Predictive Modeling of Learning Continuation in Preschool Education Using Temporal Patterns of Development Tests

Junpei Naito, Yukino Baba, Hisashi Kashima
Department of Intelligence Science
and Technology, Kyoto University
naito.junpei.45m@kyoto-u.jp
{baba, kashima}@i.kyoto-u.ac.jp

Takenori Takaki
Shimane IT Open-Innovation Center
takaki@s-itoc.jp

Takuya Funo
Shimane Industrial Promotion Foundation
tfuno@joho-shimane.or.jp

Abstract

Learning analytics applies data analysis techniques to learning data in order to support students' learning processes and to improve the quality of education. Despite the increasing attention to learning analytics for higher education, it has not been fully addressed in primary and preschool education. In this research, we apply learning analytics to preschool education to predict the continuation of learning of preschool children. Based on our hypothesis that temporal patterns in the assessment scores of development tests are effective features for prediction, we extract the temporal patterns using time-series clustering, and use them as the features of prediction models. The experimental results using a real preschool education dataset show that the use of the temporal patterns improves the predictive accuracy of future continuation of study.

Introduction

Learning analytics applies data science technologies to education in order to support learners and educators to improve learning effectiveness and to obtain insights for better education. Typical analyses include predicting future performance of learning, identifying learners who have problems, and making recommendations of appropriate materials to learners, based on various kinds of data such as learning time, teaching materials, and performance scores. In recent years, the growing adoptions of information technologies to educational fields including on-line education such as MOOCs have accelerated collection of large amount of data associated with learning, and learning analytics has been actively carried out, mainly in higher education (Siemens and Long 2011).

Despite the wide applications of learning analytics in higher education, learning analytics for early childhood education and primary education has not been addressed thoroughly. A part of the reason is that evaluation metrics such as learning time and examination scores cannot be interpreted in the same way as those in higher education. For example, it is often difficult for children to take written exams, and interventions by parents or teachers are necessary to conduct exams and to measure the ability of children, and therefore evaluation metrics are often affected by their subjective judgments. In addition, lectures using the Internet and computers are not performed very often in early childhood education and primary education, and therefore the number of collectable types of data and the amount of data are limited, which makes application of learning analytics difficult. Furthermore, when we consider dropout prediction, which is of great interest in learning analytics in higher education, the meaning of "dropout" is different in early childhood education and primary education, because decisions on whether or not a child will continue learning does not depend only on the child herself, and it is usually her parents who observe the motivation of children and make judgments on whether or not the continuation is worthwhile.

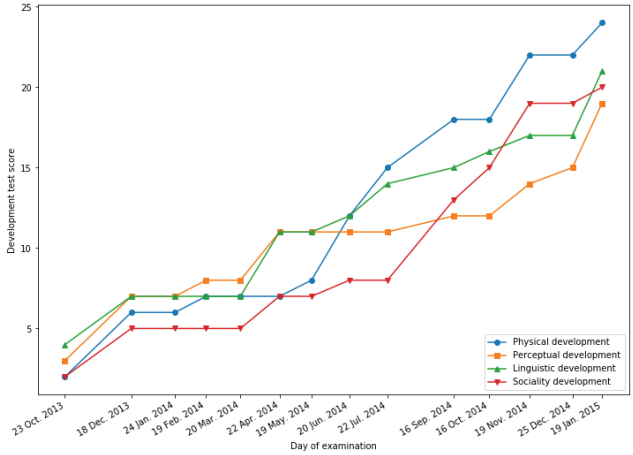


Figure 1: Time-series data of the development test results of a particular child.

Motivated by the above concerns, in this research, we analyze the learning data of preschool children and attempt predictive modeling of continuation of learning, which corresponds to the dropout prediction problem in higher education. Our dataset provided by Shichida Educational Institute includes the time series data of development test results (Fig. 1), which we expect are useful information for prediction. However, the results of the development tests are subjective evaluations by parents, and therefore are noisy and biased. To cope with the noisy and biased subjective data, we apply time-series clustering to extract robust temporal patterns of the evaluation scores, and include them as features in our prediction model.

In our experiments, we use logistic regression to predict whether each child will continue to learn after two years of age from the data of zero to two years of age children. The predictive performance is evaluated in terms of the area under the curve (AUC). The result using the temporal patterns is 0.8001 at the maximum, which outperforms the result without temporal patterns that is 0.7549. Another experiment using data from two to four years of age for children to predict whether the child will continue to learn after four years of age also shows similar results; the AUC is 0.8421 for the model using temporal patterns, and it is 0.8093 for the one without temporal patterns. Our results indicate the temporal patterns of the scores of development tests are useful for predicting the continuation of learning.

Related work

One of the main objectives of learning analytics is to predict poor performers and dropouts and to make early and effective interventions. Machine learning and statistical methods are used to find such students who need help.

Tamhane et al. (2014) made predictions of poor grades at Grade 8 using logistic regression, and showed long term grade information from Grade 1 to Grade 7 were effective for prediction. Aguiar et al. (2015) predicted dropouts and their timings for students of grades from 6 to 12, in which the prediction models predict whether or not a particular student dropouts at the end of each grade. Lakkaraju et al. (2015) developed a framework for predicting students leaving high school, where various machine learning methods such as random forest and logistic regression can be compared, and important features for prediction are visualized. They used data including class absences, late-arrival rate, economic situations as well as GPA scores. Vihavainen, Luukkainen, and Kurhila (2013) predicted poor performance in a university programming lecture based on behaviors of writing codes. Predictive features such as the amount of time required for coding or correction and edit distance of the codes before and after the modification were extracted from the modification history of the codes. He et al. (2015) applied logistic regression to prediction of learners who would not complete MOOC courses. Hlosta, Zdrahal, and Zendulka (2017) predicted whether or not a learner who had not yet submitted an assignment would submit the assignment by the final deadline based on the information of other learners who had already submitted the assignment.

Most of the existing studies target secondary education, higher education, and MOOC; however, the effectiveness of learning analytics and predictive modeling in primary and preschool education has not been fully investigated yet.

Problem settings

In this research, we analyze a dataset of preschool children to obtain a prediction model of their learning continuation. Our dataset is provided by Shichida Educational Institute consisting of the results of development tests, where each result is associated with the time stamp of the test being performed (Fig. 1). We divide the dataset into two parts at a particular point of time, that are, the ones before and after the time point, respectively. We predict learning continuation based on the first part (the dataset available before the time point). We define the learning continuation of a particular child as the existence of data of the child after the time point. For each child, we give a feature vector x using data before the time point and a label $y \in \{+1, -1\}$ that indicates whether the child continues learning $(+1)$ or not (-1). Our goal is to obtain a prediction model that predicts y given x.

Methods

The important issue in predictive modeling of continuation of learning is how to design the feature vector x for a child. For this purpose, we use both the personal information of the child and the temporal patterns in his/her results of development tests. We use time-series clustering to extract the temporal patterns from data, and construct features indicating if each pattern is observed in the results of the development tests, based on the assumption that children who show similar temporal patterns in their development test results tend to have similar tendencies in continuation of learning.

Design of feature vectors using temporal patterns

Two kinds of information is available for each child; one is static information like the date of birth of the child, and the other is time-series information like the results of development tests. The static information is easily included in the features with some appropriate coding and scaling. However, the time-series information cannot be included as it is, since they are sometimes noisy and biased as in our present case; our time-series data includes subjective evaluations by parents. Therefore, we extract temporal patterns from the time-series data using time-series clustering. Suppose the clustering results in K clusters, we create K corresponding features; each of them indicates whether the child is included in the corresponding cluster (1) or not (0). This means that only one of the K features is set to 1.

Hierarchical clustering

In principle, we can use arbitrary clustering methods for time-series data; we employ hierarchical clustering which repeats calculation of the distance between all data points and all clusters and merging two data points or clusters having the smallest distance as a new cluster until the number

of clusters reaches a predetermined number. There are several choices in the ways to calculate distances among data points and clusters; we use the single linkage method, complete linkage method, and group average method.

The single linkage method, or the nearest neighbor method, takes smallest distance between two data points from two different clusters C and C' as the distance between the two clusters:

$$d^{\mathrm{SL}}(C, C') = \min_{x \in C, x' \in C'} d(x, x'),$$

where d is a distance measure between two data points, in our case, the distance between two time-series sequences. On the other hand, the complete linkage method, or the furthest neighbor method, takes the greatest distance between two data points instead of the smallest distance used in the single linkage method:

$$d^{\mathrm{CL}}(C, C') = \max_{x \in C, x' \in C'} d(x, x').$$

The group average method uses the average of the distances between all pairs of data points from the two clusters:

$$d^{\mathrm{GA}}(C, C') = \frac{1}{|C||C'|} \sum_{x \in C, y \in C'} d(x, x').$$

Distance between time series data

In hierarchical clustering, we need a distance measure between two time-series sequences. A simplest choice would be the Euclidean distance which regards two time series sequences as two vectors and takes the 2-norm of their difference. The Euclidean distance assumes the two sequences has the same length; this is not true in our case. Therefore, we need interpolation to align their lengths, which will be discussed later in this section. Another option is to use another distance measure applicable to different-length time series. The dynamic time warping (DTW) distance (Keogh and Ratanamahatana 2005) is a typical choice of such distance. The DTW distance allows distance calculation between two asynchronous time-series data with different lengths by shifting the times of the elements in two time series to best align with each other. The DTW distance $d^{\mathrm{DTW}}(\mathbf{s}, \mathbf{s}')$ between two time-series sequences $\mathbf{s} = (s_1, s_2, \cdots, s_n)$ and $\mathbf{s}' = (s'_1, s'_2, \cdots, s'_{n'})$ is efficiently calculated by using dynamic programming (Berndt and Clifford 1994):

$$\begin{aligned} &d^{\mathrm{DTW}}(\mathbf{s}, \mathbf{s}') = \gamma(n, n'), \\ &\gamma(i, j) = \delta(s_i, s_j) + \min \begin{cases} \gamma(i-1, j-1) \\ \gamma(i-1, j) \\ \gamma(i, j-1) \end{cases}, \\ &\gamma(i, 0) = 0,\ \gamma(0, j) = 0. \end{aligned}$$

where $\delta(\cdot, \cdot)$ is the distance between two elements. We use the Euclidian distance in our experiments.

Interpolation

We need interpolation of time-series data with different length to use the Euclidian distance. In this research, we use linear interpolation which interpolates two consecutive observations with a straight line connecting them. For two data points $(t_1, s_1), (t_2, s_2)$ such that $t_1 < t_2$, the data point (t, s) for arbitrary $t_1 < t < t_2$ is given as

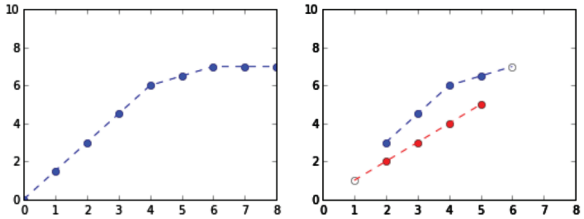
$$s = \frac{(t - t_1)s_1 + (t_1 - t)s_0}{t_2 - t_1}.$$

However, we still cannot give the Euclidian distance between two time series when the first or last elements of two time series have different time, for which we try two possible solutions.

The first solution is to introduce a virtual data point $(0, 0)$ and extend the interpolation, that is, to set the value of data at time 0 as 0 when we need to interpolate before the first data point. When we need to interpolate after the last data point, we interpolate with the same value of the last data point (See Figure 2a). We call this solution the *extension* method

The second solution is to use only the time period in which two time series overlap. Let the first time point and the last time point of a time series be t_1 and t_n, respectively. Similarly, let those for the other time series be t'_1 and $t'_{n'}$. We calculate the Euclidean distance focusing on the time period between $\min\{t_1, t'_1\}$ and $\max\{t_n, t'_{n'}\}$ (See Figure 2b). We call this solution the *overlap* method.

In both the Euclidean distance and the DTW distance, as the number of data points of the time series increases, their distance increases. For this reason, we normalize the distance with the number of data points. In our experiments, we consider both the normalized distance and the original distance.



(a) Extension method (b) Overlap method

Figure 2: Interpolation methods

Experiments

Our experiments demonstrate that the temporal patterns in the development test results improve the prediction accuracy of learning continuation in two settings: the one for predicting learning continuation after two years of age and the other for after four years of age. We also interpret several obtained temporal patterns effective for prediction.

Dataset

Our dataset provided by the Shichida Educational Institute includes personal information such as children IDs, the numbers of children in each family, and the dates of birth. In addition, it also includes the time series data of the results of

development tests provided by Shichida Educational Institute. A development test consists of four subjects: 'physical development', 'perceptual development', 'linguistic development', and 'development of sociality'. Each subject has several check items; for example in the physical development section, items like "Can hold on to a climbing pole for 5 seconds" and "Can throw a baseball ball for four meters" (Figure 3) are given.

The parents of a child complete the check items, and the score for each subject is determined based on the answers. As mentioned before, there are four categories in the check items; therefore, four scores are obtained at each time development test performed. Most of the check items are objective; however, since the evaluator is a parent of the child, there is not a small possibility that the parent's subjectivity comes in.

For our experiments, we prepare two datasets for two prediction settings. One of them is the dataset with $1,540$ zero-to-two year old children, 803 of which are children who continued learning, and 737 of which are not. Another dataset is with $1,322$ two-to-four year old children, in which 529 of them continued learning and the other 903 children did not.

Experimental procedure

Features representing a child consist of static features and temporal features. The static features include a child ID, the number of older or younger brothers and sisters, the age at the time of the first examination, the average of the age at the time of examinations, the average of the examination interval, the number of tests, and the average of the development test scores.

The temporal features are extracted from the time series of the development test scores by using the time-series clustering methods as described before. Since we have four categories of the scores, if we set the number of clusters as K, the total number of features becomes $4K$.

To extract the temporal features, we calculate the distance between the time series and create a distance matrix. For both the DTW distance and the Euclidean distance, we calculate both the original distance and the normalized distance. We also use the two types of interpolation methods in calculating the Euclidean distance. Hierarchical clustering is performed by either of the single linkage method, the complete linkage method, and the group average method. The number of clusters K is chosen from $\{10, 20, 50, 100, 200\}$. Using the clustering results, we construct $4K$-dimensional feature vectors; they are concatenated with the feature vectors of static information to obtain the final feature vectors.

The dataset is divided into the training dataset for constructing a classifier and the verification data for evaluating the prediction accuracy of the trained classifier. The ratio of the training dataset and the verification data is 80% and 20%, respectively. We train the logistic regression model using the training dataset, and make predictions for the verification dataset using the resultant classifier, and the prediction accuracy is measured in the area under the curve (AUC). We perform predictions both with and without using temporal patterns of the development tests, and compare their accuracy. In time series clustering, the results are different depending on the choice of distance between time series, the one between clusters, as well as the number of clusters. We compare the maximum prediction performance among the models with or without temporal patterns.

Table 1: Estimated feature weights in the prediction model for learning continuation after two years of age using the data of zero-to-two year old children without the temporal patterns.

Feature name	Weight
Child ID	-0.0440
Day of the first examination	-1.0622
Average day of examinations	2.3269
Average examination interval	0.2380
Number of examinations	0.3985
Average total score	0.1282
Average body growth score	0.0437
Average perceptual growth score	-0.1585
Average linguistic growth score	0.3448
Average social growth score	0.1993

Table 2: Estimated feature weights in the prediction model for learning continuation after four years of age using the data of two-to-four year old children without the temporal patterns.

Feature name	Weight
Child ID	-0.0862
Day of the first examination	-1.1626
Average day of examinations	2.7068
Average examination interval	0.1916
Average number of examinations	0.5335
Average total score	0.0028
Average body growth score	-0.1828
Average perceptual growth score	0.3219
Average linguistic growth score	0.1287
Average social growth score	-0.2765

Results

Which static features are effective? We first show the result of prediction of learning continuation after two years of age using the data of zero-to-two year old children, when the model does not use the temporal patterns. The AUC value is 0.7549, which is a reasonably high predictive performance. The weights of the features used in the estimated model are shown in Table 1. A large absolute value of a weight means a large influence of the corresponding feature on the predictions. The average day at the time of examinations has a large positive weight, which means that the more examinations a child takes at a higher age, the more likely he/she continues learning. The day at the time of the first examination has a relatively large negative value, which means that the lower the age at the time of the first examination is, the easier the child continues learning.

Similarly, Table 2 shows the weights of the features in the

Check	Item number	Check item
⋮	⋮	⋮
✓	85	Can hold on to a climbing pole for 5 sedonds
	86	Can throw a baseball ball for four meters
✓	87	Can jump and move back, forth, right, and left
✓	88	Can jump over 20-centimeter height
	89	Can ride on a swing in the upright position
	90	Can jump for 70-80 centimeters long
⋮	⋮	⋮

Figure 3: An example of checklist in the 'physical development' category (quoted from 'Development Study Book Part 3' by Shichida Educational Institute). The check marks are given by parents if they think the corresponding check items are accomplished by the child.

Table 3: Comparison of the predictive performance of the model using the temporal patterns and the one without the temporal patterns in prediction for continuation after two years old.

Prediction model	AUC
Without temporal patterns	0.7549
With temporal patterns	0.8001

Table 4: Comparison of the predictive performance of the model using the temporal patterns and the one without the temporal patterns in prediction for continuation after four years old.

Prediction model	AUC
Without temporal patterns	0.8093
With temporal patterns	0.8421

prediction model for learning continuation after four years of age using data of two-to-four years old children. We observe the similar features are effective compared with the previous results (Table 1). The AUC is 0.8093, which is a better performance than the previous setting.

Do temporal patterns improve prediction? We investigate the effectiveness of the temporal patterns. Tables 3 and 4 show the comparison of the predictive performance of the model using the temporal patterns and the one without the temporal patterns. Table 3 shows the result for prediction of the continuation after two years old, and Table 4 is for after four years old. The results with the temporal patterns show the best performance.

Table 3 indicates that, when the clustering method is appropriately selected, prediction accuracy is reasonably improved. The maximum value was obtained in the experiments when we set the number of clusters $K = 200$, and use the normalized Euclidean distance with the overlap-type interpolation in the single linkage method. Similarly, Table 4 also shows a reasonable improvement by the use of temporal patterns. The best choices were $K = 200$ clusters, and the single connection method using the DTW distance calculated without normalization.

Insights from effective temporal patterns. Let us look closely at some of the temporal patterns that turned out to be effective in prediction.

Figures 4a and 4b show two temporal patterns. The horizontal axis is the age at the time of examination and the vertical axis is the perceptual development score.

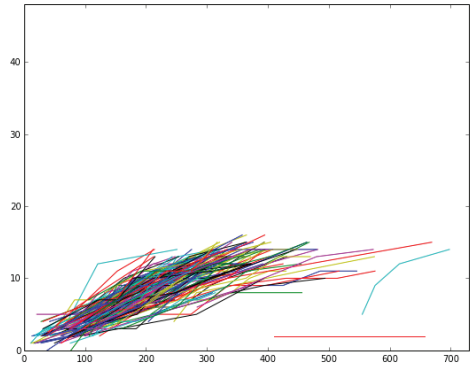
Figure 4a shows a temporal pattern correlated with children who did not continue learning, which includes 15 children who continued learning and 157 who did not. Children in this cluster took examinations at the early stage but the frequency decreased with advancing age. Figure 4b shows a temporal pattern correlated with children who continued learning, which includes 19 children who continued learning and 5 who did not. They started to take examinations recently and the scores grew rapidly. These observations are consistent with our previous observation in the model without temporal patterns that the average day of examinations positively correlates with learning continuation.

Figures 4c and 4d show other examples; Figure 4d is a temporal pattern including 34 continuing children and 133 non-continuing children, and Figure 4d is a temporal pattern including 27 continuing children and 4 non-continuing children. In Figure 4c which shows the group of a low continuation rate, they regularly took examinations but the improvement of the scores is not significant. On the other hand, Figure 4d which shows a group of a high continuation rate, they regularly took examinations and continuously improved their scores.

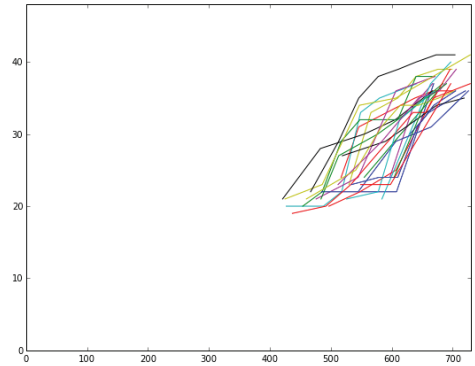
The above observations are based on the results with $K = 50$ clusters by the group average method with the DTW distance calculated without linear interpolation; similar trends were often observed for other settings.

Conclusion

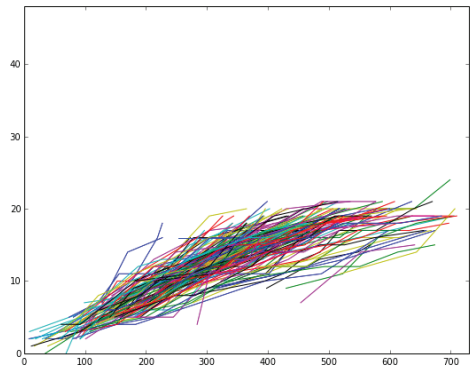
In this study, we applied learning analytics to preschool children education, especially, we predicted the continuation of learning, which is a similar problem to the dropout prediction problem in higher education. To cope with the noisy and biased time-series scores of development tests based on



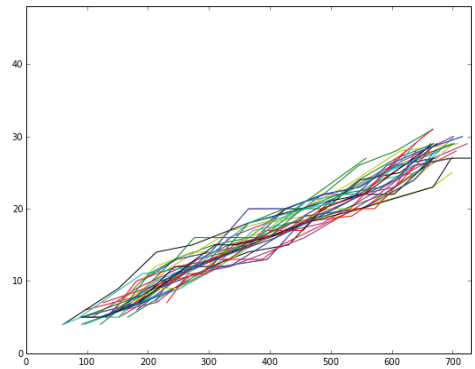
(a) An example of temporal patterns represented by non-continuing children; children in this cluster took examinations at the early stage but the frequency decreased with advancing age.



(b) An example of temporal patterns represented by continuing children; they started to take examinations recently and the scores grew rapidly.



(c) An example of temporal patterns represented by non-continuing children; they regularly took examinations but the improvement of the scores is not significant.



(d) An example of temporal patterns represented by continuing children; they regularly took examinations and continuously improved their scores.

Figure 4: Temporal patterns which turned to be effective in prediction.

subjective evaluation, we used robust temporal patterns extracted by using time-series clustering. Using experiments with a real preschool education dataset, we demonstrated that the prediction accuracy of learning continuation was improved by using the temporal patterns as a part of features for prediction. In addition, the temporal patterns which turned out to be useful for prediction characterized children who continued learning and those who did not.

Although our present study demonstrated the effectiveness of the temporal patterns, the best performing choices of the clustering methods and the other parameters depend on the datasets. Investigation for more insights about most appropriate choices for the learning continuation problem is left for the future work. In addition, we only focused on the logistic regression model, which is a simple linear prediction model, and obtained reasonable prediction accuracy; using more complex prediction models such as deep neural networks such as recurrent neural networks should further improve the performance, which would also be addressed in the future work.

References

Aguiar, E.; Lakkaraju, H.; Bhanpuri, N.; Miller, D.; Yuhas, B.; and Addison, K. L. 2015. Who, when, and why: A machine learning approach to prioritizing students at risk of not graduating high school on time. In *Proceedings of the Fifth International Learning Analytics and Knowledge Conference (LAK)*, 93–102.

Berndt, D., and Clifford, J. 1994. Using dynamic time warping to find patterns in time series. In *Proceedings of the Third ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 359–370.

He, J.; Bailey, J.; Rubinstein, B. I.; and Zhang, R. 2015. Identifying at-risk students in massive open online courses. In *Proceedings of the Twenty-Ninth AAAI Conference onArtificial Intelligence (AAAI)*, 1749–1755.

Hlosta, M.; Zdrahal, Z.; and Zendulka, J. 2017. Ouroboros: early identification of at-risk students without models based on legacy data. In *Proceedings of the Seventh International*

Learning Analytics and Knowledge Conference (LAK), 6–15.

Keogh, E., and Ratanamahatana, C. A. 2005. Exact indexing of dynamic time warping. *Knowledge Information Systems* 7(3):358–386.

Lakkaraju, H.; Aguiar, E.; Shan, C.; Miller, D.; Bhanpuri, N.; Ghani, R.; and Addison, K. L. 2015. A machine learning framework to identify students at risk of adverse academic outcomes. In *Proceedings of the Twenty-First ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1909–1918.

Siemens, G., and Long, P. 2011. Penetrating the fog: Analytics in learning and education. *EDUCAUSE Review* 46(5):31–40.

Tamhane, A.; Ikbal, S.; Sengupta, B.; Duggirala, M.; and Appleton, J. 2014. Predicting student risks through longitudinal analysis. In *Proceedings of the Twentieth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1544–1552.

Vihavainen, A.; Luukkainen, M.; and Kurhila, J. 2013. Using students' programming behavior to predict success in an introductory mathematics course. In *Proceedings of the Sixth International Conference on Educational Data Mining*, 300–303.