Identification of Prior Knowledge for the Optimization of Adaptive Fact-Learning

Malte L. Krambeer

S4251970

Faculty of Social and Behavioural Sciences, University of Groningen

Supervisor: Prof. Dr. Hedderik van Rijn

July 29, 2024

Abstract

The present study demonstrates the possibility of predicting initial item knowledge for English vocabulary and Geography facts using learning data from the user's interaction with the MemoryLab adaptive learning system. Considering accuracy, speed of forgetting, and response time as potential predictors of initial item knowledge, we evaluate inclusion rates and regression coefficients of lasso regression models fitted during a k-fold cross-validation procedure. Additional modified prediction models are examined for the potential of adaptation to specific learning situations. We conclude that user and item accuracy measures provide the highest predictive power, with measures such as speed of forgetting and response time providing limited and more domain-specific value. Preliminary findings imply the potential to optimize adaptive fact learning by removing predicted known items from the learning set, and should be confirmed by replications in a realistic setting.

Keywords: Adaptive Learning, Prior Knowledge Identification, Lasso Regression

Identification of Prior Knowledge for the Optimization of Adaptive Fact-Learning

It is safe to assume that everyone is familiar with the tediousness of having to learn large amounts of information for a specific purpose, be it to lay the foundation for more engaging knowledge, or simply to pass the next vocabulary test. To help with this common struggle, students use various study methods, from simple read-and-review to the use of flashcards, visualization techniques, or digital trainers. Especially the latter kind has seen tremendous advancements throughout recent years, with digital tools permeating and optimizing many aspects of learning (Haidari et al., 2020; Nan Cenka et al., 2023; Straatemeier, 2014). This study will take an in-depth look at one such tool, specifically an adaptive learning system, which allows for user-specific optimization through an efficient scheduling of learning materials. Going beyond its mechanisms of adaptation, we investigate its ability to predict existing learner knowledge to further improve the learning process. The present paper thus aims to answer the question: How can a user's knowledge be predicted on the basis of their previous interaction with the learning system?

Existing Research

Adaptive Learning Systems

Adaptive Learning Systems (ALS) are digital tools that enable a user-specific learning experience with the intent to optimize learning gains (Anderson, 1995; Pavlik & Anderson, 2008; Sense et al., 2016; Straatemeier, 2014; van Rijn et al., 2009). In a wide range of studies, employing a similarly wide array of approaches, such systems have been demonstrated to surpass more traditional learning in terms of knowledge retention and other learning outcomes (Jastrzembski et al., 2006; Klinkenberg et al., 2011; Pavlik & Anderson, 2008; van der Velde et al. 2021; van Rijn et al., 2009; Wilschut et al., (submitted)).

The key to the success of ALS' is their adaptive nature. By developing an internal model of the user based on their performance and interaction with the system, a learning algorithm can

predict the optimal way of rehearsing the to-be-studied material (Pavlik & Anderson, 2005; Pavlik & Anderson, 2008; Sense et al., 2016; van der Velde et al. 2021; van Rijn et al., 2009). While the optimal way of studying may take different shapes, it is often concerned with providing the learner with material of suitable difficulty. This difficulty is determined by the user's familiarity with a given information, which influences the extent to which they can retrieve answers to questions or cues. Information is learned at the point when a user can reliably and correctly retrieve it from memory. The goal of an ALS, then, is to present new material that should be learned, while reinforcing the knowledge of material that has not yet been mastered.

Non-adaptive learning systems are known to result in frustrated or disinterested learners as they do not account for the variation in ability within their user base (Eggen and Verschoor, 2006; Hamari et al., 2016; Kennedy et al., 2014). Ambitious users of non-adaptive systems may run the risk of studying information that exceeds their current level of competence, resulting in a knowledge gap and an insufficient foundation. Conversely, more proficient students may disengage from learning when they sense a lack of challenge or progress (Eggen and Verschoor, 2006; Hamari et al., 2016; Kennedy et al., 2014). Adaptive learning systems can ameliorate these issues as they are able to select appropriate materials that a specific learner should study at a given time. They thus gain an advantage over more traditional methods, leading to better learning gains and a more engaging experience for users (Jastrzembski et al., 2006; Klinkenberg et al., 2011; van der Velde et al. 2021; van Rijn et al., 2009; Wilschut et al., (submitted)).

The goal of providing material of appropriate difficulty for a learner implies a secondary, yet vital, function of the tutor, which is the accurate assessment of the learner's capabilities. An adaptive learning system thus contains methods from the field of computerized adaptive testing (CAT), which provides widely used procedures for the precise measurement of user ability through digital means (Wainer et al., 2006). As a consequence, the goals of CAT closely align with those of adaptive learning systems, which extend this functionality only through their aim

of teaching the user the given material. An ALS should thus be able to develop a continuous assessment of the learner's ability, adapting to changes that occur directly through exposure to the material.

The MemoryLab Algorithm

Generally, an adaptive learning system adapts to each of its users individually, solely through the interaction of the two (Klinkenberg et al., 2011; van der Velde et al. 2021; van Rijn et al., 2009; Wilschut et al., (submitted)). Keeping score of user performance through answer correctness or reaction time, an algorithm may schedule the future learning process such that harder items are prioritized above already-mastered ones. An example of this is the MemoryLab algorithm (Sense et al., 2016; van der Velde et al. 2021; van Rijn et al., 2009), which will be the focus of this paper.

The MemoryLab algorithm is an extension of the adaptive spacing model by Pavlik and Anderson (2005 and 2008), which fits a computational cognitive model to a user's interaction with an item learning system (van der Velde et al. 2021; van Rijn et al., 2009). The model, in turn, is based on the ACT-R theory of declarative memory (Anderson, 2007), where each learned item is represented internally by a *memory chunk*. Each chunk has an activation strength associated with it, which signifies the item's probability of retrieval at a given time. This strength is a function of a user's encounters with the item at previous times and a decay parameter, indicating the speed at which an item's activation decreases. Through this decay parameter, items can differ in learnability, since it determines an estimate of the speed at which information will be forgotten by the user. Using this information, the MemoryLab algorithm can predict the time at which an item will be forgotten and cues it beforehand to strengthen its activation. In this manner, items with low activations can be rehearsed, while other, more well-learned items may be de-emphasized. New items are introduced once all others have been learned to a sufficient degree (van Rijn et al., 2009).

At the core of this mechanism lies the calculation of an item's decay. It is derived from an item's previous activation value, the time that has passed since the last occurrence, and a *decay intercept alpha*. This alpha value has been labeled the *speed of forgetting* (SoF) and is modified throughout a user's learning trajectory. When a user's response to an item is correct and unexpectedly fast, the SoF associated with that item will be adjusted down, resulting in slower decay and a prediction of better future performance. Conversely, when a user responds to an item incorrectly, or slower than estimated by the model, the SoF for that item will be increased, since the model incorrectly predicted a higher activation. As a result, the MemoryLab algorithm gradually establishes an optimal learning schedule, through which the learning of easier items is de-emphasized in favor of rehearsing the ones that are forgotten more quickly.

This form of adaptive learning system has already been shown to significantly improve learning outcomes in a range of applications (van der Velde et al. 2021; van Rijn et al., 2009, Wilschut et al., 2021; Wilschut et al., (submitted)). Additionally, its underlying model and parameters show a consistent relation to learning outcomes, while simultaneously allowing for performance prediction (Sense et al., 2016; Sense et al., 2018; Sense et al., 2021; van der Velde et al., 2020; van der Velde et al., 2023). For instance, interaction data of undergraduate students with the algorithm previously allowed Sense and colleagues (2021) to predict student's exam grades on a course involving the learned material. We may therefore expect a similar opportunity for the prediction of a learner's knowledge of specific items, resulting from the data that can be collected through their use of the system. This prediction could then be used to optimize the learning process even further by identifying and removing items that are likely to be known by the user.

Learning Optimization

Before investigating the possibility of user knowledge prediction, other means of optimization have already been researched. One deals with the problem of a *cold start*, which occurs at the very beginning of a learning session, when the algorithm has not yet adapted to the

user and its predictions are still inaccurate. This initial mismatch may be alleviated by considering prior information about the learner or the materials. Each may help create a pre-configured system that is, to some degree, already adapted to the expected characteristics of a learner (van der Velde et al., 2020; van der Velde et al., 2023). Specifically, information on the material may be considered to identify more difficult facts sooner, while learner-specific information may help to configure the algorithm's internal model ahead of time. Previous work by van der Velde (2021 and 2023) suggests that considering both these aspects helps to generate a more suitable item schedule, resulting in better learning and recall performance when a given set of materials varies in difficulty.

Another way of enhancing the adaptability of such an algorithm is to improve the adjustments made during the learning process. This can be done by gathering more performance data, such as item category-specific information, or by making certain assumptions about the material or learner, such that optimization may be accelerated (Wilschut et al., 2023; van der Velde et al., 2020; van der Velde et al., 2023). The latter point is especially relevant when working with large sets of material, which a learner may be expected to study in a short period of time. In the regular MemoryLab algorithm, each item must be encountered to create a perfectly adapted learning schedule. However, it may be possible to predict a user's knowledge of an item when enough performance data on other items is available. When the learner's ability level can be estimated with sufficient accuracy, an item that is predicted to be known can be discarded from the learning set, allowing even greater focus on more challenging material.

Research Goal

We observe that multiple methods of optimizing the learning process are indeed conceivable. All these methods are concerned with the identification of prior knowledge in the learner, which enables an ALS to emphasize more challenging material more efficiently. Continuing in this endeavor, this study sets out to investigate the possibilities of identifying a user's existing knowledge of specific items. To that end, we will consider item- and user-specific

information contained in existing learning data and create a regression model to identify predictors of specific item knowledge for a given user. Learning about the predictive potential of a user's interaction with the ALS may allow for further adaptation when unknown and challenging items can be identified and emphasized during learning. Ultimately, we will report on the theoretical and practical implications of our findings, and the possibility of integrating such predictions into adaptive learning systems.

Methods

In this section, we will outline the steps of the analysis process. Beginning with a description of the data and its underlying samples, we will move on to illustrate the use of the MemoryLab algorithm in data collection before describing other materials and software used. Ultimately, we will outline the specific analysis procedure and regression model used.

Data

Our analysis of the predictors of initial knowledge will be based on two datasets gathered in past studies involving the MemoryLab system. The first comprises learning data on country outlines (CO) of 40 highly populous or otherwise notable countries, previously collected in Wilschut (submitted). This data was gathered when teaching students the topological shape of countries, and we will refer to it as the *CO data* set. The second, more recent, set was gathered in a subsequent study by the same author (unpublished), and it comprises around 450 Dutch-English vocabulary items stemming from the *Engels in het Basisonderwijs* (EIBO) list of English terms. This list suggests vocabulary words that graduates of Dutch elementary schools who are going to lower secondary school are often expected to know. We will refer to this set as the *EIBO data* set.

The participants who contributed to the collection of the CO data were sampled from the population of first-year psychology students at the University of Groningen. A total of 104 students participated in a within-subjects experiment and used the MemoryLab system to study the outlines of 20 countries drawn from a pool of 40. A detailed description of the sample and item characteristics can be found in Wilschut (submitted). The EIBO data was gathered from the learning behavior of Dutch students who previously scored low on an English exam in their first year of high school. Their interaction with the system was recorded over four sessions and involved a 150-word subset of the previously mentioned *Engels in het Basisonderwijs* list. A detailed description of the data collection method sample is available in Wilschut (unpublished).

As outlined above, the MemoryLab system optimally and adaptively schedules to-be-learned items, as a result of the learner's previous responses. However, the algorithm and data collection methods for the CO and EIBO sets have the additional feature of including an *initial attempted retrieval* condition (Wilschut et al., (submitted); Wilschut et al., (unpublished)). This condition prompts users to respond to an item before teaching them the correct response. Consequently, it allows for the *dropping* of items, where items that were answered correctly upon their first encounter were removed from the learning set under the assumption that initially known items need no further teaching. The design of these conditions is the core reason for the suitability of these data sets for our analysis, since without the initial attempted retrieval, there would exist no information on prior knowledge levels of the users.

For the present analysis, the user's initial item knowledge (IIK), defined as the user's response to an item at its first encounter, will present the dependent variable. We will predict IIK with the help of parameters contained in the MemoryLab system and the learning data that is gathered throughout its use. Through a successful prediction, and the potential for dropping known items even before their first encounter, future learning sessions may be optimized further.

Analyses

The data processing and statistical analysis were run using R 4.4.2 (R Core Team, 2024). For graphical illustration, the ggplot2 package (Wickham et al., 2007) was used, while the glmnet package served to create the generalized mixed effects models (Friedman et al., 2008). We will begin our analysis with an exploration of our variable of interest: a user's initial knowledge of a specific learning item. Consequently, we will investigate the relationships between IIK and potential predictors for both user and item. An overview and definition of all potential predictors of IIK can be found in Table 1. It is important to note that all variables have been collected for item-user pairs, such that user variables reflect the characteristics of the user excluding the item to be predicted. Conversely, item variables reflect item characteristics

without consideration of the user whose knowledge will be predicted. As an example, a user's accuracy was calculated considering all items seen by the user, except for the item for which IIK is to be predicted. Using these predictors, we will conduct a k-fold cross-validation to evaluate multiple lasso regression models (Tibshirani, 1996) to investigate the predictive power of the independent variables listed in Table 1, and create a single prediction model that we will evaluate for practical uses.

Table 1

Collected Learning Data of the MemoryLab System and Potential Predictors of IIK

Domain	Variable	Description
User	Initial accuracy	The mean percentage of items that this user answered correctly upon first encounter
User	Speed of Forgetting (SoF)	The mean alpha value associated with items that this user encountered during learning, where higher alpha indicates a shorter retention time
User	Response time (RT)	The median amount of time a user spent on responding to items
Item	Initial accuracy	The mean proportion of users that responded correctly to this item upon first encounter
Item	Speed of Forgetting (SoF)	The mean average alpha value of this item across the different users that encountered it
Item	Response time (RT)	The median amount of time different user's took to respond to this item

Note. The dependent variable requires a pair of user and item, while all predictors are specific to either a user (excluding item information) or an item (excluding user information).

Descriptive Analysis

For our descriptive analysis, we will first provide an overview of the dependent variable, initial item knowledge. IIK is measured as the user's answer correctness to a learning item and therefore takes a binary value - correct or incorrect - for each user-item pair. For each data set,

we will report the mean and standard deviation of initial item knowledge. Additionally, we provide a similar overview of the independent variables, which constitute the predictor variables of IIK in our later regression model. As can be seen in the list of predictors in Table 1, these variables will be reported for both items and users, since item and user characteristics may hold different predictive values. Lastly, we will determine the correlation coefficients and their significance between all pairs of variables to provide an initial overview of the underlying relations that may inform predictive power.

Lasso Regression Model

Initially proposed by Tibshirani, (1996), *Least Absolute Shrinkage and Selection Operator* (Lasso) regression is a common machine learning approach to handle high-dimensional data through regularization. The method adds a penalty equal to the residual sum of squares of the regression coefficients. This penalty encourages the model to shrink some coefficients to zero, effectively performing variable selection. In doing so, it simplifies the model, enhances interpretability, and can improve prediction accuracy by mitigating overfitting. Additionally, it handles multicollinearity between predictors by similarly reducing coefficients of correlated, but less predictive, variables (Hastie et al., 2009, as cited in Sense et al., 2021). Therefore, we see lasso regression as a suitable means of predicting initial item knowledge.

In addition to its use as a prediction model, lasso regression also facilitates an evaluation of the predictive power of our variables. Since the standardization of predictors is necessary when creating a lasso regression model, the resulting coefficients directly indicate the weight of each predictor on the final result. Previous research successfully employed this feature of lasso regression when predicting students' exam grades using assessments of their learning performance within MemoryLab (Sense et al., 2021). We will adapt their analysis methods for the purpose of IIK prediction, and extend their research by also investigating the predictive power of item characteristics.

Specifically, we will implement a k-fold cross-validation procedure and generate 300 lasso regression models to achieve a robust estimate of prediction performance. In each of the 300 folds, learning data will be split into training (80%) and evaluation (20%) sets. A lasso model will be fit to the training data and a decision threshold of 0.5 will be applied to the model responses in order to generate binary predictions, which can then be compared against the evaluation set. The model's coefficients and prediction performance will be saved. After performing the cross-validation, we can investigate three outcomes: (1) the overall predictive performance, measured in terms of accuracy, precision, recall, and F1-score, (2) the coefficient values and their resulting relative impact on the prediction, and (3) the inclusion rate of the predictors in the model. The latter outcome results from the lasso model reducing some coefficients to zero, essentially removing predictors from the model. This feature allows us to further evaluate the importance of predictors for estimating initial item knowledge.

Single Prediction Model

Lastly, a single prediction model will be trained using lasso regression to evaluate precision, recall, and F1 scores. In doing so, we can evaluate a realistic model more in-depth, and explore potential adjustments for practical application. Since predictions of the model fall in the interval [0, 1], a decision criterion can be implemented to mitigate the model's type-I or type-II error rates. In varying this decision threshold when predicting IIK, we demonstrate the creation of more conservative or more lenient model variants. Confusion matrices for these variations will be provided to illustrate the model's potential applications in educational settings.

Results

Descriptive Analysis

In the following sections, we will outline the analysis results of the EIBO and CO data sets. In order to provide an overview of the variables in question, we will first describe both the

outcome and predictor variables and investigate their correlations: In both datasets, the dependent variable is the user's initial item knowledge of a learning fact in the initial attempted retrieval condition. In the EIBO data set, the mean for IIK was 0.445 (sd = 0.497), while it was 0.434 (sd = 0.496) in the CO data set. These values indicate a similar level of domain knowledge with high variability between individuals for English vocabulary and country outlines in both samples tested.

In both data sets, we assessed the user's learning performance through the collection of a variety of metrics, which constitute our independent variables, or predictors (see Table 1). An overview of the descriptive statistics for these variables can be found in Table 2. The results show that users in both studies have a similar level of existing domain knowledge, but that the country outline facts are less known than English vocabulary. In addition, the higher SoF values in the CO data indicate that these facts are forgotten more quickly than English vocabulary. Finally, the time needed to respond is lower and varies much less for EIBO users and items, compared to users and items of the CO data.

Table 2

Descriptive statistics for dependent variables in the EIBO and CO data

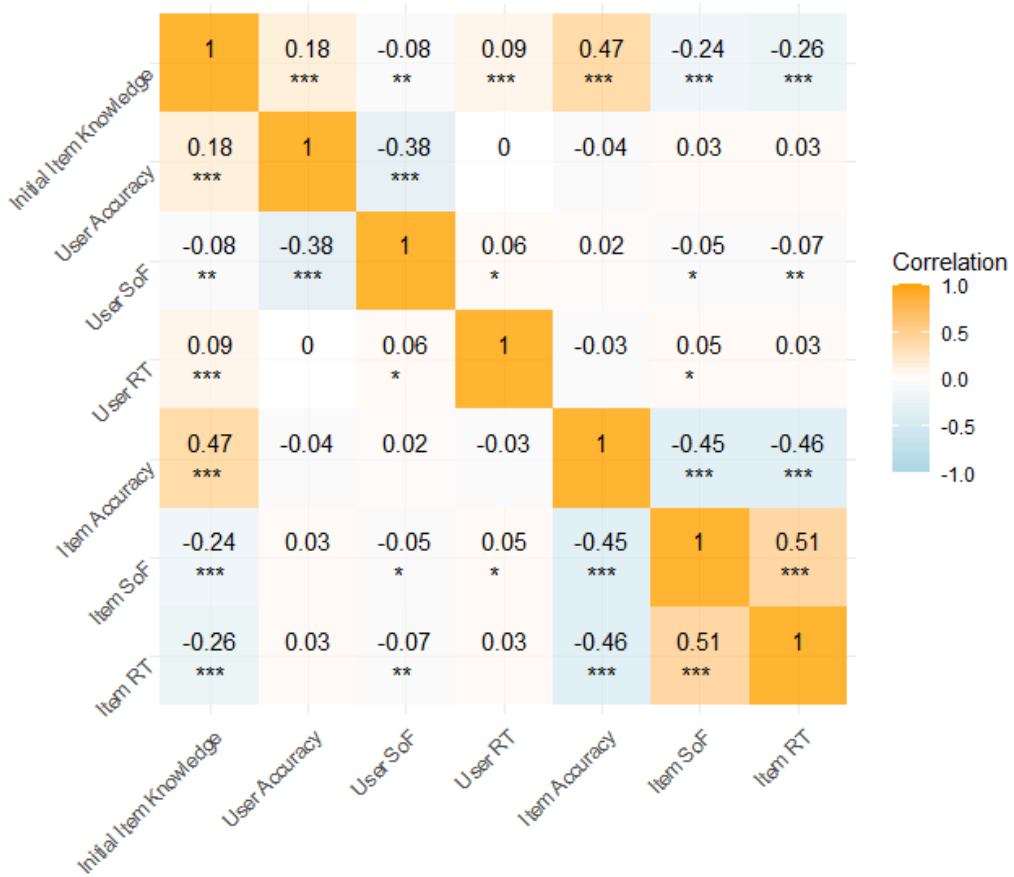
		EIBO Data		Country Outlines Data	
Domain	Predictor	Central tendency	Variability	Central tendency	Variability
User	Initial accuracy (Mean, SD)	0.433	0.172	0.405	0.254
User	SoF (Mean, SD)	0.263	0.024	0.347	0.026
User	RT (Median, IQR)	1993.125	636.625	2576.223	1628.598
Item	Initial accuracy (Mean, SD)	0.447	0.297	0.335	0.241
Item	SoF (Mean, SD)	0.252	0.035	0.347	0.019
Item	RT (Median, IQR)	1814.5	433.25	2365.66	1379.497

To investigate the relationships between IIK and its predictors, as well as between the predictors themselves, we assess their correlations for significance. An overview of the correlations between variables from the EIBO data can be found in Figure 1, and an overview of the correlations between variables from the CO data can be found in Figure 2. For the EIBO data, the results show significant correlations between IIK and all predictors, with stronger correlations for item characteristics than for user characteristics. For the CO data, the results show significant correlations between IIK and all predictors, but stronger correlations are observed in user and item accuracy. Interestingly, the correlations between IIK and user response time (RT) show opposing values: 0.09 in the EIBO data and -0.19 in the CO data. This indicates that users who spent more time answering questions about country outlines were less likely to answer correctly, while the opposite is true for English vocabulary items.

Since multiple predictors correlate with IIK, we can expect them to possess predictive potential as well. For instance, given that a high overall item accuracy is associated with a user's initial answer being correct, we anticipate a high inclusion rate and non-zero coefficient for item accuracy as a factor in the lasso regression model. However, Figures 1 and 2 also show significant multicollinearity between the predictors. For instance, an item's SoF correlates negatively with IIK, but also with item accuracy. This leads us to believe that their combined predictive power is less than the sum of their individual powers, meaning that the lasso regression model is likely to penalize one of the predictors.

Figure 1

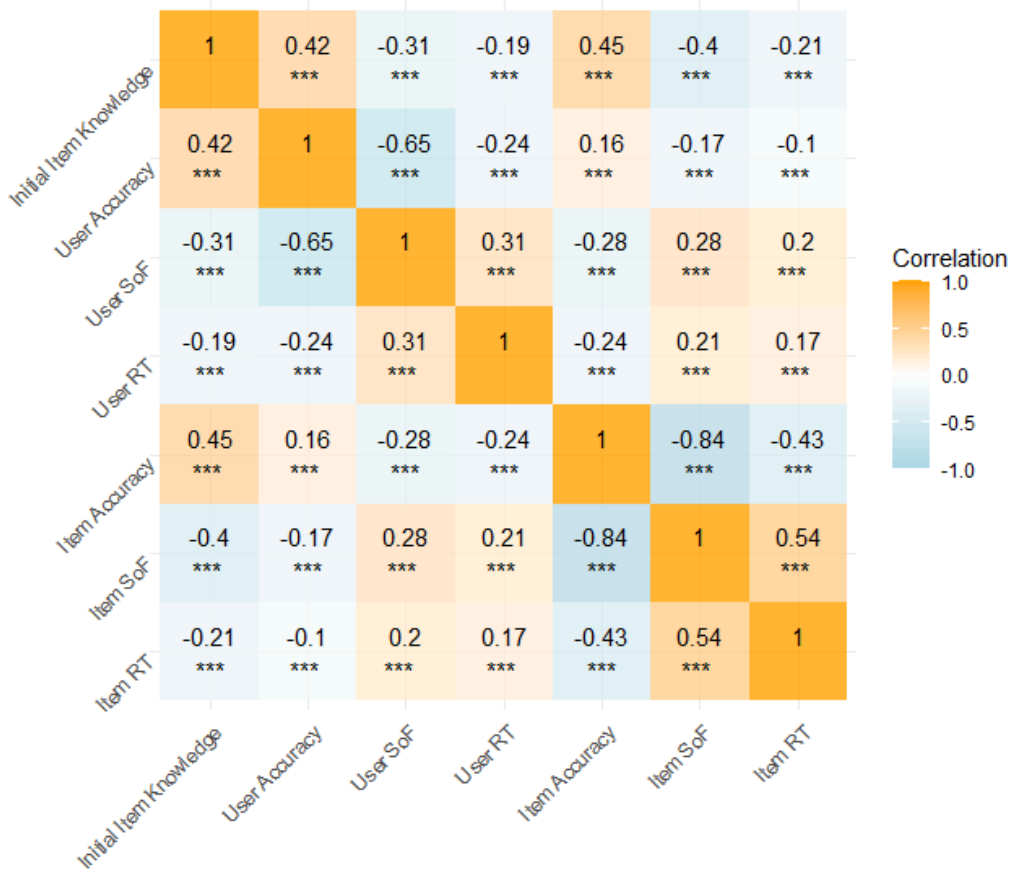
Correlation coefficients and significance of variables in the EIBO data



Note. Asterisks '***' indicate a p-value < 0.001; '**' a p-value < 0.01, and '*' a p-value < 0.05.

Figure 2

Correlation coefficients and significance of variables in the country outlines data



Note. Asterisks '***' indicate a p-value < 0.001; '**' a p-value < 0.01, and '*' a p-value < 0.05.

Lasso Regression Model

For the purpose of training a lasso regression model, two participants in the EIBO study and one participant in the CO study were removed due to incomplete data. During 300 cross-validation runs, the data was split into training and validation subsets. Individual lasso regression models were first fit to the training data, then evaluated on the validation data. In doing so, we collected regression coefficient values, inclusion rates, and model performance metrics from each fold. Figure 3 shows the density distribution for each coefficient used to

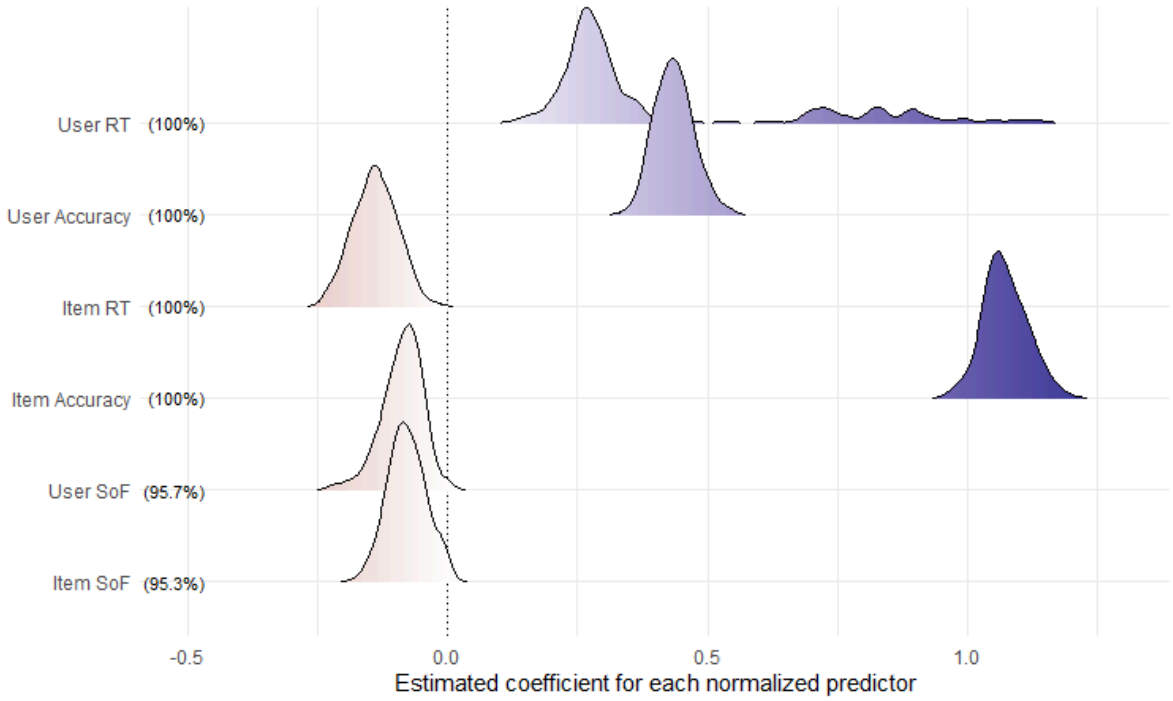
predict initial item knowledge in the EIBO data, ordered by inclusion rate. Figure 4 shows the same information for the CO data. Model performance metrics can be found in Table 3.

Regarding the EIBO data, Figure 3 shows that all predictors were included in the regression models more than 95% of the time. This indicates that each predictor adds value to the model, even if some of their coefficient values were penalized to account for multicollinearity. In particular, we observe significant predictive power in item and user accuracy, which have the most consistent distribution of coefficients and were included in all prediction models. Additionally, user RT has a 100% inclusion rate, but its coefficient value varies strongly between models. This is likely due to the strongly right-skewed distribution of response times in the user data. A chance overrepresentation of extreme values in the training data may have led some regression models to overestimate the predictive power of user RT, leading to the distribution of coefficients shown in Figure 3.

Less predictive variables are item RT, as well as user SoF and item SoF. While still included in most models, their coefficients lie close to zero, indicating that their predictive power overlaps in part with that of other variables. Looking at the correlation matrix in Figure 1, we see that each of these predictors is strongly correlated with at least one of the stronger predictors mentioned above. It can therefore be assumed that the model was regularized to shrink the coefficients of predictors with less explanatory power.

Figure 3

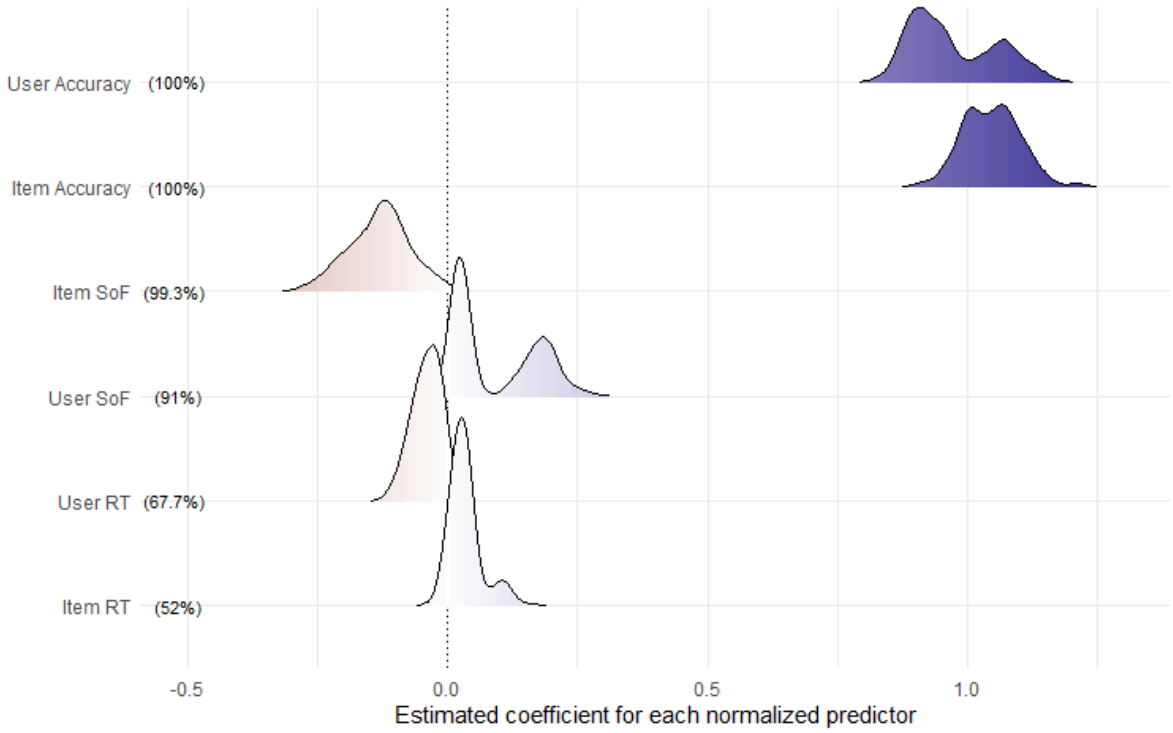
Density plot and inclusion rates of the estimated coefficients of lasso regression models trained on the EIBO data



Regarding the CO data, Figure 4 shows that all predictors except user RT and item RT were included in more than 90% of models. Out of the included predictors, only user accuracy and item accuracy show coefficient values that are clearly distinct from zero, indicating that they possess the highest predictive power. Item SoF and user SoF were included in 99.3% and 91% of models, respectively. However, their coefficients lie close to zero, indicating that they were likely penalized during model fitting to account for their correlation with the accuracy terms, which are shown in Figure 2. Interestingly, User SoF received a positive coefficient in many regression models, meaning that higher SoF was found to be predictive of IIK in such models. Lastly, user RT and item RT were included in 67.7% and 52% of models, respectively. The distribution of their coefficients is centered close to zero, indicating that they do not provide much value over other variables when predicting IIK.

Figure 4

Density plot and inclusion rates of the estimated coefficients of lasso regression models trained on the country outlines data



For each set of data, the generated regression lasso models demonstrate above-chance classification accuracy. At the same time, the precision and recall scores for both model sets indicate good discriminability between positives (initially known items) and negatives. A consistently higher precision in both sets shows the pronounced ability to avoid false positives specifically. However, this outcome may also be owed to a high proportion of negatives in the data. Nonetheless, the F1-score averages for both sets of models indicate a reasonable balance between capturing positive instances and keeping their identification reliable. Finally, the low standard deviations across all metrics collected suggest a very stable model performance across data splits.

Table 3

Performance metrics for the lasso regression models predicting initial item knowledge

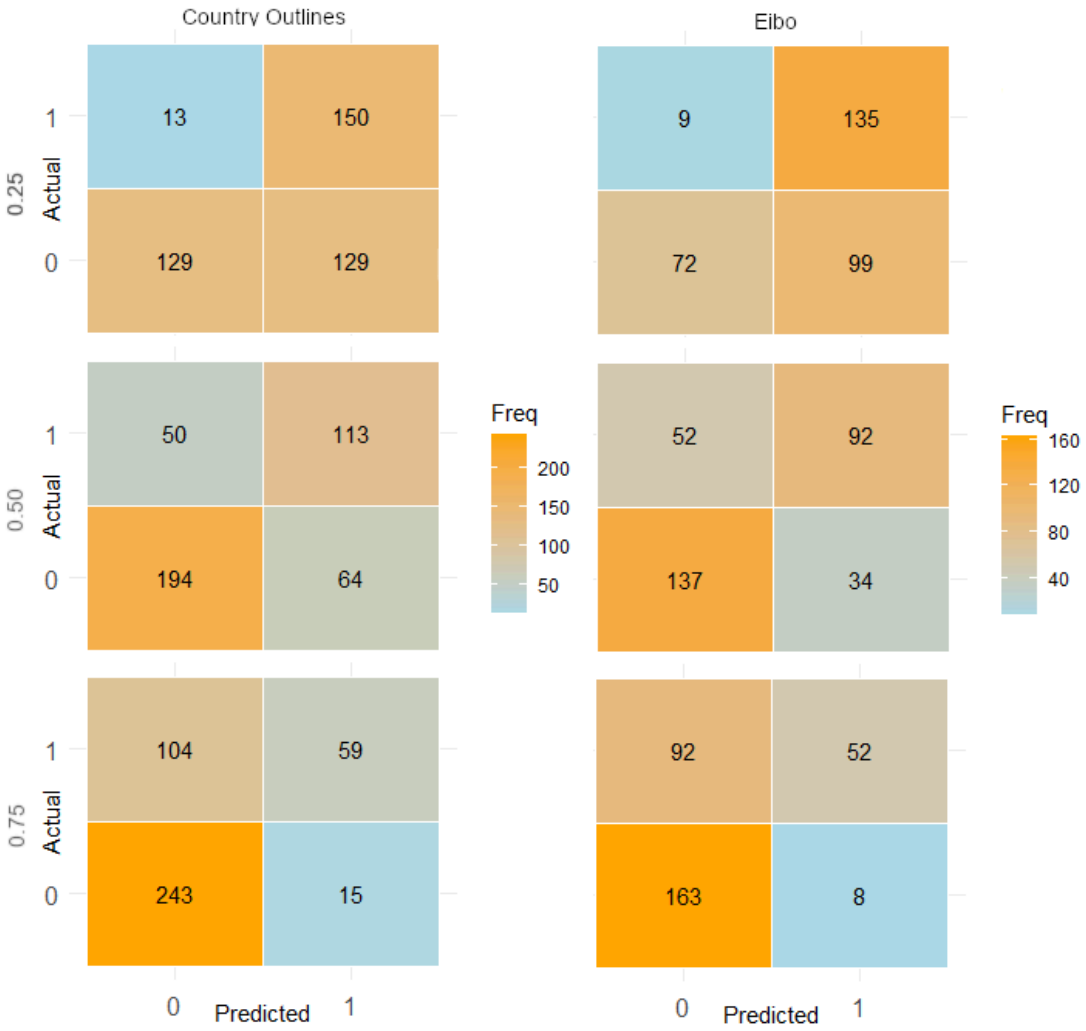
	Eibo Data		Country Outlines Data	
Metric	Mean	SD	Mean	SD
Accuracy	0.732	0.023	0.738	0.02
Precision	0.714	0.038	0.703	0.032
Recall	0.652	0.042	0.67	0.038
F1-Score	0.68	0.029	0.685	0.026

Single Predictor Model

In the final part of our analysis, we trained a single lasso regression model in the same manner as previously in each of the folds during cross validation. However, in addition to the previously used decision threshold of 0.5, other values were applied to explore model variants that emphasize precision or recall over pure accuracy. As shown in Figure 5 (top row), a lower decision threshold reduces the proportion of false negatives, but increases the risk of wrongly identifying items as known. Conversely, raising the decision threshold (bottom row), increases the proportion of false negatives, while reducing the risk of misclassifying unknown items as known. For instance, lowering the decision threshold for the EIBO prediction model in Figure 5 to 0.25, we are able to raise its recall to 0.94, meaning that out of all initially known items, 94% were correctly identified as such. Raising its decision threshold to 0.75, we raise its precision to 0.87, meaning that 87% of all predicted known items were truly known. Compared to the baseline model with a decision threshold of 0.5, these modifications allow for increases of 31% in recall when lowering the threshold and 14% in precision when raising the threshold. However, the modifications come at the cost of raising the Type I and Type II error rates respectively. Nonetheless, these possibilities allow for a more nuanced applicability of the model for practical purposes.

Figure 5

Confusion matrices for model predictions with varying decision thresholds



Note. Models use a decision threshold of 0.25, 0.50, and 0.75 (from top to bottom)

Discussion

Adaptive learning systems have significantly improved the effectiveness of fact-learning in the past (Anderson, 1995; Haidari et al., 2020; Nan Cenka et al., 2023; Straatemeier, 2014). As an ALS, the MemoryLab system optimizes study time by scheduling learning items adaptively, such that more difficult items are rehearsed more often, while mastered items are de-emphasized (Sense et al., 2016; van der Velde et al. 2021; van Rijn et al., 2009). However, optimization currently requires full traversal of the item set, resulting in long study times for larger item sets, even with optimal scheduling. Aiming to improve on this limitation, the current study investigated the possibility of predicting a user's initial item knowledge based on measurements of their previous learning behaviour and item characteristics. Items that can successfully be identified as known, even before their first encounter, may then be removed from the learning data. In doing so, material that is difficult to the learner can be identified and emphasized during study. Learning analytics from two datasets containing English vocabulary and country outline facts enabled both evaluation and comparison of their predictive potential.

After running a k-fold cross validation which fitted and evaluated 300 lasso regression models on subsets of EIBO and CO data, findings point to a significant but variable predictive power in all learning measures. Out of the six predictors assessed, item and user accuracy serve as the most predictive variables. Their importance for the prediction of IIK is highlighted by the consistently and distinctly non-zero distributions of regression coefficients, showing that each of the generated models included the accuracy predictors in its regression equations. Specifically, item accuracy proved to be highly predictive of IIK in both data sets, whereas user accuracy is more predictive of CO knowledge than of English vocabulary. The role of user and item accuracy also makes intuitive sense, since users that previously demonstrated their item knowledge are likely to also know other items of the same domain, while items that are known by many users are likely to be known by others as well. These findings are in line with results by Wilschut and

colleagues (2023) and Sense and colleagues (2021) who previously demonstrated the potential of accuracy metrics for the prediction of learning outcomes.

User and item speed of forgetting proved to be predictive for IIK in both datasets. Being included in more than 90% of all prediction models, its regression coefficient distributions are centered at low negative values away from zero, indicating slight predictive potential. One exception is user SoF as a predictor of initial CO item knowledge, since it shows a split distribution of regression coefficients in the positive range, despite negatively correlating with IIK. A possible explanation for this pattern is that only a subset of SoF values were predictive of IIK, in which case faster speeds of forgetting in users actually predict initial knowledge. However, this would contradict previous research on the parameter, since it is usually associated with less item knowledge (Pavlik and Anderson, 2005; Sense et al., 2018; Sense et al., 2021; van Rijn et al., 2009). It is therefore possible that the observed effect is an artifact of the data, rather than representing real predictive potential and the result should be considered with caution.

Finally, the predictive role of user and item response times differs between item sets. In the domain of English vocabulary, both RT measures are slightly predictive of IIK, albeit in opposing directions. Lower median response times to learning items predicted their knowledge in all regression models, while higher median response times in users predicted their IIK. This is in line with the positive correlation of user RT and IIK in the EIBO data and shows that users who spent more time responding to newly introduced cues were more likely to answer correctly. In the CO data, the RT measures were included in about half of the regression models, with coefficients close to zero. This indicates that response times to country outline items is hardly predictive of initial knowledge, possibly since fundamental differences in response strategies exist between individuals: some may answer correctly after thinking for longer, while others either know the answer immediately or not at all.

Despite differences in the predictive power of SoF and RT measures, prediction models consistently achieved above-chance classification accuracy and reasonable precision and recall.

Considering the relatively low predictive power of the non-accuracy measures, the similarity of prediction performances between both types of models may predominantly rest on the predictive power of the accuracy measures. We thus observe a need for the investigation of additional learning measures that may inform IIK prediction. Additional, or more detailed, learning analytics could increase prediction accuracy and provide more insight into determinants of knowledge in users and domains. More detailed predictors may be computed for subdomains of the learning sets, providing accuracy measures, for instance, for verbs and nouns separately, when predicting knowledge of vocabulary items. Additional predictors could be found in learning data from other domains of knowledge, investigating a transfer of IIK between, for instance, geographical and vocabulary items.

An additional limitation of our findings lies in the fact that prediction models have been fitted using complete sets of learning data, which facilitates the prediction of items that users encountered. In realistic applications, a prediction model would be trained during the learning process to create a running estimate of IIK for each item-user pair, and would have substantially less information to work with initially. It is therefore critical to investigate the applicability of such models in situations where data becomes available over time. While initial prediction accuracies may be low, it can be expected that additional information will further improve performance. Especially in learning systems that host a variety of lessons, an abundance of user and item data may greatly benefit prediction models, even across domains and users.

Even for limited sets of training data, modified prediction models have the potential to aid in learning optimization. Since the MemoryLab algorithm is applied in an educational setting, it may be desirable to minimize Type I errors by preventing the incorrect prediction of existing item knowledge. This is especially relevant when users study smaller sets of items, of which a large percentage may be important for a later test. Conversely, when large sets of items are studied, or a quick identification of the most difficult material is critical, a more lenient prediction model may be desirable. To this end, the adjustment of the decision threshold allows

for a fine-tuning of the model to favor either recall or precision. However, the usability of such prediction models should be validated in realistic settings, where the need to train a new model may initially preclude the benefit of IIK prediction.

In summary, the present study demonstrates the potential of predictive models to estimate initial item knowledge in adaptive learning systems, using measures of user and item accuracy, speed of forgetting, and response times. The results highlight the role of accuracy metrics in predicting IIK, while the predictive power of SoF and RT remains limited and varies between domains. Despite their theoretical nature and the need for further research into additional predictors and realistic applicability, the findings suggest that the prediction of IIK can optimize adaptive learning systems. By modifying prediction models for specific learning situations, adaptive systems such as MemoryLab can target unknown knowledge more effectively, and ultimately improve the learning experience and outcomes for its users.

References

Anderson, J., Corbett, A., Koedinger, K., & Pelletier, R. (1995). Cognitive Tutors: Lessons Learned. *Journal of the Learning Sciences*, 4, 167–207. https://doi.org/10.1207/s15327809jls0402_2

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

Eggen, T. J. H. M., & Verschoor, A. J. (2006). Optimal Testing With Easy or Difficult Items in Computerized Adaptive Testing. *Applied Psychological Measurement*, *30*(5), 379–393. https://doi.org/10.1177/0146621606288890

Friedman, J., Hastie, T., Tibshirani, R., Narasimhan, B., Tay, K., Simon, N., & Yang, J. (2008). glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models (p. 4.1-8) [software]. https://doi.org/10.32614/CRAN.package.glmnet

Haidari, S. M., Baysal, S., & Kanadli, S. (2020). THE IMPACT OF DIGITAL TECHNOLOGY-MEDIATED FOREIGN LANGUAGE INSTRUCTION ON VOCABULARY LEARNING: A META-ANALYTIC REVIEW. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi*, *20*(1), 236–251. https://doi.org/10.17240/aibuefd.2020.20.52925-552769

Hamari, J., Shernoff, D., 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*. https://doi.org/10.1016/j.chb.2015.07.045

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

Jastrzembski, T., Gluck, K., & Gunzelmann, G. (2006). Knowledge tracing and prediction of future trainee performance. 1498–1508.

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

Nan Cenka, B. A., Santoso, H. B., & Junus, K. (2023). Personal learning environment toward lifelong learning: An ontology-driven conceptual model. *Interactive Learning Environments*, *31*(10), 6445–6461. https://doi.org/10.1080/10494820.2022.2039947

Pavlik Jr, P., & Anderson, J. (2005). Practice and Forgetting Effects on Vocabulary Memory: An Activation‐Based Model of the Spacing Effect. *Cognitive Science*, 29, 559–586. https://doi.org/10.1207/s15516709cog0000_14

Pavlik Jr, P., & Anderson, J. (2008). Using a Model to Compute the Optimal Schedule of Practice. *Journal of Experimental Psychology. Applied*, 14, 101–117. https://doi.org/10.1037/1076-898X.14.2.101

R Core Team. (2024). A Language and Environment for Statistical Computing [Computer software]. *R Foundation for Statistical Computing*. www.R-project.org

Rijn, H., Maanen, L., & Woudenberg, M. (2009). Passing the test: Improving learning gains by balancing spacing and testing effects.

Sense, F., Behrens, F., Meijer, R. R., & Van Rijn, H. (2016). An Individual's Rate of Forgetting Is Stable Over Time but Differs Across Materials. *Topics in Cognitive Science*, *8*(1), 305–321. https://doi.org/10.1111/tops.12183

Sense, F., Meijer, R. R., & Van Rijn, H. (2018). Exploration of the Rate of Forgetting as a Domain-Specific Individual Differences Measure. *Frontiers in Education*, 3, 112. https://doi.org/10.3389/feduc.2018.00112

Straatemeier, M. (2014). Math Garden: A new educational and scientific instrument. https://dare.uva.nl/search?metis.record.id=417091

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Van Der Velde, M., Sense, F., Borst, J. P., & Van Rijn, H. (2023). Large-scale evaluation of cold start mitigation in adaptive fact learning: Knowing "what" matters more than knowing "who." https://doi.org/10.31234/osf.io/z3vtn

Velde, M., Sense, F., Borst, J., & Rijn, H. (2020). Alleviating the Cold Start Problem in Adaptive Learning using Data-Driven Difficulty Estimates. https://doi.org/10.31234/osf.io/hf2vw

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., & Mislevy, R. J. (2014). Computerized Adaptive Testing: A Primer (2nd ed.). *Routledge*. https://doi.org/10.4324/9781410605931

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., Dunnington, D., & Van Den Brand, T. (2007). ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics (p. 3.5.1) [software]. https://doi.org/10.32614/CRAN.package.ggplot2

Wilschut, T., Sense, F., van der Velde, M., Fountas, Z., Maaß, S. C., & van Rijn, H. (2021). Benefits of Adaptive Learning Transfer From Typing-Based Learning to Speech-Based Learning. *Frontiers in Artificial Intelligence*, *4*. https://doi.org/10.3389/frai.2021.780131

Wilschut, T., van der Velde, M., Sense, F., Finn, B., Arslan, B., & van Rijn, H. (submitted). Attempted Retrieval Benefits are Limited in Realistic Learning Settings, Unless Used for Prior-Knowledge Based Personalisation