Alleviating the Cold Start Problem in Adaptive Learning using Data-Driven Difficulty Estimates

Maarten van der Velde¹ · Florian Sense¹ · Jelmer Borst² · Hedderik van Rijn¹

Accepted: 14 February 2021 / Published online: 15 March 2021
© The Author(s) 2021

Abstract

An adaptive learning system offers a digital learning environment that adjusts itself to the individual learner and learning material. By refining its internal model of the learner and material over time, such a system continually improves its ability to present appropriate exercises that maximise learning gains. In many cases, there is an initial mismatch between the internal model and the learner's actual performance on the presented items, causing a “cold start” during which the system is poorly adjusted to the situation. In this study, we implemented several strategies for mitigating this cold start problem in an adaptive fact learning system and experimentally tested their effect on learning performance. The strategies included predicting difficulty for individual learner-fact pairs, individual learners, individual facts, and the set of facts as a whole. We found that cold start mitigation improved learning outcomes, provided that there was sufficient variability in the difficulty of the study material. Informed individualised predictions allowed the system to schedule learners' study time more effectively, leading to an increase in response accuracy during the learning session as well as improved retention of the studied items afterwards. Our findings show that addressing the cold start problem in adaptive learning systems can have a real impact on learning outcomes. We expect this to be particularly valuable in real-world educational settings with large individual differences between learners and highly diverse materials.

Keywords Adaptive fact learning · Cold start problem · Memory · ACT-R · Bayesian modelling

Introduction

The process of learning increasingly takes place in digital environments. Adaptive learning systems, such as intelligent tutoring systems or cognitive tutors, allow students to learn in a digital environment tailored to them, offering a personalised approach that is often not possible in a traditional classroom setting (VanLehn 2006). By analysing a learner's actions, an adaptive learning system can provide error-specific feedback at an appropriate level of detail, select suitable practice problems to try next, and provide insight into the learner's mastery of the required skills. Adaptive learning systems have been used in many domains, such

as learning country names on a map (Mettler et al. 2016), solving simple arithmetic problems (Klinkenberg et al. 2011) and algebra exercises (Corbett et al. 1997), writing computer code (Reiser et al. 1985), and responding to cardiac issues (Eliot et al. 1996).

The defining advantage of adaptive learning systems over non-adaptive systems is their ability to adjust to individual learners' continually evolving knowledge and skills. Nevertheless, if there is no prior information about the capabilities of a learner or about the difficulty of a set of materials, even an adaptive system can initially be poorly attuned to the needs of a learner. In this so-called cold start scenario, adaptation can only occur once a learner starts using the system. It takes some time to identify the skill level of a new learner or to establish the difficulty of a new problem. During this time, the system can be just as poorly adapted to the individual as a non-adaptive learning system, by presenting exercises that are too easy or too difficult, or by providing inadequate or excessive feedback. Such a cold start may discourage new students from continuing to use the system. Studies have shown that in order to stay engaged, students require a learning experience that is

✉ Maarten van der Velde
m.a.van.der.velde@rug.nl

¹ Department of Experimental Psychology,
University of Groningen, Groningen, the Netherlands

² Bernoulli Institute for Mathematics, Computer Science
and Artificial Intelligence, University of Groningen,
Groningen, the Netherlands

challenging but has a difficulty that is proportionate to their own skill level—the so-called balance hypothesis (Shernoff et al. 2003; Kennedy et al. 2014; Hamari et al. 2016). With its potential to create an unbalanced learning experience, the cold start is therefore an important problem to address.

Different adaptive learning systems cope with the cold start problem in different ways. For instance, a system that faces a cold start may initially make large adjustments to its internal model following each response, and then gradually decrease the size of the adjustment with each subsequent response. The dynamic K factor in the Elo rating system used by Klinkenberg et al. (2011) is an example of this: it scales the change in an ability estimate as a function of that estimate's uncertainty. Alternatively, a system might adapt its presentation schedule to reduce uncertainty (Wauters et al. 2010). This can be done by prioritising items that are most informative for determining a learner's skill level, rather than optimising for learning or motivation (Chen et al. 2000). However, such methods attempt to deal with the consequences of the cold start problem, rather than addressing the problem at its root: inaccurate initial estimates of item difficulty and learner ability.

Inspiration for addressing the root cause can be found in the field of recommender systems, where an analogous problem exists. Recommender systems make recommendations that are tailored to individual users, typically relying on user-generated data to do so. Well-known examples include the personalised recommendations for products to buy on Amazon, automatically generated playlists on Spotify, and movie suggestions on Netflix. The cold start problem in recommender systems has received significant attention, and various solutions have been put forth. Broadly, these fall into one of three categories (Lika et al. 2014). Firstly, there are collaborative filtering methods, which generate recommendations for a new user on the basis of other users' preferences (e.g. recommending movies based on others' ratings, Bobadilla et al., 2012). These methods work well when little or nothing is known about a user. However, because their recommendations are crowd-sourced, such methods do rely on users having similar interests, an assumption that does not always hold. Secondly, there are content-based methods, which make recommendations to existing users based on their profiles and previous interactions with the system (e.g. recommending news articles based on previously read articles; see Lops et al. (2011) for a review). Content-based methods are geared towards the individual, so while they can work well in a heterogeneous population, they miss out on commonalities between users. Finally, there are hybrid methods that combine collaborative filtering and content-based methods in some way, with the aim of overcoming their respective shortcomings (e.g. recommending scientific

literature using information about both users and items, Popescul et al., 2001).

Because recommender systems and adaptive learning systems share the problem of creating an accurate user model, cold start mitigation strategies used in recommender systems can also be applicable to adaptive learning systems. For instance, Pardos and Heffernan (2010) used the accuracy of students' first step in solving an algebra problem to decide on their assumed level of prior knowledge. This stereotype-based approach, a content-based method, entailed quickly making assumptions about a learner's skill level based on their initial interaction with the system. More recently, Nedungadi and Remya (2014) took a hybrid approach to the cold start problem. They identified clusters of similarly skilled learners, based on previously collected data, and used learners' membership of these clusters to predict their performance on new problems. In a similar vein, Park et al. (2018) used background information about new learners, such as their grade level and gender, to assign the learners to groups. The learning data collected from other group members was then used to generate an initial ability estimate.

While all of these studies demonstrated ways to mitigate the cold start problem in adaptive learning systems, they did so in post hoc simulations performed on data that had already been collected. To our knowledge, the proposed mitigation strategies have not yet been scientifically tested in an applied setting. The current study addresses this issue by implementing several cold start mitigation strategies in an adaptive learning system, and investigating their effects on users' performance during a learning session, as well as on subsequent learning outcomes. The study includes collaborative filtering, content-based, and hybrid methods, enabling us to compare their efficacy in a practical application.

We used an adaptive fact learning system to test these mitigation strategies. The system lets students memorise declarative information through retrieval practice. It has been validated in experimental settings and applied successfully in the classroom, helping students learn facts in many domains, including foreign vocabulary, biopsychological terms, and geography (van Rijn et al. 2009; Sense et al. 2016; Sense et al. 2018; van den Broek et al. 2019). The adaptive nature of the system lies in its scheduling of presentations within a learning session. In each trial, the system quizzes the learner on the fact for which the expected learning gains are largest at that point in time. This selection is based on a computational model of the learner's memory in which each fact is represented. The memory model keeps track of the estimated difficulty of each fact for a particular learner, operationalised in a continuous-valued *rate of forgetting* estimate. The memory model adjusts its rate of forgetting estimates for individual

facts on the basis of the accuracy and latency of the learner's responses. To avoid getting stuck on a single fact, the system does not allow more than two successive presentations of any fact. Facts are introduced one at a time, with new facts only being added when the model deems all previously encountered facts to be encoded sufficiently well. The average rate of forgetting associated with a particular learner provides an index of that learner's (domain-specific) ability to memorise facts (Sense et al. 2018).

When a learner first uses the adaptive fact learning system, or begins learning a new set of facts, the memory model starts out with the assumption that all facts are equally difficult and therefore have the same (default) rate of forgetting. This means that there is a cold start, during which difficult facts are not repeated often enough, while easy facts are presented too often. The current study attempts to alleviate this cold start by harnessing known individual differences in rate of forgetting between facts and between learners—some facts are consistently easier than others, and some learners are consistently better at memorisation than others (see e.g. Sense et al. 2016; Zhou et al. 2020)—for changing the initial estimates of the memory model, creating a “warm start” instead. Informed initial estimates are derived from observations of other learners studying the same fact (a collaborative filtering approach), from an earlier session in which the same learner studied different facts (a content-based approach), or from a combination of the two (a hybrid approach). In all cases, the previously observed rates of forgetting are used to update a Bayesian model that yields a new predicted rate of forgetting.

This paper reports on two experiments. In both experiments, participants memorised a set of facts using the adaptive learning system, modified to use one of the cold start mitigation strategies outlined above. We measured participants' performance during the learning session and on a subsequent delayed recall test. Both experiments were pre-registered, and the first was replicated in an online sample. Experiment 1 found that all mitigation strategies were better at predicting the rate of forgetting inferred at the end of the learning session than the system's default starting estimate, but there was no difference between the mitigation strategies in this regard. Exploratory analysis showed higher response accuracy in learning sessions in which mitigation strategies were used, suggesting that informed initial estimates did indeed lead to a presentation schedule more conducive to successful retrieval. The expected improvement in performance on the delayed recall test did not materialise, however. We hypothesised that this was due to the generally high difficulty of the learning material, which reduced opportunities for introducing new facts and thereby limited the potential positive effect of a more efficient presentation schedule. For this reason, we conducted

a second experiment in which participants learned a more naturalistic set of facts that included both easy and difficult items, with a simplified design. This experiment did indeed find higher accuracy on the delayed recall test, along with higher accuracy during the learning session. Together, these experiments demonstrate that mitigating the cold start in an adaptive learning system can benefit learners' performance, both during *and* after a learning session, especially when heterogeneous item sets are used.

Adaptive Fact Learning System

The adaptive fact learning system used in this study relies on a computational cognitive model of human memory to create an optimal repetition schedule for facts during a learning session. The system is an extension of the one developed by Pavlik and Anderson (2005, 2008). Learners study a set of paired associates in a sequence of trials, each time typing the answer corresponding to the prompt shown on screen. New items are introduced once, with an initial presentation of both the prompt and the associated response. A learner's responses are used to inform the system's individualised estimate of the difficulty of each fact. The repetition schedule is adjusted accordingly, so that more study time is allocated to difficult facts and less time to easy facts. The system aims to capitalise on both the spacing effect (longer spacing between repetitions improves retention; e.g. Dempster, 1988), by spreading repetitions as far apart as possible, and the testing effect (active, effortful recall improves retention; e.g. van den Broek et al., 2016), by repeating facts at a moment when they can still be recalled.

The cognitive model that underpins the adaptive learning system represents each fact by a memory chunk with a certain activation. This activation, which corresponds to the strength of the declarative representation of the chunk, is boosted whenever the chunk is (re)created. It then decays over time. To calculate a chunk's activation at a particular time, the model uses the memory strength equation from the ACT-R cognitive architecture (Anderson 2007). The activation A of a chunk x at time t , given n previous encounters at $t_1, \dots, t_n$ seconds ago, is:

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

Before the start of each trial, the model chooses a fact to rehearse, based on the activations of the facts that have been encountered so far. It calculates the projected activation of all facts fifteen seconds from the current time and selects whichever fact has the lowest value, since that fact is expected to be forgotten soonest. If, however, this

fact still has an activation higher than a predefined retrieval threshold, the model randomly picks a new fact to practice instead. In this way, new facts will be introduced whenever none of the current facts need urgent rehearsal. The system also limits the number of successive presentations of any single fact to two. This means that, if selecting the least active fact would constitute a third successive presentation, the system repeats the next best candidate or, if none is available, introduces a new fact.

Some facts are easier to remember than others. The model accounts for these differences by allowing the rate at which the activation decays over time to vary between facts. It assumes that the activation of a difficult fact decays more rapidly than that of an easy fact, resulting in earlier and more frequent repetitions. This principle is illustrated in Fig. 1. The decay $d_x(t)$ of a chunk x at time t is a function of the activation of the chunk at the time of its most recent encounter, together with a chunk-specific *rate of forgetting* α_x :

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

By default, α_x is initialised to a value of 0.3. Over the course of a learning session, the model uses the speed and accuracy of the learner's responses to update this value whenever the associated fact is rehearsed. At the moment a fact is presented to the learner, the model calculates an expected response time using the ACT-R equation for retrieval time (Anderson 2007), adding a fixed amount of time t_0 for perceptual and motor processes:

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

The discrepancy between expected and observed response time determines how the model adjusts its rate

of forgetting estimate for the fact. A faster-than-expected response signals that activation has not yet decayed to the predicted level, so the true rate of forgetting must be lower than assumed. Conversely, an unexpectedly slow or incorrect response implies that activation has decayed further than anticipated, meaning that the current rate of forgetting estimate is too low. An incorrect response is recorded in the model as a slow response (1.5 times the maximum expected response time for a fact with an activation equal to the retrieval threshold). To prevent overcompensation on the basis of a single discordant response, the model takes the five most recent responses for a fact into account when updating its rate of forgetting estimate. No adjustments are made until a fact has been presented at least three times. To update the rate of forgetting estimate, the model performs a binary search within a range of 0.05 below or above the current estimate. The value that minimises the mismatch between the expected and observed response times becomes the new estimate.

Previous work has shown this adaptive fact learning system to work in controlled laboratory settings (van Rijn et al. 2009; Nijboer 2011; Sense et al. 2016), as well as in applied educational settings (Sense et al. 2018). However, since updates to the rate of forgetting estimate only begin after the third repetition and are constrained to a relatively narrow range around the current value, it can take quite some time to correct an inaccurate starting estimate. Rate of forgetting estimates typically range from about 0.1 to about 0.5. Accurately representing items at the edges of this range already requires them to be presented six times. Having better initial estimates may therefore lead to a significantly more effective presentation schedule.

Fig. 1 Activation of a declarative chunk over time, at three different rates of forgetting (α ; a chunk-specific offset to the decay). In the adaptive learning system, an item is presented for restudy when its

activation reaches the retrieval threshold, indicated here by a dashed line. As α increases, an item requires more frequent repetition to stay above the threshold

Bayesian Prediction of Rate of Forgetting

We evaluated four methods for predicting initial rate of forgetting estimates from previously collected learning data. These methods differ in terms of their granularity. The most granular is *Fact & Learner*, a hybrid method, which makes an individual prediction $\alpha_{F_i \otimes L_j}$ for the rate of forgetting value of every learner-fact pair. Less specific are *Fact*, a collaborative filtering method that predicts a single value α_{F_i} for each fact, and *Learner*, a content-based method that predicts a single value α_{L_j} for each learner. The least granular prediction method is *Domain*, a collaborative filtering method that only predicts a single value α_D for all learners and facts by taking the mean of all fact-level predictions. For comparison, we also included a *Default* condition in which the predicted rate of forgetting $\alpha_\emptyset$ is always the default value of 0.3.

All prediction methods rely on a Bayesian model to arrive at a predicted rate of forgetting. This model assumes that the rate of forgetting α_i for a fact or learner i is normally distributed with some unknown mean μ_i and unknown precision (the reciprocal of the variance) λ_i :

$$\alpha_i \sim \mathcal{N}(\mu_i, \lambda_i^{-1}) \quad (4)$$

Using the conjugate prior of this distribution, the joint Normal-Gamma distribution, the Bayesian model can simultaneously infer the mean and precision (Murphy 2007):

$$p(\mu_i, \lambda_i) = p(\mu_i | \lambda_i) * p(\lambda_i) \quad (5)$$

$$p(\mu_i | \lambda_i) = \mathcal{N}(\mu_0, \kappa_0 \lambda_i^{-1}) \quad (6)$$

$$p(\lambda_i) = \mathcal{G}(\alpha_0, \beta_0) \quad (7)$$

Since the prior is conjugate, the posterior is also a Normal-Gamma distribution, and can be found analytically rather than through computationally expensive Markov chain Monte Carlo, which is an important consideration for our implementation in the Web browser. The posterior predictive for the next observation, after having seen n data points, follows a t -distribution (see Murphy, 2007, for a derivation):

$$p(x|D) = t_{\alpha_n} \left(x | \mu_n, \frac{\beta_n(\kappa_n + 1)}{\alpha_n \kappa_n} \right) \quad (8)$$

We selected a weakly informative prior, with $\mu_0 = 0.3$, $\kappa_0 = 1$, $\alpha_0 = 3$, and $\beta_0 = 0.2$. This particular prior was chosen because it reflects our assumption that the rate of forgetting is normally distributed around 0.3, which previous studies have shown to be a reasonable average across materials and learners (e.g. van Rijn et al., 2009, Sense et al., 2016), and because it yields a sensible prior predictive distribution. To arrive at a single predicted value, we take the mode of the posterior predictive distribution,

which represents the rate of forgetting with the highest probability of being observed, given the model and the data.

The model is updated using rate of forgetting estimates obtained from prior learning data. Estimates are only included when they are based on at least three presentations, so that the adaptive model has had at least one opportunity for adjustment¹. Figure 2 illustrates the process of making fact-level, learner-level, and fact- and learner-level predictions on the basis of previous observations.

In the *Fact & Learner* prediction, the predicted rate of forgetting for a given fact when studied by a particular learner is based on the combination of the *Fact* prediction (which uses previously estimated rates of forgetting of that fact among other learners) and the *Learner* prediction (which uses previously estimated rates of forgetting of the learner on other facts). First, a separate posterior predictive distribution is generated for each prediction, following the process outlined above. The two distributions are then combined into a single posterior predictive distribution p_{LOP} , using logarithmic opinion pooling (Genest et al. 1984) with $k = 2$ equal weights w :

$$p_{LOP}(x|D) = \frac{\prod_{i=1}^k p_i(x|D)^{w_i}}{\int \prod_{i=1}^k p_i(x|D)^{w_i} dx} \quad (9)$$

$$= \frac{p_{fact}(x|D)^{0.5} * p_{learner}(x|D)^{0.5}}{\int p_{fact}(x|D)^{0.5} * p_{learner}(x|D)^{0.5} dx} \quad (10)$$

The mode of this pooled distribution becomes the predicted rate of forgetting of the fact for this learner. By using logarithmic pooling, we ensure that the resulting distribution is always unimodal (given the two unimodal inputs), even when the separate predictive distributions are very different. Because this method weights the two predictive distributions equally, their relative contribution to the final prediction is determined by their uncertainty. If one distribution is relatively spread out—reflecting less agreement in the data, or fewer observations—it exerts a smaller pull on the mode of the combined distribution than the other, more peaked distribution. In this way, the *Fact & Learner* prediction is biased towards whichever of its two components is more certain.

Experiment 1: Comparison of Mitigation Strategies

The first experiment consisted of an *estimation phase*, in which rate of forgetting estimates were collected for facts and learners, and a *comparison phase*, in which those

¹It is possible that a rate of forgetting estimate has not yet fully adjusted to the appropriate value within the available study time. If anything, this makes the resulting prediction more conservative, as the prediction will be closer to the default starting estimate of 0.3.

Fig. 2 The process of predicting rate of forgetting from observations. Left pane: Over the course of a learning session, the adaptive fact learning model refines its rate of forgetting (α) estimate for every fact a learner encounters, ultimately resulting in a final rate of forgetting estimate for each observed learner-fact pair. Right pane: Depending on the type of prediction, a different subset of the observed final rates of

forgetting is used to train the Bayesian model. The predicted value is the mode of the posterior predictive distribution of the model. A combined fact- and learner-level prediction combines the two posterior predictive distributions using logarithmic opinion pooling and takes the mode of the resulting distribution as its predicted value

estimates were used to mitigate the cold start problem in a new learning session via one of the mitigation strategies (see Fig. 3). In the estimation phase, the system

obtained rate of forgetting estimates for a set of facts in one participant sample. In addition, using comparable but different facts, it determined individual rate of forgetting

Fig. 3 Design of experiment 1. Rate of forgetting (α) estimates for facts and learners, obtained in the estimation phase, were applied in the comparison phase as starting values in a new learning session using one of the cold start mitigation strategies. The bottom label and colour

of each learning session box indicate the type of prediction used for cold start mitigation; the middle of each box visualises the process by which α was predicted. Learning performance was measured during the learning session and on a delayed recall test

estimates for learners in a second participant sample. During the comparison phase, predictions based on the collected rate of forgetting estimates were used as starting values for the adaptive algorithm in a new learning session, done by the participants from the second sample. We performed a between-subject comparison of five different prediction types in the comparison phase: fact-level prediction, learner-level prediction, fact- and learner-level prediction, domain-level prediction (the average of all fact-level predictions), and, as a control condition, the fixed value used by the current system. We compared learning performance, as measured during the learning session and on a subsequent delayed recall test, and model-based measures between these five conditions.

Methods

Materials

Participants learned the names of two sets of thirty small towns across the contiguous USA, shown as dots on a map (see Fig. 4). To minimise the influence of existing knowledge, all towns, which were selected from the Simplemaps US cities database², had no more than 5000 inhabitants, and each town shared its name with at least four other towns in the database. All towns had single-word names with no special characters and no more than eight letters. A complete list of the towns in both sets can be found in the online supplement³.

Participants

We recruited 241 participants from the research participant pool for first-year psychology students at the University of Groningen. They were split into two samples, one for each phase of the experiment. There were 82 participants in the first sample and 159 in the second. Data collection was conducted in the lab and participants were compensated with course credits.

For the online replication, which we conducted to examine how well our findings would extrapolate to a different population, an additional 217 participants were recruited in the USA via Amazon's MTurk platform. Of these participants, 85 were in the first sample, with the remaining 132 being in the second sample. The MTurk participants were selected to be similar to the student population tested in the lab in terms of age and education. To be eligible for participation, they were required to have a HIT approval rate of 95% or higher, be located in the USA, be born between 1992 and 1999 (age range, 19 to 27 years),

Fig. 4 Example stimulus, to which one would respond by typing the name of the highlighted town (here, *buchanan*) into the text box

and have a US high school diploma. The MTurk participants completed the experiment online and were compensated in accordance with the US federal minimum wage.

The experiment was approved by the Ethical Committee of Psychology at the University of Groningen (study codes: 18215-SO and 18216-SO). All participants gave informed consent.

Procedure

Estimation Phase The task was implemented in jsPsych (de Leeuw 2015) and ran in a Web browser. A minimum browser window size of 1280 × 768 pixels was required to display the experiment. For participants in the lab, the experiment was presented in full screen on a desktop computer. Since we could not fully control the circumstances under which MTurk participants completed the task, we asked them to ensure that they were in a quiet environment in which they would not be interrupted by phone calls, notifications, or other distractions. In addition to this, they were warned that they could not click away to another tab or window during the experiment. If they did, the task would be terminated. To assess participants' basic typing proficiency, and to acclimatise them to the response and feedback format, sessions started with a six-item typing test in which participants typed a word shown on screen. They received feedback about their accuracy. All participants were found to have sufficient typing ability.

The typing test was followed by a 10-min learning session using the adaptive fact learning system. The two participant samples studied different sets of facts, but the procedure was otherwise the same. The order in which new facts were introduced was randomised. The adaptive fact learning system determined when and how often each fact was repeated. An example trial is shown in Fig. 4. In each trial, participants saw a map of the USA on which one town was highlighted in red. The twenty-nine other towns in the

²<https://simplemaps.com/data/us-cities>

³<https://osf.io/snfyz/>

set were marked with smaller, grey dots. Participants typed the name of the highlighted town in lowercase in a text box below the map, confirming their response by pressing the enter key. If they did not recall the name of the town, they were instructed to simply press enter. New towns were introduced through a study trial in which the town's name was displayed directly above the text box. Feedback was always shown below the text box directly after a response had been made. The answer that the participant had given remained visible in the text box. If a participant's response matched the correct answer exactly, the feedback *Correct* was shown below the text box for 600 ms. A response with a Damerau-Levenshtein distance of 1 from the correct answer was considered almost correct (the assumption being that the participant did in fact know the correct answer, but made a single typing error), and was treated as correct by the model. In this case, the text *Almost correct*, along with the correct answer, appeared below the text box for 1.2 s. Finally, a response with a Damerau-Levenshtein distance from the correct answer of two or more was considered incorrect, and prompted the text *Incorrect*, along with the expected answer, to appear below the text box for 4 s.

Participants in the first sample were done with the task after completing the 10-min learning session. They saw a debriefing screen showing the number of trials they had completed, the number of towns they had encountered, and their overall response accuracy. Participants in the second sample, however, took a delayed recall test following a 5-min Tetris game that served as a filler task. In the recall test, facts that had been studied were presented one-by-one in the same manner as before. Feedback was withheld until the end of the test.

Comparison Phase Participants in the second sample completed another block, once again consisting of a 10-min learning session, 5 min of Tetris, and a delayed recall test of the studied facts. This time, participants studied the other set of facts, and the adaptive fact learning system used one of the cold start mitigation strategies. Participants were randomly assigned to one of the five prediction types, which are described in the next section. The block ended with the debriefing screen.

Analysis

Before starting data collection in the second sample, we preregistered a set of analyses at <https://osf.io/vwg6u/>. The preregistration stated that we would collect data from a minimum of 25 participants per condition by a specified date. We did not reach this minimum by the deadline, so we continued data collection. After the data had been collected, we realised that several of the planned analyses were ill-conceived. The analyses reported in this paper therefore

deviate from our preregistration in two ways: they include data collected after the deadline, and in some cases use different methods. An online supplement containing the data and full analysis scripts, including analyses conducted only on the subset of the data collected before the deadline as well as the originally planned analyses, is available at <https://osf.io/snfyz/>.

We used Bayesian statistics for all analyses because Bayesian methods allow for the quantification of evidence in favour of the null model in the event that no difference is found between conditions (Gallistel 2009). Following Jeffreys (1961), an effect was considered meaningful if the data provided at least moderate evidence in its favour, i.e., a Bayes factor of 3 or higher. All analyses were performed in R (version 3.6.3; R Core Team, 2018). Bayesian ANOVAs and *t*-tests were done using the BayesFactor package with default settings (version 0.9.12-4.2; Morey and Rouder, 2018).

Bayesian regression models were fitted with the brms package (version 2.12.0; Bürkner, 2017). In cases with multiple measurements per participant and/or item, the models included the appropriate random intercepts. In each analysis, we fitted a maximal model with main effects of prediction type (*Default*, *Domain*, *Fact*, *Learner*, *Fact & Learner*) and population (*Lab*, *MTurk*) plus the interaction of these terms. Both this maximal model and models with a simpler fixed effects structure were compared to a baseline intercept-only model (see the rows in Table 1) using the bridgesampling package (version 0.7-2; Gronau et al., 2020). Following Rouder and Morey (2012), fixed effects had weakly informative Cauchy(0,1) priors. All model runs used 4 Markov chains with 10,000 iterations each, including 5000 warm-up samples. Other options were left at their default setting.

Results

Figure 5 provides an overview of performance in the comparison phase of experiment 1. Statistical analyses associated with each subfigure are summarised in Table 1, giving an overview of the effects across various outcome measures at a glance. Further details regarding the type of model used in each analysis as well as the specific findings are given in the following sections.

Rate of Forgetting Prediction

To assess the accuracy of the rate of forgetting predictions—made using the final rates of forgetting observed during the estimation phase—we compared them to the final rates of forgetting estimated at the end of the learning session (Fig. 5a). Prediction accuracy was quantified as the root-mean-square error (RMSE) of the predicted rate

Table 1 Bayesian model comparisons

Model	A Prediction accuracy	B Facts studied	C Learning accuracy	D Test score	E Test accuracy
Intercept only	1	1	1	1	1
Prediction type	44	0.0	5.9	0.0	0.0060
Population	14	0.034	3857	0.025	0.20
Prediction type + Population	320	0.0	14,881	0.0	0.0012
Prediction type $\times$ Population	36	0.0	57	0.0	0.000020

Columns show different outcome measures related to rate of forgetting prediction (**A**), learning session performance (**B**, **C**), and delayed recall test performance (**D**, **E**). Rows show candidate models with different predictors. Cells contain Bayes factors, expressing the evidence for each model relative to the intercept-only null model. The best model for each outcome measure is shown in bold

of forgetting relative to the rate of forgetting observed at the end of the session. The RMSE was calculated within-subject, and then averaged across subjects within each condition. A Bayesian ANOVA confirmed main effects of prediction type and population on prediction accuracy (Table 1A). The next best model included only an effect of prediction type, suggesting that this factor contributed more to the strength of the winning model than the effect of population. That said, follow-up pairwise comparisons of prediction accuracy between conditions were generally inconclusive. Only a one-sided Bayesian *t*-test comparing RMSE in all four predictive conditions ($M = 0.084$, $SD = 0.025$) to the *Default* condition ($M = 0.102$, $SD = 0.035$) found strong evidence for an improvement in prediction accuracy ($BF = 1.0 \times 10^3$). The main effect of population indicated that RMSE was slightly lower in the lab population ($M = 0.083$, $SD = 0.023$) than in the MTurk population ($M = 0.093$, $SD = 0.033$).

Learning Session Performance

The number of distinct facts studied during the learning session did not differ between conditions (Fig. 5b). We modelled the data using a Bayesian Poisson regression. Model comparison showed the data to be most likely under the intercept-only null model (Table 1B). According to this model, participants encountered 17.7 distinct facts on average (95% CI [17.2, 18.2]). In contrast, an analysis of response accuracy during the learning session did find a difference between conditions (Fig. 5c). A comparison of Bayesian logistic mixed-effects models showed that the data were most likely under a model that included main effects of prediction type and population (Table 1C). This model found similarly sized improvements in response accuracy over the *Default* condition in all four cold start mitigation conditions: on average, response accuracy in these conditions was about 7.3 percentage points higher (95% CI [3.9, 11.0]). In addition, participants in the lab

outperformed MTurk participants by about 5.5 percentage points (95% CI [3.1, 8.0]).

Delayed Recall Test Performance

There was no difference between conditions in the number of correctly recalled facts on the delayed recall test (Fig. 5d). A comparison of Bayesian Poisson regression models showed strong evidence in favour of the intercept-only null model (Table 1D), which found an average test score across conditions of 15.0 (95% CI [14.5, 15.4]). We obtained a similar result when looking at response accuracy on just the items that participants had encountered in the preceding learning session (Fig. 5e): the Bayesian logistic mixed-effects model that was most likely to have generated the data was the intercept-only null model (Table 1E). This model showed that overall accuracy on studied items was 87.9% (95% CI [85.2%, 90.3%]) across conditions.

Discussion

In this experiment, we attempted to improve the initial difficulty estimates of the cognitive model underlying our adaptive fact learning system, with the aim of reducing the cold start problem and thereby improving learning outcomes. We predicted the rate of forgetting of individual facts for individual learners by training a Bayesian model on previous data from the learning system. Several prediction methods with different granularity were compared to the fixed starting estimate currently used by the system.

Although the rate of forgetting predictions made by the Bayesian model were closer to observed values than the default prediction, this did not translate into an improvement in preregistered learning outcomes. Exploratory analysis did suggest that response accuracy increased during the learning session, indicating that facts were repeated on a more appropriate schedule than they would have been otherwise.

Rate of forgetting prediction

Learning session performance

Delayed recall test performance

Fig. 5 Experiment 1 comparison phase results. **a**: Predicted versus observed rate of forgetting per condition. The predicted values were used as initial rate of forgetting estimates in the learning session; the observed values were the final estimates. The black lines show the best linear fit to the data. **b**: Number of distinct facts encountered during

the learning session. **c**: Response accuracy during the learning session. **d**: Number of items answered correctly on the delayed recall test. **e**: Response accuracy on studied items on the delayed recall test. The box plots show the median and the first and third quartile

We expected fine-grained predictions to be more accurate than coarser ones, but comparisons between prediction types were largely inconclusive. It appears that domain-level predictions were already able to capture most of the relevant difficulty information, and predictions at finer levels of detail did not provide sufficient additional information to make a meaningful difference. Fine-grained predictions did show a larger spread than domain-level predictions (Fig. 5a), suggesting that there were facts that were consistently easy or difficult and learners with consistently higher or lower rates of forgetting, in line with earlier work (e.g. Sense et al., 2016). It should be noted that both the participant samples and the set of materials were fairly homogeneous, especially when compared to the more diverse educational settings found in real life. In settings where there are greater differences in skill between students, or differences in difficulty between facts, one would expect a domain-level prediction to fall short.

The stimulus material was also more difficult than expected: in both populations, predicted rates of forgetting were generally higher than the default value, which was found to be a reasonable average in previous studies. A higher rate of forgetting leads to more frequent repetitions of an item during the learning session. The learning system would therefore have had less opportunity to introduce new facts during the learning session than if some of the facts had been easier. The difficulty of the material meant that, despite having more accurate starting estimates, the system could not really capitalise on the differences between facts.

In short, while the use of informed starting estimates in the adaptive learning system may have improved response accuracy during the learning session, test scores remained unaffected, possibly as a result of having a uniformly difficult stimulus set. To test whether this was indeed the case, we conducted a second experiment with a more diverse set of facts that included both easy and difficult items but with a similarly homogeneous participant sample. For simplicity, this experiment focused on fact-level predictions.

Experiment 2: More Variable Fact Difficulty

Figure 6 shows the design of experiment 2. As in experiment 1, the experiment consists of an estimation phase, in which we collect the rate of forgetting estimates that form the basis for the difficulty predictions, and a comparison phase, in which those predictions are used as starting estimates in the adaptive learning system. Experiment 2 uses a more varied set of stimuli, with the aim of letting the system capitalise on the larger differences in difficulty between facts. For this reason, we only use fact-level predictions. These predictions are derived from the learning sessions of one set of participants during the estimation phase, and

subsequently applied in the learning sessions of another set of participants during the comparison phase. To see if cold start mitigation does improve learning outcomes under these circumstances, we perform a within-participant comparison of learning performance between a cold start session that uses default predictions and a warm start session that uses fact-level predictions.

Methods

Materials

As in experiment 1, participants learned the names of places on a map. To increase the variability in difficulty, we created a set of sixty towns or cities, half of which were small towns and half of which were larger cities that participants were more likely to know already. The small towns had populations of 5000 people or fewer and shared their name with at least four other such towns (as in experiment 1), while the larger cities had at least 100,000 inhabitants. All places had single-word names that were no more than eight letters long and did not contain any special characters. The places were spread evenly over the contiguous USA.

Participants

We tested 197 participants from the research participant pool for first-year psychology students at the University of Groningen, none of whom had participated in Experiment 1. They were divided over two samples that consisted of 128 and 69 participants, respectively. Both sample sizes were determined by preregistered stopping rules. Data collection in the first sample continued until each of the sixty facts in the stimulus set had been studied by at least thirty participants (analyses on data from experiment 1 have shown that we need about thirty observations per fact to get a stable rate of forgetting prediction that does not change substantially with extra observations). For the second sample, we used a Sequential Bayes Factors (SBF) design, which allows stopping based on an evidence criterion specified in advance (Schönbrodt et al. 2017). Details of the stopping rule are given in the section “[Analysis](#)”. Data were collected in the lab and participants received course credits for taking part.

The experiment was approved by the Ethical Committee of Psychology at the University of Groningen (study codes: 18215-SO and 18216-SO). All participants gave informed consent.

Procedure

Estimation Phase Participants in the first sample completed a single 10-min learning session using the default adaptive

Fig. 6 Design of experiment 2. Rate of forgetting (α) estimates for facts, obtained during the estimation phase, were applied as starting values in a new learning session during the comparison phase. Learning performance was measured during the learning session and on a delayed recall test

fact learning system. The format of the learning session was identical to that of the learning sessions in experiment 1 in terms of stimulus presentation and the response and feedback mechanism, but using the new stimuli. All sixty places were marked on the map, but each participant was only presented a subset of thirty. To optimise the data collection process, we tracked how many participants had encountered each place. Whenever a new place was to be introduced during a session, it was randomly selected from the least practiced places. A debriefing screen was shown at the end of the learning session, containing the number of completed trials, the number of places studied, and the overall response accuracy.

Comparison Phase Participants in the second sample completed two 10-min learning sessions with the adaptive fact learning system, each followed by a 5-min game of Tetris and a delayed recall test of the studied items. To allow for a within-subject comparison between conditions, one of the learning sessions used the default rate of forgetting predictions, while the other session used fact-level predictions derived from the observations in the estimation phase. The condition order was counterbalanced (see Fig. 6). The sixty facts were divided over the two learning sessions in a semi-random manner. First, the facts were sorted by their predicted rate of forgetting and split

into an “easy” and a “difficult” half. For each participant, we then randomly sampled fifteen facts from the easy half and fifteen from the difficult half to be presented in the first learning session. The remaining facts were used in the second session.

Analysis

Bayesian mixed-effects regression models were fitted using the same procedure as in experiment 1, but this time the maximal model had main effects of prediction type (*Default*, *Fact*) and experimental block (1, 2), along with the interaction of these terms. The simpler alternatives to which this model was compared are listed in Table 2.

We preregistered one analysis (delayed recall test accuracy; Fig. 7f) at <https://osf.io/w4gtd/>. The other reported analyses are exploratory. An online supplement containing the data and full analysis scripts is available at <https://osf.io/snfyz/>.

The preregistration included a stopping rule to determine when to halt data collection in the second sample. As planned, once an initial set of data had been collected from twenty participants per condition order, we used a Bayesian stopping rule (Rouder 2014) that we evaluated after each day of data collection: stop data collection when there is sufficiently strong evidence—a Bayes factor of

Table 2 Bayesian model comparisons

Model	A Prediction accuracy	C Facts studied	D Learning accuracy	E Test score	F Test accuracy
Intercept only	1	1	1	1	1
Prediction type	408	0.10	1.1×10^{25}	0.042	229
Block	13	9.3	2.3×10^9	857	66
Prediction type + Block	6579	0.95	2.7×10^{33}	37	13, 624
Prediction type $\times$ Block	26, 498	0.45	7.7×10^{32}	44	6384

Columns show different outcome measures related to rate of forgetting prediction (**A**), learning session performance (**C**, **D**), and delayed recall test performance (**E**, **F**). Rows show candidate models with different predictors. Cells contain Bayes factors, expressing the evidence for each model relative to the intercept-only null model. The best model for each outcome measure is shown in bold

10 or higher—for or against the hypothesis that using fact-level predictions improves test accuracy on studied items, or when data have been collected from eighty participants per condition order. Per the preregistration, the evidence for a difference in test accuracy was evaluated in a Bayesian logistic mixed-effects model with main effects of prediction type and block, and random intercepts for participants and items. We used a Savage-Dickey density ratio test (Wagenmakers et al. 2010) to quantify the evidence for the alternative hypothesis (a non-zero coefficient for prediction type) over the null hypothesis (a coefficient of 0, represented by a weakly informative Cauchy(0,1) prior):

$$BF_{10} = \frac{p(D | \mathcal{H}_1 : \beta_{\text{prediction type}} \neq 0)}{p(D | \mathcal{H}_0 : \beta_{\text{prediction type}} = 0)} \quad (11)$$

The evidence criterion was satisfied after 3 days of data collection ($BF_{10} = 181.9$). We excluded the data from one participant in the second sample for failing to follow the task instructions.

Results

Figure 7 shows the performance in the comparison phase of experiment 2. Table 2 provides a summary of the statistical analyses associated with each subfigure. Further details about the analyses and findings are given in the following sections.

Rate of Forgetting Prediction

Figure 7a compares the predicted rates of forgetting, which were generated from the data of the estimation phase, to the final rate of forgetting estimates at the end of the learning session. Like in experiment 1, prediction accuracy (measured as RMSE) differed between conditions. A Bayesian ANOVA found that the data were most likely under a model that assumed main effects of prediction type and block, as well as their interaction. A comparison of

RMSE values shows that predictions were less accurate in the *Default* condition than in the *Fact* condition, and less accurate in the first block than in the second, though the difference between the two conditions was smaller in the second block (*Default*₁: $M = 0.120$, $SD = 0.044$; *Default*₂: $M = 0.091$, $SD = 0.023$; *Fact*₁: $M = 0.085$, $SD = 0.023$; *Fact*₂: $M = 0.082$, $SD = 0.024$).

The second experiment used a mix of small towns and larger well-known cities as stimuli, in an attempt to create a more balanced difficulty distribution that included more easy facts. Figure 7b shows the predicted rates of forgetting for both categories of fact. As expected, big cities tend to have a lower rate of forgetting than small towns, but the average predicted rate of forgetting in both distributions is still higher than 0.3, the default value. This means that, although the stimulus material is easier than in experiment 1, on average it will still be slightly harder for the adaptive fact learning system to introduce new facts during the learning session when using these predictions than when using the default starting estimate.

Learning Session Performance

The number of distinct facts that participants encountered during the learning session was similar between conditions (Fig. 7c). As Table 2C shows, the best Bayesian mixed-effects Poisson regression model included only a main effect of block. According to this model, the number of distinct facts increased by about 2.3 (95% CI [0.98, 3.7]) from the first to the second block, regardless of prediction type.

As Fig. 7d shows, participants did respond more accurately in the *Fact* condition. The best Bayesian logistic mixed-effects model included main effects of prediction type and block, but no interaction term (Table 2D). This model found that, across blocks, response accuracy was about 7.7 percentage points higher (95% CI [6.2, 9.3]) in the *Fact* condition. Accuracy also increased from the first to the second experimental block, by about 4.8 percentage points (95% CI [3.3, 6.3]).

Rate of forgetting prediction

Learning session performance

Delayed recall test performance

Fig. 7 Experiment 2 comparison phase results. **a** Predicted versus observed rate of forgetting per condition. The predicted values were used as initial rate of forgetting estimates in the learning session; the observed values were the final estimates. The black line shows the best linear fit to the data. **b** Predicted rate of forgetting in the *Fact* condition,

by type. **c** Number of distinct facts encountered during the learning session. **d** Response accuracy during the learning session. **e** Number of items answered correctly on the delayed recall test. **f** Response accuracy on studied items on the delayed recall test. The box plots show the median and the first and third quartiles

Delayed Recall Test Performance

The absolute number of items that participants were able to recall on the delayed recall test was similar between conditions (Fig. 7e). As Table 2E shows, the observed test scores were most likely under a Bayesian mixed-effects Poisson regression model with only a main effect of block. According to this model, test scores increased by about 2.7 points (95% CI [1.5, 3.9]) from the first block to the second, regardless of prediction type.

That said, participants' response accuracy on items that they had encountered during the preceding session was higher if that session used fact-level predictions (Fig. 7f). In accordance with our preregistration, we used a Bayesian logistic mixed-effects model to assess the evidence for this effect. The model included fixed effects for prediction type

and block, and random intercepts for items and participants. A post hoc model comparison showed this model to be preferred over the alternatives (Table 2F). The model indicated that, on the whole, response accuracy was about 6.8 percentage points higher (95% CI [3.4, 10.5]) on the test following the *Fact* block than on the test after the *Default* block. Test accuracy was also found to increase with task experience, as learners scored about 6.3 percentage points higher (95% CI [2.8, 9.9]) on the second test than on the first.

Exploratory: Factors Explaining Improved Performance

To better understand where the gains in response accuracy came from, we conducted several exploratory follow-up analyses.

Stronger Differentiation in Presentation Schedule The memory model (see Eqs. 1 and 2) predicts that a higher rate of forgetting causes a fact to decay faster, reducing the probability of it being successfully recalled in the future. The adaptive fact learning system should compensate for these differences in decay by changing the rehearsal schedule: difficult facts are repeated sooner and more frequently so that they remain sufficiently active in memory. The more successful rehearsal and retention of difficult items in the *Fact* condition indicate that the adaptive fact learning system was able to schedule item repetitions at more appropriate times when initialised with fact-level predictions. We fitted a Bayesian mixed-effects model to the number of presentations of a fact in the session (modelled as a Poisson distribution). The best model included main effects of prediction type (*Default* or *Fact*), predicted rate of forgetting, and their interaction, along with a random intercept for participants. As the model fit in Fig. 8a shows, there was indeed

stronger differentiation between facts in the *Fact* condition. In both conditions, the number of presentations increased as facts became more difficult, showing that the online adaptation works as expected. This increase was stronger in the *Fact* condition. Initialising the system with fact-level predictions allowed it to present difficult items at a higher rate from the very start of the session, giving participants more opportunities to learn these items. The model also suggests that the easiest facts may have been presented slightly less often in the *Fact* condition, which would have freed up some rehearsal time.

Better Retention of Difficult Items In an ideal world, we would know each fact's rate of forgetting exactly and be able to adapt the rehearsal schedule accordingly. Recall success—at least during the learning session—would then be equally high for all facts, regardless of their difficulty. In reality, of course, we can never have perfect knowledge

Fig. 8 Effects of fact difficulty on scheduling and performance. **a** Number of presentations of a fact during the learning session as a function of predicted fact-level rate of forgetting. **b** Response accuracy during the learning session and on the subsequent delayed recall

test. In both figures, lines and shaded areas show the posterior medians and 95% credible intervals of the best-fitting Bayesian mixed-effects regression model (details provided in the text). Individual points show the mean across participants (± 1 SE) for each fact

of fact difficulty, and this leads to lower recall accuracy for more difficult facts. We fitted separate Bayesian logistic mixed-effects models to response accuracy during the learning session and on the test. In both cases, a model comparison using bridge sampling yielded a model with main effects of the predicted rate of forgetting, block (1 or 2), and the interaction between predicted rate of forgetting and prediction type (*Default* or *Fact*), as well as a random intercept for participants. The model fits are presented in Fig. 8b. (For simplicity, the intercept difference between blocks was left out.) The models suggest that response accuracy did decrease as facts became more difficult, regardless of prediction type, but that this effect was weakened in the *Fact* condition. While participants were more or less equally successful in recalling easy facts, the use of fact-level predictions in the *Fact* condition gave participants a better chance of retaining the difficult facts too. As can be seen in the figure, this pattern carried over from the learning session to the test.

Discussion

The second experiment showed that learning outcomes can be improved through mitigation of the cold start experienced with adaptive learning systems. Using a more realistic distribution of easy and difficult facts, we found that participants' response accuracy on a delayed recall test was higher when they had studied the items with the learning system informed by fact-level predictions. Exploratory follow-up analyses suggest that our cold start mitigation strategy was effective: the informed system gave participants more opportunities to rehearse difficult facts, resulting in more successful retrievals of these facts during the learning session and higher accuracy on the subsequent test.

Since the more difficult facts were identified as such from the start, the system was able to dedicate more rehearsal time to them, giving participants a better chance of retaining these facts. In addition, being immediately aware of the easiest facts may have saved the system some time it would have otherwise spent on their rehearsal, without hurting retention.

General Discussion

We implemented and tested a set of cold start mitigation strategies in an adaptive fact learning system. In two experiments, we found that the tested mitigation strategies made more accurate predictions of difficulty than the fact learning system's default prediction, allowing the system to dedicate more time to difficult items without sacrificing the learning of easy items. The effect that

this improved item scheduling had on learning outcomes depended on there being sufficient variability in fact difficulty. Provided a sufficient number of easy facts, the use of individualised difficulty predictions created a more pronounced differentiation in the scheduling of easy and difficult items, resulting in better retention of items studied during the learning session.

Limitations

These results showed that better predictions do not invariably lead to better learning, but rather that their effect depends on the dynamics of the adaptive system in which they are used. In addition, they remind us that designers of adaptive systems should be mindful of the system's dynamics when deciding on appropriate metrics for measuring success. In the case of our fact learning system, high predicted difficulty of the facts being studied can make the scheduling algorithm more inclined to repeat previously rehearsed facts, at the expense of introducing new facts. This shift towards a more conservative presentation schedule is a sign of the system working correctly, but has the unintended consequence of reducing the number of facts presented during the learning session, thereby also lowering the maximum score achievable on a recall test.

We found no difference between the tested mitigation strategies. All were equal in terms of prediction accuracy and learning outcomes, which suggests that fine-grained prediction methods are not necessarily better than simpler, less specific methods. It is possible that the additional benefit of a more fine-grained method only becomes apparent when individual differences between learners are more pronounced than in the current sample, or when the study material is more varied in its difficulty. With our relatively homogeneous participant sample learning facts of fairly similar difficulty, it is perhaps unsurprising that population-level predictions were comparable to individualised predictions. The same principle likely holds for learners: the ability to make a priori assumptions about individual learners' skills becomes more relevant in a heterogeneous population, such as when the target population includes learners at different stages in their education (Klinkenberg et al. 2011) or with learning disorders like dyslexia and dyscalculia (Pliakos et al. 2019).

Implications and Future Directions

As the success of our method shows, we were able to exploit systematic differences in the rate of forgetting of facts to make useful difficulty predictions for new learners encountering these facts. Not only did these differences generalise over participants within a sample, they were also

fairly consistent across populations. Experiment 1, in which participants learned the names of small towns on a map of the USA, was conducted in two populations, one in the Netherlands and one in the USA. Participants in the USA may have been helped by their higher familiarity with the US map, allowing them to integrate the new information with their existing knowledge more easily (Anderson 1981). Nevertheless, facts that were more difficult for one population tended to be more difficult for the other population, too. The ability to find a “universal” rate of forgetting for a fact aligns with the idea that facts are made more or less memorable by their features (see, e.g. Madan, 2019, Celikkale et al., 2013, Broers and Busch, 2019). In the case of places on a map, these could include spatial features like proximity to borders or other places, as well as semantic and lexical features of the place names.

We were also able to make rate of forgetting predictions at the level of the individual learner which were about as accurate as fact-level predictions. The fact that a learner’s previously observed rates of forgetting are predictive of their rate of forgetting in a future session (albeit on a relatively short timescale in the experiments reported here) is in line with the view of rate of forgetting as a stable, learner-specific trait (Sense et al. 2016; Zhou et al. 2020).

Experiment 2 showed that using difficulty predictions created a more strongly differentiated repetition schedule that improved participants’ response accuracy, particularly on difficult facts. One could trade some of this additional accuracy for more opportunities to introduce new facts, by lowering the fact learning system’s retrieval threshold (the activation value at which a fact is repeated). A change in the threshold would represent a shift in task difficulty: the lower the threshold is set, the more difficult the learning session becomes. With a lower threshold, the system waits longer before it repeats a fact, and it can fill in any gaps by introducing new facts. As such, learners may be able to cram in a few more facts, albeit at a slightly higher risk of forgetting them. The appropriate retrieval threshold is likely to differ between learners, and giving learners the ability to adjust it to their desired level of difficulty may increase their motivation to use the system (Metcalf and Kornell 2005; Kennedy et al. 2014).

Aside from varying the retrieval threshold, the adaptive fact learning system may benefit from using the uncertainty inherent in the rate of forgetting predictions made by the Bayesian model. Currently, this uncertainty, which is expressed through the width of the posterior predictive distribution, plays no role once the model has arrived at a prediction, since that prediction is only a point estimate. However, the Bayesian approach could be extended to the updating of the rate of forgetting during the learning session, thereby integrating the uncertainty of the prediction in the process. The adaptive learning system would then

more readily adjust its difficulty estimates on the basis of new evidence for facts that it is less certain about (e.g. because the initial prediction is based on fewer or more varied observations), while requiring more evidence to significantly change estimates in which it has higher confidence. Such an extension to the system could provide more control over how much belief is placed in predictions, while reducing the impact of unlikely response times on difficulty estimates.

The cold start mitigation methods presented here are not limited to the specific fact learning system used in this paper, as they do not rely on changing any part of the system itself. Indeed, other adaptive learning systems which represent difficulty or skill by a continuous-valued model parameter and are susceptible to the cold start problem could use a similar approach to mitigate the cold start. This includes learning systems based on different underlying models, such as the Elo rating system (Klinkenberg et al. 2011), and learning systems teaching procedural rather than declarative knowledge (e.g. Anderson et al., 1995). A benefit of starting a learning session with informed parameter estimates is that it can reduce the need to make sweeping parameter adjustments in the beginning of the session, based on few, and possibly noisy, responses.

Conclusion

This work has shown that mitigating the cold start problem in adaptive learning systems using data-driven difficulty predictions can improve learners’ performance. We found that starting a learning session with individualised predictions derived from prior learning data increased the adaptive system’s ability to capitalise on differences in item difficulty, which translated to higher response accuracy during a learning session and on a delayed recall test.

Acknowledgements The authors thank Valeria Cernei and Sanne van Dam for their help with data collection.

Author Contribution All the authors contributed to the conceptualisation of the study, the design of the experiments, the interpretation of the results, and the final manuscript. MvdV performed the experiments, analysed the data, and drafted the original manuscript.

Availability of data and material The data and stimulus materials are available at <https://osf.io/snfyz/>.

Code Availability The code required to reproduce the analyses is available at <https://osf.io/snfyz/>.

Declarations

Ethics Approval The experiments were approved by the Ethical Committee of Psychology at the University of Groningen (study codes: 18215-SO and 18216-SO).

Consent to Participate All participants gave informed consent for participation in the study.

Consent for Publication All authors consent to the publication of this manuscript.

Conflict of Interest An earlier version of the adaptive learning system discussed in this manuscript is licensed to Noordhoff Publishers by the University of Groningen. This project was partially funded by these license fees. However, the publishing house had no involvement in this study.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Anderson, J.R. (1981). Effects of prior knowledge on memory for new information. *Memory & Cognition*, 9(3), 237–246. <https://doi.org/10.3758/BF03196958>.
- Anderson, J.R. (2007). *How Can the Human Mind Occur in the Physical Universe?* Oxford, United Kingdom: Oxford University Press.
- Anderson, J.R., Corbett, A.T., Koedinger, K.R., Pelletier, R. (1995). Cognitive tutors: Lessons learned. *Journal of the Learning Sciences*, 4(2), 167–207. https://doi.org/10.1207/s15327809jls0402_2.
- Bobadilla, J., Ortega, F., Hernando, A., Bernal J. (2012). A collaborative filtering approach to mitigate the new user cold start problem. *Knowledge-Based Systems*, 26, 225–238. <https://doi.org/10.1016/j.knsys.2011.07.021>.
- Broers, N., & Busch, N. (2019). The effect of intrinsic image memorability on recollection and familiarity (Preprint). PsyArXiv. <https://doi.org/10.31234/osf.io/hvzyw>.
- Bürkner, P.C. (2017). Brms: An R package for bayesian multilevel models using stan. *Journal of Statistical Software*, 80(1), 1–28. <https://doi.org/10.18637/jss.v080.i01>.
- Celikkale, B., Erdem, A., Erdem, E. (2013). Visual Attentionx-Driven Spatial Pooling for Image Memorability. In *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops* (976–983). OR, USA: IEEE. <https://doi.org/10.1109/CVPRW.2013.142>.
- Chen, S.-Y., Ankenmann, R.D., Chang, H.-H. (2000). A comparison of item selection rules at the early stages of computerized adaptive testing. *Applied Psychological Measurement*, 24(3), 241–255. <https://doi.org/10.1177/01466210022031705>.
- Corbett, A.T., Koedinger, K.R., Anderson, J.R. (1997). Intelligent Tutoring Systems. In Helander, M.G., Landauer, T.K., Prabhu, P.V. (Eds.) *Handbook of human-computer interaction*. 2nd edn. (pp. 849–874). Amsterdam: Elsevier.
- de Leeuw, J.R. (2015). jsPsych: A javascript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, 47(1), 1–12. <https://doi.org/10.3758/s13428-014-0458-y>.
- Dempster, F.N. (1988). A case study in the failure to apply the results of psychological research. *American Psychologist*, 43(8), 627–634.
- Eliot, C.R., Williams, K.A., Woolf, B.P. (1996). An intelligent learning environment for advanced cardiac life support. In *Proceedings of the AMIA Annual Fall Symposium* (pp 7–11) Washington, DC, American Medical Informatics Association.
- Gallistel, C.R. (2009). The importance of proving the null. *Psychological Review*, 116(2), 439–453. <https://doi.org/10.1037/a0015251>.
- Genest, C., Weerahandi, S., Zidek, J.V. (1984). Aggregating opinions through logarithmic pooling. *Theory and Decision*, 17(1), 61–70. <https://doi.org/10.1007/BF00140056>.
- Gronau, Q.F., Singmann, H., Wagenmakers, E.J. (2020). Bridgesampling: An R package for estimating normalizing constants. *Journal of Statistical Software*, 92(10), 1–29. <https://doi.org/10.18637/jss.v092.i10>.
- Hamari, J., Shernoff, D.J., Rowe, E., Coller, B., Asbell-Clarke, J., Edwards, T. (2016). Challenging games help students learn: An empirical study on engagement, flow and immersion in game-based learning. *Computers in Human Behavior*, 54, 170–179. <https://doi.org/10.1016/j.chb.2015.07.045>.
- Jeffreys, H. (1961). *Theory of probability*, Third ed. Oxford: Oxford University Press.
- 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. <https://doi.org/10.1016/j.concog.2014.07.007>.
- Klinkenberg, S., Straatemeier, M., van der Maas, H. (2011). Computer adaptive practice of Maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers & Education*, 57(2), 1813–1824. <https://doi.org/10.1016/j.compedu.2011.02.003>.
- Lika, B., Kolomvatsos, K., Hadjiefthymiades, S. (2014). Facing the cold start problem in recommender systems. *Expert Systems with Applications*, 41(4), 2065–2073. <https://doi.org/10.1016/j.eswa.2013.09.005>.
- Lops, P., de Gemmis, M., Semeraro, G. (2011). Content-based Recommender systems: state of the art and trends. In Ricci, F., Rokach, L., Shapira, B., Kantor, P.B. (Eds.) *Recommender systems handbook* (pp 73–105) Boston, MA, Springer US.
- Madan, C.R. (2019). How well do different word properties explain item free recall? (Preprint). PsyArXiv. <https://doi.org/10.31234/osf.io/kfh87>.
- Metcalfe, J., & Kornell, N. (2005). A region of proximal learning model of study time allocation. *Journal of Memory and Language*, 52(4), 463–477. <https://doi.org/10.1016/j.jml.2004.12.001>.
- Mettler, E., Massey, C.M., Kellman, P.J. (2016). A comparison of adaptive and fixed schedules of practice. *Journal of Experimental Psychology: General*, 145(7), 897–917. <https://doi.org/10.1037/xge0000170>.
- Morey, R.D., & Rouder, J.N. (2018). BayesFactor: computation of bayes factors for common designs.
- Murphy, K.P. (2007). *Conjugate Bayesian Analysis of the Gaussian Distribution (Technical Report)*. Vancouver: BC:University of British Columbia.
- Nedungadi, P., & Remya, M.S. (2014). Predicting students' performance on intelligent tutoring system – Personalized clustered BKT (PC-BKT) model. In *2014 IEEE Frontiers in Education Conference (FIE) Proceedings 2014 Proceedings* (pp 1–6). Madrid, Spain: IEEE. <https://doi.org/10.1109/FIE.2014.7044200>.
- Nijboer, M. (2011). *Optimal Fact Learning: Applying Presentation Scheduling to Realistic Conditions (Unpublished Master's Thesis)*. Groningen: University of Groningen.
- Pardos, Z.A., & Heffernan, N.T. (2010). Modeling Individualization in a Bayesian Networks Implementation of Knowledge Tracing.

- In *Proceedings of the 18th International Conference on User Modeling, Adaptation and Personalization* (pp 255–266), Big Island, Hawaii. <https://doi.org/10.1007/978-3-642-13470-8-24>.
- Park, J.Y., Joo, S.H., Cornillie, F., van der Maas, H.L.J., van den Noortgate, W. (2018). An explanatory item response theory method for alleviating the cold-start problem in adaptive learning environments. *Behavior Research Methods*, 51(2), 895–909. <https://doi.org/10.3758/s13428-018-1166-9>.
- Pavlik, P.I., & Anderson, J.R. (2005). Practice and forgetting effects on vocabulary memory: an activation-based model of the spacing effect. *Cognitive Science*, 29(4), 559–586. <https://doi.org/10.1207/s15516709cog000014>.
- Pavlik, P.I., & Anderson, J.R. (2008). Using a model to compute the optimal schedule of practice. *Journal of Experimental Psychology: Applied*, 14(2), 101–117. <https://doi.org/10.1037/1076-898X.14.2.101>.
- 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. <https://doi.org/10.1016/j.compedu.2019.04.009>.
- Popescul, A., Ungar, L.H., Pennock, D.M., Lawrence, S. (2001). Probabilistic Models for Unified Collaborative and Content-Based Recommendation in Sparse-Data Environments. In *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence* (pp 437–444). San Francisco, CA.
- R Core Team (2018). R: a language and environment for statistical computing. r foundation for statistical computing. Vienna, Austria.
- Reiser, B.J., Anderson, J.R., Farrell, R.G. (1985). Dynamic Student Modelling in an Intelligent Tutor for LISP Programming. In *Proceedings of the 9th international joint conference on artificial intelligence* (pp 8–14). Los Angeles, CA.
- Rouder, J.N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, 21(2), 301–308. <https://doi.org/10.3758/s13423-014-0595-4>.
- Rouder, J.N., & Morey, R.D. (2012). Default Bayes factors for model selection in regression. *Multivariate Behavioral Research*, 47(6), 877–903.
- Schönbrodt, F.D., Wagenmakers, E.J., Zehetleitner, M., Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, 22(2), 322–339. <https://doi.org/10.1037/met0000061>.
- Sense, F., Behrens, F., Meijer, R.R., van Rijn, H. (2016). An individual's rate of forgetting is stable over time but differs across materials. *Topics in Cognitive Science*, 8(1), 305–321. <https://doi.org/10.1111/tops.12183>.
- Sense, F., van der Velde, M., van Rijn, H. (2018). Deploying a Model-based Adaptive Fact-Learning System in University Courses. In *Proceedings of the 16th International Conference on Cognitive Modeling* (pp 136–137), Madison, WI.
- Shernoff, D.J., Csikszentmihalyi, M., Schneider, B., Shernoff, E.S. (2003). Student Engagement in High School Classrooms from the Perspective of Flow Theory. In *Applications of Flow in Human Development and Education* (pp 19). Dordrecht: Springer. <https://doi.org/10.1007/978-94-017-9094-9-24>.
- van den Broek, G., Segers, E., van Rijn, H., Takashima, A., Verhoeven, L. (2019). Effects of elaborate feedback during practice tests: Costs and benefits of retrieval prompts. *Journal of Experimental Psychology: Applied*.
- van den Broek, G., Takashima, A., Wiklund-Hörnqvist, C., Karlsson Wirebring, L., Segers, E., Verhoeven, L. (2016). Neurocognitive mechanisms of the “testing effect”: a review. *Trends in Neuroscience and Education*, 5(2), 52–66. <https://doi.org/10.1016/j.tine.2016.05.001>.
- 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 on Cognitive Modeling* (pp 110–115). Manchester, United Kingdom.
- VanLehn, K. (2006). The Behavior of Tutoring Systems. *International Journal of Artificial Intelligence in Education*, 16, 227–265.
- Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the savage-dickey method. *Cognitive Psychology*, 60(3), 158–189. <https://doi.org/10.1016/j.cogpsych.2009.12.001>.
- Wauters, K., Desmet, P., van den Noortgate, W. (2010). Adaptive item-based learning environments based on the item response theory: Possibilities and challenges: Adaptive ITSs based on IRT. *Journal of Computer Assisted Learning*, 26(6), 549–562. <https://doi.org/10.1111/j.1365-2729.2010.00368.x>.
- Zhou, P., Sense, F., van Rijn, H., Stocco, A. (2020). Reflections of idiographic long-term memory characteristics in resting-state neuroimaging data (Preprint). bioRxiv. <https://doi.org/10.1101/2020.04.18.047662>.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.