Deep (Un)Learning: Using Neural Networks to Model Retention and Forgetting in an Adaptive Learning System

Jeffrey Matayoshi[1]*, Hasan Uzun[1], and Eric Cosyn[1]

McGraw-Hill Education/ALEKS Corporation, Irvine, CA, USA
{jeffrey.matayoshi,hasan.uzun,eric.cosyn}@aleks.com

Abstract. ALEKS, which stands for "**A**ssessment and **LE**arning in **K**nowledge **S**paces", is a web-based, artificially intelligent, adaptive learning and assessment system. Previous work has shown that student knowledge retention within the ALEKS system exhibits the characteristics of the classic Ebbinghaus forgetting curve. In this study, we analyze in detail the factors affecting the retention and forgetting of knowledge within ALEKS. From a dataset composed of over 3.3 million ALEKS assessment questions, we first identify several informative variables for predicting the knowledge retention of ALEKS problem types (where each problem type covers a discrete unit of an academic course). Based on these variables, we use an artificial neural network to build a comprehensive model of the retention of knowledge within ALEKS. In order to interpret the results of this neural network model, we apply a technique called permutation feature importance to measure the relative importance of each feature to the model. We find that while the details of a student's learning activity are as important as the time that has passed from the initial learning event, the most important information for our model resides in the specific problem type under consideration.

Keywords: Forgetting curves · Neural networks · Knowledge space theory · Adaptive learning · Permutation feature importance

1 Introduction

ALEKS, which stands for "**A**ssessment and **LE**arning in **K**nowledge **S**paces", is a web-based, artificially intelligent, adaptive learning and assessment system [18]. The artificial intelligence of ALEKS is a practical implementation of knowledge space theory (KST) [5, 7, 8], a mathematical theory that employs combinatorial structures to model the knowledge of learners in various academic fields of study including math [14, 22], chemistry [12, 26] and even dance education [31].

Understanding the behavior of retention and forgetting within adaptive systems is an important area of research, as it has been shown that student models can be significantly improved when these aspects of learning are accounted for

* Corresponding author.

[21, 27]. Furthermore, some previous results have emphasized the importance of identifying the variables that affect forgetting [28, 29], while others have shown that personalized interventions and review schedules can improve students' long-term retention of knowledge [16, 30].

Motivated by these previous works, in this study we analyze in detail the factors that affect the forgetting and retention of knowledge within the ALEKS system. Given that the retention of knowledge within ALEKS exhibits the characteristics of the famous Ebbinghaus forgetting curve [1, 6, 17], we begin with some exploratory data analysis of the factors affecting this curve. Based on these results, we then build a comprehensive model of forgetting and retention within ALEKS using an artificial neural network. Finally, by combining our exploratory data analysis with an application of permutation feature importance [2, 24, 25], we are able to get a clearer understanding of the relative importance of each of the features to our final neural network model.

2 Background

In KST, an *item* is a problem type that covers a discrete unit of an academic course. Each item contains many examples called *instances*, and these examples are carefully chosen to be equal in difficulty and to cover the same content. A *knowledge state* in KST is a collection of items that, conceivably, a student at any one time could know how to do.

Another concept important to our study is the *inner fringe* of a knowledge state. An item is contained in the inner fringe of a knowledge state when the item can be removed from the state and the remaining set of items forms another knowledge state. An inner fringe item can be viewed as being at the edge of a student's knowledge, as complete knowledge of the item is not required to know any of the other items in the knowledge state.

Within the ALEKS system, the student is guided through a course via a cycle of learning and assessments. In an assessment, a student is presented an item for which they can attempt to answer, or they can respond "I don't know" if they, presumably, have little knowledge of how to solve the problem. If the student attempts to answer the item, the response is classified as either correct or incorrect. A course begins with an *initial assessment*, the goal of which is to accurately measure the starting knowledge of the student. Then, in the learning mode, the student is presented items based on her knowledge state, with the system tracking the student's performance and continually updating the student's knowledge state. Each subsequent *progress assessment* is given to a student after some time has been spent in the learning mode, and the process continues. The purpose of these progress assessments is to verify the student's recent learning, as well as to act as a mechanism for enforcing spaced practice and retrieval.

For the purposes of this study, we define *retention* as the act of answering an item correctly on a progress assessment at a point in time after the item is learned in ALEKS. We can then define the *retention rate* as the correct answer rate to these assessment questions. For our analyses, we gather data from over 3.3

million ALEKS progress assessment questions that are drawn from 10 different math and chemistry courses. The questions we use are restricted to items that are contained in the inner fringe of the student's knowledge state. In looking only at inner fringe items, we are attempting to reduce any bias from students reinforcing the core knowledge of an item by working on related content. We partition the data into a training set of $2,989,835$ questions, along with validation and test sets each consisting of $166,102$ questions. In addition to being used to train our neural network models, we also use the training data to perform all our exploratory data analysis. The validation set is used to test different neural network architectures and tune the hyperparameters, while the test set is used for the final model evaluation.

3 Forgetting Curve

As shown in [17], the retention rate of an inner fringe item in ALEKS changes as a function of the time since the item was learned (with an item being "learned" after a certain amount of success on the item has been demonstrated in the learning mode). To see this, for each assessment question in our training data we compute the number of days from the time the student learned the item to the time the item appeared in the progress assessment, and then we group these questions based on the outcome (correct, incorrect, or "I don't know"). The results are shown in Figure 1, where the solid curve (the proportion of corrects) can be considered a forgetting curve [1, 6]. Note that this curve is analogous to the curve first shown in [17].

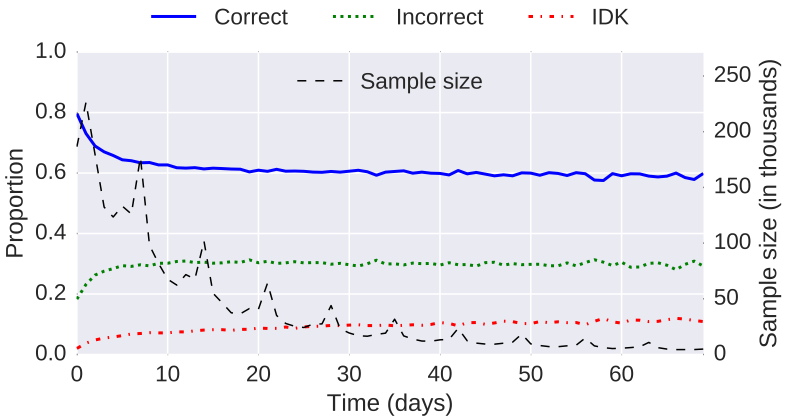

Fig. 1. Proportions of responses as a function of the time (in days) since the inner fringe item appearing as a progress assessment question was learned.

At this point it should be emphasized that inner fringe items are at a very specific, and critical, place in a student's knowledge state. The overall retention rate on these items is relatively modest, with the average correct rate for our dataset being 0.64. Since an inner fringe item has recently been learned by a student, without any learning that reinforces the skill(s) contained in the item, the relatively low correct rate is not unexpected. Thus, predicting the retention of inner fringe items is a difficult task, as the items that are most likely to be retained (i.e., the items with highest correct rates) would not generally be found in the inner fringe. (See Figure 1 in [4] for an example of how the correct rate increases for items "deeper" in the knowledge state.) That being said, there are many factors that can affect the inner fringe correct rate, and it is important to identify these factors when building models of retention [28, 29]. Thus, in the next section we take a look at these factors in more detail.

4 Exploratory Data Analysis

Now that we have established a baseline forgetting curve, we can look at what factors, or variables, affect this curve. The first variable we discuss is the knowledge of the student at the beginning of the course, which is measured by what we call the student's initial score; this is simply the proportion of the items in the course that are in the student's knowledge state at the end of the initial assessment. The results are shown in Figure 2, which compares the forgetting curves for students in the first decile (in terms of the initial score) and in the tenth decile. We can see that there is a relatively large gap between the two correct answer curves. Additionally, the "I don't know" curves show an interesting difference, in that the students in the first decile show an increasing rate of "I don't know" answers over time, while the students in the tenth decile have a constant rate after about a week.

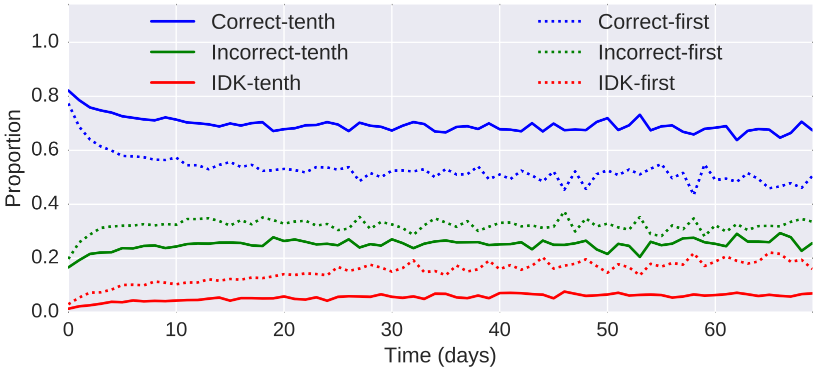

Fig. 2. Proportions of responses conditioned on the student being in the first decile (in terms of initial score) or the tenth decile. The top set of lines (blue) represents the correct responses, the middle set (green) represents the incorrect responses, and the bottom set (red) represents the "I don't know" responses.

The next factor we consider is the classification of the learned item after the student's initial assessment. An ALEKS assessment finishes with each item classified into one of three distinct categories.

- Items that are most likely in the student's knowledge state (in-state)
- Items that are most likely not in the student's knowledge state (out-of-state)
- The remaining items (uncertain).

The learned items in our dataset are exclusively composed of items classified as either out-of-state or uncertain after the initial assessment. The out-of-state items are items that the ALEKS system, at the conclusion of the initial assessment and based on the student's responses to the assessment questions, strongly believes the student does not know. On the other hand, the uncertain items are those for which the system does not have enough information to make a confident classification of either in-state or out-of-state. Thus, it stands to reason that a good portion of these uncertain items are actually items that the student already knows, in which case the forgetting curve and retention may take a different form. The results in Figure 3 support this conjecture, where there is a clear separation between the forgetting curves for the uncertain items and the out-of-state items, with the uncertain items being retained at a higher rate.

We next look at how retention is affected by a student's *learning sequence*, which is the sequence of events taken by the student when learning an item. The possible events in a learning sequence are (a) submitting a correct answer, (b) submitting an incorrect answer, and (c) viewing an explanation of the current instance. If an item is deemed uncertain, a student can demonstrate mastery of the item in the learning mode by correctly answering the first two given instances of the item. Intuitively, if the first two instances of an uncertain item are answered correctly, this would appear to be strong evidence that the student does actually know the item, and that the ALEKS system simply lacked the information to give this classification after the initial assessment (or, at the very least, it is evidence that the student has a strong grasp of the material in the item). These learning sequences are labeled as "CC" in Figure 3, where we can see that the retention rate is even higher than the rate for the uncertain items. Thus, by taking into account the specific answer pattern of a student while learning an item, we can extract even more information about the likelihood that the item will be retained successfully.

Continuing with our analysis of the learning sequence, we next look at the length of the learning sequence (i.e., the number of events it contains). A first guess would be that longer learning sequences give more practice, which would help to improve the retention rate. However, as shown in Figure 4, the length of the learning sequence is actually negatively correlated with the retention of a learned item. A moment's thought shows that this is not actually that surprising. Given that the learning sequence ends when the ALEKS system decides the student has shown mastery of an item, there is a selection effect when partitioning the items by the learning sequence length. More specifically, the shorter sequences tend to involve simpler items for which it is easier to demonstrate mastery (and for which it is also easier to retain the knowledge), or are from

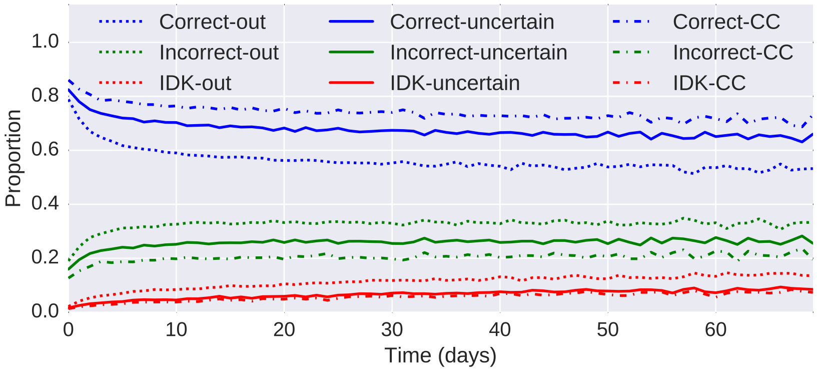

Fig. 3. Proportions of responses conditioned on whether the ALEKS initial assessment classified the item as out-of-state or uncertain (which are mutually exclusive categories), or if the item has a CC learning sequence (which is a subset of the uncertain items). The top three lines (blue) represent the correct responses, the middle three lines (green) represent the incorrect responses, and the bottom three lines (red) represent the "I don't know" responses.

students who have a stronger grasp of the material (again leading to a higher retention rate). On the other hand, the longer learning sequences either involve noisy and difficult items (for which we would expect a lower retention rate) or students who are struggling with the material (in which case we might again expect a lower retention rate).

The last variable we consider is the item itself. While the majority of ALEKS items have a similar open-ended answer format, the actual content, as well as the intrinsic difficulty, can vary. To get a sense of these differences, Figure 5 shows a histogram of the item correct rates in the training set, restricted to the 1664 items with at least 200 data points each. The mean and median correct rates are 0.64 and 0.65, respectively, with a standard deviation of 0.11. While the majority of the items cluster around the mean, there are certain items with somewhat extreme behavior. For example, the maximum and minimum values for retention are 0.92 and 0.17, respectively, with 88 items having a rate above 0.8 and 168 having a rate below 0.5. Thus, the specific characteristics of an item appear to have a significant effect on its retention rate.

5 Retention Models

To make use of the information discussed in the previous section, we next develop a model of retention using an artificial neural network. This neural network model takes the form of a classifier that attempts to predict whether or not a student will give a correct answer when presented an inner fringe item during a progress assessment. The following features are used to build this model.

- ALEKS course: categorical variable with 10 values

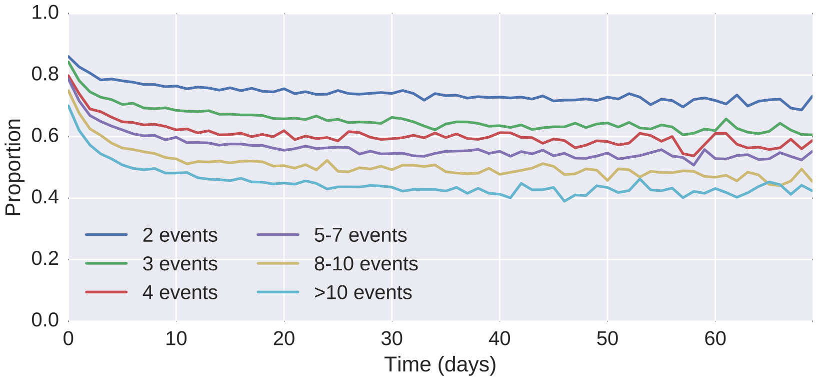

Fig. 4. Proportion of correct responses conditioned on the number of events in the associated learning sequence. Note that, as grouped, the correct rate is a decreasing function of the number of events.

- Item: categorical variable with 2190 values
- Initial score: continuous variable with values in $[0, 1]$
- Time in days since item was learned: discrete variable with values in $[0, 399]$
- Learning sequence: responses encoded as a sequence of categorical variables, each with three values corresponding to the student's action (correct answer, incorrect answer, reading the explanation)

The learning sequence variable is fed to a recurrent neural network (RNN), a type of neural network that is well-suited to handling sequential data [11]. The output from this RNN is then concatenated with the original set of features, and this combined set of features is then fed to a multilayer perceptron (MLP). For the hidden units of our RNN, we evaluate two different recurrent units on our validation set: gated recurrent units (GRU) [3] and long short-term memory (LSTM) units [13]. Additionally, the learning rate, number of hidden layers, and number of units in each hidden layer are also tuned on the validation set. In all cases we use batch normalization [15] while training, and we also apply early stopping [20] and dropout [9, 23] to help prevent overfitting.

Our best performing model on the validation set, which we evaluate in detail in the next section, is comprised of an RNN containing four layers of LSTM units. The output from the last LSTM layer is then combined with the other features and fed to an MLP. The MLP consists of an initial hidden layer with 800 units and 2 subsequent hidden layers with 400 units each. Lastly, each hidden unit of the MLP uses a rectified linear unit (ReLU) as the activation function.

6 Model Evaluation and Feature Importance

One use of an accurate model of retention and forgetting would be to optimize the set of items that are chosen to be tested in an ALEKS progress assessment.

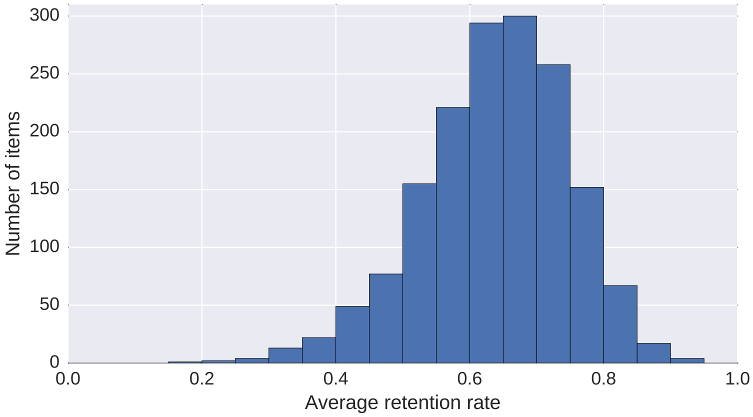

Fig. 5. Histogram of average inner fringe retention rate by item.

That is, if it is very likely that the student will answer a learned item correctly on a progress assessment, it may be more beneficial to the student's learning if a different item, one that the student is struggling to retain, be tested instead. In this case, the student would gain the benefits of retrieval and spaced practice focused on the more troublesome items [16]. Under this implementation, an effective model is one that can correctly identify items that are very likely to be retained; thus, a natural measure of this ability is precision. Additionally, the model must also identify a large enough subset of these items to be effective, which can be measured by the true positive rate or recall. To that end, Figure 6 shows the precision-recall curve on the data in the test set. For comparison, we also give the results for a baseline forgetting curve that uses only time as a parameter and is fit to the correct rate data in Figure 1 (specifically, we use the power function model that is discussed at length in [1]).

One common criticism of neural networks is that they are difficult to interpret, and that in some cases a simpler model such as a logistic regression may be preferable because of this. However, if the goal is to have an idea of the relative importance of each feature to the model (which is typically the argument for using a regression model where, in theory, the coefficients can be interpreted), this can be accomplished using a technique called permutation feature importance [2, 24, 25]. The idea behind permutation feature importance is the following. Given a metric to evaluate the performance of our classification model, we first compute the score for the classifier on our test set. Then, to determine the relative importance of a feature (or, set of features), we randomly shuffle the values for that feature (or, again, set of features) across all the data points in our test set; importantly, however, while doing this we leave the order of the rest of the features untouched. We then run this modified test set through our classifier, extract the predicted probabilities, and then recompute the score of our chosen metric. Comparing this score to the score on the unshuffled test set gives an

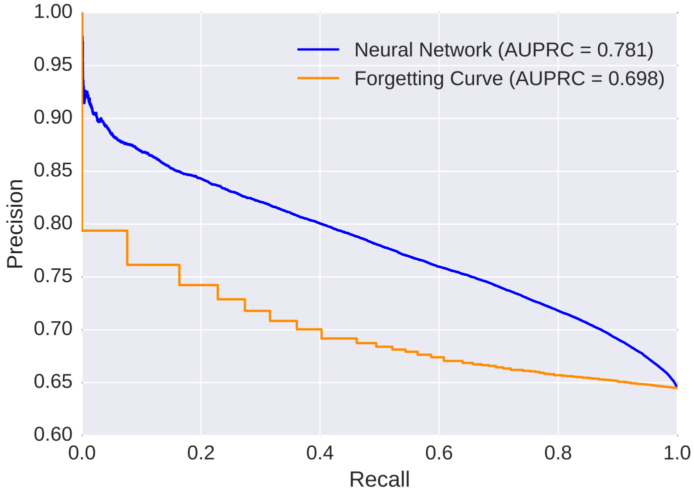

Fig. 6. Comparison of precision-recall curves for the neural network classifier and the single-parameter forgetting curve model.

idea of how "important" this feature is to the performance of the model; if the feature is very important, we can expect a large negative effect on the metric's score on the shuffled test set, while a minor change in the score indicates that the feature is not as crucial to the performance of the model. While some other measures of feature importance may exhibit a bias towards categorical variables with many values, permutation feature importance does not suffer from these same shortcomings [25], and thus is well-suited to our neural network model.

The results from applying permutation feature importance to our classifier are shown in Table 1, where we display the area under the curve for both the precision-recall (PR) and receiver operating characteristic (ROC) curves, averaged over 10 trials (i.e., each set of features is randomly shuffled 10 times, and the average scores over these 10 trials are reported). We can see that the variable with the greatest effect on the scores is the item categorical variable. Taking the histogram in Figure 5 into account, this makes intuitive sense given that some of the items vary widely in their overall retention rates. Additionally, while not quite as impactful as the item variable, both the time since the item was learned, and the information from the learning sequence, are important to the model. In the latter case, this is supported by Figures 3 and 4, which show large differences in retention based on the properties of the learning sequence. On the other hand, the course variable and the initial score are the least important variables. Regarding the initial score, while the differences shown in Figure 2 are significant, these differences are from a comparison of the most extreme decile groups. The initial score has less of an effect when looking at other deciles, which most likely explains the smaller importance of this variable to the final model.

Table 1. Area under the precision-recall (PR) and receiver operating characteristic (ROC) curves using permutation feature importance.

Permuted feature	**PR (Change)**	**ROC (Change)**
None (optimal classifier)	0.781	0.680
ALEKS product	0.765 (-0.016)	0.658 (-0.022)
Item	0.713 (-0.068)	0.591 (-0.089)
Initial score	0.760 (-0.021)	0.653 (-0.027)
Time	0.753 (-0.028)	0.641 (-0.039)
Learning sequence	0.749 (-0.032)	0.634 (-0.046)

7 Discussion and Future Work

In this paper we give a detailed study of how the retention of knowledge works within the ALEKS system. By aggregating data from a large number of ALEKS assessments, we are able to look at the effects of several different variables on this retention. Based on these results, we then build a neural network model of retention within ALEKS. This neural network combines the sequential data from the student learning sequences with several other (non-sequential) variables to make predictions of the likelihood an item will be retained, improving upon the basic one-dimensional forgetting curve model. Furthermore, to help address a common criticism that neural network models are difficult to interpret, we show that an application of permutation feature importance to our neural network model, combined with our exploratory analysis of the data, gives a coherent picture of the relative importance of these variables to our model. Both the learning sequence of the student, and the time since an item was learned, are more informative to our model; on the other hand, the starting knowledge of the student and the specific ALEKS course being used have relatively smaller effects. However, the most influential information came from the categorical variable representing the items, an indication that being able to differentiate between the items is important when building an accurate model of retention. This last result is seemingly consistent with studies that have shown improvements in Bayesian knowledge tracing (BKT) models when item-specific information is taken into account [10, 19, 32].

Given the importance of the item variable when predicting retention, it would be of interest to explore this topic further. For example, are there certain skills and content that characterize, or are inherent to, hard or easy to retain items? Alternatively, it is possible that the outsized influence of the item variable is due to something specific to ALEKS. As an example, a low retention rate could be an indication that an item is placed at a suboptimal position within the ALEKS system, and in such a case a student would benefit from seeing additional prerequisite material before learning the item. Thus, it is not a stretch to think that the information contained within the item variable may be due to factors such as this. Answering these questions would give an even more complete picture of how retention works within ALEKS.

References

1. Averell, L., Heathcote, A.: The form of the forgetting curve and the fate of memories. Journal of Mathematical Psychology **55**, 25–35 (2011)
2. Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001)
3. Cho, K., van Merrienboer, B., Gülçehre, Ç., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. CoRR **abs/1406.1078** (2014), `http://arxiv.org/abs/1406.1078`
4. Doble, C., Matayoshi, J., Cosyn, E., Uzun, H., Karami, A.: A data-based simulation study of reliability for an adaptive assessment based on knowledge space theory. International Journal of Artificial Intelligence in Education (2019). https://doi.org/10.1007/s40593-019-00176-0
5. Doignon, J.P., Falmagne, J.C.: Spaces for the assessment of knowledge. International Journal of Man-Machine Studies **23**, 175–196 (1985)
6. Ebbinghaus, H.: Memory: A Contribution to Experimental Psychology. Originally published by Teachers College, Columbia University, New York (1885; translated by Henry A Ruger and Clara E Bussenius (1913))
7. Falmagne, J.C., Albert, D., Doble, C., Eppstein, D., Hu, X. (eds.): Knowledge Spaces: Applications in Education. Springer-Verlag, Heidelberg (2013)
8. Falmagne, J.C., Doignon, J.P.: Learning Spaces. Springer-Verlag, Heidelberg (2011)
9. Gal, Y., Ghahramani, Z.: A theoretically grounded application of dropout in recurrent neural networks. In: Advances in Neural Information Processing Systems 29 (2016)
10. González-Brenes, J., Huang, Y., Brusilovsky, P.: General features in knowledge tracing to model multiple subskills, temporal item response theory, and expert knowledge. In: Proceedings of the 7th International Conference on Educational Data Mining. pp. 84–91 (2014)
11. Graves, A.: Supervised Sequence Labelling with Recurrent Neural Networks. Springer-Verlag, Heidelberg (2012)
12. Grayce, C.: A commercial implementation of knowledge space theory in college general chemistry. In: Falmagne, J.C., Albert, D., Doble, C., Eppstein, D., Hu, X. (eds.) Knowledge Spaces: Applications in Education, chap. 5, pp. 93–114. Springer-Verlag (2013)
13. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**, 1735–1780 (1997)
14. Huang, X., Craig, S., Xie, J., Graesser, A., Hu, X.: Intelligent tutoring systems work as a math gap reducer in 6th grade after-school program. Learning and Individual Differences **47**, 258–265 (2016)
15. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning. pp. 448–456 (2015)
16. Lindsey, R.V., Shroyer, J.D., Pashler, H., Mozer, M.C.: Improving students long-term knowledge retention through personalized review. Psychological science **25**(3), 639–647 (2014)
17. Matayoshi, J., Granziol, U., Doble, C., Uzun, H., Cosyn, E.: Forgetting curves and testing effect in an adaptive learning and assessment system. In: Proceedings of the 11th International Conference on Educational Data Mining. pp. 607–612 (2018)
18. McGraw-Hill Education/ALEKS Corporation: What is ALEKS? `https://www.aleks.com/about_aleks`

19. Pardos, Z.A., Heffernan, N.T.: KT-IDEM: Introducing item difficulty to the knowledge tracing model. In: Proceedings of the 19th International Conference on User Modeling, Adaption, and Personalization. pp. 243–254. UMAP'11, Springer-Verlag (2011)
20. Prechelt, L.: Early stopping – but when? In: Montavon, G., Orr, G., Müller, K. (eds.) Neural Networks: Tricks of the Trade, Lecture Notes in Computer Science, vol. 7700. Springer, Berlin, Heidelberg (2012)
21. Qiu, Y., Qi, Y., Lu, H., Pardos, Z.A., Heffernan, N.T.: Does time matter? modeling the effect of time with Bayesian knowledge tracing. In: Proceedings of the 4th International Conference on Educational Data Mining. pp. 139–148 (2011)
22. Reddy, A., Harper, M.: Mathematics placement at the University of Illinois. PRIMUS **23**, 683–702 (2013)
23. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research **15**, 1929–1968 (2014)
24. Strobl, C., Boulesteix, A.L., Kneib, T., Augustin, T., Zeileis, A.: Conditional variable importance for random forests. BMC Bioinformatics **9**(1), 307 (2008)
25. Strobl, C., Boulesteix, A.L., Zeileis, A., Hothorn, T.: Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics **8**(1), 25 (2007)
26. Taagepera, M., Arasasingham, R.: Using knowledge space theory to assess student understanding of chemistry. In: Falmagne, J.C., Albert, D., Doble, C., Eppstein, D., Hu, X. (eds.) Knowledge Spaces: Applications in Education, chap. 6, pp. 115–128. Springer-Verlag (2013)
27. Wang, Y., Heffernan, N.: Towards modeling forgetting and relearning in ITS: Preliminary analysis of ARRS data. In: Proceedings of the 4th International Conference on Educational Data Mining. pp. 351–352 (2011)
28. Wang, Y., Beck, J.E.: Incorporating factors influencing knowledge retention into a student model. In: Proceedings of the 5th International Conference on Educational Data Mining (2012)
29. Xiong, X., Li, S., Beck, J.E.: Will you get it right next week: Predict delayed performance in enhanced its mastery cycle. In: The Twenty-Sixth International FLAIRS Conference (2013)
30. Xiong, X., Wang, Y., Beck, J.B.: Improving students' long-term retention performance: A study on personalized retention schedules. In: Proceedings of the Fifth International Conference on Learning Analytics And Knowledge. pp. 325–329. ACM (2015)
31. Yang, Y., Leung, H., Yue, L., Deng, L.: Automatic dance lesson generation. IEEE Transactions on Learning Technologies **5**, 191–198 (2012)
32. Yudelson, M.: Individualizing Bayesian knowledge tracing. Are skill parameters more important than student parameters? In: Proceedings of the 9th International Conference on Educational Data Mining (2016)