Information Engineering Express
International Institute of Applied Informatics
2018, Vol.4, No.1, P.11 - 21

Difference Between Successful and Failed Students Learned from Analytics of Weekly Learning Check Testing

Hideo Hirose *

Abstract

One of the crucial issues in universities where a variety of enrolled students shall be educated to a level of university diploma policy is to identify students at risk for failing courses and/or dropping out early, to take care of them, and to reduce their risks. Using the recently developed follow-up program system aimed at helping students who need basic learning and aimed at assisting teachers who have to engage in teaching a variety of educational students, we can analyze the accumulated testing results in detail because the testings are performed every week to all the first-year undergraduate students. We have found that those who failed in the final examination show the much steeper decreasing trend of correct answer rates in the learning check testing comparing to those who were successful in the final examination. Although the subjects dealt with in this paper are limited to mathematics (calculus and linear algebra), this kind of system will easily be applied to other subjects such as STEM.

Keywords: learning analytics, learning check testing, correct answer rate, odds ratio.

1 Introduction

Nowadays, a variety of students are enrolled in universities (see [21], e.g.), and we teachers have to educate them to a level of universities' diploma policy. In such a situation, to identify students at risk for failing courses and/or dropping out early is crucial, as pointed out in [17, 19]. However, the more rich in variety, the more we need assisting systems because conventional methods may not work with the limited number of staffs in such a condition. Thus, a new assisting system will be required to solve this difficulty.

To overcome such a difficulty, we recently established a follow-up program (FP) system aimed at helping students who need basic learning skills and aimed at assisting teachers who have to engage in teaching a variety of educational students. The FP system consists of the learning check testing (LCT), follow-up program testing (FPT), and collaborative work testing (CWT). All the students take the LCT; some students cannot show good results in the LCT then they are recommended to attend the follow-up program class (FPC) to cover the lacked knowledges using the CWT with the aid of supporters. To perform all the testings online, we provided WiFi circumstances in every regular class. The system has been

* Data Science Research Center, Hiroshima Institute of Technology, Hiroshima, Japan

successfully operating ([10], [11]), and the computational results of other corresponding cases were well investigated ([12], [13], [18], [16]).

As indicated in [2], [3], and [17], the immediate research attention for learning analytics is crucial to make a sustainable impact on the research and practice of learning and teaching. Since the FP system adopts online testings, it is not so difficult to collect a large-scale of learning data; we should actively tackle with the collected data to find the optimal strategies for better learning methods. It is also important to analyze the data theoretically (see [20]). This paper is aimed at obtaining good learning strategies for students at risk for failing courses and/or dropping out.

The LCT consisting of the small number of questions is performed at the beginning or at the end of each lecture in regular classes for about ten minutes to check if they comprehend the content of the lecture. Analysis Basic (i.e., Calculus) and Linear Algebra are two fundamental subjects that mathematics teachers are involved in. In our university, more than 1,000 students are to be enrolled as freshman students, and we have two semesters (15 weeks lectures). Thus, we deal with four subjects and less than 15 testing results are obtained. Although all the students must take the LCT in the first semester, almost half of the students do not necessarily take the LCT in the second semester. In this paper, we mainly analyze the accumulated LCT testing results in detail using the first semester case.

The item response theory (IRT) provides us the difficulties of the test items (problems) and the examinees' abilities together ([1], [4], [14]), resulting in evaluating the examinees' abilities accurately and fairly. In addition, adaptive testing using the IRT selects the most appropriate items to examinees automatically, resulting in more accurate ability estimation and more efficient test procedures ([5], [6], [7], [8], [9], [15]). Thus, we incorporated the IRT evaluation method into the LCT, FPT and CWT. However, we consider the correct answer rate (CAR) to questions for the student ability evaluation in this paper because it would be easier to be understood than the ability evaluation used in the IRT.

2 Learning Check Testing, LCT

The LCT is a kind of mini test using less than ten questions in each LCT. All the students in regular classes take the LCT for about ten minutes using their personal computers; that is, the test is taken online. All the questions are the same to each student, but the order to each question to a student is sorted in a different order from the next student.

Wi-fi systems are equipped in every lecture room to assist the network connection in the campus. After a teacher in a class admits accesses to the LCT to all the attendees in the class, students can begin to take the examination. After the students finish the examination or after pre-setting testing time duration is elapsed, the system computes the students' scores, and send them to the portfolio system in the university. The questions and answers for the LCT are not open to the public.

Each problem in ten items consists of multiple small questions. Students select appropriate answers to each small question from many choices. For example, problem one consists of three questions and each question has ten choices, then the probability to answer the problem correctly by random selections will be reduced to 0.001. Thus, we adopts the two-parameter logistic function $P(\theta_i; a_j, b_j)$ shown below instead of the three-parameter logistic function including pseudo-guessing parameter for the item response theory to esti-

mate the item difficulties.

$$P(\theta_i; a_j, b_j) = \frac{1}{1+\exp\{-1.7a_j(\theta_i - b_j)\}},$$

where θ_i expresses the ability for student i, and a_j, b_j are constants in the logistic function for item j, and they are called the discrimination parameter and the difficulty parameter, respectively; the larger the value of a_j, the more discriminating the problem is (i.e., the better problem), and the larger the value of b_j, the more difficult the problem is.

3 Analytics of the LCT

3.1 Item response matrix

Figure 1 shows an item response matrix for Analysis Basic in the first semester in 2016. The rows and columns mean the questions and students; red and green colors represent the correct answer and incorrect answer, respectively; vacant elements mean that the answers were not given. Nine LCTs were performed; there were six questions in the first six LCTs, but from the seventh to ninth LCTs, ten questions were provided. Thus, 66 questions were used; that is, the number of columns is 66. We have 1160 students in this year; that is, the number of rows is 1160. In the figure, we divided the whole matrix into four because the number of rows is too large to express the matrix in a single one. The number of rows in each matrix is 290 ($290 \times 4 = 1160$). In faculty D, we observe a block of vacant elements; this means that there were no testing in this class. Although all the elements are not occupied by 0/1 responses we can estimate the students' abilities and item difficulties by using the EM-type IRT (see [6]). However, we deal with only the CAR here.

At first sight, students solved the problems well, but gradually the CAR decreases as lectures go forward, because we see that red colors were predominant first then gradually green colors were mixed.

Bar charts for the CAR values to each question are shown in Figure 2 in the cases of Analysis Basic and Linear Algebra in the first and second semesters. In the figure, the abscissa axis represents the question id. For example in the case of Analysis Basic in the first semester, question numbers 1, 2, 3, ... , 6 are used to the first section, and question numbers 7, 8, 9, ... , 12 are used to the second section, and etc. Thus 66 question cases were shown. Although the total number of questions in the second semester is different from that in the first semester, we can see that the CAR decreasing phenomena to each subjects. For the CAR decreasing phenomena, two reasons are imagined: one is that questions of the low numbers were easy because already-known problems were given, and the other is that students payed full attention to tackle the problems at first, but they gradually lost tense. This tendency will be discussed later again.

3.2 Failed student CAR versus successful student CAR in the final examination

In order to identify students at risk for failing, we pay attention to the difference of the LCT scores between students who failed in the final examination and students who were successful. The final examinations are taken at the end of each semester using paper tests. The duration is 70-80 minutes; the questions in the final examinations are more difficult than those in the LCTs. To grasp the difference at first sight, we compared the response

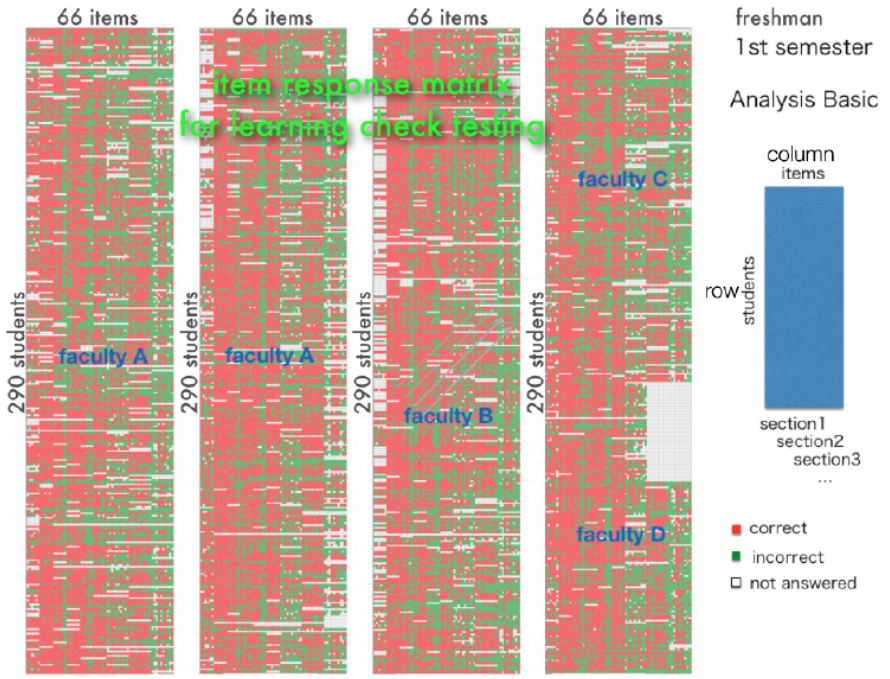

Figure 1: Response matrix (Analysis Basic in the first semester).

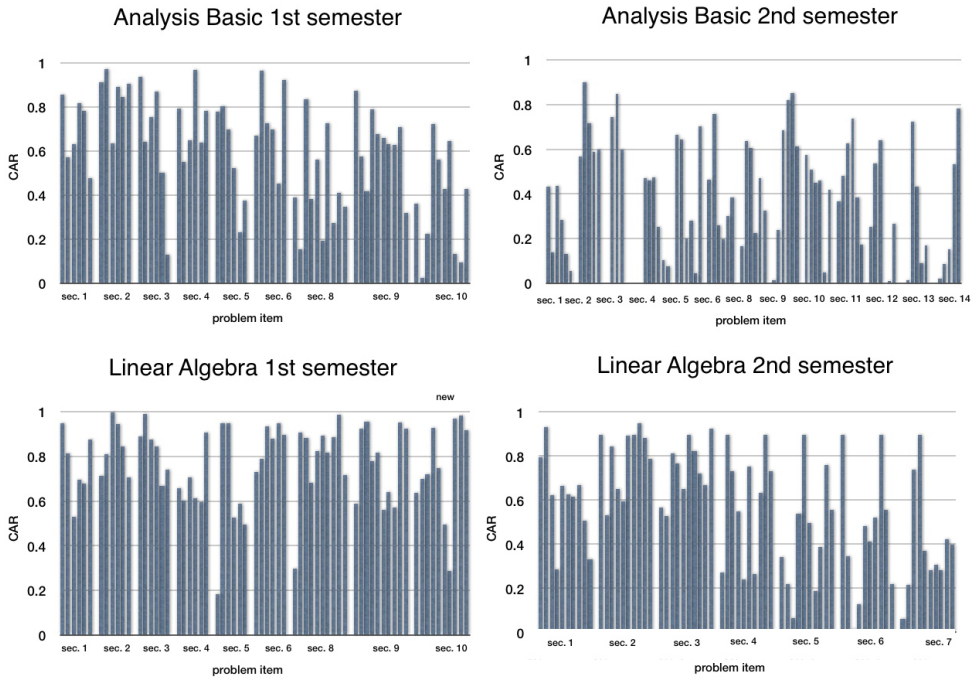

Figure 2: Bar charts for the values of the CAR to each question.

matrices between the two groups of successful students and failed students. Figure 3 shows four cases to each subject and to each semester. On the right in the figure, expanded parts of the two matrices are seen; blue lines drawn on the right of the matrices mean the failed student group and yellow lines the successful student group. We can roughly distinguish the difference between the two groups such that the LCT CARs of the failed students are lower than those of the successful students. However, we cannot estimate whether a certain student belongs to the failed group or the successful group.

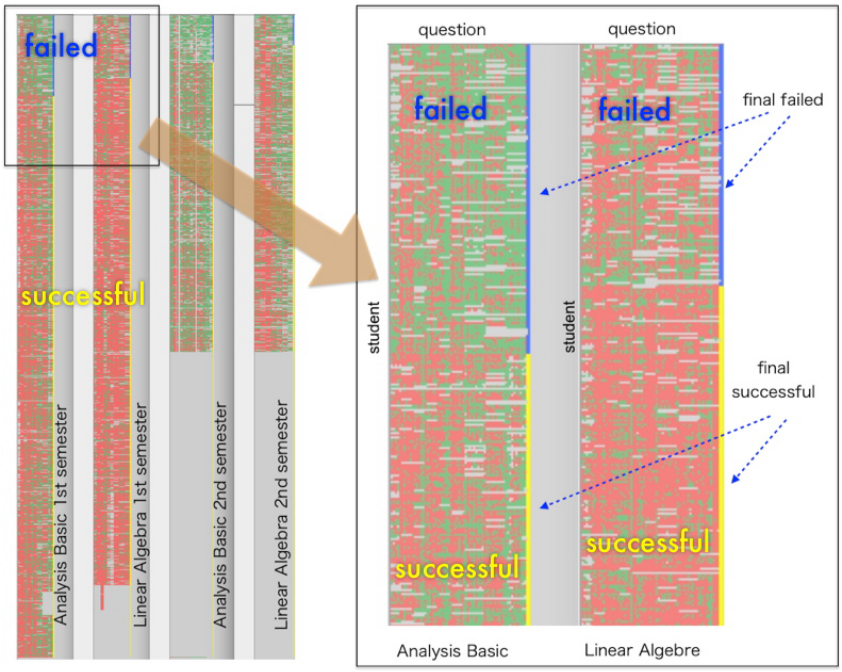

Figure 3: Response matrices of two groups of successful students and failed students. On the left in the right box: Analysis Basic in the first semester, On the right in the right box: Linear Algebra in the first semester.

To see the difference of the two groups much clearer, we provided two bars to each question in the LCTs in the CAR chart; one is for failed group and the other is successful group in the final examination. Figure 4 shows the bar charts for the cases of Analysis Basic and Linear Algebra in the first and second semesters. Intuitively, we observe the clear difference between the two groups. In addition, we can find how extent the differences are between the two groups. To find this, we have compared the failed student CAR values with the successful student CAR values as shown in Figure 5. Looking at the figure, we recognize the difference among the subjects and semesters. For example, the CAR value around 0.5 (to some question id) in the failed group is around 0.8 in the successful group in Analysis Basic and Linear Algebra in the first semester, but the CAR value around 0.5 in the failed group is around 0.7 in Analysis Basic and Linear Algebra in the second semester. However, this could be found after the final examinations were performed. We want to detect such a risk sign earlier.

We next arranged the frequency distributions (histograms) for the CAR to each LCT to both the groups together in the same charts. Figure 6 shows the histograms of the CAR

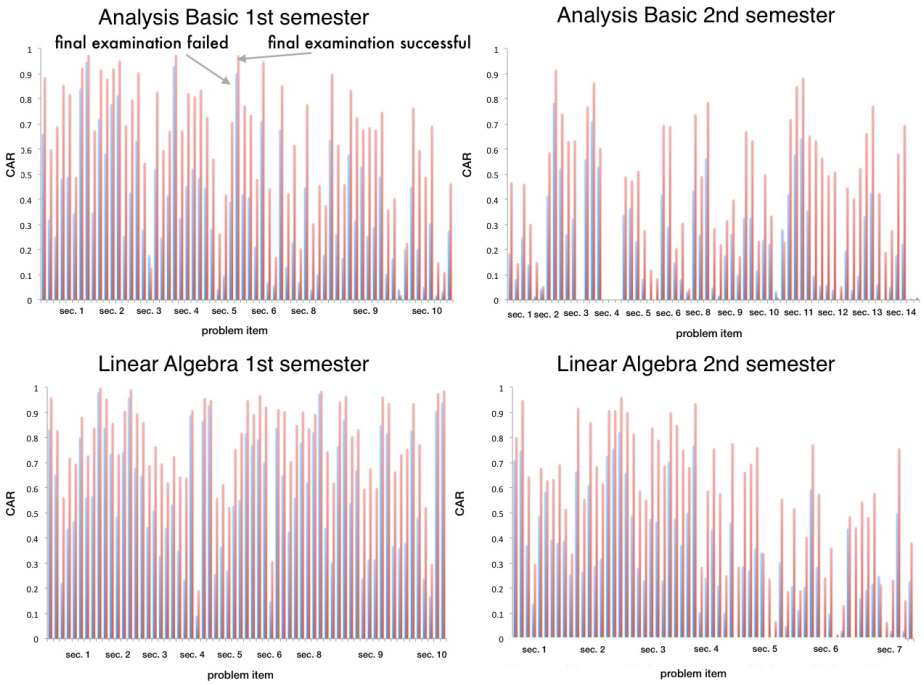

Figure 4: Bar charts of the final examination successful group CAR values and failed group CAR values.

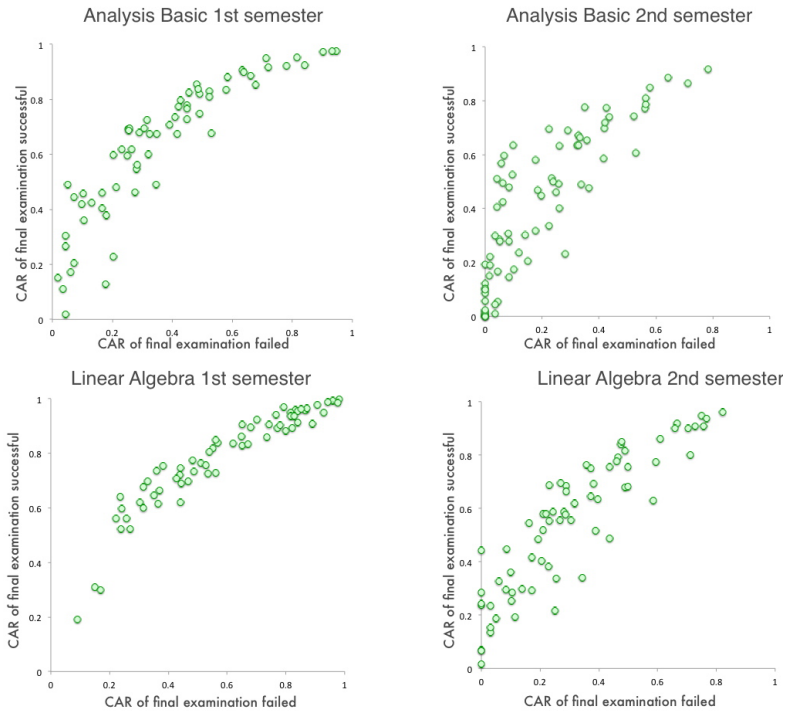

Figure 5: Comparison between the failed student CAR values and the successful student CAR values.

values to each question in the LCT for final examination successful and failed students for four cases. Dark parts indicate the failed group CAR values and bright parts the successful group. Frequencies of 0-0.2, 0.2-0.4, ⋯, 0.8-1.0 CAR values are seen. The histograms tell us that the failed students in the final examination have tendencies of higher frequencies to the lower CAR values. Comparing to this phenomenon, the successful students in final examination show the opposite tendencies in many LCT cases; it may depend on the difficulties of the problems. Moreover, we can comprehend the trends of the CAR distribution differences between the two groups as time goes on. The lower CAR values of failed students seem to be increasing as the LCT proceeds. This tendency seems to be grasped to some extent, but it is not clear.

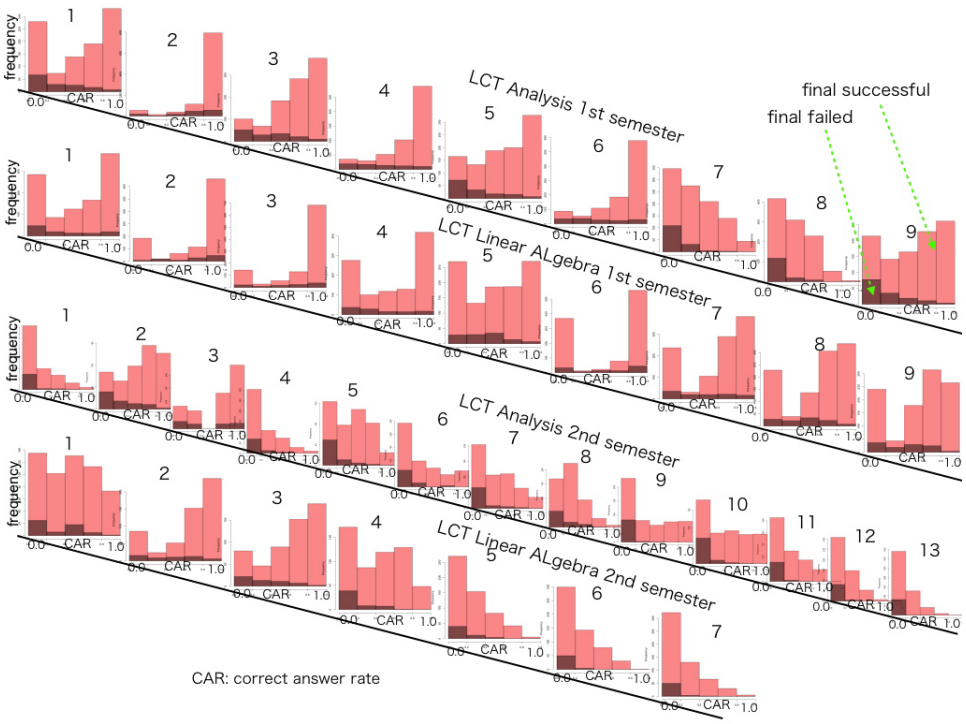

Figure 6: Histogram of the CAR to each question in the LCT for final examination successful and failed students.

3.3 Trend of odds ratio of CAR value of the failed group to that of the successful group in the final examination

To see the time-dependence of the difference between the failed group CAR values and the successful group CAR values, we have provided the odds ratio (*odds ratio = CAR value in the failed group / CAR value in the successful group*) to each question in the LCTs, as shown in Figure 7. We can see that the odds ratios are decreasing in all the four cases.

To comprehend the time-dependence of the CAR values to each LCT, computing the mean values of the CARs in each LCT (i.e., to each lecture) are useful. Figure 8 shows the trends of the mean values of the CARs in the failed group and the successful group. In the figure, we can see a clear difference between the two groups. Considering the large numbers of samples of students in these two groups, say 1000 and 100 in detailed values, there is a

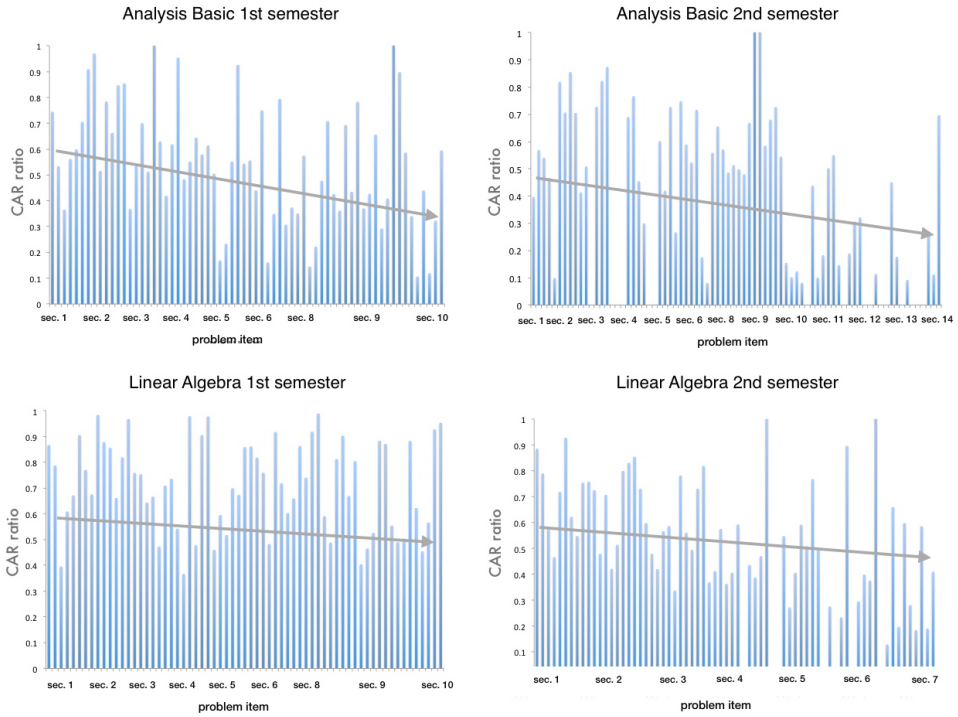

Figure 7: Trend of odds ratio of CAR value of the failed group to that of the successful group.

significant difference between the two groups without having to make a t-test (p-value is extremely small). T values in Welch t-test is approximately $0.5/\sqrt{(0.5^2/100)} = 10$, then p-value is less than $\Phi(10)$ in normal distribution. When the number of samples is larger than 100, a t distribution tends to a normal distribution.

Next, we show the trends of the mean odds ratios of CAR values of the failed group to that of the successful group. Figure 9 indicates intriguing phenomena regarding the trends of odds ratios of mean values of failed cases to those of successful cases in the final examination. First, we can again find the decreasing trends to the odds ratios. This means that the gap of the CARs between the failed students and the successful students becomes large as lectures go forward. The failed students in the final examination must have felt depressing as time goes on.

By applying the linear regression to these trends such that

$$R(t) = \beta_0 + \beta_1 L(t) \ (t = 1, 2, \ldots,)$$

where, $R(t)$ denotes the estimated odds ratio at lecture id = t, and the $L(t)$ is the observed mean odds ratio at lecture t. The tangents (β_1) of decreasing lines for four cases are -0.0277, -0.0273,-0.0062, and -0.0061 for Analysis Basic in the first semester, Analysis Basic in the second semester, Linear Algebra in the first semester, and Linear Algebra in the second semester. The standard deviations using the bootstrap method are 0.0010, 0.0010, 0.0012, and 0.0013, respectively. The 95% confidence intervals are $(-0.0297, -0.0257)$,$(-0.0293, -0.0253)$,$(-0.0086, -0.0038)$, and $(-0.0087, -0.0035)$, respectively. Thus, the assumed tendencies that the odds ratios are decreasing are not rejected with 5% significance level. See Figure 9.

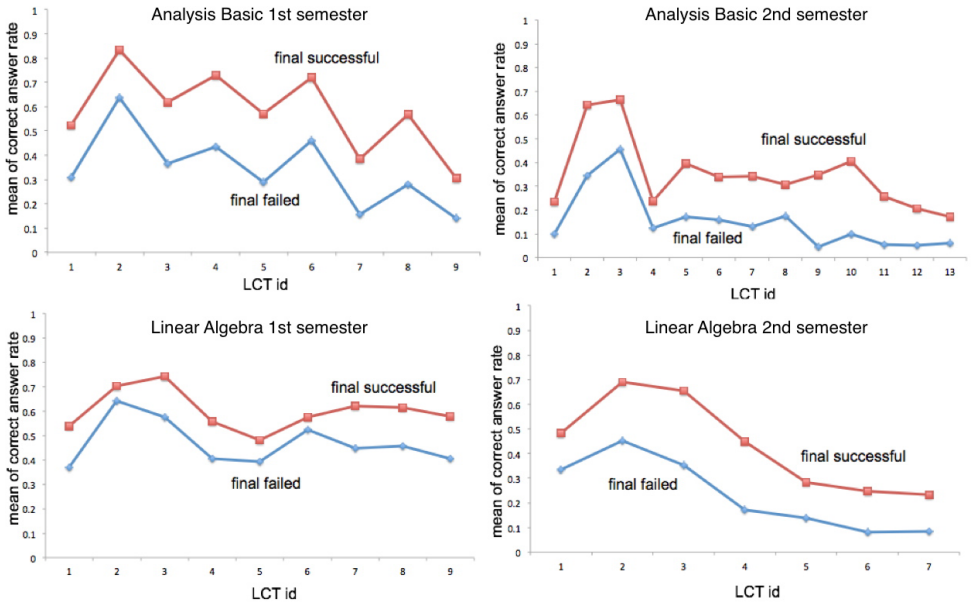

Figure 8: Trends of the mean CAR values of the failed group and those of the successful group.

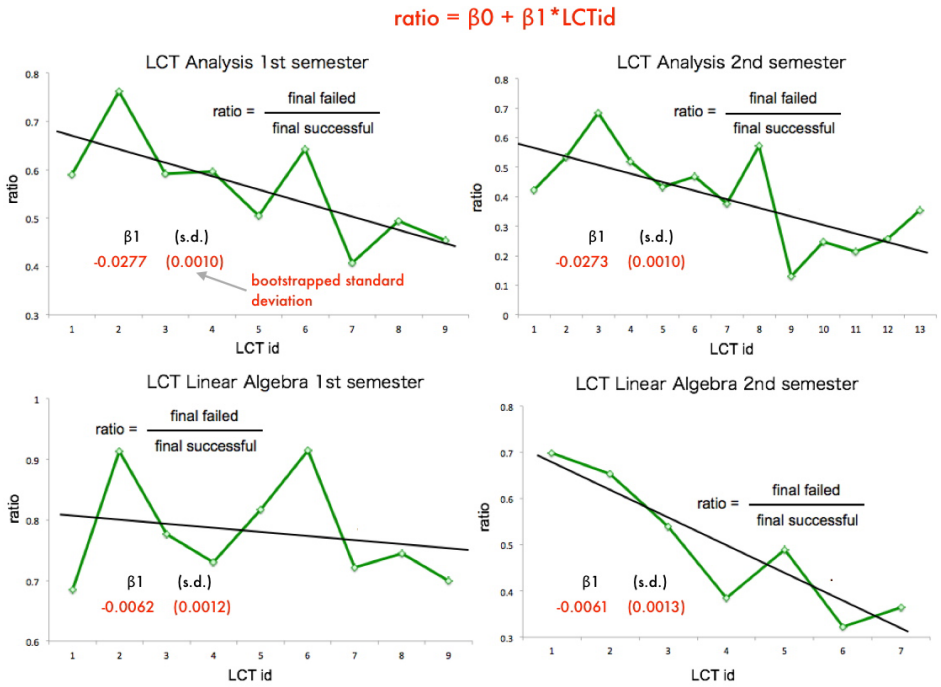

Figure 9: Trends of the mean odds ratios of CAR values of the failed group to that of the successful group.

The declining trends may be affected by the contents of the final examinations and scores. They may be also affected by the threshold scores dividing the successful and failed students. The smaller the number of failed students is, the much clearer the declining trends are seen, because scores of successful students are not so affected by the threshold variations but those of failed students are strongly affected due to the ratios of the relative variation numbers to the total numbers. In our case, about 10% to 15% to all the students failed, which is commonly seen. Thus, as long as a very large number of students will not fail, the declining trend may be commonly seen.

4 Concluding Remarks

To overcome the crucial issue in universities that we identify students at risk for failing courses and/or dropping out early, we have analyzed the detailed analytics for the learning check testings results performed in each lecture. We have found that those who failed in the final examination show the much steeper decreasing trend of correct answer rates in the learning check testing comparing to those who were successful in the final examination. This seems obvious empirically. However, the important point is that the fact is again confirmed statistically by using meticulously accumulated large scale testing results.

Acknowledgments

The author would like to thank mathematical staffs at Hiroshima Institute of Technology. This work was supported by JSPS KAKENHI Grant Number 17H01842.

References

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

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

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

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

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

[15] 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.

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

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

[18] 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.

[19] 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.

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

[21] Fundamental Statistics of Japanese Higher Education, http://www.mext.go.jp/b_menu/toukei/chousa01/kihon/ 1267995.htm, 2016.