International Journal of Learning Technologies and Environments
International Institute of Applied Informatics
2021, Vol. 4, No. 1, 21 – 42

Current Failure Prediction for Final Examination via Nearest Neighbor Method using Past Trends of Weekly Online Testing

Hideo Hirose [*] [†]

Abstract

We showed previously that we can predict the success/failure status for the final examination to each student at early stages in courses using the current trends of estimated abilities to the learning check testing in terms of the item response theory, where we used the same testing results in prediction and in construction of the mathematical model. However, such a treatment may cause the overfitting effect. In this paper, we have shown that we can still predict the current success/failure status for the final examination using the past trends of estimated abilities to the learning check testing and the past final examination results. In prediction, we applied the nearest neighbor method for determining the similarity in the trends of estimated abilities to the learning check testing.

Keywords: current failure prediction, past trends, item response theory, nearest neighbor, similarity, online testing, learning analytics.

1 Introduction

Since it is crucial to identify students at risk for failing courses and/or dropping out as early as possible in educating many students altogether as pointed out in [34, 36], we established online testing systems aimed at helping students who desire further learning skills for mathematics education, including the learning check testing, the LCT, for every class to check if students comprehend the contents of lectures or not (see [15–18, 20–23, 27, 28, 32, 35]), resulting in the importance of learning analytics as suggested in [6, 8, 37]; regarding learning analytics, see also [20–23, 27].

In the previous paper [19], we showed that we predicted the success/failure status for the final examination to each student at early stages in courses using the current trends of estimated abilities to the learning check testing in terms of the item response theory (e.g., see [1, 9, 29]). However, we used the same testing results in prediction and in construction of the mathematical model. In such a situation, actually, we cannot predict the final examination status in the middle of the semester term because we have not yet obtained the

[*] Kurume University, Fukuoka, Japan
[†] This work was supported by JSPS KAKENHI Grant Number 17H01842.

teacher data to refer to. In addition, if we dare to use the same testing results in prediction, the proposed method may cause the overfitting effect.

In this paper, we will show that we can still predict the current success/failure status for the final examination using the past trends of estimated abilities of the LCT and the past final examination results. In machine learning, when we want to obtain prediction accuracy, it seems that it is common to prepare two types of data, learning data and test data, from one observation data. Even in such a case, we are using the same data in constructing the mathematical model and in testing the model validity. However, this paper deals with two totally different data sets: one is a certain fiscal year academic data set, and the other is another fiscal year academic data set. Such a treatment in machine learning is uncommon. This point is new. Moreover, in order to improve the prediction accuracy, we have made the testing data prepessed in some ways. This is the standardization method in different fiscal years, and this preprocessing is also new.

The subject we deal with in this paper is analysis basic as a typical case, although we have been performing two subjects of analysis basic and linear algebra. However, the methodology we propose in this paper can be applied to other subjects.

2 Weekly Online Testing Scheme

In every lecture of fundamental mathematics classes, all the enrolled students have been taking online testings since 2016. The subjects were analysis basic (i.e., calculus) and linear algebra. Testing time duration is ten minutes, and m questions using multiple choice are provided to each testing; in 2017 and 2018 semesters, $m = 5$ and $m = 7$, respectively. The testings to grasp the comprehension of each unit are incorporated into regular classes. The numbers of online testings to analysis basic and linear algebra in 2017 are 14 and 13, and in 2018, they are 12 to both subjects.

For example in 2018 semester to analysis basic, if we denote K as the number of opportunities that students take the LCT, then $K = 12$. In addition, we define the number of freshman students to be registered in this subject as N, then N is 1,230. Thus, we have user-item response matrix sized of $N \times mK = 1230 \times 84$ to this subject at the end of the semester.

3 Ability Evaluation Method Using IRT

Since the item response theory (IRT) provides us the difficulties of the test items (problems) and the examinees' abilities together, we incorporated the IRT evaluation method into the online testing systems. This method results in evaluating the examinees' abilities accurately and fairly (see [10–14, 31]). In this paper, we deal with the cases of the standard IRT evaluation using the two-parameter logistic function $P(\theta_i; a_j, b_j)$ shown below.

$$\begin{aligned} P(\theta_i; a_j, b_j) &= \frac{1}{1+\exp\{-1.7a_j(\theta_i - b_j)\}}, \\ &= 1 - Q_{i,j}(\theta_i; a_j, b_j), \end{aligned} \tag{1}$$

where θ_i expresses the ability for student i, and a_j, b_j are constants in the logistic function for item j called the discrimination parameter and the difficulty parameter, respectively. The corresponding likelihood function for all the examinees, $i = 1, 2, \ldots, N$, and all the items,

$j = 1, 2, \ldots, n$, will become

$$L = \prod_{i=1}^{N} \prod_{j=1}^{n} \left(P_{i,j}^{\delta_{i,j}} \times Q_{i,j}^{1-\delta_{i,j}} \right), \tag{2}$$

where $\delta_{i,j}$ denotes the indicator function such that $\delta = 1$ for success and $\delta = 0$ for failure in answering a question. We can obtain the maximum likelihood estimates $\hat{\theta}_i$ and $\hat{a}_j, \hat{b}_j$ for parameters θ_i and a_j, b_j by maximizing the likelihood function (2). When student i misses a LCT, we regard $\delta_{i,j} = 0$ in that LCT.

4 Trend of Estimated Ability Using Cumulative Unit Response Matrix in IRT

We define $\theta_1(i,k)$ as student i's ability using the response results from the 1st LCT to kth LCT, that is, the response matrix becomes a $N \times km$ size matrix. Figure 1 shows two trends of estimated abilities $\theta_1(i,k)$ using the corresponding response matrices, one for successful, and the other for failed students, in the final examination of analysis basic in 2017.

Looking at the figures, we can see that the estimated ability to each student tends to reach a certain value as lectures go forward from 1 to 14. Moreover, the two trends indicate a clear difference between the successful and failed students, i.e., the former shows increasing tendency and the latter shows decreasing tendency of ability values. The figure also tells us that the estimated abilities show rather large variations around 0 value initially, but later the variation to each student becomes lower as lectures go forward, which suggests that the estimates become more stable and accurate as lectures go forward.

Such a characteristic leads us to use these trends in discriminating between the successful students and failed students.

5 Discrimination Method of Failed Students

In the previous paper [19], we proposed the similarity via the nearest neighbor using the estimated ability trends in order to identify successful/failed students with much higher reliability than the one simple decision tree result using the full response matrix in prediction. In using the similarity, we used $\theta_1(i,k)$ by incorporating the tentative response matrices $M_{m,k}(N, mk),\ k = 1, \ldots, K$ using LCT no.1 to no.k, in contrast to the use of the full matrix of $M_{m,K}(N, mK)$ in the decision tree.

We defined the similarity of the two ability trends (i and j) by the following formula $S_{i,j,1,1}^{k}$ such that

$$S_{i,j,1,1}^{k} = \sqrt{\frac{1}{k} \sum_{l=1}^{k} (\theta_1(j,l) - \theta_1(i,l))^2},\ (i \neq j), \tag{3}$$

where, two trends were chosen from the same database, i.e., from the database in 2017 both to $\theta_1(j,l)$ and $\theta_1(i,l)$.

Sorting $S_{i,j,1,1}^{k}$ in ascending order in terms of j such as $S_{i,(1),1,1}^{k} \leq \ldots, \leq S_{i,(N-1),1,1}^{k}$, $S_{i,(j),1,1}^{k}$ expresses the ordered statistics of $\{S_{i,(j),1,1}^{k}\}$. We selected the 10 least $S_{i,(j),1,1}^{k}$ (i.e., $S_{i,(1),1,1}^{k}, \ldots, S_{i,(10),1,1}^{k}$), and obtained the mean value $\mu(i,k,1,1)$ of these final examination's

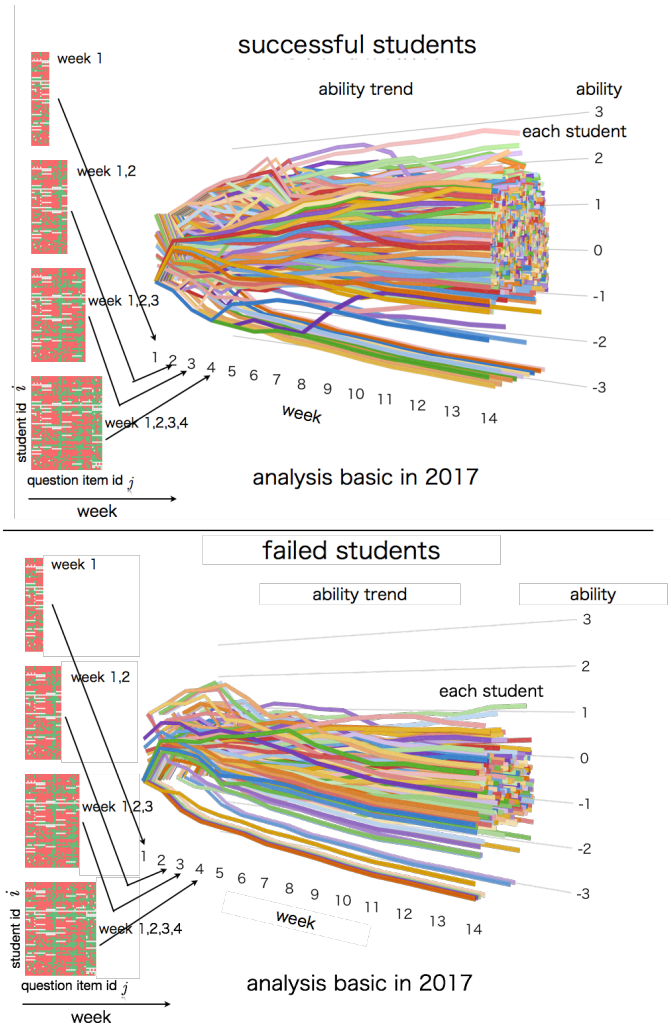

Figure 1: Trends of estimated abilities $\theta_1(i,k)$ for successful and failed students (analysis basic in the first semester in 2017).

success/failure indicator functions $\delta^k_{i,(j),1,1}$, i.e., 1 for success and 0 for failure from (j)th final success/failure results. Thus, $\mu(i,k,1,1)$ will take values of $0, 0.1, \ldots, 0.9, 1$. Then, $\mu(i,k,1,1)$ expressed the predicted value for success in the final examination.

However, we used the same testing results in prediction and in construction of the mathematical model. In such a situation, actually, we cannot predict the final examination status in the middle of the semester term because we have not yet obtained the teacher data to refer to. In addition, if we dare to use the same testing results in prediction, the proposed method may cause the overfitting effect. We should choose a different reference database from the target database in order to obtain the more accurate similarities.

We have now a 2018 database of trends, and we want to use this database to predict the 2018 success/failure prediction. Figure 2 shows two trends of estimated abilities $\theta_2(i,k)$ using the corresponding response matrices, one for successful, and the other for failed students, in the final examination of analysis basic in 2018. We can see that successful cases in Figures 1 and 2 show similar trends, and failed cases in Figures 1 and 2 also show similar trends. Regarding how these two figures are similar to each other, we investigate this in the discussion section 7.3 later.

Thus, we may use the 2018 database of trends for target and the 2017 database of trends for reference to avoid the overfitting. Here, the similarity $S^k_{i,j,2,1}$ is now defined by

$$S^k_{i,j,2,1} = \sqrt{\frac{1}{k}\sum_{l=1}^{k}(\theta_2(j,l) - \theta_1(i,l))^2}, \ (i \neq j), \tag{4}$$

where, $\theta_2(j,l)$ are taken from the 2018 database. The mean value $\mu(i,k,2,1)$ of the final examination's success/failure indicator functions $\delta^k_{i,(j),2,1}$ is also redefined.

6 Identifying Successful/Failed Students Using Similarity

As was shown in the previous paper [19], we dealt with typical three cases in using the LCT response results: 1) from LCT no.1 to LCT no.4, 2) from LCT no.1 to LCT no.7, 3) from LCT no.1 to LCT no.11. We will deal with the same cases treated similarity in [19].

Tables 1-3 show the confusion matrix for these three patterns of response results and three probability cases using the target database as 2018 database and the referred database as 2017 database; Table 1 corresponds to the case of LCT no.1 to LCT no.4, Table 2 to LCT no.1 to LCT no.7, and Table 3 to LCT no.1 to LCT no.11. In these tables, we see $p \geq 0.3$, $p \geq 0.4$, and $p \geq 0.5$, where e.g., $p \geq 0.3$ means that the successful probability to the final examination is larger than or equal to 0.3, in the case of analysis basic. More concretely, in Table 1, "LCT #1-#4 $p \geq 0.3$" is corresponding to the case of "$\mu(i,4,2,1) \geq 0.3$ using LCT from no.1 to LCT no.4. If a student i catches ten students showing very similar trends and the mean value for these ten success/failure values is 0.5 ($p = 0.5$), then this student i is predicted to be successful, and this case is counted in the cases of $p \geq 0.3$. In the table, 938 students were predicted to be successful and 292 students were predicted to be failed in such a manner. Of these 292 students, 204 students were actually successful and 88 students were actually failed. This explanation is much easier to be understood when combined with Figure 3 shown later; in the figure, upper green parts express the observed successful number of students, and lower orange parts express the observed failed number of students.

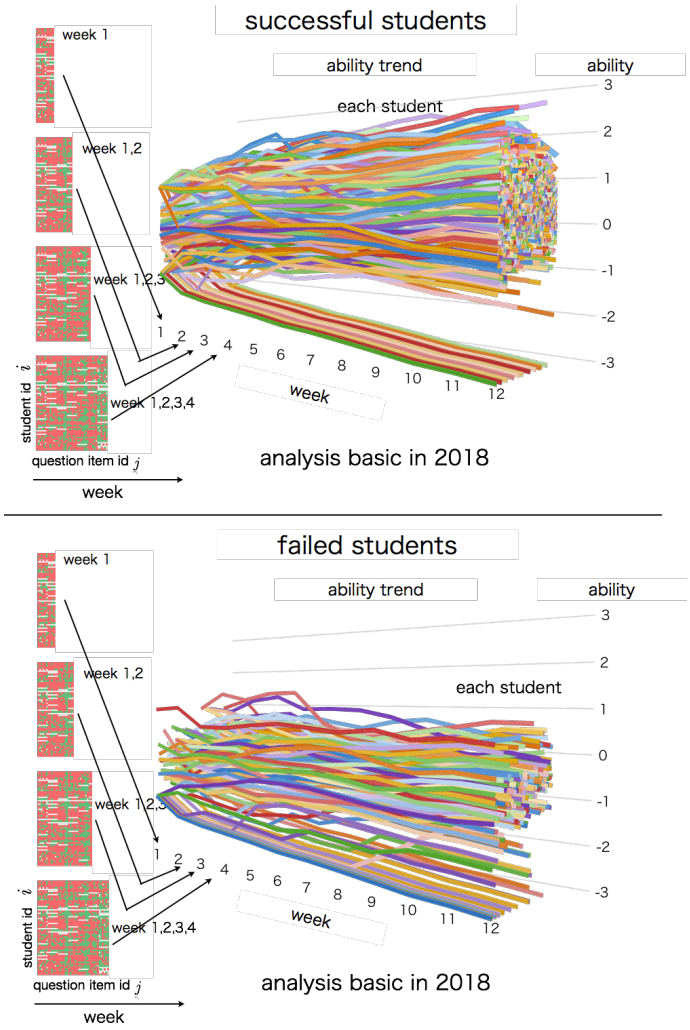

Figure 2: Trends of estimated abilities $\theta_2(i,k)$ for successful and failed students (analysis basic in the first semester in 2018).

Table 4 shows the misclassification rates to these cases; the smaller the numbers, the more accurate the prediction accuracy. Table 5 shows the hit rates for failed cases, i.e., the ratio of the number of actually failed students to the number of predicted failed students; the larger the numbers, the more reliable the prediction accuracy.

Table 1: Confusion matrix determined by the nearest neighbor (analysis basic LCT #1-#4)

LCT #1-#4	$p \geq 0.3$	predicted		
		successful	failed	total
observed	successful	831	204	1035
	failed	107	88	195
	total	938	292	1230

LCT #1-#4	$p \geq 0.4$	predicted		
		successful	failed	total
observed	successful	908	127	1035
	failed	134	81	195
	total	1042	208	1230

LCT #1-#4	$p \geq 0.5$	predicted		
		successful	failed	total
observed	successful	970	65	1035
	failed	157	38	195
	total	1127	103	1230

Figures 3 shows the bar charts for the predicted number of students to be failed in the final examination in the first semester in 2018, in the case of analysis basic. That is, these are corresponding to the hit rates in Table 5. As explained before, upper green parts express the observed successful number of students, and lower orange parts express the observed failed number of students. In the figure, we see a notation of $p \geq 0.3$, e.g., which is the same as $\mu(i,4,2,1) \geq 0.3$ when using LCT no.1 to LCT no.4, and other notations are expressed in a similar manner.

7 Discussions

7.1 Failure Probability We Should Adopt

We have shown nine cases of confusion matrices, misclassification rates, and hit rates to the subject of analysis basic. The misclassification rates shown here are all small enough compared to that in the case that we used the single full response matrix using the decision tree method as shown in [19], where the misclassification rate in the case of analysis basic in 2017 was 0.34, although the misclassification rates using the same database in 2017 to the target and to the reference are smaller to some extent than those computed in this paper (the target is 2018 trends and the reference is 2017 trends). We see that the proposed method to use the previous year database for reference works well.

To determine which case is recommended to use among the three cases of $p \geq 0.3$, $p \geq 0.4$, and $p \geq 0.5$, we plotted the Receiver Operating Characteristic (ROC) curve in Figure 4.

Table 2: Confusion matrix determined by the nearest neighbor (analysis basic LCT #1-#7)

LCT #1-#7	$p \geq 0.3$	predicted		
		successful	failed	total
observed	successful	826	209	1035
	failed	80	115	195
	total	906	324	1230

LCT #1-#7	$p \geq 0.4$	predicted		
		successful	failed	total
observed	successful	960	75	1035
	failed	142	53	195
	total	1102	128	1230

LCT #1-#7	$p \geq 0.5$	predicted		
		successful	failed	total
observed	successful	1003	32	1035
	failed	167	28	195
	total	1170	60	1230

Table 3: Confusion matrix determined by the nearest neighbor (analysis basic LCT #1-#11)

LCT #1-#11	$p \geq 0.3$	predicted		
		successful	failed	total
observed	successful	803	232	1035
	failed	51	144	195
	total	854	376	1230

LCT #1-#11	$p \geq 0.4$	predicted		
		successful	failed	total
observed	successful	963	72	1035
	failed	115	80	195
	total	1078	152	1230

LCT #1-#11	$p \geq 0.5$	predicted		
		successful	failed	total
observed	successful	991	44	1035
	failed	144	51	195
	total	1135	95	1230

Table 4: Misclassification rates corresponding to Tables 1-3 (analysis basic)

	LCT #1-#4	LCT #1-#7	LCT #1-#11
$p \geq 0.3$	0.25	0.24	0.23
$p \geq 0.4$	0.21	0.18	0.15
$p \geq 0.5$	0.18	0.16	0.15

Table 5: Hit rates for failures corresponding to Tables 1-3 (analysis basic)

	LCT #1-#4	LCT #1-#7	LCT #1-#11
$p \geq 0.3$	0.30	0.36	0.38
$p \geq 0.4$	0.39	0.41	0.53
$p \geq 0.5$	0.37	0.47	0.54

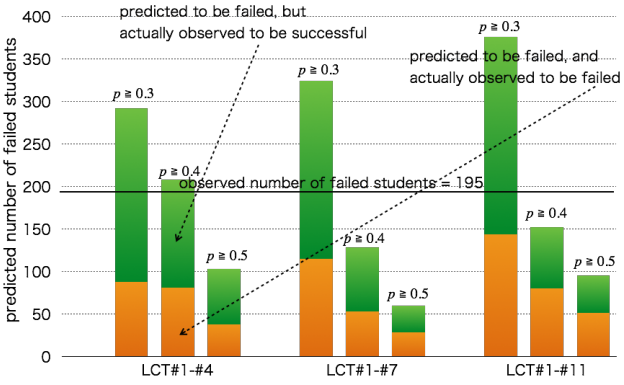

Figure 3: Numbers of successful/failed students using the similarity of the trends of estimated students' abilities (analysis basic in the first semester).

That means what probability value P we should adopt among cases $p \geq P$. Our concern is to know students faced to risk of drop-out, we have set that the false positive rate is corresponding to the actually failed but estimated to be successful students, and true positive rate is corresponding to the actually failed and estimated to be failed students.

If we assume the cost of true positive is four times higher than that of false positive, then the tangent becomes to be 4. Thus, by looking at the figure, we may choose the appropriate case which could be $p \geq 0.4$. In addition, it would be convenient that the prediction time is earlier, thus, we choose the case of using LCT no.1 to LCT no.7 in prediction. In such a case, the misclassification rate is 0.18 in the analysis basic case. As for the hit rate of failed students, it is 0.41 in the analysis basic case. In the case we used 2017 database for both target and reference, the misclassification rate was 0.21, and the hit rate was 0.40 in the case of analysis basic; they are comparable to the results we have proposed in this paper. Therefore, we can say that the proposed method works very well.

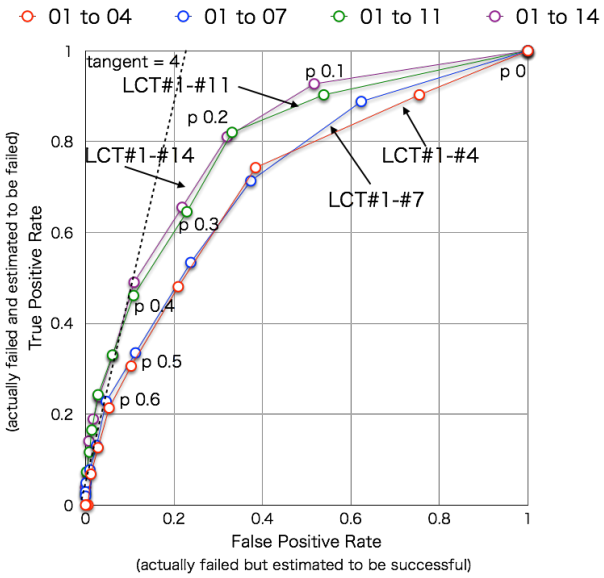

Figure 4: ROC curve (analysis basic in the first semester in 2017).

7.2 When Can We Predict the Success/Failure?

By looking at Figure 3 and Table5, the prediction result for success/failure using LCT no.1 to LCT no.7 is similar to that using LCT no.1 to LCT no.11. Thus, the prediction could be performed at the timing of LCT no.7, just in the middle of the semester. By looking at Figure 4, the ROC curve using LCT no.1 to LCT no.7 is similar to that using LCT no.1 to LCT no.4. Therefore, we may set the prediction time when LCT no.4 is done, even if the hitting rates are smaller to some extent comparing to the case using LCT no.1 to LCT no.7. Then, we can predict the final stage status even before the middle of the semester. This shows that we can identify students at risk for failing courses and/or dropping earlier.

7.3 Effect of Preprocessing to the Testing Data

As mentioned earlier, we have made the testing data preprocessed in some ways in order to improve the prediction accuracy. The reason why we have performed such a trial is that the estimated abilities in 2017 and those in 2018 were not standardized. There may exist additional errors due to this non-standardization. Moreover, there were some students who were not taking the testings at all, and we think that the information from these students may disturb the accurate estimation.

Therefore, we performed additional three cases to the original data case; these are 1) CAR=0 cases removed in 2017 LCT from the full LCT data (2017 CAR=0 removed), 2) LCT values are standardized with mean=0 and standard deviation=1 (standardized), 3) both 2017 CAR=0 removed and standardized.

Figure 5 shows the comparison of histograms of the 2018 failure probabilities to the final examination between using 2018 LCT#1-#4 data and 2017 LCT#1-#4 data, Figure 6 shows those using 2018 LCT#1-#7 and 2017 LCT#1-#7, and Figure 7 shows those using 2018 LCT#1-#11 and 2017 LCT#1-#11. In any figures, we can see no obvious differences among original case and preprocessed cases.

Figure 8 shows the comparison of cumulative distribution functions of the 2018 failure probabilities to the final examination between using 2018 LCT#1-#4 data and 2017 LCT#1-#4 data, Figure 9 shows those using 2018 LCT#1-#7 and 2017 LCT#1-#7, and Figure 10 shows those using 2018 LCT#1-#11 and 2017 LCT#1-#11. In any figures, we can see no obvious differences among original case and preprocessed cases.

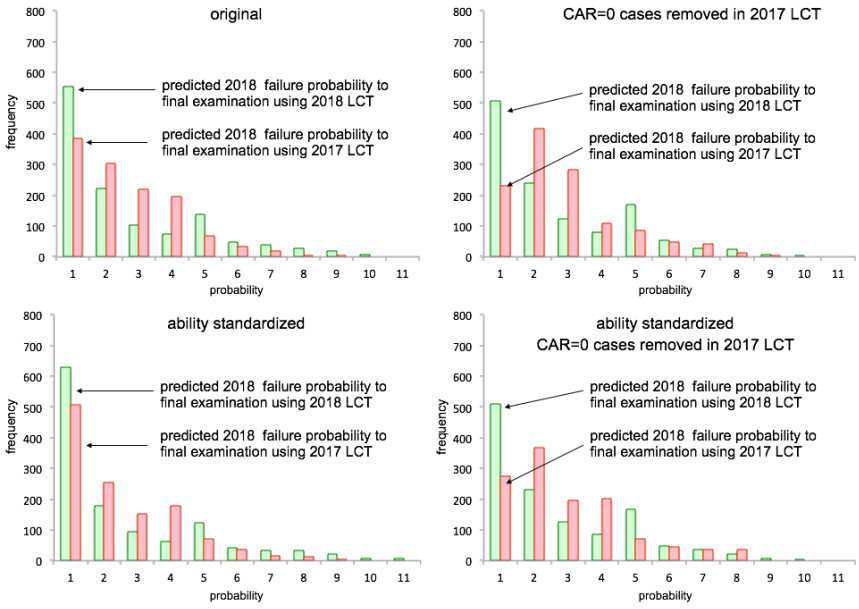

Figure 5: Comparison of histograms of the 2018 failure probabilities to the final examination between using 2018 LCT#1-#4 data and 2017 LCT#1-#4 data.

Then, we have compared the misclassification rates and the hit rates for failure prediction among four cases: 1) original (no-preprocessing), 2) 2017 CAR=0 removed (CAR=0 cases removed in 2017 LCT from the full LCT data), 3) standardized (LCT values are standardized with mean=0 and standard deviation=1), and 4) standardized & 2017 CAR=0 removed.

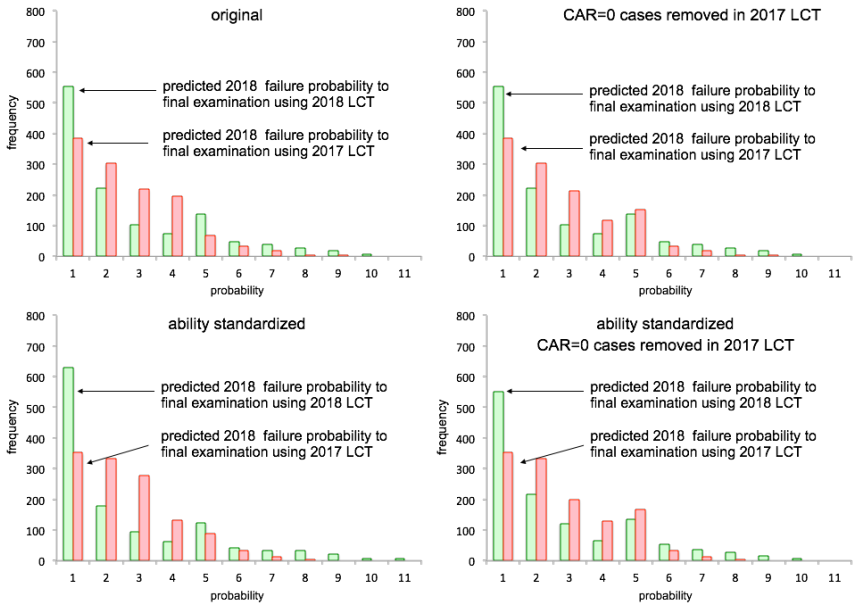

Figure 6: Comparison of histograms of the 2018 failure probabilities to the final examination between using 2018 LCT#1-#7 data and 2017 LCT#1-#7 data.

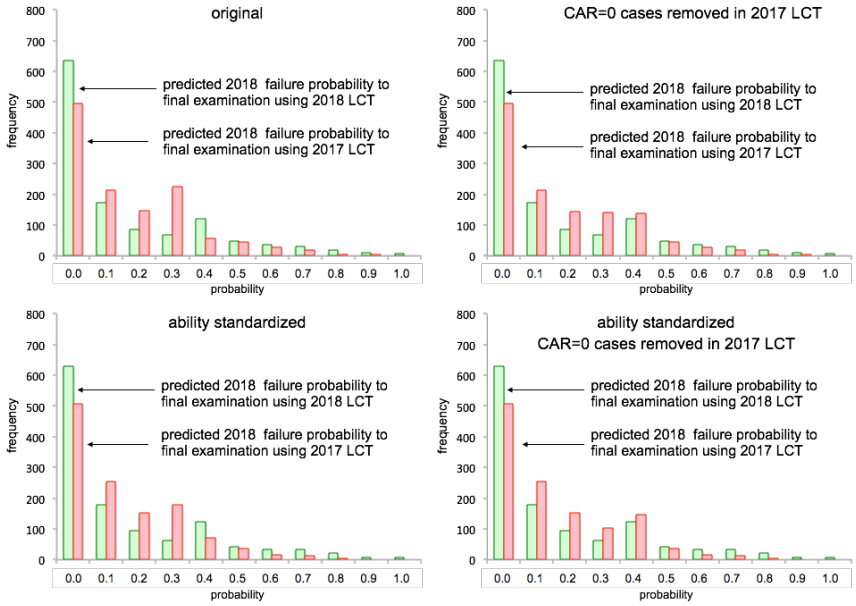

Figure 7: Comparison of histograms of the 2018 failure probabilities to the final examination between using 2018 LCT#1-#11 data and 2017 LCT#1-#11 data.

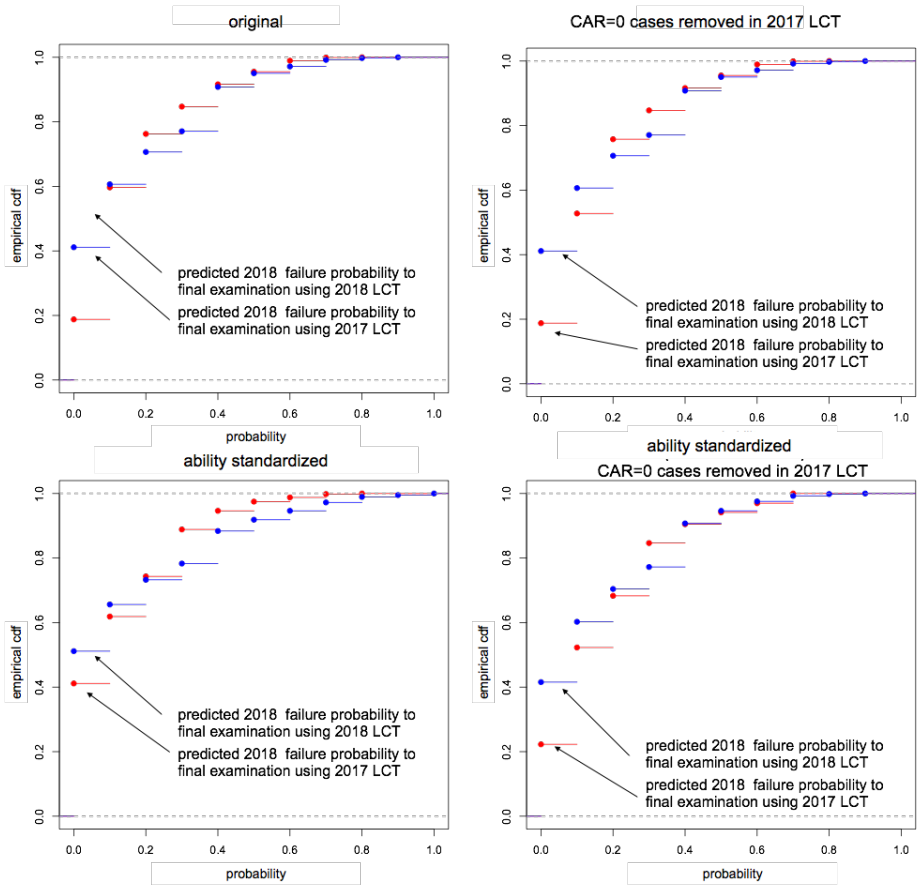

Figure 8: Comparison of cumulative distribution functions of the 2018 failure probabilities to the final examination between using 2018 LCT#1-#4 data and 2017 LCT#1-#4 data.

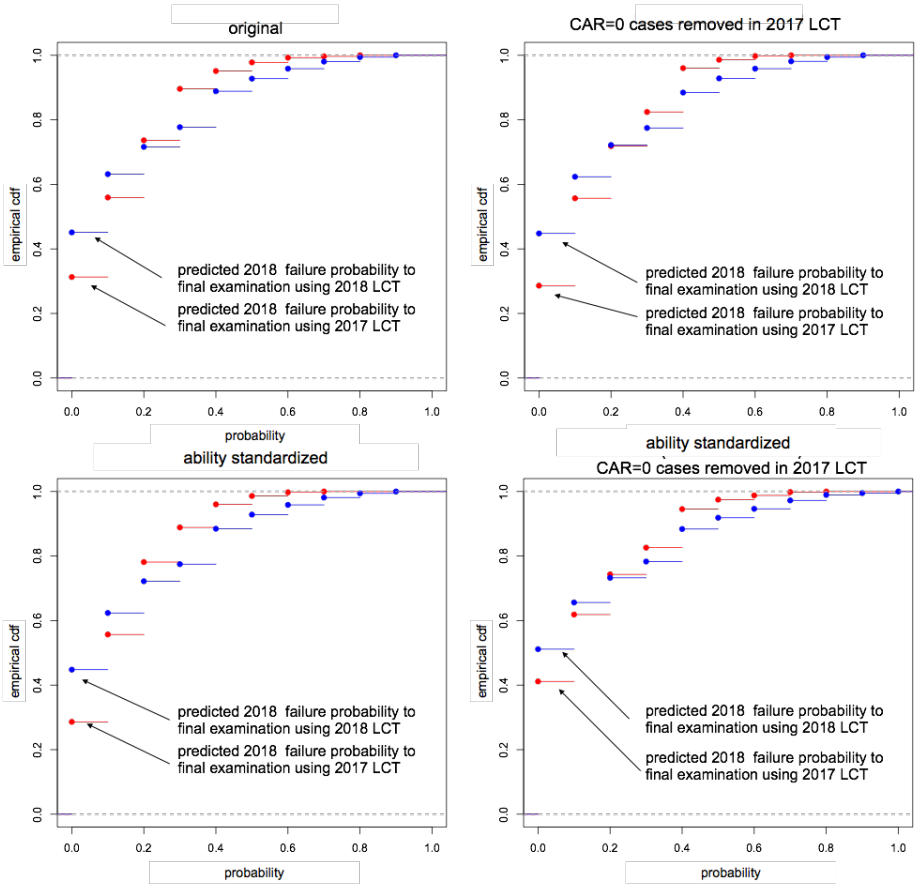

Figure 9: Comparison of cumulative distribution functions of the 2018 failure probabilities to the final examination between using 2018 LCT#1-#7 data and 2017 LCT#1-#7 data.

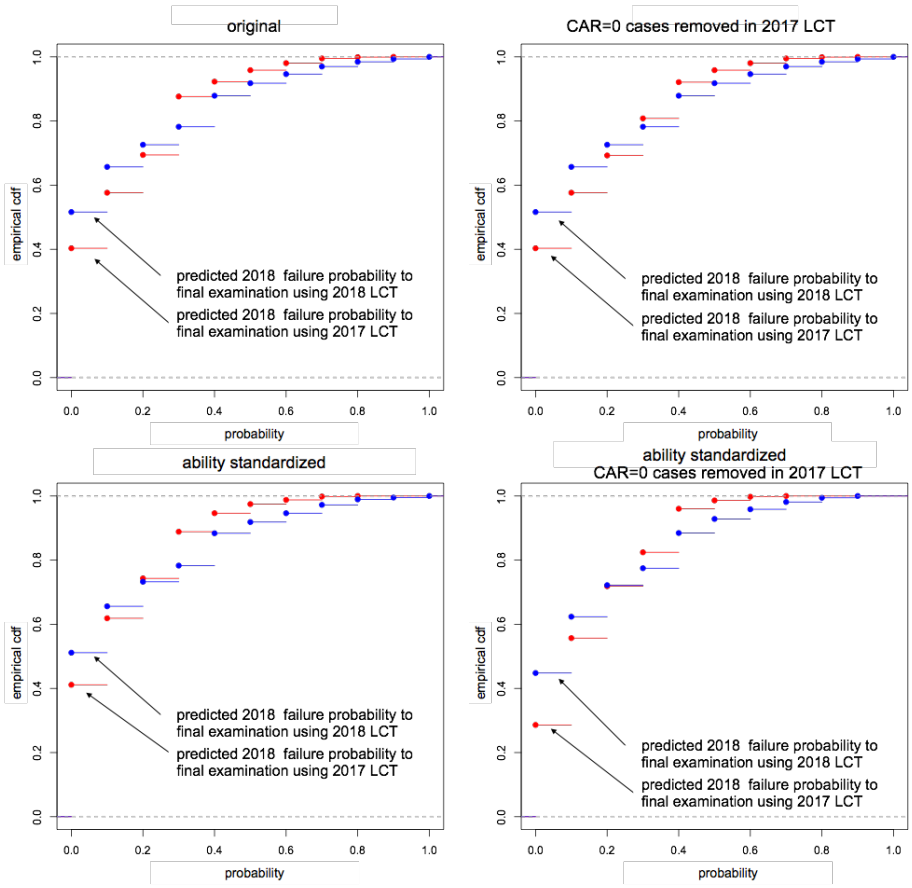

Figure 10: Comparison of cumulative distribution functions of the 2018 failure probabilities to the final examination between using 2018 LCT#1-#11 data and 2017 LCT#1-#11 data.

Table 6 shows the misclassification rates prediction result using 2017 LCT data, and Table 7 shows those using 2018 LCT data. In general, the former results are a little bit larger than the latter results due to overfitting effects. However, the differences are small.

Table 8 shows the hit rates for failure prediction result using 2017 LCT data, and Table 9 shows those using 2018 LCT data. In general, the former results are a little bit smaller than the latter results due to overfitting effects. However, the differences are small.

In conclusion, we have not seen any obvious improvements even though some data preprocessing were performed. Therefore, we can use the proposed method such that we use the 2018 LCT results as the target database and we use 2017 LCT results as the reference database. Looking at the results in Tables 6-9, standardization in 2017 LCT abilities could be a possible selection among the data preprocessing cases.

Table 6: Misclassification rates prediction result using using 2017 LCT data (analysis basic)

		$p \geq 0.3$	$p \geq 0.4$	$p \geq 0.5$
LCT #1-#4	original	0.253	0.212	0.180
	2017 CAR=0 removed	0.249	0.212	0.180
	standardized	0.233	0.189	0.167
	standardized & 2017 CAR=0 removed	0.233	0.189	0.167
LCT #1-#7	original	0.235	0.178	0.162
	2017 CAR=0 removed	0.231	0.197	0.162
	standardized	0.202	0.163	0.154
	standardized & 2017 CAR=0 removed	0.230	0.189	0.154
LCT #1-#11	original	0.230	0.152	0.153
	2017 CAR=0 removed	0.228	0.175	0.146
	standardized	0.196	0.148	0.154
	standardized & 2017 CAR=0 removed	0.198	0.150	0.150

original: using full LCT data
2017 CAR=0 removed: CAR=0 cases removed in 2017 LCT from the full LCT data
standardized: LCT values are standardized with mean=0 and standard deviation=1

7.4 Significance and Effectiveness of the Research

Regarding the early prediction of drop-out in academic examinations using learning analytics, similar studies have been published (see [2, 4, 7, 33, 38]). Among them, the review research by Liz-Dominguez et al. [4] was versatile and profound. They picked up 1382 papers by using IEEE Xplore, ACM Digital Library, Elsevier ScienceDirect, Wiley Online, SpringerLink, Emerald, Taylor & Francis, Scopus, and Web of Science; then, fully relevant 67 papers to early prediction of drop-out were selected and analyzed; of these, they gave brief explanations to 13 papers. The paper by Hirose [19] was included in these and it is directly related to this article. The prediction algorithms they investigated were Naive Bayes, logistic regression, tree, random forest, nearest neighbor, support vector machine and neural networks. He mentioned in his paper that Hirose [19] dealt with the IRT in evaluating students' abilities combined with a machine learning method of k-nearest neighbor, and the misclassification rate in classifying the success/failed examinees was low.

Table 7: Misclassification rates prediction result using using 2018 LCT data (analysis basic)

		$p \geq 0.3$	$p \geq 0.4$	$p \geq 0.5$
LCT #1-#4	original	0.237	0.211	0.167
	2017 CAR=0 removed	0.237	0.211	0.167
	standardized	0.240	0.209	0.168
	standardized & 2017 CAR=0 removed	0.240	0.209	0.168
LCT #1-#7	original	0.221	0.191	0.156
	2017 CAR=0 removed	0.221	0.191	0.156
	standardized	0.214	0.193	0.157
	standardized & 2017 CAR=0 removed	0.221	0.191	0.156
LCT #1-#11	original	0.195	0.170	0.128
	2017 CAR=0 removed	0.195	0.170	0.128
	standardized	0.185	0.169	0.133
	standardized & 2017 CAR=0 removed	0.185	0.169	0.133

original: using full LCT data
2017 CAR=0 removed: CAR=0 cases removed in 2017 LCT from the full LCT data
standardized: LCT values are standardized with mean=0 and standard deviation=1

Table 8: Hit rates for failures prediction result using using 2017 LCT data (analysis basic LCT)

		$p \geq 0.3$	$p \geq 0.4$	$p \geq 0.5$
LCT #1-#4	original	0.301	0.324	0.367
	2017 CAR=0 removed	0.312	0.324	0.369
	standardized	0.401	0.482	0.551
	standardized & 2017 CAR=0 removed	0.351	0.402	0.453
LCT #1-#7	original	0.355	0.414	0.467
	2017 CAR=0 removed	0.365	0.412	0.467
	standardized	0.401	0.482	0.551
	standardized & 2017 CAR=0 removed	0.373	0.412	0.551
LCT #1-#11	original	0.383	0.526	0.537
	2017 CAR=0 removed	0.365	0.412	0.621
	standardized	0.427	0.547	0.551
	standardized & 2017 CAR=0 removed	0.425	0.537	0.585

original: using full LCT data
2017 CAR=0 removed: CAR=0 cases removed in 2017 LCT from the full LCT data
standardized: LCT values are standardized with mean=0 and standard deviation=1

Table 9: Hit rates for failures prediction result using using 2018 LCT data (analysis basic LCT)

		$p \geq 0.3$	$p \geq 0.4$	$p \geq 0.5$
LCT #1-#4	original	0.366	0.387	0.451
	2017 CAR=0 removed	0.366	0.387	0.451
	standardized	0.363	0.389	0.447
	standardized & 2017 CAR=0 removed	0.363	0.389	0.447
LCT #1-#7	original	0.390	0.427	0.511
	2017 CAR=0 removed	0.390	0.427	0.511
	standardized	0.401	0.422	0.507
	standardized & 2017 CAR=0 removed	0.390	0.427	0.511
LCT #1-#11	original	0.433	0.474	0.624
	2017 CAR=0 removed	0.433	0.474	0.624
	standardized	0.450	0.476	0.608
	standardized & 2017 CAR=0 removed	0.450	0.476	0.608

original: using full LCT data
2017 CAR=0 removed: CAR=0 cases removed in 2017 LCT from the full LCT data
standardized: LCT values are standardized with mean=0 and standard deviation=1

The author thinks that this combination approach is unique and scholarly significant in prediction methodology. The use of similarity in dealing with the nearest neighbor method is also new. In addition, Figueroa-Canas et al. [7] mentioned in their paper that Hirose [19] dealt with the ROC curve to find the optimal point for segregating the failed examinees in the final examination.

Although the subject we have treated is mathematics in undergraduate courses, the proposed prediction method can also be applied to other subjects such as English and physics, as Azuma [2] suggested.

7.5 Possibility to the Online Testing for the Final Examination

In 2020, COVID-19 has totally changed the learning manner worldwide from face-to-face to online. All the teachers and students were forced to accept lectures online. However, many teachers may be wondering whether the final examination should be taken by face-to-face style to evaluate the students' scores fairly and accurately.

We have been experiencing issues that could arise surround computer based testing until now. Internet crashes, glitches in programs, internet connection issues, data security are among them. However, with advances in information technology, they will be overcome in the future. Rather, it is much more important that many students preferred testing on computers rather than with a pencil and paper (see [5]). This is true also in our case.

The principal issue in the online testing may be the prevention of cheating. Chirumamilla et al. report such aspects (see [3]). They considered cases of impersonation, forbidden aids, peeking, peer collaboration, outside assistance, and student-staff collusion.

According to questionnaires and interviews, both students and teachers perceived cheating as easier with e-exams, and especially with bring student's own device. Here, e-exam means computer based testing. Thus, it will be crucial to prevent cheating in online testing

from now on.

If we adopt multiple choice type testing rather than description type testing, much fairer and much more accurate student's ability evaluation could be achieved with teacher's evaluation bias free and without cheating. From a statistical viewpoint, this is also supported by comparing paper based testing and computer based testing using the IRT (see [30]).

As long as we can prevent cheating, the results of this paper suggest the possibility to the online testing to the official final examination. How we proceed the online testing fairly and accurately is the future work to be resolved to the online education era.

8 Concluding Remarks

We showed previously that we can predict the success/failure status for the final examination to each student at early stages in courses using the current trends of estimated abilities in terms of item response theory for online testing, using the same testing results in prediction and in construction of the mathematical model. However, we used the same testing results in prediction and in construction of the mathematical model. We may not predict the final examination status in the middle of the semester term. In addition, the proposed method may cause overfitting errors in prediction. We should choose a different reference database from the target database in order to obtain the more accurate similarities.

Thus, in this paper, we have proposed to use the different databases in reference and in target in prediction. i.e., we use the 2018 LCT results as the target database and we use 2017 LCT results as the reference database.

We have investigated whether the proposed method works well by using the subject of analysis basic as a typical case, and we have shown that we can still predict the current success/failure status for the final examination using the past trends of estimated abilities of the online testing and the past final examination results.

To assess whether some data preprocessing will work or not, we have investigated additional such cases. In conclusion, we have not seen any obvious improvements even though some data preprocessing were performed. Comparing with the results in data preprocessing cases, standardization in 2017 LCT abilities could be a possible selection among them.

Although the subject we have dealt with is mathematics for undergraduate students, the methodology shown here can be applied to other subjects such as physics and English.

References

[1] R. de Ayala, The Theory and Practice of Item Response Theory. Guilford Press, 2009.

[2] R. Azuma, Effectiveness of Comments on Self-reflection Sheet in Predicting Student Performance, 10th International Conference on Operations Research and Enterprise Systems, 2021, pp. 394-400.

[3] A. Chirumamilla, G. Sindre, A. Nguyen-Duc, Cheating in e-exams and paper exams: the perceptions of engineering students and teachers in Norway, Assessment & Evaluation in Higher Education, 45, 2020, pp. 940-957.

[4] M. Liz-Dominguez, M. Caeiro-Rodriguez, M. Llamas-Nistal, F.A. Mikic-Fonte, Systematic Literature Review of Predictive Analysis Tools in Higher Education, Applied Sciences, 9, 5569, 2019, pp.1-26.

[5] S. Gonzalez, The Pros and Cons of Computer-Based Standardized Testing for Elementary Students, Capstone Projects and Master's Theses. 853, 2020.

[6] N. Elouazizi, Critical Factors in Data Governance for Learning Analytics, Journal of Learning Analytics, 1, 2014, pp. 211-222.

[7] J. Figueroa-Canas, T. Sancho-Vinuesa, Early Prediction of Dropout and Final Exam Performance in an Online Statistics Course, IEEE Revista Iberoamericana de Tecnologias del Aprendizaje, 15, 2020, pp. 86-94.

[8] D. Gasevic, S. Dawson, and G. Siemens, Let's not forget: Learning analytics are about learning, TechTrends, 59, 2015, pp. 64-71.

[9] R. Hambleton, H. Swaminathan, and H. J. Rogers, Fundamentals of Item Response Theory. Sage Publications, 1991.

[10] H. Hirose and T. Sakumura, Test evaluation system via the web using the item response theory, in Computer and Advanced Technology in Education, 2010, pp.152-158.

[11] H. Hirose, T. Sakumura, Item Response Prediction for Incomplete Response Matrix Using the EM-type Item Response Theory with Application to Adaptive Online Ability Evaluation System, IEEE International Conference on Teaching, Assessment, and Learning for Engineering, 2012, pp.8-12.

[12] H. Hirose, Yu Aizawa, Automatically Growing Dually Adaptive Online IRT Testing System, IEEE International Conference on Teaching, Assessment, and Learning for Engineering, 2014, pp.528-533.

[13] H. Hirose, Y. Tokusada, K. Noguchi, Dually Adaptive Online IRT Testing System with Application to High-School Mathematics Testing Case, IEEE International Conference on Teaching, Assessment, and Learning for Engineering, 2014, pp.447-452.

[14] H. Hirose, Y. Tokusada, A Simulation Study to the Dually Adaptive Online IRT Testing System, IEEE International Conference on Teaching, Assessment, and Learning for Engineering, 2014, pp.97-102.

[15] H. Hirose, Meticulous Learning Follow-up Systems for Undergraduate Students Using the Online Item Response Theory, 5th International Conference on Learning Technologies and Learning Environments, 2016, pp.427-432.

[16] H. Hirose, M. Takatou, Y. Yamauchi, T. Taniguchi, T. Honda, F. Kubo, M. Imaoka, T. Koyama, Questions and Answers Database Construction for Adaptive Online IRT Testing Systems: Analysis Course and Linear Algebra Course, 5th International Conference on Learning Technologies and Learning Environments, 2016, pp.433-438.

[17] H. Hirose, Learning Analytics to Adaptive Online IRT Testing Systems "Ai Arutte" Harmonized with University Textbooks, 5th International Conference on Learning Technologies and Learning Environments, 2016, pp.439-444.

[18] H. Hirose, M. Takatou, Y. Yamauchi, T. Taniguchi, F. Kubo, M. Imaoka, T. Koyama, Rediscovery of Initial Habituation Importance Learned from Analytics of Learning

Check Testing in Mathematics for Undergraduate Students, 6th International Conference on Learning Technologies and Learning Environments, 2017, pp.482-486.

[19] H. Hirose, Success/Failure Prediction for Final Examination Using the Trend of Weekly Online Testing, 7th International Conference on Learning Technologies and Learning Environments, 2018, pp.139-145.

[20] H. Hirose, Attendance to Lectures is Crucial in Order Not to Drop Out, 7th International Conference on Learning Technologies and Learning Environments, 2018, pp.194-198.

[21] H. Hirose, Time Duration Statistics Spent for Tackling Online Testing, 7th International Conference on Learning Technologies and Learning Environments, 2018, pp.221-225.

[22] H. Hirose, Prediction of Success or Failure for Examination using Nearest Neighbor Method to the Trend of Weekly Online Testing, International Journal of Learning Technologies and Learning Environments, 2, 2019, pp.19-34.

[23] H. Hirose, Relationship Between Testing Time and Score in CBT, International Journal of Learning Technologies and Learning Environments, 2, 2019, pp.35-52.

[24] H. Hirose, Current Failure Prediction for Final Examination using Past Trends of Weekly Online Testing, 9th International Conference on Learning Technologies and Learning Environments, 2020, pp.142-148.

[25] H. Hirose, More Accurate Evaluation of Student's Ability Based on A Newly Proposed Ability Equation, 9th International Conference on Learning Technologies and Learning Environments, 2020, pp.176-182.

[26] H. Hirose, Difference Between Successful and Failed Students Learned from Analytics of Weekly Learning Check Testing, Information Engineering Express, Vol 4, 2018, pp.11-21.

[27] H. Hirose, Key Factor Not to Drop Out is to Attend Lectures, Information Engineering Express, 5, 2019, pp.11-21.

[28] H. Hirose, Dually Adaptive Online IRT Testing System, Bulletin of Informatics and Cybernetics Research Association of Statistical Sciences, 48, 2016, pp.1-17.

[29] W. J. D. Linden and R. K. Hambleton, Handbook of Modern Item Response Theory. Springer, 1996.

[30] H. Retnawati, The Comparison of Accuracy Scores on the Paper and Pencil Testing vs. Computer- Based Testing, The Turkish Online Journal of Educational Technology, 14, 2015, pp.135-142.

[31] T. Sakumura and H. Hirose, Making up the Complete Matrix from the Incomplete Matrix Using the EM-type IRT and Its Application, Transactions on Information Processing Society of Japan (TOM), 72, 2014, pp.17-26.

[32] T. Sakumura, H. Hirose, Bias Reduction of Abilities for Adaptive Online IRT Testing Systems, International Journal of Smart Computing and Artificial Intelligence, 1, 2017, pp.57-70.

[33] R. Sekiya, S. Oyama, M. Kurihara, User-Adaptive Preparation of Mathematical Puzzles Using Item Response Theory and Deep Learning, Advances and Trends in Artificial Intelligence. From Theory to Practice, 2019, pp.530-537.

[34] G. Siemens and D. Gasevic, Guest Editorial - Learning and Knowledge Analytics, Educational Technology & Society, 15, 2012, pp.1-2.

[35] Y. Tokusada, H. Hirose, Evaluation of Abilities by Grouping for Small IRT Testing Systems, 5th International Conference on Learning Technologies and Learning Environments, 2016, pp.445-449.

[36] R. J. Waddington, S. Nam, S. Lonn, S.D. Teasley, Improving Early Warning Systems with Categorized Course Resource Usage, Journal of Learning Analytics, 3, 2016, 263-290.

[37] A.F. Wise and D.W. Shaffer, Why Theory Matters More than Ever in the Age of Big Data, Journal of Learning Analytics, 2, 2015, pp. 5-13.

[38] J. Xiao, O. Bulut, Evaluating the Performances of Missing Data Handling Methods in Ability Estimation From Sparse Data, Educational and Psychological Measurement, 80, 2020, pp.932-954.