AFLS to create a personalised learning environment, it requires information about its users. For example, it needs to be fed information on a user's response accuracy in previous trials to make an estimate of their learning ability. With each new trial, more is discovered about the user's learning ability, prompting the AFLS's estimate to be refined. In cases where nothing is known about the user, such as when someone uses the AFLS for the first time, the AFLS will not be able to provide a personalised learning experience until enough responses have been given for the user's learning ability to be estimated. This problem is known as the 'cold start problem'. An AFLS that suffers from the cold start problem cannot account for any individual differences between users nor any differences in difficulty between facts before it has 'warmed up', that is, collected sufficient information to construct a user model. During initial encounters with new users and facts, an AFLS's ability to improve learning outcomes is thus compromised (for more detail, see chapter: 'The SlimStampen AFLS'), leaving it as effective (or ineffective) as a regular fact learning system. The current study aimed to mitigate the cold start problem by predicting an AFLS parameter that described fact difficulty and students' learning ability. In consequence, the starting values of that parameter should be closer to parameter values observed after personalisation through adaptation, even if the AFLS did not encounter those facts or students before.

In recent work that inspired the current study, Van der Velde, Sense, Borst, and van Rijn (2021) investigated whether they could mitigate the cold start problem by predicting the 'rate of forgetting', a parameter used by the cognitive model at the base of their AFLS that represents the speed at which a memory of a fact decays. To illustrate, a difficult fact that is harder to memorise will have a higher rate of forgetting than an easy fact. Similarly, fast learning students will have an overall lower rate of forgetting than students that have trouble with memorising facts. In their lab study, Van der Velde and colleagues (2021) found that using a predicted rate of forgetting instead of a default value as the starting rate of forgetting resulted in increased response accuracy. Indeed, they found that predicted fact-related starting rates of forgetting based on earlier performances on the same facts by other users resulted in higher response accuracy than when a default starting rate of forgetting was used. They did not, however, find evidence for an increase in learning outcomes when starting rates of forgetting were user-related. This concerned predicted starting rates of forgetting per user based on prior performances as well as predicted staring rates of forgetting that were predicted using a *hybrid* method combining fact and user-related predictions. Van der Velde and colleagues (2021) explained their finding by arguing that their participant sample was too homogeneous, as they had found that increased heterogeneity of fact difficulty did increase the effectiveness of fact-related predictions on learning outcomes. Van der Velde and colleagues (2021) hence proposed that predictions based on the differences between users would lead to better learning outcomes if they were made in a more heterogeneous participant sample.

Other efforts using post-hoc simulation instead of an experimental approach investigated whether the cold start problem of an AFLS could be mitigated. Park, Joo, Cornillie, van der Maas, and Van den Noortgate (2019) tried to mitigate the cold start problem by predicting student learning ability for maths. Predictions were made by employing an item response theory model informed by previous performances of students and student characteristic. Park and colleagues (2019)'s simulation showed that learning outcomes would have increased when a student learning

ability level were a predicted value at the start of a learning session instead of a non-personalised default. In a similar vein, Pliakos and colleagues (2019) were able to predict the learning ability and response accuracy of new students and, in doing so, mitigate the cold start problem. Predictions were based on a combination of an item response theory model and an implementation of random forests.

The current study means to follow-up on the experiment of Van der Velde and colleagues (2021) by investigating the ability of their prediction methods to mitigate the cold start problem in a post-hoc simulation with a very large naturalistic sample. The fact that the sample is naturalistic provides greatly added value to the field of AFLS research, as using real data of this magnitude is rare and can give more insight into the use of AFLSs in an applied setting. Because of this, the current study's findings will closely reflect the effects of mitigating the cold start problem in the actual operating environment of AFLSs. Moreover, using the same AFLS as Van der Velde and colleagues (2021) granted the unique opportunity to compare lab findings with findings from the field. The use of such a sample puts this study into a position where it can take advantage of the natural diversity in secondary school students and the diversity in their teaching material. This way, Van der Velde and colleagues (2021)'s suggestion that student-related predictions could help mitigate the cold start problem given a diverse student sample could be falsified.

The AFLS that was used suffered from the cold start problem when it used the *default* method to determine the starting values of the rate of forgetting parameter. Namely, with the *default* method, all new facts or student were assigned the same rate of forgetting independent of the difficulty of the fact or the student's learning ability. To try to mitigate the cold start problem, five Bayesian prediction methods were applied in a post-hoc simulation. Four of these were also used by Van der Velde and colleagues (2021): the *domain*, *fact-level*, *student-level* and *hybrid* prediction methods (for detailed descriptions see section: '*Bayesian prediction*' under *Methods*). A *demographic* prediction method was added to find whether it was possible to make meaningful predictions for demographic groups based on prior performances of the students in each group.

It is expected that least one of these prediction methods should be able to mitigate the cold start problem. In other words, at least one method can predict the rates of forgetting and response times so accurately that they are closer to the observed rates of forgetting and response times following adaptation than the default starting value. Furthermore, it is expected that there is a relationship between a method's prediction accuracy and its granularity, whereby the more fine-grained the prediction method, the more accurate it is. For context, 1) the *default* method is the least fine-grained, 2) followed by the *domain* method, 3) *demographic*, 4) *fact-level* and *student-level*, 5) *hybrid*. The *hybrid* method is thus expected to result in the most accurate predictions for both rates of forgetting and response times. Lastly, it is expected that prediction accuracy will be higher if the prediction was informed by more observations.

In the following sections of this paper will be described: 1) the workings of the AFLS used in this study; 2) the primary and secondary hypotheses of this study; 3) information on the sample data in addition to the methods used. These include pre-processing, the post-hoc simulation, the various Bayesian prediction methods used and, finally, the analyses used to acquire the results; 4) the results of the analyses; 5) A discussion of the results regarding our hypotheses, a comparison with the findings of Van der Velde and colleagues (2021) and other studies, the implications of our findings as well as a note on future improvements and research.

The SlimStampen AFLS

The AFLS that was used to perform the current study was SlimStampen (see Sense, Behrens, Meijer, and van Rijn (2016), and Van Rijn, van Maanen, and van Woudenberg (2009)). The SlimStampen AFLS employs a cognitive model that is based on the ACT-R architecture Anderson (2009) to model a student's declarative memory. Specifically, the model keeps track of memory chunks that each represent a fact learned by a student. These chunks have a certain activation, which reflects how strongly they are represented in declarative memory. Whenever a given fact is encountered, the activation of the corresponding chunk increases. Over time, how-

ever, the activation decays. The activation A of chunk x at time t, with n previous encounters from $t_1, \ldots, t_n$ seconds ago, accounting for decay d is given by:

$$A_x(t) = ln\left(\sum_{j=1}^{n} t_j^{-d_x(t)}\right) \tag{1}$$

Before a new trial starts, the SlimStampen AFLS chooses which fact to present by calculating the activation of each chunk in declarative memory fifteen seconds in the future. Whichever fact's activation is lowest at that point in time will be presented in the upcoming trial. However, if no chunks are estimated to decay below a set retrieval threshold, a new, unknown fact will be presented instead. Furthermore, a fact cannot be presented in more than two successive trials. Through this procedure, a presentation schedule is built up whereby each fact's activation is kept above the retrieval threshold.

As some facts may be harder to memorise than others, the rate at which activation decays varies accordingly. The greater the decay, the faster activation decreases. The decay $d_x(t)$ of chunk x at time t is given by the chunk's activation during its most recent encounter and its rate of forgetting α:

$$d_x(t) = c * e^{A_x(t_{n-1})} + \alpha_x \tag{2}$$

In the current SlimStampen AFLS, the rate of forgetting of each fact starts at 0.3, following the *default* method. After three repetitions of a given fact, its rate of forgetting is adapted according to the response time and the correctness of a student's answer. Namely, the observed response time is compared with an expected response time for that trial. It is given by the ACT-R equation for retrieval time Anderson (2009), where t_0 denotes a fixed amount of time needed for perception of the prompt and the motor processes involved in giving the answer:

$$\mathbb{E}(RT) = e^{-A_x} + t_0 \tag{3}$$

The further the observed response time is removed from the expected response

time, the more strongly the rate of forgetting will be adapted. If the observed response time in a given trial is lower than expected, the true activation of the prompted fact in declarative memory must be higher than the model had thus far assumed. The rate of forgetting is then adapted downward in order to achieve more accurate response time and activation estimates in future trials. Vice versa, when the observed response time is longer than expected, the rate of forgetting is adjusted upward. For incorrect answers, the observed response time was taken as 1.5 times the expected response time. Since rate of forgetting adapts according to the performance on each fact, it gives an indication of each fact's difficulty as experienced by the student. Note that this adaptation takes into account the last five encounters with a fact to reduce the influence of outlier observations.

Through the cognitive model described above, the SlimStampen AFLS is able to harness two proven effects that improve student learning outcomes. Firstly, the testing effect Van den Broek et al. (2016) is invoked as the prompts of the AFLS stimulate active recall of facts. Compared to passive learning methods, for instance reading both the word and its translation from a list, active recall leads to better learning outcomes. However, the testing effect is less prominent when recall is unsuccessful van den Broek, Segers, Takashima, and Verhoeven (2014). The SlimStampen AFLS therefore makes sure all facts are encountered again before their activation crosses the retrieval threshold, increasing the chances of successful recall.

Secondly, the AFLS makes use of the spacing effect Dempster (1988); Ebbinghaus (1885) which states that longer periods between recall of a given fact reinforces its place in declarative memory. Namely, the AFLS repeats facts at the latest possible moment where the student is still likely to know the answer given, before the memory activation of a fact is expected to decay below the retrieval threshold. It should once again be noted that the adaptiveness of the AFLS only comes into play after a fact has been encountered at least three times. A default value of $\alpha = 0.3$ is used as the starting rate of forgetting for new facts irrespective of their difficulty or a student's learning ability. Naturally, this default starting rate of forgetting makes it so that easy facts are presented too often while difficult facts are presented too

rarely during the first few encounters with a fact. This prevents optimal use of the spacing effect. Furthermore, the balance between a student's skill and the challenge they experience from the AFLS is likely distorted, leading to lower learning performance Engeser and Rheinberg (2008); Kennedy, Miele, and Metcalfe (2014). Here, we tried to mitigate the effect of the cold start problem of the AFLS introduced by the starting rate of forgetting by predicting starting rates of forgetting using various prediction methods (see section: 'Bayesian prediction' under Methods).

Methods

Dataset

177,074,411 trials of fact learning performance data from 135,105 Dutch secondary school students on 36,469 distinct facts were used to inform and validate our predictions. In total, the students performed 1,084,130 distinct learning sessions were performed where they answered 165 questions on average (SD = 86.7). The data were collected by Noordhoff, a publisher of teaching materials in the Netherlands, who distributed the SlimStampen AFLS to secondary schools during the academic years between 2018 and 2020. The AFLS was provided as an additional learning resource for students between age 11 and 17 from years 1 to 4 of Dutch secondary school. Different school levels were represented in the data with 48.1% of students from pre-vocational education (VMBO), 33.8% from general secondary education (HAVO) and 18.1% from pre-university education (VWO). Students used the AFLS to learn vocabulary for their courses in English, French and German, which made up 69.8%, 29.6% and 0.69% of trials, respectively. Due its relatively small presence in the dataset, data on German vocabulary were excluded, leaving English and French.

The privilege of having large naturalistic dataset provided us with the unique opportunity to validate or falsify previous research with findings based on an insurmountable amount of trials performed in an applied setting. Furthermore, post-hoc simulation enabled us to test all five Bayesian prediction methods against the *default* method on the same data, ruling out any variable influence of trial-to-trial noise and

environmental factors during data acquisition.

Pre-processing

The data was 'performance data' in the sense that it included variables such as response time, correctness and the answer given by a student. However, it did not contain the actual correct answer. In case of incorrect responses (7.2% of the data) it was thus unclear which fact had been questioned. For the most part, the fact which had been questioned could be deduced through unique within-session IDs of facts. Occasionally however, these IDs were ambiguous even within sessions, making it impossible to deduce what fact was questioned with absolute certainty. In such cases, the facts with ambiguous IDs were excluded from the data. In total, around 395,158 trials, or 0.34% of the data, were filtered out for this purpose.

The performance data also did not include any model parameters of the AFLS, such as activation or rate of forgetting. The AFLS needed to be re-run with the performance data as input to acquire these parameters. The model parameters from the AFLS were calculated as they were during the original learning sessions, using the *default* method. This process resulted in a dataset detailing the activation as well as the rate of forgetting in every single trial. Notably, the rates of forgetting from the last repetition of each fact in a learning session were distilled. These 'final' rates of forgetting represented the rates of forgetting of each fact that approached the true rate of forgetting of each fact, following the adaptations they underwent during a learning session. They were the input of the Bayesian models used in the prediction methods described in the following section.

Bayesian prediction

As a continuation of Van der Velde and colleagues (2021) research, the current study investigated the effect of various prediction methods to resolve the cold start problem, now in a large and varied naturalistic sample. In total, five prediction methods were tested to determine whether they could provide more suitable starting rates of forgetting compared to the *default* method's starting rate of forgetting of

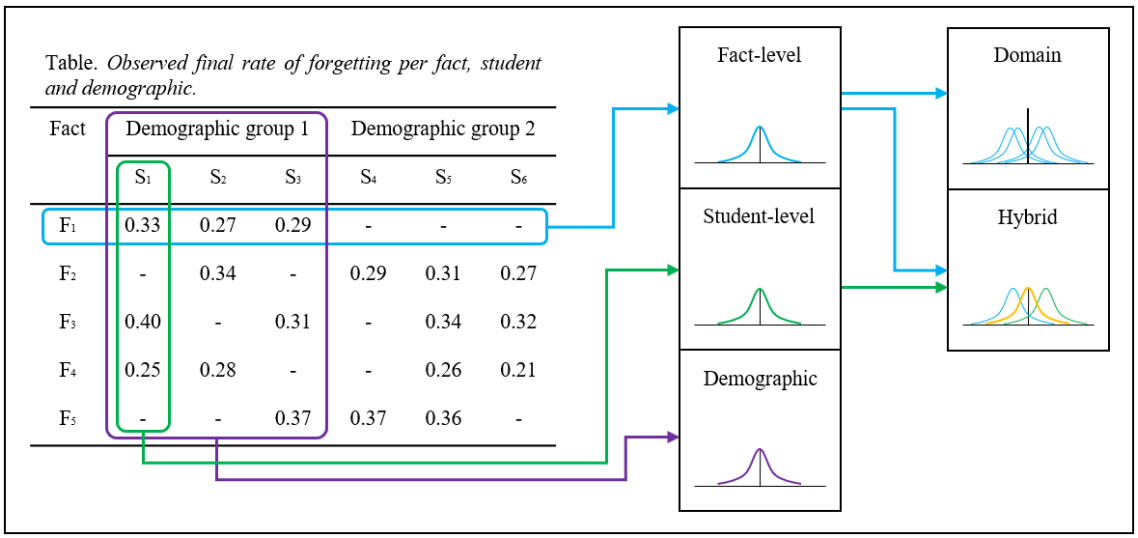

Figure 1: *(Left) The observed rates of forgetting from the final repetition of a given fact in a learning session, also known as final rates of forgetting, were grouped per fact, student or demographic for training the fact-level, student-level or demographic prediction methods respectively. (Right) For Bayesian prediction posterior predictive distributions were formed based on the aforementioned groups of final rates of forgetting. First, the mode of these distributions (indicated by the black vertical line) was taken as the predicted starting rate of forgetting for facts, students or demographic groups. Second, the domain prediction was obtained by taking the mean of all fact-level predictions. Hybrid predictions for each fact-student pair were obtained by taking the mode of a distribution gained from logarithmic opinion pooling the posterior distributions of a given fact and student.*

0.3 and compared to each other. All prediction methods, apart from *demographic* prediction, were also used in the study of Van der Velde and colleagues (2021). The prediction methods are described below in ascending order of granularity. Figure 1 provides a visual summary of the prediction methods and the Bayesian prediction process.

Default

A starting rate of forgetting of 0.3 is used for all facts and students. This method was used by the AFLS when the performance data was collected. The 0.3 default represented an average, commonly observed rate of forgetting Van Rijn et al. (2009). The AFLS is known to suffer from the cold start problem with this method.

Domain

The average of all predicted starting rates of forgetting for facts based on the previous performances of all students on a given fact (i.e. all *fact-level* predictions) was taken as a singular starting rate of forgetting for all facts. The *domain* prediction method aimed to capture learning performance with the domain of foreign language learning in a singular starting rate of forgetting. In this respect it is similar to the *default* method, since only one starting rate of forgetting is used for all facts. However, because it is based on previous observations through *fact-level* predictions, the *domain* prediction method should provide a more fitting starting rate of forgetting than the *default* method's 0.3.

Demographic

Groups separated on school level and grade received distinct starting rates of forgetting. The starting rates of forgetting were given by the modes of the posterior distributions of a Bayesian model (for more detail, see section 'Procedure' below) trained on the previous performances of all students in a given demographic. The *demographic* prediction method was added to the methods used by Van der Velde and colleagues (2021) because the demographic variation in the dataset allowed for

it and because it lies in-between *domain* prediction and *fact-level* and *student-level* prediction in terms of granularity. The *demographic* prediction method should highlight the differences in learning ability between demographic groups.

Fact-level

A distinct starting rate of forgetting was predicted for each fact based on the previous performances of all students on a given fact. *Fact-level* predictions should highlight the differences in difficulty between facts.

Student-level

Each student received a distinct starting rate of forgetting based on their previous performances. *Student-level* predictions should highlight the individual differences in learning ability between students.

Hybrid

Each student-fact pair received a distinct starting rate of forgetting based on a combination of the *fact-level* and *student-level* predictions for that particular student and fact. This method should result in more accurate predictions than the *fact-level* and *student-level* methods in case of an interaction between students' learning ability and fact difficulty.

Procedure

The abovementioned prediction methods were applied by means of a Bayesian prediction model. The model is grounded in the assumption that the rate of forgetting is normally distributed and that its mean and precision (reciprocal of the variance) are unknown. This distribution has a conjugate prior that follows a Normal-Gamma distribution. Following findings from earlier AFLS research Sense et al. (2016); Van Rijn et al. (2009), it was appropriate to let the prior distribution be weakly informative; normally centred around $\mu_0 = 0.3$ (with $\kappa_0 = 1$, $\alpha_0 = 3$, and $\beta_0 = 0.2$).

Because the prior is conjugate, the posterior also follows a Normal-Gamma distribution. Hereby the parameters of the posterior distribution may be inferred analytically through Bayesian statistics. This inference is computationally inexpensive, contrary to predicting methods that rely on sampling such as Markov chain Monte Carlo. Crucially, this would allow the AFLS to run smoothly for students when they were to use it in class or at home.

Inference occurred iteratively, as the most recent rate of forgetting from n data points together with the corresponding prior were taken to obtain a posterior distribution. With the distribution of parameters of the posteriors obtained through Bayesian inference, a posterior predictive distribution was formed. The posterior predictive distribution represents the probability that any rate of forgetting is observed on the next iteration of the model. The mode of this distribution represents the value that is most likely to be observed next. As such, the mode gives a rate of forgetting that should fit better than the current one, given the distribution of parameters. This is how the predicted starting rate of forgetting is obtained by the *demographic*, *student-level* and *fact-level* prediction methods. In the case of the *hybrid* prediction method, the posterior predictive distribution from both the *student-level* and the *fact-level* prediction methods were combined using logarithmic opinion pooling Genest (1984). The predicted *hybrid* rate of forgetting was the mode of the combined distribution.

The prediction process was 20-fold. In each fold the data was split in a training set and a test set of 95% and 5% of the data, respectively, in such a way that all data had been part of the test set once. The splits were made by randomly distributing all learning sessions over the test sets of the 20 folds. This way, accurate estimates of activation and response time could be made using the predicted starting rates of forgetting, for which whole learning sessions were required. Additionally, overlap of facts and students between the training and tests sets was more likely with a random order than a chronological order.

Predictions were informed only by final rates of forgetting from facts that had been seen more than three times by a student in a given learning session. This

was because final rates of forgetting from facts that had been shown three or fewer times would have had one or no opportunity to adapt. Such rates of forgetting are unlikely to approach a value that reflected the true rate of forgetting. However, very easy-to-learn facts oftentimes needed only three repetitions to be memorised. By excluding final rates of forgetting from facts with few repetitions, the training data is skewed towards a greater average fact difficulty. The decision to use a limit of three repetitions is a compromise in the dilemma where we cannot assume that final rates of forgetting based on few repetitions have approached the true rate of forgetting though many easy facts would be excluded otherwise. In total, 10,336,100 final rates of forgetting were used to inform the predictions.

Analyses

For analysis, the predictions made in all of the 20 folds were pooled together to reform full dataset. Because all prediction methods were simulated on the same performance data, direct comparison between the prediction methods was possible and any influences of environmental factors during data acquisition could be ruled out. The prediction methods were compared on their accuracy, which was measured by 1) the error between predicted starting rates of forgetting and observed final rates of forgetting and 2) the error between predicted response times and observed response times. All analyses were performed using R (version 3.6.3; R Core Team, 2020).

Bayesian linear regression models were fit to both metrics with prediction method included as a main effect using the brms R package (version 2.16.1; Bürkner, 2017). Both models were compared to an intercept-only variant using bridge sampling Gronau, Singmann, and Wagenmakers (2017). The resulting Bayes factors were interpreted following Jeffreys (1998), meaning a Bayes factor greater than 3 indicated substantial evidence for a difference between two models. Above 100, evidence for difference was considered to be very strong Jarosz and Wiley (2014). The superior model was inspected for the specific relationships between its factors.

As a follow-up, pairwise Bayesian t-tests were performed using the BayesFac-

tor R package (version 0.9.12-4.2; Morey, Rouder, Jamil, and Morey, 2015). The aim of the tests was to find whether any individual prediction method(s) performed better than the *default* method, and to find which method(s) achieved the most accurate predictions. The Bayesian t-tests compared the prediction methods within deciles of the observed rate of forgetting and the observed response time to gain additional insight into the methods' ability to predict both common and extreme values. Because this process consisted of many pairwise comparisons, Westfall's correction for multiplicity was applied Westfall, Johnson, and Utts (1997). Uncorrected Bayes factors that resulted from the Bayesian t-tests were transformed into posterior odds. Similar to Bayes factors, posterior odds were interpreted following Jeffreys (1998). Accordingly, posterior odds greater than 3 indicated substantial belief in a difference in predictive performance between methods. Whichever prediction method comparison favoured was determined by which method achieved the smallest mean pairwise difference in absolute prediction error. Note that all Bayes factors or posterior odds represented evidence, or belief for evidence, in favour of a difference between prediction methods.

For the above analyses on rate of forgetting prediction error, only the final repetition of a fact in a learning session was used. Entries with a final repetition of three or below were excluded because the final observed rate of forgetting could not be assumed to have approached the true rate of forgetting within the first three repetitions. Entries with a final repetition of 25 or higher were also excluded since students should always be able to fully memorise a word within 25 repetitions. They were assumed to not have actively tried to learn if they required that many repetitions. In all, 11,509,857 trials, each with six distinct predicted rates of forgetting, were part of the analyses.

For the analyses on response time prediction error, particular interest went out to the response time prediction error during the first three repetitions of a given fact. Since the starting rate of forgetting had not undergone any adaptation up to that point, there had been no chance for it to converge with the predictions of other methods. As such, the influence of the prediction methods would be the most

noticeable in the response times in first three repetitions. Unfortunately however, the cognitive model underlying the AFLS was unable to estimate the activation of a fact in declarative memory nor predict the response time from just one repetition. Moreover, observed response times in the second repetition were found to be unusually low, while response times in the third repetition were rather high. It turned out that the second repetition of a given fact almost always followed the first repetition or the trial after. It was assumed that students were subject to the recency effect Murre and Dros (2015) in the second repetition, whereby they were able to give the correct answer by retrieving it from their working memory instead of their declarative memory. In the third repetition, students were assumed to have retrieved the answer from declarative memory for the first time, with a higher average response time as a result. The analysed response time prediction errors were therefore only taken from the third repetition of each fact.

Furthermore, for the analyses on response time prediction error, incorrect trials, and trials where the observed response time was below 300 ms or above 25,000 ms, were excluded from the analyses. The lower response time limit reflected the minimal reading time instituted by the AFLS. Responses given faster than 300 ms were thus assumed to have been given without reading the prompt. Response times above 25,000 ms were assumed to indicate that the student was distracted. In total, 8,682,756 trials, each with six distinct predicted response times, were used in the analyses.

Finally, Bayesian regression models were fit on response time prediction error with main effects of prediction method and the number of final rates of forgetting that informed a prediction, as well as an interaction between these factors. A stepwise model comparison strategy using bridge sampling was employed to find the best fitting model. The coefficients of the superior model were then inspected for the specific relationships between its factors. Note that the Bayesian regression models on response time prediction error were constructed on a subset a hundredth of the size of the full dataset because the sheer size of the full dataset exceeded the available computing capacity. The contents of the subset were sampled randomly from the

full dataset. The subset was inspected to ascertain the relevant distributions in the full dataset were of similar make-up in the subset.

Results

Rate of forgetting prediction

As is visible in Figure 2A, there was less variance in the predicted starting rates of forgetting than in the observed rates of forgetting. Naturally, the *default* method had no variance at all, as all its predictions were 0.3. The *domain* method's predictions varied very little between folds, with predictions ranging between 0.3208 and 0.3209. In accordance with its level of granularity, the *demographic* method made more varied predictions than the *default* and *domain* methods, but less varied than the *fact-level*, *student-level* and *hybrid* methods. Predicted rates of forgetting were found to be lower as school level increased, with average starting rates of forgetting of 0.335, 0.319 and 0.311 for pre-vocational education, general secondary education and pre-university education, respectively. Further investigation using Bayesian linear regression revealed that there was no evidence for such a relationship, however, as model comparison using bridge sampling with an intercept-only model resulted in a Bayes factor of 0.049. Finally, the *fact-level*, *student-level* and *hybrid* methods made more varied predictions. The *fact-level* method's predictions displayed the largest variance, followed by the *hybrid* method.

Two Bayesian regression models were fitted on rate of forgetting prediction error. One model included prediction method as a predictor of rate of forgetting prediction error while the other was a simple intercept-only model. The two models were then compared using bridge sampling to find whether there was evidence for a relationship between rate of forgetting prediction error and prediction method. The model with prediction method as a predictor for rate of forgetting prediction error was found to better fit the data than an intercept-only model. Model comparison using bridge sampling resulted in a Bayes factor in favour of the model with prediction method that was so high that it was approximated to infinity. Model estimates

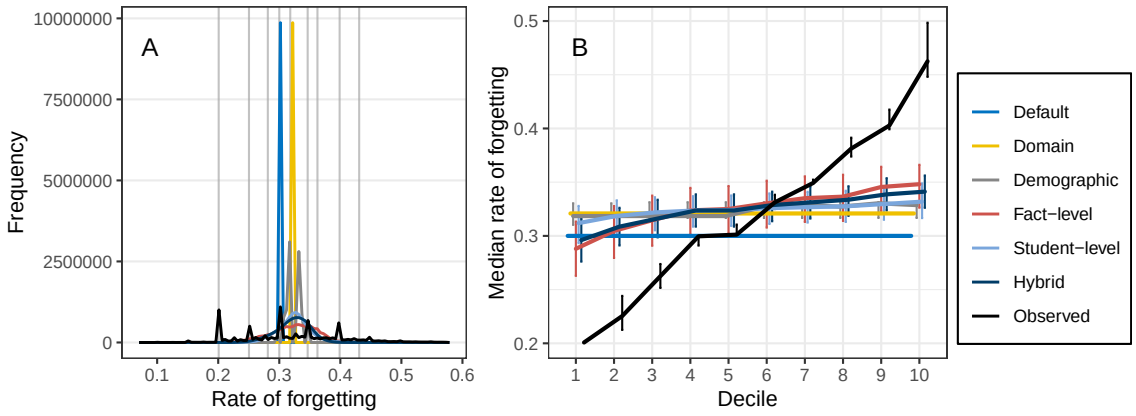

Figure 2: *A) Frequency distributions of the observed rate of forgetting and the predicted rates of forgetting per prediction method. The vertical lines indicate the decile borders. B) Line graph that displays the median observed rate of forgetting and the median predicted rates of forgetting per prediction method in each decile. The error bars denote interquartile ranges.*

and their posterior distributions are shown in Figure 3 (for the exact estimates see Table A1 in 'Appendix').

According to the Bayesian regression model, the absolute rate of forgetting prediction error of the *domain* method was significantly smaller than that of the *default* method. In turn, the *demographic* method had a significantly lower absolute prediction error than the *domain* method, and so on for the *student-level* method, the *hybrid* method and lastly the *fact-level* method. The *fact-level* and the *hybrid* method both showed the large deviations from the absolute prediction error of the *default* method, with estimates of -8.55e -03 (95% CI: [-8.59e -03, -8.51e -03]) and -8.39e -03 (95% CI: [-8.43e -03, -8.35e -03]), respectively, from an average prediction error of 0.0672 of the *default* method.

Further investigation using Bayesian t-tests uncovered individual differences in rate of forgetting prediction error between the prediction methods within ten deciles of the observed rate of forgetting. Bayes factors were corrected for multiple testing using the Westfall correction. The Bayes factors were transformed to posterior odds through multiplication with a prior odds value of 0.086 for 150 pairwise comparisons de Jong (2019); Westfall et al. (1997). Belief of evidence was found to be very strong for all individual differences in rate of forgetting prediction error. For almost all

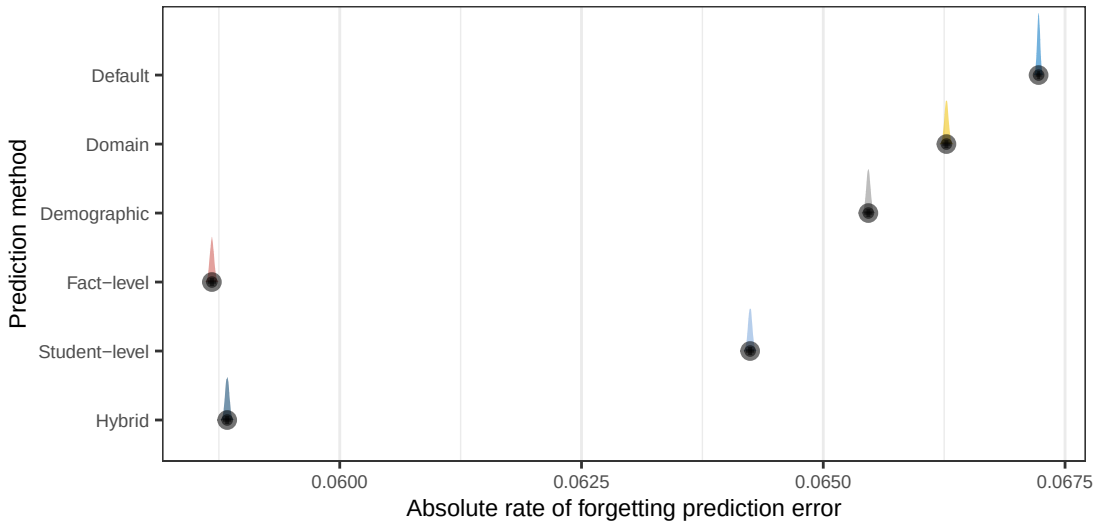

Figure 3: *Posterior distributions of the regression coefficients from the model on rate of forgetting prediction error that included prediction method as a predictor. The median of each distribution is indicated by a black dot. Confidence intervals are too small to be visible.*

comparisons the posterior odds were so high that they were approximated to infinity. The lowest posterior odds were recorded for the difference between the *domain* and *student-level* methods, at 2.03e +21, which still signified belief of very strong evidence. Going by the mean difference between the pairwise absolute prediction, the favoured prediction method could be determined for each comparison.

Overall, the median rate of forgetting prediction error of each prediction method was: 0.0531 (*default*), 0.0591 (*domain*), 0.0570 (*demographic*), 0.0496 (*fact-level*), 0.0554 (*student-level*), 0.0502 (*hybrid*). Within deciles, the prediction error was larger for the outer deciles as given by the distance (visible in Figure 2B) of the prediction methods' predictions to the observed rates of forgetting. Furthermore, Figure 2B shows how the *default* methods' predictions are generally the closest to the observed rates of forgetting in the first five deciles, but the farthest in the latter deciles. Moreover, the *fact-level* and *hybrid* methods, and to a lesser extent the *student-level* method, can be seen following the upwards trend of the observed rate of forgetting.

For further insight into the differences in rate of forgetting prediction error

Table 1: *Rankings of the methods' predictive performance on rate of forgetting per decile, as well as average of all decile rankings.*

Method	Deciles										Average
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	rank
Default	3	1	1	1	1	6	6	6	6	6	3.70
Domain	6	5	4	2	2	1	5	5	5	5	4.00
Demographic	5	6	6	3	3	2	2	4	4	4	3.90
Fact-level	1	2	2	6	6	5	3	1	1	1	2.80
Student-level	4	4	5	5	5	4	4	3	3	3	4.00
Hybrid	2	3	3	4	4	3	1	2	2	2	2.60

between prediction methods within deciles, however small they may be, a ranking of the methods is shown in Table 1. Most notable finding is that the *fact-level* method made relatively accurate predictions in the outer deciles, but it was relatively inaccurate in the middle deciles 4, 5 and 6. The opposite was true for the *domain* method. Moreover, the *hybrid* and *demographic* methods showed a similar, although weakened, trend to the aforementioned methods, respectively. Keep in mind, however, that the numeric differences between prediction methods are larger in the outer deciles. As was also made clear by their median prediction errors, the *hybrid* and *fact-level* methods performed much better than the other methods overall.

Response time prediction

Another two Bayesian regression models were fitted, this time on response time prediction error. Once again, one model included prediction method as a predictor and the other model only included the intercept. Model comparison using bridge sampling revealed that the model with prediction method as a predictor fit the data better, evidenced by an infinitely high Bayes factor in favour of it over the intercept-only model. Inspection of the model showed that, compared to an average response time error of 1814 ms of the *default* method, the *demographic* (b = -30.95, 95% CI: [-57.23, -4.37]), *fact-level* (b = -89.45, 95% CI: [-116.5, -62.20]), *student-level* (b = -41.97, 95% CI: [-68.36, -15.15]) and *hybrid* (b = -80.47, 95% CI: [-107.0, -53.68]) methods all resulted in significantly lower response time errors (see also Figure 5).

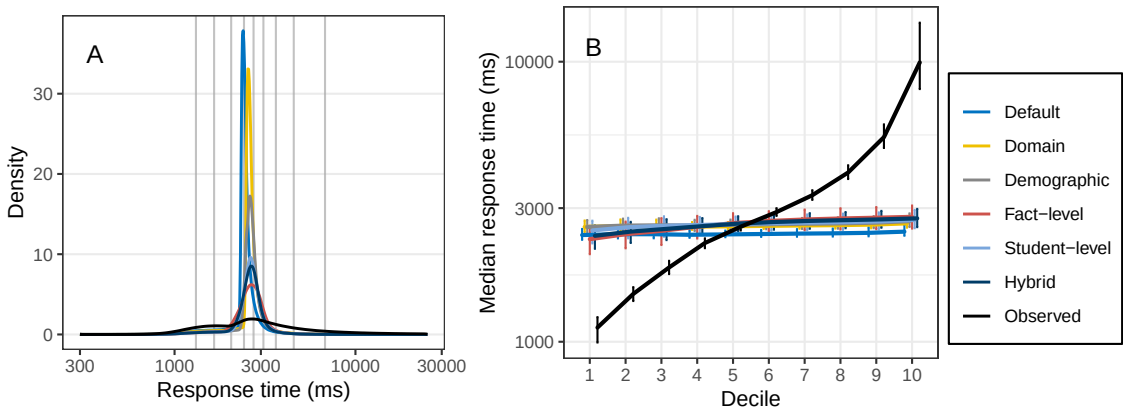

Figure 4: *A) Density functions of the observed response time and the predicted response times per prediction method. The vertical lines indicate the decile borders. B) The line graph displays the median observed response time and the median predicted response times per prediction method in each decile. The error bars denote interquartile ranges.*

The *fact-level* and *hybrid* methods were also found to have a significantly lower response time prediction error than the other methods.

Figure 4A shows that the observed response times varied more than any method's predicted response times, similar to the distributions of the predicted and observed rates of forgetting depicted in Figure 2A. Moreover, the *fact-level* method once again had the largest variance in predictions, followed by the *hybrid* method and then the *student-level* method. Figure 4B also shows many similarities with Figure 2B, as the *default* method's predictions generally are lower than those of other methods and both the *fact-level* and *hybrid* methods slightly follow the upward trend of the observed response times. Bayesian pairwise t-tests were carried out to find evidence for differences in absolute response time prediction error between the prediction methods in each quantile. The Bayes factors were transformed according to the Westfall correction for multiplicity into posterior odds through multiplication with a prior odds value of 0.086 for 150 pairwise comparisons. The Bayesian t-tests showed that there was very strong belief for all differences except for the difference between the *domain* method and the *demographic* method in decile 2 and the difference between the *domain* method and the *student-level* method in decile 3.

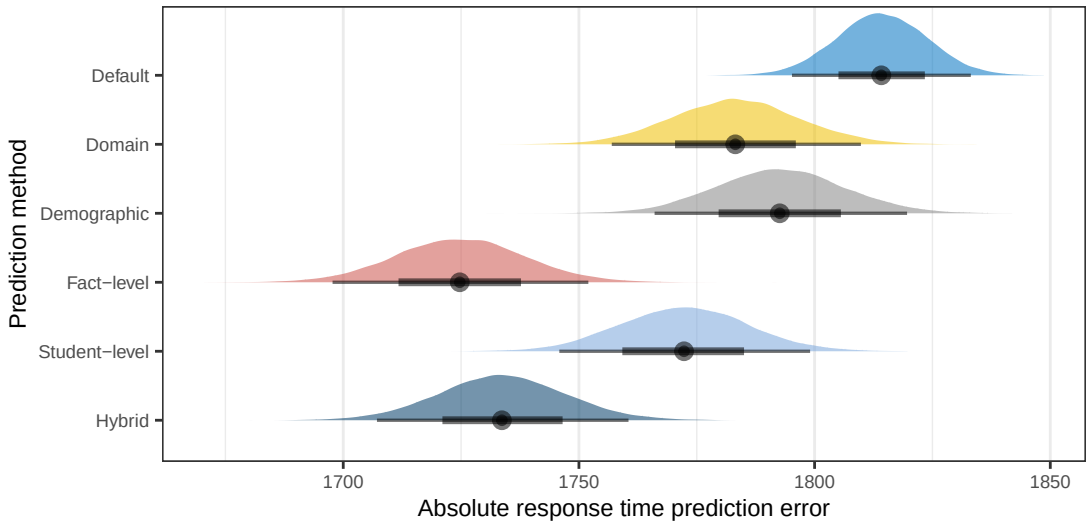

Figure 5: *Posterior distributions of the regression coefficients from the model on response time prediction error that included prediction method as a predictor. The median of each distribution is indicated by a black dot. The thick and thin black lines indicate the 66% and 95% confidence intervals, respectively. Note that compared to the model on rate of forgetting prediction error, this model was constructed on a subset of 1/100 of the data.*

Table 2: *Rankings of the methods' predictive performance on response time per decile, as well as average of all decile rankings.*

Method	Deciles										Average
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	rank
Default	2	1	1	1	1	6	6	6	6	6	3.80
Domain	6	6*	4*	2	1	4	5	5	5	5	4.30
Demographic	5	5*	6	4	2	2	4	4	4	4	4.05
Fact-level	1	2	2	5	6	5	1	1	1	1	2.50
Student-level	4	4	5*	6	5	3	3	3	3	3	3.85
Hybrid	3	3	3	3	4	1	2	2	2	2	2.50

*No belief for evidence was found for the difference between the two marked methods. For the average rank, both methods received the average of their rankings in that decile(e.g. both methods get 4.5 counted towards their average rank if there was no belief for a difference between their ranks 4 and 5).

Based on the results from the Bayesian t-tests a ranking could be made for each decile to show how the differences in absolute response time prediction error were reflected in the data. These rankings are shown in Table 2. Once again, the *default* method was found to perform relatively well in the first five decile but poorly in the last five. The *fact-level* method predicted accurate response times in the outer deciles, but not in the inner deciles. The opposite behaviour of the *domain* method as well as the similar, but weaker respective trends of the *hybrid* and *demographic* methods were also found in the ranking of response time errors. On average, the *fact-level* and *hybrid* methods performed equally well and far better than the other prediction methods.

The finding that the *student-level* method generally did not result in more accurate predictions than the *default* method prompted a follow-up investigation into the variance in learning ability between students in the current study and in the study of Van der Velde and colleagues (2021). Distributions of student learning ability, given by their average rate of forgetting, from both studies are visualised in Figure 6. Surprisingly, the variance in learning ability between students was lower in the current study ($\sigma^2_{current} = 8.44\text{e}-04$; $\sigma^2_{VanDerVelde} = 1.41\text{e}-03$), even though our sample consisted of secondary school students from distinct school levels and years while Van der Velde and colleagues (2021)'s sample consisted of first-year psychology students.

Amount of data per prediction and prediction error

Additional Bayesian regression analyses were performed to find whether response time predictions made based on small amounts of data were less accurate than predictions based on many observations. A Bayesian regression model with prediction method and the amount of data used to make each prediction as main effects and an interaction effect between the two was fit on response time prediction error. The amount of data that was used to make each prediction was given by the number of final rates of forgetting that informed each rate of forgetting prediction which in turn influenced response time predictions in trials with respective facts, students,

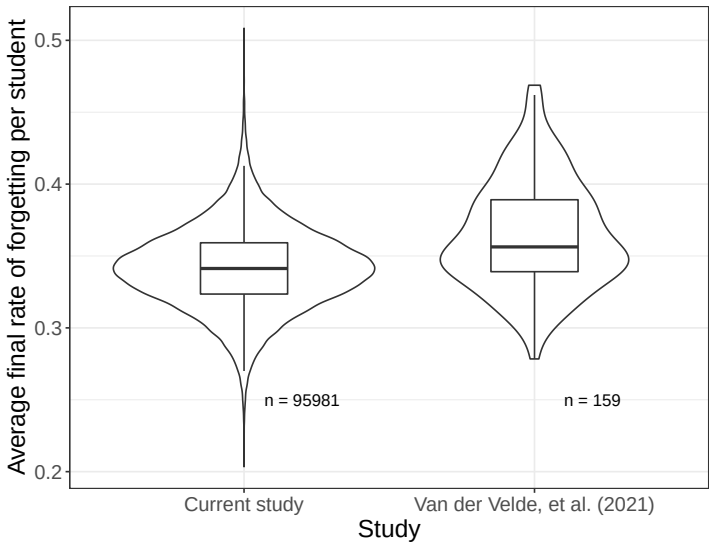

Figure 6: *Boxplot with distributions of student learning ability, given by their average rate of forgetting, in the current study and in Van der Velde, et al. (2021).*

domains and demographics.

The distribution of the number of final rates of forgetting that informed each prediction varied greatly between prediction methods (e.g. the median number of final rates of forgetting that informed a prediction was 149 for the *student-level* method, 1264 for the *hybrid* method and over 1.3 million for the demographic method; see also Figure 7). To accommodate for this and so facilitate a more interpretable comparison of the prediction methods, the number of final rates of forgetting that informed a prediction was standardised within each prediction method. The *default* method was excluded from the analysis because it does not make predictions and hence was not informed by any number of datapoints. Stepwise model comparison using bridge sampling revealed that the full model and a model that included only the main effect of prediction method and the interaction between prediction method and the amount of data used fit the data better than less complex models. Namely, there was very strong evidence in favour of the full model and the model without the main effect of the amount of data used over an intercept-only model, given by Bayes factors of 4.44e +07 and 4.53e +07, respectively. The Bayes factor in favour of the full model over the model without the main effect of the amount of data used was 0.997, indicating there was no evidence for a difference in

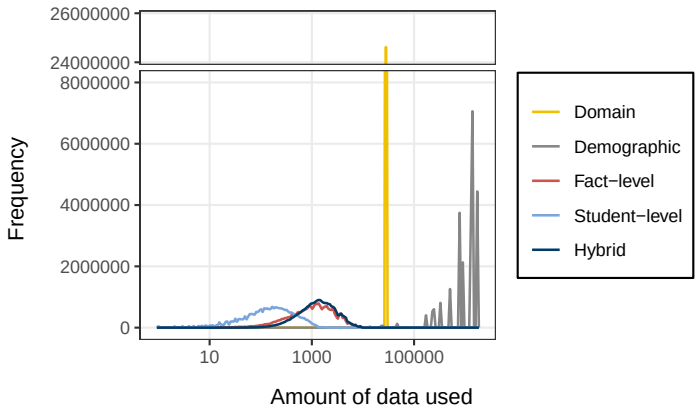

Figure 7: *Frequency distributions of the number of observations that informed each prediction per prediction method*

fit between the models. Thus, according to Occam's razor, the model without the main effect of the amount of data used to make each prediction was preferred over the full model.

Inspection of the model itself revealed that only the *fact-level* and *hybrid* methods had a significantly lower response time prediction error than the *domain* method. Moreover, they were also the only prediction methods that showed a significant interaction effect with the amount of data used to make each prediction. The interaction coefficient of the *fact-level* method (-20.77; 95% CI: [-39.39, -2.04]) indicated that the absolute response time prediction error of the *fact-level* method decreased by an estimated 20.77 ms for every standard deviation of datapoints used (SD = 1301) over the mean amount of data used by the *fact-level* method (m = 1405). For the *hybrid* method, the interaction coefficient (-20.99; 95% CI: [-39.73, -2.49]) meant that the absolute response time prediction error of the *hybrid* method decreased by an estimated 20.99 ms for every standard deviation of datapoints used (SD = 1319) over the mean amount of data used by the *hybrid* method (m = 1632). The regression coefficients of the best fitting model are given in Table 3 and visualised with their posterior distributions in Figure 8.

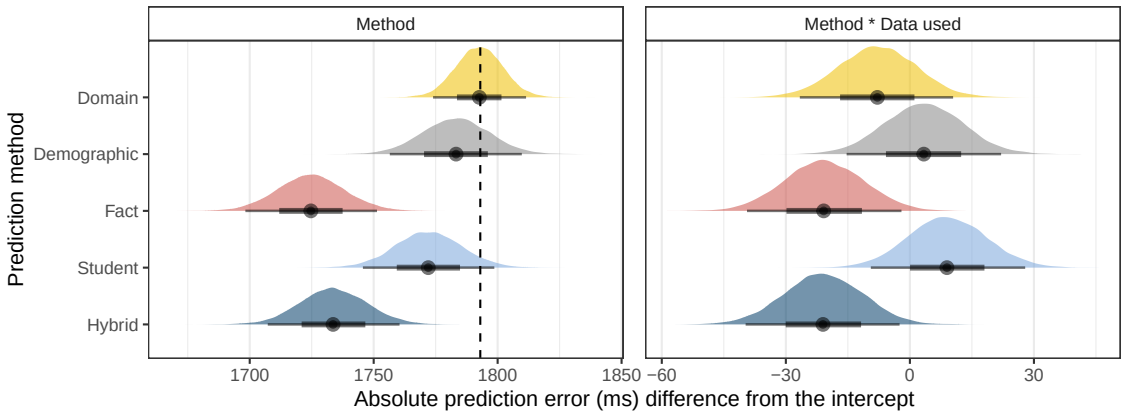

Figure 8: *Posterior distributions of the regression coefficients from the best fitting model. The median of each distribution is indicated by a black dot. The thick and thin black lines indicate the 66% and 95% confidence intervals, respectively. The coefficients of the main effect of prediction method are shown on the left-hand side and the interaction coefficients on the right-hand side. The coefficients on the right are relative to the intercept given by the main effect coefficient of the domain prediction method.*

Discussion

Five distinct Bayesian prediction methods were trained on a large naturalistic sample to predict AFLS parameters in order to try to mitigate the cold start problem of the AFLS. Currently, the AFLS employs a *default* method that assigns a rate of forgetting of 0.3 to every fact a student is yet to learn, regardless of its difficulty or the student's learning ability. The prediction methods were compared with the *default* method as well as each other on their prediction accuracy of the AFLS's rate of forgetting parameter and response times. The aim of the current study was to find evidence that one or more of the five proposed prediction methods approached observed response times and the rates of forgetting observed at the end of a learning session more closely than the *default* method and, in doing so, could mitigate the cold start problem. Further investigation into the relationship between the granularity of a prediction method and its prediction accuracy as well as the relationship between prediction accuracy and the number of observations that informed each prediction was done to find which factors influenced the performance of a prediction method.

Taken together, the results suggest that the cold start problem can likely be mitigated when AFLS parameters are predicted with Bayesian prediction methods informed by parameter values obtained through prior use of the AFLS. Specifically, the rate of forgetting parameter of the SlimStampen AFLS could be predicted so that the rate of forgetting at the start of a learning session closer to the true rate of forgetting than a default value of 0.3 was. Furthermore, several prediction methods could also predict response times more accurately than the *default* method. However, only the *fact-level* method and the *hybrid* method consistently achieved higher rate of forgetting prediction accuracy and response time prediction accuracy than the *default* method in all analyses. By employing a *fact-level* or a *hybrid* prediction method instead of the *default* method, the starting rate of forgetting assigned to a new fact or a new student thus more accurately reflected the difficulty of that fact and, in case of the *hybrid* method, also the learning ability of that student. The higher accuracy would allow the AFLS to provide its users with a personalised learning experience from the start of a learning session.

Looking at their performance within rate of forgetting deciles as well as response time deciles, the *fact-level* and the *hybrid* methods were found to achieve relatively good predictions for extreme observations, although they were relatively poorly in predicting more common observations. When pondering over this finding, one should keep in mind that the absolute difference in prediction error for both rate of forgetting and response times was much larger for extreme observations than for common observations. For the purpose of mitigating the cold start problem it is favourable to accurately predict extreme observations because that is where the greatest gain lies. For example, a default starting rate of forgetting of 0.3 might need only one adaptation to reach the true rate of forgetting of 0.32, but it will need at least five to reach 0.52. A *fact-level* method might make a worse prediction of 0.37 for a common observation of 0.32, requiring two adaptations. The extreme observation will be predicted better at, say, 0.46, also requiring two adaptations. This scenario illustrates how *fact-level* predictions lead to faster approximation of

the true rate of forgetting, as it is roughly accurate overall rather than only accurate for common observations. Using a *fact-level* or *hybrid* prediction method, an AFLS will be able to personalise and optimise the learning experience faster.

Even though it was impossible to measure the effect of cold start mitigation on learning outcomes due to the inflexible nature of the post-hoc simulation, extrapolation from earlier findings suggests that by applying a Bayesian *fact-level* or *hybrid* prediction method to an AFLS may lead to greater learning outcomes (Van der Velde et al., 2021). Furthermore, user frustration with the learning system may be reduced whereby abandonment of the system becomes less likely Pliakos et al. (2019). Mitigating the cold start problem may also increase users' motivation to learn Wauters, Desmet, and Van Den Noortgate (2010), which could be essential for students who are less receptive to traditional teaching methods Prensky (2010).

The question remains whether the most fine-grained prediction method was the most accurate. As the method in question, the *hybrid* method was generally found to perform on par with the more coarse-grained *fact-level* method. Both were found to predict rates of forgetting and response times better than the *default* method. Furthermore, both methods achieved the same average rank on response time prediction error divided by decile. On rate of forgetting prediction error divided by decile, the *hybrid* method achieved a marginally higher average rank. Analyses investigating the interaction between prediction method and the number of observations that informed each prediction revealed that both the *fact-level* method and the *hybrid* method were equally superior to the *domain* method in terms of response time prediction error. Both methods also interacted similarly with the number of observations that informed each prediction.

To break the stalemate, the way in which the two prediction methods operate should be considered. Namely, the *hybrid* method makes its predictions by taking the mode of a logarithmically pooled distribution made up of the posterior of a given *fact-level* and the posterior of a given *student-level* prediction. From figures 2A and 4A as well as the ranking tables becomes clear that *hybrid* predictions are less varied

than *fact-level* predictions. Since *hybrid* predictions are drawn from pooled *fact-level* and *student-level* posteriors, it is plausible that *hybrid* predictions reflect the more extreme *fact-level* predictions that have tempered by *student-level* predictions. For the purpose of mitigating the cold start problem, however, accurate predictions of extreme observations are very valuable. Arguably then, the *hybrid* method's good performance may largely be attributed to *fact-level* predictions. Accordingly, it was concluded that, since the *fact-level* method is more coarse-grained than the *hybrid* method, the hypothesis that the most fine-grained prediction method would make the most accurate predictions was rejected.

These results are congruent with the findings of Van der Velde and colleagues (2021). They found that the cold start problem could be mitigated if predictions were made following a *fact-level* method, be it only when the variance of fact difficulty was large enough.

The *fact-level* method is recommended for future application for the above reason. Moreover, the *fact-level* method is much less computationally intensive compared to the *hybrid* method. This practical advantage makes it the ideal candidate for cold start mitigation in future AFLSs.

The *domain*, *demographic* and *student-level* methods had a lower absolute prediction error on rate of forgetting than the *default* method, although they achieved a similar average rank to the *default* method when comparisons were made within deciles of the observed rates of forgetting. Additionally, the *demographic* and *student-level* methods were revealed to have a lower absolute prediction error on response times than the *default* method, though they again achieved a similar average rank to the *default* and *domain* methods over deciles of the observed response times.

Within deciles, the *domain*, *demographic* and *student-level* methods made less accurate predictions than the *default* method for the lower half of the observed rates of forgetting as well as the lower half of the observed response times, though they were more accurate for both upper halves. The main reason for this finding is that

all prediction methods predicted higher median rates of forgetting and response times than the *default* method in all deciles, except for the *fact-level* and *hybrid* methods in decile 1. This upwards shift from the median predictions of the *default* method may be attributed to the exclusion of final rates of forgetting from facts that had been repeated less than four times from the training data, whereby the predictions were biased towards a higher difficulty of facts, expressed in a higher rate of forgetting and response time.

However, that does not explain why the *domain*, *demographic* and *student-level* methods performed worse than the *fact-level* and *hybrid* methods. Firstly, the lack of rate of forgetting prediction accuracy of the *domain* method may be because it is too shallow. Since only one prediction was made for all facts, the *domain* method could not account for, nor take advantage of, the variety in the dataset.

Secondly, the *demographic* method's underperformance may be attributed to the fact that there was no traceable difference in the rate of forgetting between demographic groups. Surprisingly, learning ability did not seem to increase as students advanced to higher school years. In contrast to the *domain* method, there was no variety for the *demographic* method to take advantage of.

Thirdly, the *student-level* method's underperformance could be explained by the observation that there were less observations to inform each prediction compared to the *fact-level* method. Yet no evidence was found for that increasing the number of observations that informed a *student-level* prediction would increase prediction accuracy. Further investigation revealed, however, that although the current study had access to data from more than 135,000 Dutch secondary school students from multiple school levels and grades and Van der Velde and colleagues (2021) tested 159 first-year psychology students, there was less variance in learning ability between students in the current study. Our findings thus agree with Van der Velde and colleagues (2021)'s suggestion that insufficient variance in the participant sample prevented a *student-level* prediction method from making meaningful predictions. Crucially however, the current study revealed that even with a large naturalistic pool of participants from various backgrounds there was insufficient vari-

ance in learning ability for the *student-level* method to capitalise on, suggesting that it might be impossible for a *student-level* method to mitigate the cold start problem in a natural setting.

Overall, the results revealed close similarities in how a method performed on rate of forgetting prediction and how they performed on response time prediction. It is not surprising that there would be a relationship between these parameters as response time predictions took into account the activation of the fact in declarative memory, and thereby by extent the rate of forgetting. Crucially however, this relationship shows that when the latent model parameter rate of forgetting can be predicted more accurately, the more tangible parameter response time can also be predicted with greater accuracy, which validates the cognitive model underlying the AFLS.

Finally, it was investigated whether response time predictions partially derived from predicted rates of forgetting based on many observations in the training set were more accurate than those derived from rate of forgetting predictions made that were based on few observations. Logically, predictions that were informed by many observations are more robust against outliers and predictions that are heavily influenced by outliers are likely to be inaccurate. The results showed that there was evidence for an interaction effect between prediction method and the number of observations each method used to make its predictions. Concretely, only the *fact-level* and *hybrid* methods were found to predict significantly more accurate response times the more observations had informed their predictions. However, for both methods for the absolute response time error to go down by 1.21% the number of observations for a prediction had to almost double. This marginal gain suggests that this interactive relationship is not practically relevant here. Following the law of diminishing returns, predictions based on much smaller datasets would likely show a more pronounced effect. For future reference, none should strive to amass larger amounts of data in the hope of increasing prediction accuracy when data is available to the

same order of magnitude as in the current study.

The current study also recognises the limitations that were introduced by the methodological procedures employed. The use of a post-hoc simulation as opposed to experimental testing meant that all training data was gathered by the SlimStampen AFLS while it employed the *default* method. This meant that the sequence in which facts were presented to students during data acquisition was set in stone. For example, a student is presented with a difficult fact for the first time (say, the true rate of forgetting is 0.4). An AFLS using the *default* method would apply a starting rate of forgetting of 0.3, whereas it may have applied a starting rate of forgetting of 0.35 if it used the *fact-level* method. With a higher starting rate of forgetting the difficult fact is repeated more often in the early stages of the learning session. Not only does this mean that the presentation schedule is altered, the student may also fully memorise the fact in fewer repetitions because of the better fitting presentation schedule. Because of the property of post-hoc simulation that the original presentation schedule cannot be changed, any potential side effects of cold start mitigation could not be measured. Pliakos and colleagues (2019) already suggested that the advantages of mitigating the cold start problem with accurate predictions are more plentiful than increased personalisation learning environment at the start of a learning session.

Another limitation was introduced by the luxury of having a very large dataset. Namely, even for the smallest of differences between prediction methods there often was overwhelming evidence. While corrections accounting for the number of analyses were performed to prevent type 1 errors, very small differences in prediction errors between methods should be adopted critically. For the reason that some of these effects may not be replicable, we focused on effects which were consistently clear from all analyses.

Ultimately, the findings of the current study and those of Van der Velde and colleagues (2021) may be used as a foundation for future research into mitigating

the cold start problem of AFLSs. The current findings pave the way for a long-term applied experiment where the *fact-level* prediction method is applied in a naturalistic setting. Using the *default* method as a control condition, the effects of mitigating the cold start problem by predicting fact difficulty may be investigated beyond the limitations of the current study. Namely, while the post-hoc simulation lent itself to perform analyses of the predictive performance of five distinct prediction methods on an unprecedented scale, it was not possible to measure any side effects when these prediction methods were used to mitigate the cold start problem. Investigations into the effect of cold start mitigation through Bayesian prediction on learning outcomes, learning motivation and system abandonment could emphasize the advantages of AFLSs. Moreover, we suggest a theory for improvement of the cold start mitigation achieved in the current study. Namely, the prediction accuracy of a mitigating prediction method like the *fact-level* method is expected to increase when it is informed by observations from a system using such a prediction method. As there is less need for large adaptations when the cold start problem is mitigated, an AFLS may be able to hone in more precisely on the actual value of the AFLS's latent parameter. Consequentially, observations that more accurately reflect this actual value are expected to produce more accurate predictions. In this way, there is potential for a positive feedback loop, increasing prediction accuracy up to its maximum, given the natural shifts and unpredictability of cognition.

Conclusion

The cold start problem hampers an AFLS's ability to provide users with a personalised learning experience that increases their learning outcomes at the start of a learning session. Through a post-hoc simulation on a large naturalistic dataset, the current study found that the cold start problem could be mitigated by predicting the starting rate of forgetting parameter value of the AFLS's underlying cognitive model prior to the start of a learning session. Five distinct Bayesian prediction methods were tested on their prediction accuracy for rates of forgetting and response times. From these five, the *fact-level* and *hybrid* prediction methods were found to consis-

tently make more accurate predictions than the *default* method, which the AFLS employed when it suffered from the cold start problem. Out of the two, the more coarse-grained *fact-level* was deemed the best candidate to mitigate the AFLS's cold start problem in an applied setting. Based on previous studies, the application of this prediction could lead to a further increase in learning outcomes facilitated by AFLSs, as well as an increase in learning motivation. In all, the current study understates the importance of the use of advanced digital learning aids next to traditional teaching methods to improve the quality of education.

References

Alshammari, M., Anane, R., & Hendley, R. J. (2016). Usability and effectiveness evaluation of adaptivity in e-learning systems. In *Proceedings of the 2016 chi conference extended abstracts on human factors in computing systems* (pp. 2984–2991).

Anderson, J. R. (2009). *How can the human mind occur in the physical universe?* Oxford University Press.

Bürkner, P.-C. (2017). brms: An r package for bayesian multilevel models using stan. *Journal of statistical software*, *80*(1), 1–28.

Dahlstrom, E., Brooks, D. C., & Bichsel, J. (2014). The current ecosystem of learning management systems in higher education: Student, faculty, and it perspectives.

de Jong, T. (2019). A bayesian approach to the correction for multiplicity.

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.

Ebbinghaus, H. (1885). *Über das gedächtnis: untersuchungen zur experimentellen psychologie*. Duncker & Humblot.

Engeser, S., & Rheinberg, F. (2008). Flow, performance and moderators of challenge-skill balance. *Motivation and Emotion*, *32*(3), 158–172.

Ennouamani, S., & Mahani, Z. (2017). An overview of adaptive e-learning systems. In *2017 eighth international conference on intelligent computing and*

information systems (icicis) (pp. 342–347).

FitzPatrick, T. (2012). Key success factors of elearning in education: A professional development model to evaluate and support elearning. *Online Submission*.

Genest, C. (1984). A characterization theorem for externally bayesian groups. *The Annals of Statistics*, 1100–1105.

Gronau, Q. F., Singmann, H., & Wagenmakers, E.-J. (2017). bridgesampling: An r package for estimating normalizing constants. *arXiv preprint arXiv:1710.08162*.

Jarosz, A. F., & Wiley, J. (2014). What are the odds? a practical guide to computing and reporting bayes factors. *The Journal of Problem Solving*, *7*(1), 2.

Jeffreys, H. (1998). *The theory of probability*. OUP Oxford.

Jeong, H. I., & Kim, Y. (2017). The acceptance of computer technology by teachers in early childhood education. *Interactive Learning Environments*, *25*(4), 496–512.

Kennedy, P., Miele, D. B., & Metcalfe, J. (2014). The cognitive antecedents and motivational consequences of the feeling of being in the zone. *Consciousness and cognition*, *30*, 48–61.

Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of educational psychology*, *106*(4), 901.

Morey, R. D., Rouder, J. N., Jamil, T., & Morey, M. R. D. (2015). Package 'bayesfactor'. *URLh http://cran/r-projectorg/web/packages/BayesFactor/BayesFactor pdf i (accessed 1006 15)*.

Murre, J. M., & Dros, J. (2015). Replication and analysis of ebbinghaus' forgetting curve. *PloS one*, *10*(7), e0120644.

Nkambou, R., Mizoguchi, R., & Bourdeau, J. (2010). *Advances in intelligent tutoring systems* (Vol. 308). Springer Science & Business Media.

Park, J. Y., Joo, S.-H., Cornillie, F., van der Maas, H. L., & Van den Noortgate, W. (2019). An explanatory item response theory method for alleviating the cold-start problem in adaptive learning environments. *Behavior research methods*,

51(2), 895–909.

Pavlik Jr, 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.

Pliakos, K., Joo, S.-H., Park, J. Y., Cornillie, F., Vens, C., & Van den Noortgate, W. (2019). Integrating machine learning into item response theory for addressing the cold start problem in adaptive learning systems. *Computers & Education*, *137*, 91–103.

Prensky, M. R. (2010). *Teaching digital natives: Partnering for real learning*. Corwin press.

R Core Team. (2020). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

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.

Steenbergen-Hu, S., & Cooper, H. (2014). A meta-analysis of the effectiveness of intelligent tutoring systems on college students' academic learning. *Journal of educational psychology*, *106*(2), 331.

Travers, C. (2017). Current knowledge on the nature, prevalence, sources and potential impact of teacher stress. *Educator Stress*, 23–54.

Van den Broek, G., Takashima, A., Wiklund-Hörnqvist, C., Wirebring, L. K., Segers, E., Verhoeven, L., & Nyberg, L. (2016). Neurocognitive mechanisms of the "testing effect": A review. *Trends in Neuroscience and Education*, *5*(2), 52–66.

van den Broek, G. S., Segers, E., Takashima, A., & Verhoeven, L. (2014). Do testing effects change over time? insights from immediate and delayed retrieval speed. *Memory*, *22*(7), 803–812.

van der Velde, M., Sense, F., Borst, J., & van Rijn, H. (2021). Alleviating the cold start problem in adaptive learning using data-driven difficulty estimates.

Computational Brain & Behavior, *4*(2), 231–249.

VanLehn, K. (2006). The behavior of tutoring systems. *International journal of artificial intelligence in education*, *16*(3), 227–265.

Van Rijn, H., van Maanen, L., & van Woudenberg, M. (2009). Passing the test: Improving learning gains by balancing spacing and testing effects. In *Proceedings of the 9th international conference of cognitive modeling* (Vol. 2, pp. 7–6).

Wauters, K., Desmet, P., & Van Den Noortgate, W. (2010). Adaptive item-based learning environments based on the item response theory: Possibilities and challenges. *Journal of Computer Assisted Learning*, *26*(6), 549–562.

Webb, S., & Nation, P. (2017). *How vocabulary is learned*. Oxford University Press.

Westfall, P. H., Johnson, W. O., & Utts, J. M. (1997). A bayesian perspective on the bonferroni adjustment. *Biometrika*, *84*(2), 419–427.

Appendix

Table A1: *Bayesian regression coefficients of the model fit on absolute rate of forgetting prediction error that included a main effect of prediction method.*

Method	Estimate	95% Confidence interval
Default	0.0672	[0.0672, 0.0673]
Domain	-0.000953	[-0.000995, -0.000911]
Demographic	-0.00176	[-0.00180, -.000172]
Fact-level	-0.00855	[-0.00859, -0.00851]
Student-level	-0.00298	[-0.00303, -0.00294]
Hybrid	-0.00839	[-0.00843, -0.00835]

Table A2: *Bayesian regression coefficients of the model fit on absolute response time prediction error that included a main effect of prediction method.*

Method	Estimate	95% Confidence interval
Default	1814	[1795, 1833]
Domain	-21.54	[-48.14, 5.42]
Demographic	-30.95	[-57.23, -4.37]
Fact-level	-89.45	[-116.5, -62.20]
Student-level	-41.97	[-68.36, -15.15]
Hybrid	-80.47	[-107.0, -53.68]

Code availability

The code required to reproduce the prediction process and analysis is available upon request. Please contact M. van der Velde, Msc. (Experimental Psychology, University of Groningen) at: m.a.van.der.velde@rug.nl