International Journal of Learning Technologies and Learning Environments
International Institute of Applied Informatics
2019, Vol. 2, No. 1, 35 – 52

Relationship Between Testing Time and Score in CBT

Hideo Hirose * †

Abstract

By looking at the relationships between the numbers of correct answers and the time durations that students spend in taking tests, we have found that there are typical three patterns. The patterns of the time durations spent in taking the test to each number of correct answers depends on the difficulties of the questions. To easy problems to solve, some smart students can use less time to solve the problems and students with low academic ability need much time to solve. To moderate problems to solve, every student requires the similar time duration to solve the problems. To difficult problems to solve, many students tend to use full time to the pre-specified time duration, but some students with low ability may give up tackling the problem soon.

Keywords: time duration to solve, problem difficulty, online testing, correct answer rate, item response theory.

1 Introduction

There may be a hypothesis: the more competent a student is, the shorter the time duration is in solving a problem; see Figure 1, as an illustrative example. This hypothesis seems to make sense and to be acceptable in general, and this could also be expanded in a situation in taking a test. That is, smarter students will spend less time than common students would spend. This paper aims at investigating such a hypothesis can be commonly assumed or not by using abundant testing data.

Historically, such kind of investigations can be seen in the literature. For example, Bridges [3] describes that no linear or curvilinear relationships between completion speed and performance on individual tests exist. Landrum et al. [16] found that test completion time was sometimes negatively correlated with test performance, but not consistently so. Terranova [20] mentions that although no significant linear relationships were found, significant curvilinear regressions of time on score were found. Other references are also seen in [2], [6] and etc. Although many researchers pursue this theme in many fields, it seems that the consistent relationship between the testing time and the performance have not yet been found, in particular, in the field of mathematics education for undergraduate students. This paper focuses on this point.

* Hiroshima Institute of Technology, Hiroshima, Japan
† This work was supported by JSPS KAKENHI Grant Number 17H01842.

Until recently, to do such an investigation, it would be considered to take time due to the lack of rich accumulated online testing response data. However, as the online testing systems become reality in these decades, we are now able to use such the data in doing learning analytics, as indicated in [3], [4], and [19]. That is, the abundant logging data can be used to investigate the relationship between the testing times and the scores of students performed. As is said (see [23]), we may be able to actively tackle the collected data to find the optimal strategies for better learning methods. It is also important to analyze the data theoretically.

In 2013, in the literature [17], it is suggested that using the logging data such a hypothesis raised at the beginning of this section is not necessarily valid. They said, using a small size of testing result, that there is no deterministic relationships between the time duration in finding solutions by a smart student and that by an ordinary student. Figure 2 shows the ability estimation trends of 67 high school students, where the ability is referred to the item response theory (IRT) with Bayes theorem (regarding the IRT, see [1] and [5], e.g.). On the top of the figure, 33 cases are shown where time durations in tackling five problems are shorter than 400 seconds, and on the bottom, 34 cases are shown where time durations are longer than 400 seconds. Each trend of dots indicates a series of successively estimated abilities using the IRT. We used the adaptive online testing system such that a successive question is selected to be best fitted to the student estimated ability. We can see that there are no typical relationships between the time durations and the estimated final abilities.

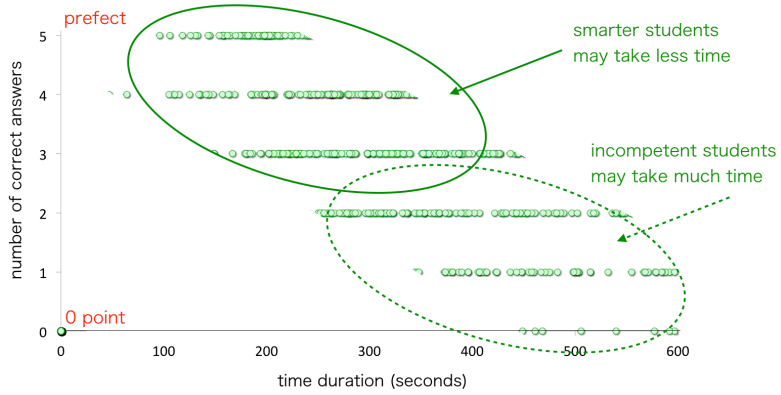

Figure 1: Hypothetical relationships between the time durations spent in taking tests and the numbers of correct answers.

In response to recent learning analytics movements, we have established online testing systems aimed at helping students who desire further learning skills for mathematics education. In such systems, we included the learning check testing, the LCT, for every class to check if students comprehend the contents of lectures or not. The system has been successfully operating (see [7], [8]), and some computational results were reported (see [10]). In addition, other relevant cases were well investigated (see [9], [11], [12], [13], [18], [21]).

In such testing systems, we can measure the time durations that students spend in taking a test. The number of students is more than 1000, and the number of tests is more than ten. Then, we may find much more precise relationships between student skills and the time durations spent in taking tests. Actually, by looking at the relationships

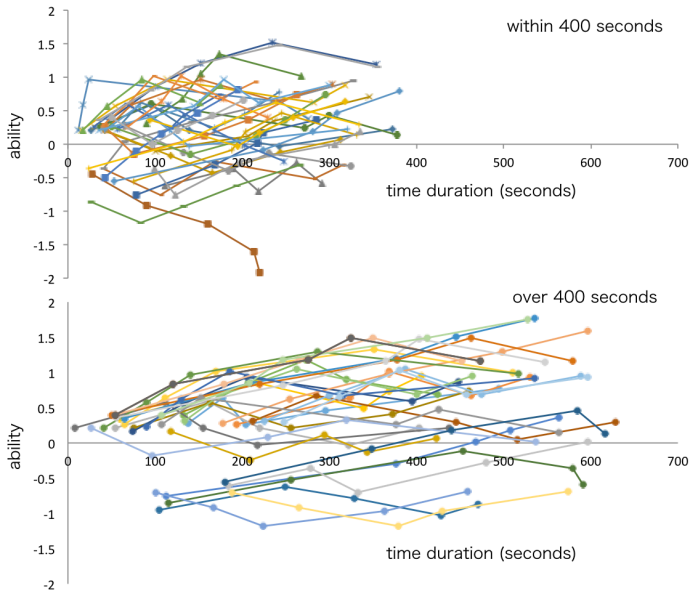

Figure 2: Ability estimation trends of 67 high school students. 33 time durations in tackling five problems cases are shorter than 400 seconds, and 34 time durations are longer than 400 seconds.

between the two terms mentioned above via many testing results, we have found that there may be typical patterns.

2 Online Testing and Its System

The learning check testing, LCT, is a kind of mini test, but the evaluation method for the LCT adopts the IRT partially in which the difficulty values are provided in advance, unlikely to the common IRT method where difficulty values and ability values are unknown simultaneously (see [7], e.g.). The standard IRT uses the two-parameter logistic function $P(\theta_i; a_j, b_j)$,

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

where θ_i expresses the ability for student i, and a_j, b_j are constants in the logistic function for item j called the discrimination parameter and the difficulty parameter, respectively.

The number of questions is five or seven, and the time duration in a test is ten minutes in each LCT. All the students in regular classes take the LCT using their personal computers, that is, the test is taken online. All the questions are the same to each student, but sorted in different order. We made effort that the levels of the questions are designed to be distributed from difficult one to easy one, that is, all levels of difficulties are covered. In such a situa-tion, every level of student learning skill may find the most appropriate level of question in the sequence of the provided questions [14].

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 schedule elapsed, the system computes the students' abilities by using the IRT, and send the scores transformed from the ability values to the portfolio system. The portfolio system presents the scores to students, class teachers, department advisors, department tutors, and remedial class teachers.

The subjects for the LCT are analysis basic (calculus) and linear algebra; in the first semester, analysis basic A (ABA) and linear algebra A (LAA), in the second semester, analysis basic B (ABB) and linear algebra B (LAB) are performed. We have two years online response data since 2016, but we use the latest year's data in this paper, that is the LCT results in 2017.

In this paper, although the ability results using the IRT are obtained, we use the number of correct answers in the test because they are intuitively understood. In addition, the number of correct answers, which corresponds to the test score directly, and the abilities are highly correlated as seen in the literature [15].

3 Relationships Between the Time Durations Spent in Taking Tests and the Numbers of Correct Answers

We have investigated the relationships between the time durations spent in taking tests and the numbers of correct answers. Although the student skills can be measured by the abilities in the IRT evaluation as shown in the literature [17] because the testing system was adaptive, we use the numbers of correct answers to the numbers of provided questions or correct answer rate (CAR) because the testing system was not adaptive. Exactly the same problems are provided to all the students. Adopting the CAR could be intuitively understood to many researchers.

We provided all the patterns of the relationships between the time durations and the CAR for 13 LCT to four cases: 1) analysis basic in the first semester, 2) linear algebra in the first semester, 3) analysis basic in the second semester, and 4) linear algebra in the second semester in the appendix. Here, we mention why we regard the CAR as the difficulty value in the IRT. This is the same as we regard the number of correct answers as the difficulty of the questions. Figure 3 shows the relationship between the CAR and the difficulty value in the IRT for all the questions given in the LCT in the first and second semesters in analysis basic and linear algebra. We see that there is a strong relationship between the CAR and the difficulty value in the IRT. Thus, we can understand that the less the number of correct answers in the test, the more difficult the questions are.

We pick up three typical cases as shown in Figures 4-6 from a variety of patterns presented in the appendix. Figure 4 is the case of ABA02 (the second LCT result of ABA in the first semester), and this case represents the results to easy questions to solve. Each dot shows one student result (horizontal axis means the time duration that the student spent to solve the questions, and vertical axis means the number of correct answers). Figure 5 is the case of ABA11 (the 11th LCT result of ABA in the first semester), and this case represents the results to the moderate questions to solve. Figure 6 is the case of LAB10 (the 10th LCT result of LAB in the second semester), and this case represents the results to the difficult questions to solve.

Although we can see some triangle pattern distributions in Figures 4 and 6 to the results to easy questions to solve and those to difficult questions and some uniform pattern distribution in Figure 5 to the result to the moderate questions to solve, we cannot grasp

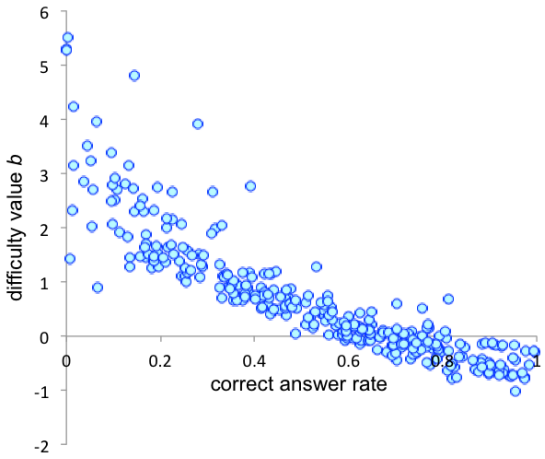

Figure 3: Relationship between the CAR and the difficulty value in the IRT (for all the questions given in the LCT in the first and second semesters in analysis basic and linear algebra).

what happened in these cases. Then, we provided other figures to see the properties from an another aspect.

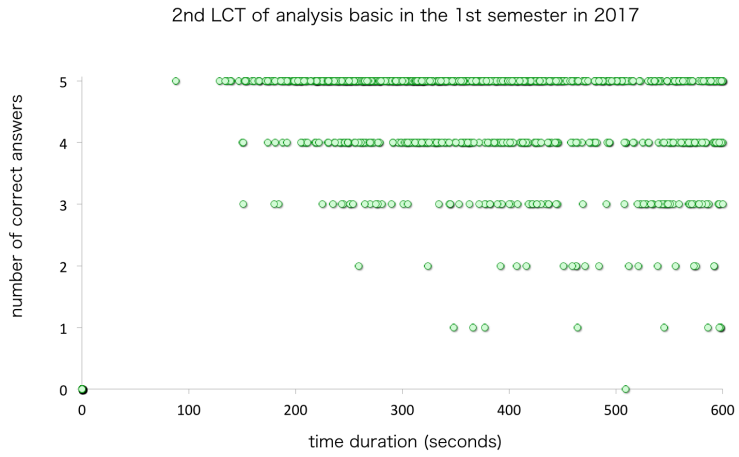

Figure 4: Relationships between the time durations spent in taking tests and the numbers of correct answers (the case of ABA02 (the second LCT result of ABA)).

Figure 7 shows the boxplots of the CAR to each LCT result in addition to the tables of the distributions of the numbers of students who answered the questions correctly in the case of ABA. On the left of the figure, the second LCT result of ABA is indicated, and on the right, the 11th LCT result of ABA is indicated. These are corresponding to Figures 4 and 5. This can be interpreted that the number of dots in the same number of correct answers are summed up and expressed in the table in Figure 7. In Figure 7, in addition to the table, corresponding boxplots to each LCT result are provided. Similarly, Figure 8 shows those corresponding to Figure 6.

To take a look at the difficulties of all the questions to each LCT corresponding Figures 4 and 5, we provided Figure 9 which shows the bar chart for the correct answer rate (CAR) to each LCT result in the case of ABA in the first semester.

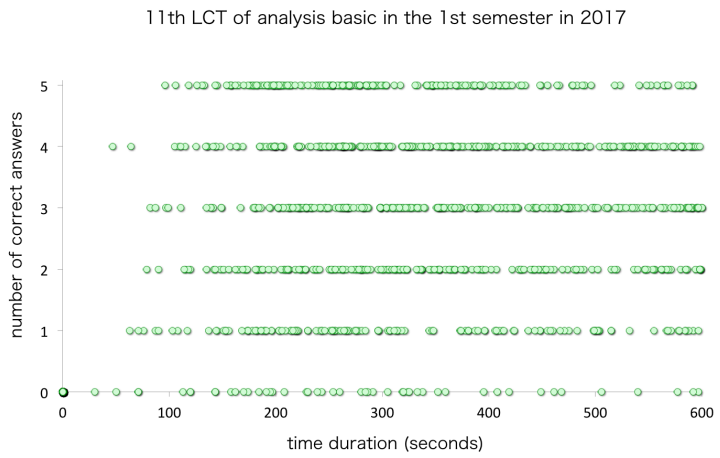

Figure 5: Relationships between the time durations spent in taking tests and the numbers of correct answers (the case of ABA11 (the 11th LCT result of ABA)).

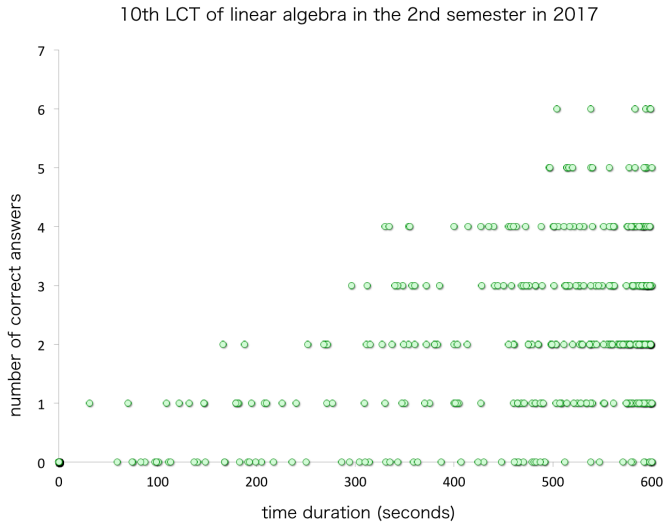

Figure 6: Relationships between the time durations spent in taking tests and the numbers of correct answers (the case of LAB10 (the 10th LCT result of LAB)).

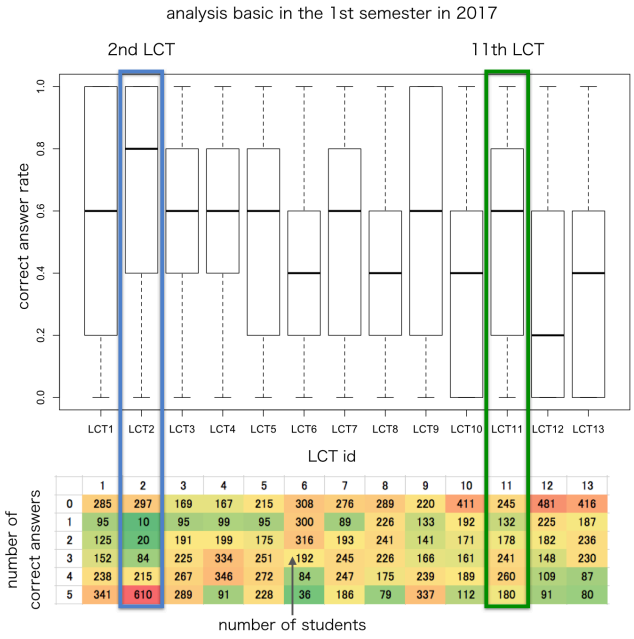

Figure 7: The boxplots of the CAR to each LCT result in addition to the tables of the distributions of the numbers of students who answered the questions correctly in the case of ABA. On the left, the second LCT result of ABA is indicated, and on the right, the 11th LCT result of ABA is indicated.

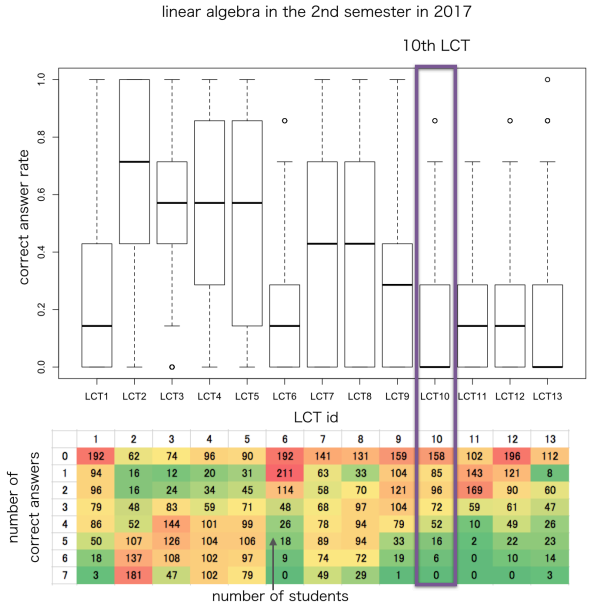

Figure 8: The boxplots to each LCT result in addition to the distributions of the numbers of students who answered the questions correctly in the case of LAB. On the right, the 10th LCT result of LAB is indicated.

Similarly, we provided Figure 10 corresponding Figure 6 in the case of LAB in the second semester. We see that ABA02 case is the easiest case in ABA LCT, but ABA11 shows similar CAR values to ABA10, ABA12, and ABA13. We see that LAB10 case is the difficult case in LAB LCT, but LAB06, LAB10, and LAB12 shows the similar CAR values to LAB10.

Therefore, we have not yet understood the properties for Figures 4-6. That is, we have not yet found the relationships between student skills and the time durations spent in taking tests precisely.

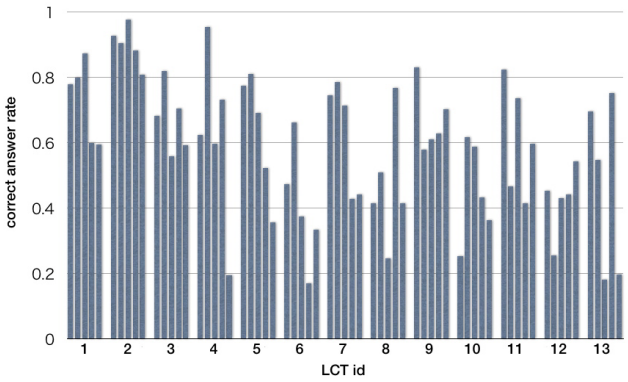

Figure 9: The correct answer rate (CAR) to each LCT result in the case of ABA in the first semester.

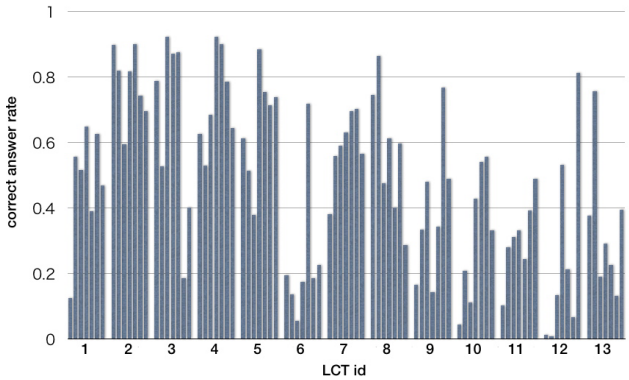

Figure 10: The correct answer rate (CAR) to each LCT result in the case of LAB in the second semester.

4 Histograms to the Time Durations Spent in Taking the Test to Each Number of Correct Answers

Since we cannot find the distributions of time durations spent in taking the test precisely with only Figures 4-6, we have provided the histograms to each number of correct answers. Figures 11-13 are such figures corresponding to Figures 4-6.

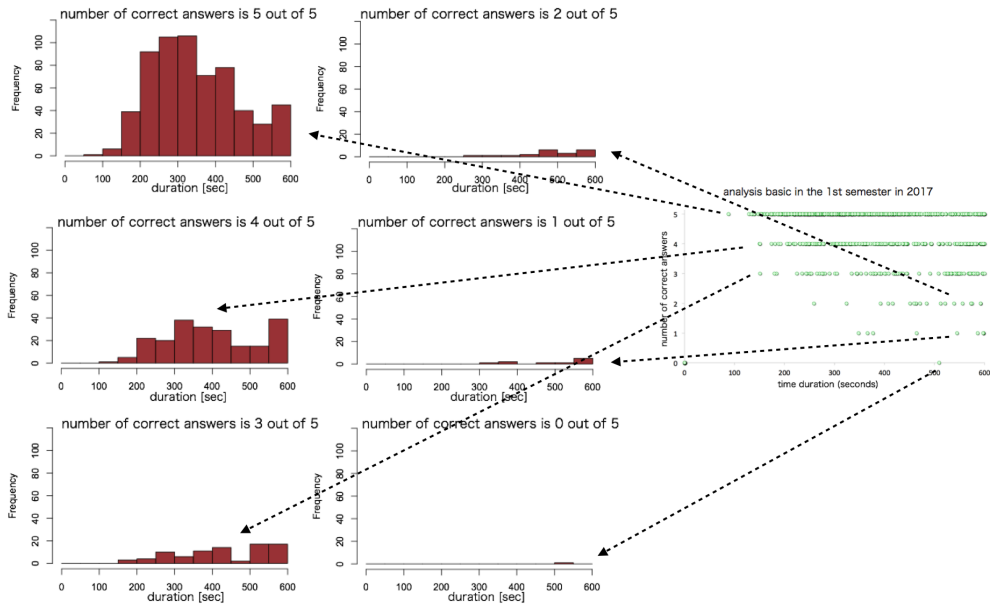

Figure 11: Histograms to the time durations spent in taking the test to each number of correct answers in the case of ABA02.

Looking at Figure 11, we see that mean values of the time durations spent in taking the test to each number of correct answers are shifting from smaller values to larger values, which suggests that the correlation coefficient value between the time duration and the number of correct answers is negative; the correlation coefficient is −0.301. This is the same as the intuition mentioned above in the hypothesis: the more competent a student is, the shorter the time duration is in solving a problem.

Looking at Figure 12, we see that mean values of the time durations spent in taking the test to each number of correct answers are located around the center of the time duration to each number of correct answers, which suggests that the correlation coefficient value between the time duration and the number of correct answers is close to zero; the correlation coefficient is 0.046.

Looking at Figure 13, we see that the distribution of the time durations spent in taking the test to each number of correct answers are shifting from larger values to smaller values, which suggests that the correlation coefficient value between the time duration and the number of correct answers is positive; the correlation coefficient is 0.405. This value is intriguing.

It is noteworthy that in the cases of the number of correct answers is zero and one, the distributions are flat (i.e., the uniform distribution) typically in the case of former case. This suggests that some students with low academic ability give up continuing taking the test because the problems are too difficult to them. On the other hand, many students with high academic ability try to continue to solve the problems by the pre-specified testing time limit.

Therefore, the patterns of the time durations spent in taking the test to each number of correct answers depends on the difficulties of the questions. To easy problems to solve, some smart students can use less time to solve the problems and students with low academic ability need much time to solve. To moderate problems to solve, every student requires the similar time duration to solve the problems.

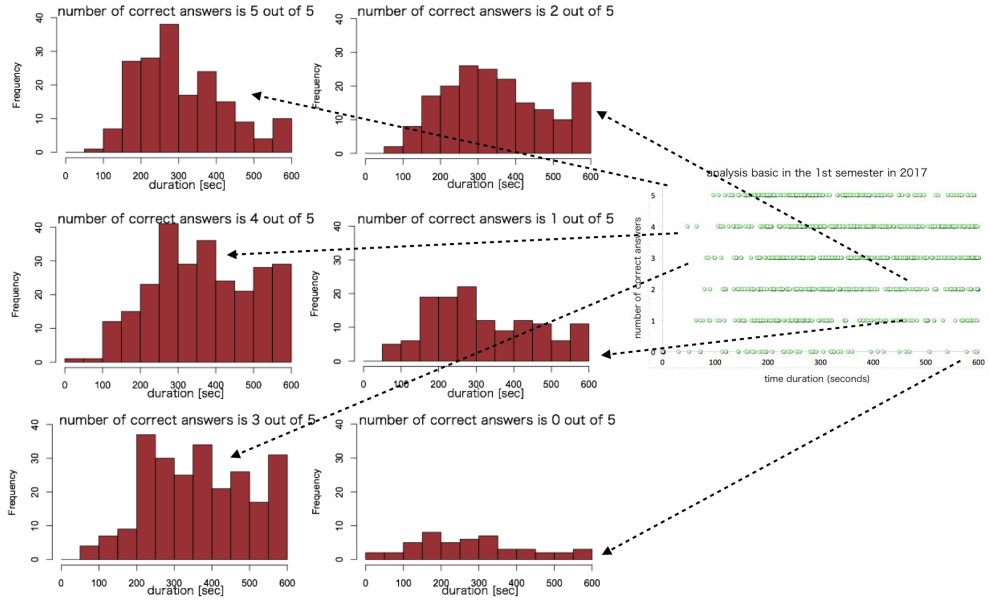

Figure 12: Histograms to the time durations spent in taking the test to each number of correct answers in the case of ABA11.

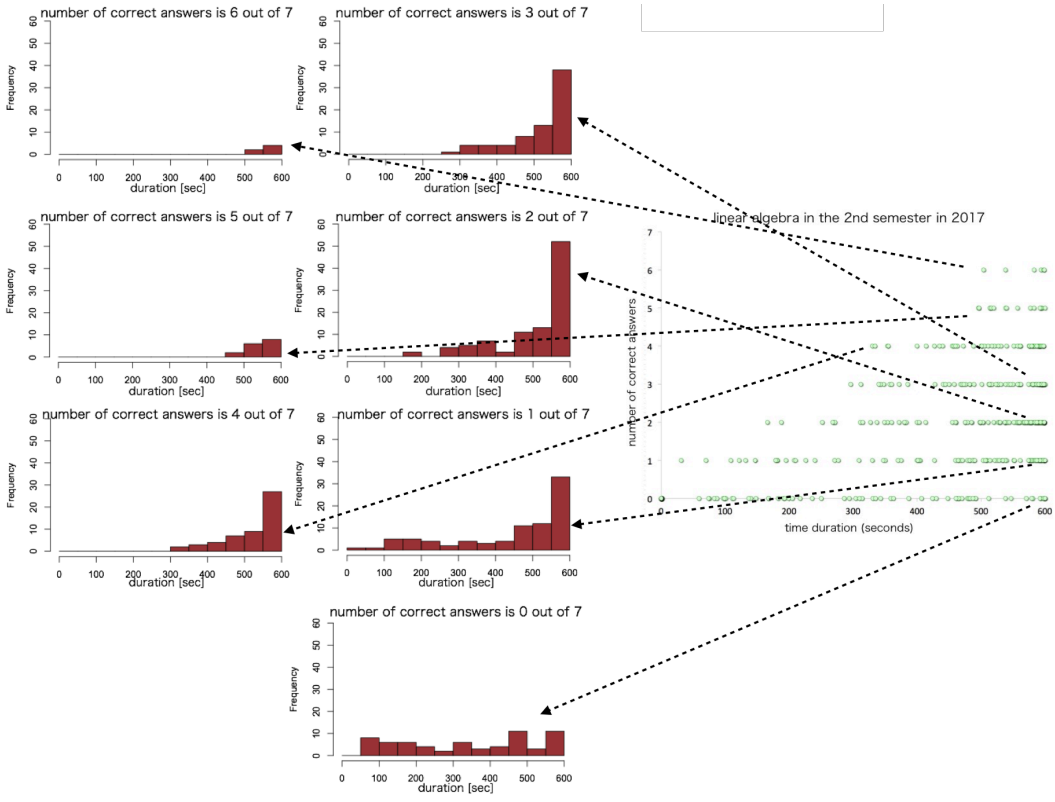

Figure 13: Histograms to the time durations spent in taking the test to each number of correct answers in the case of LAB10.

To difficult problems to solve, many students tend to use full time to the pre-specified time duration, but some students with low ability are reluctant to tackle the problem. They may give up solving the problem soon. Figure 14 shows the typical three patterns of the relationships between the time durations spent in taking the test and the number of correct answers.

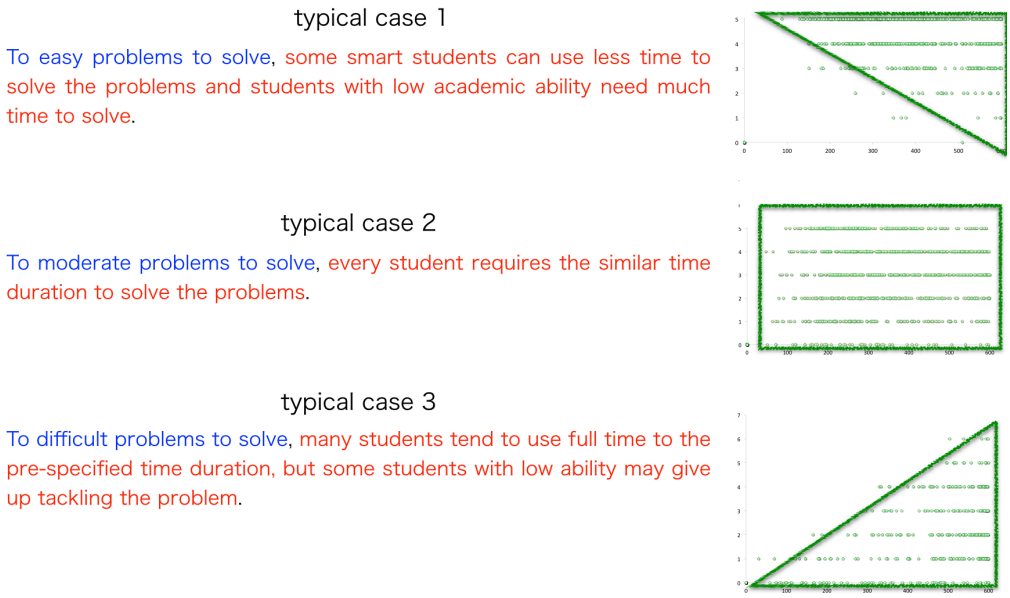

Figure 14: Typical three patterns of the relationships between the time durations spent in taking the test and the number of correct answers.

5 Discussions

Looking at Figure 11 where questions are easy to solve, we see that many smarter students use short time with exceptions of much time users, while incompetent students use much time. This pattern somewhat resembles the pattern in Figure 1. Thus, we may say that the hypothesis that the more competent a student is, the shorter the time duration is in solving a problem is correct in a sense.

Watching the number of correct answers in Figure 12 to each histogram, the mean values to tackle the problem are moving from left to right as the number of correct answers becomes smaller. This phenomenon is similar to the hypothesis above mentioned. However, we find a flat pattern of time durations to the incompetent students. This tendency is different from the hypothesis, and this could be explained as the following.

As explained, the problems provided in Figure 13 were difficult even for smarter students. We can assume that many students use much time to solve the problems. However, there are some students who are reluctant to tackle the problems. They seem to give up success in passing the LCT because they did not solve any items. Very short time in solving the problem indicates the phenomenon.

Therefore, the patterns of the time durations spent in taking the test to each number of correct answers depends on the difficulties of the questions. To easy problems to solve,

smart students use less time to solve the problems and students with low academic ability need much time to solve. To moderate problems to solve, every student requires the similar time duration to solve the problems. To difficult problems to solve, many students tend to use full time to the pre-specified time duration, but some students with low ability are reluctant to tackle the problem. They may give up tackling the problem soon.

Figures 15 shows the similar histograms to Figures 11-13 for 13th testing result to LAA where examinees were successful to the final examination. Figures 16 shows the similar histograms to Figure 15 where examinees failed to the final examination. Figure 16 indicates the examinees' behaviors in taking the examination; that is, they are inclined to give up success soon.

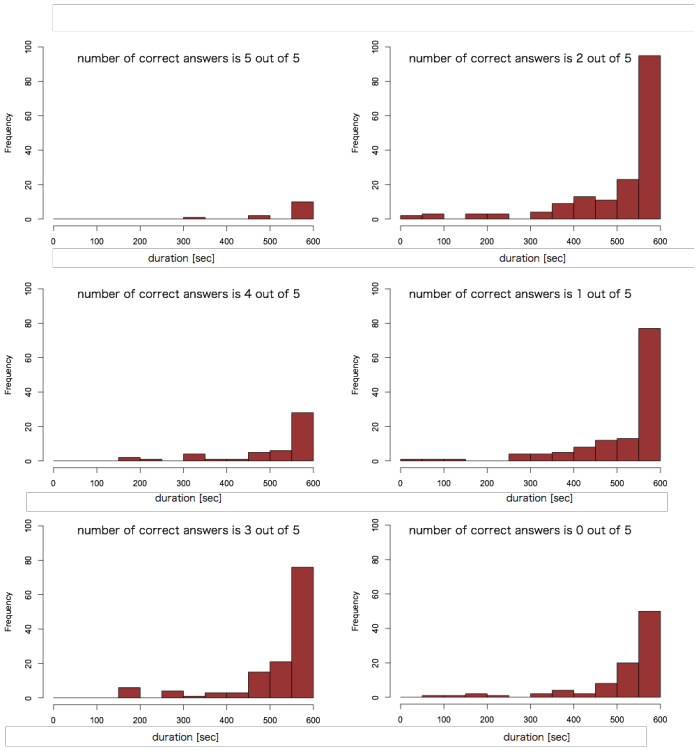

Figure 15: Histograms to the time durations spent in taking the test to each number of correct answers in the case of LAA13. Examinees were successful to the final examination.

We would like to mention suggestions/implications or limitations from this research findings briefly, here. When automatic logging systems were not available, the testing time may be measured by the start time and the end time of the test. It would be rather inaccurate. However, we can now memorize the log-in time and log-out time to each question, resulting in the much more accurate time data accumulation with big data. Actually, although we could not see the clear relationship between the ability in the IRT and the tackling time to questions using smaller data sets previously, we are able to figure out such relationships more accurately in detail. Therefore, we may expect further discovery in the future if a large scale of data accumulation. These would not be simple statistical interpretations but rather inclusion of psychological human behavior interaction via the testing, although many interpretations may still appear as we described the historical way of the research in this field.

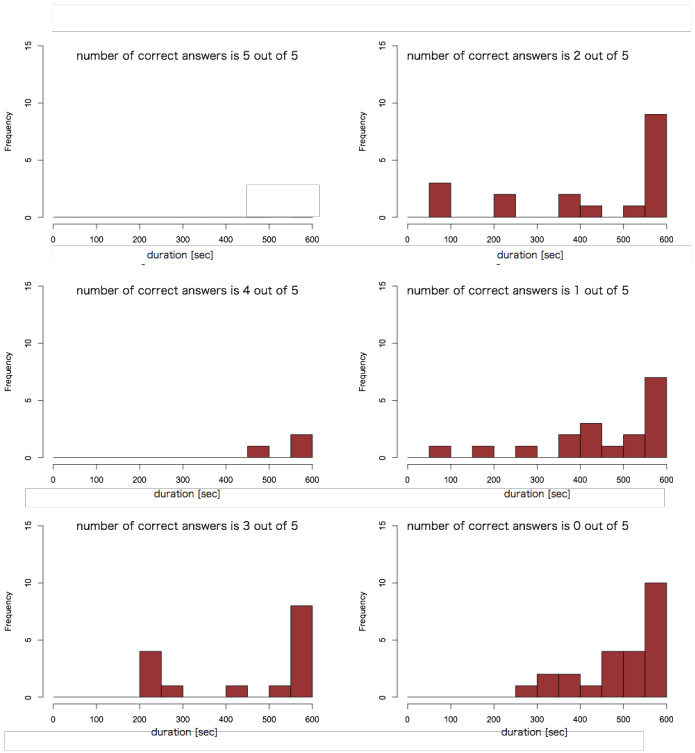

Figure 16: Histograms to the time durations spent in taking the test to each number of correct answers in the case of LAA13. Examinees failed to the final examination.

6 Concluding Remarks

By using the online testing systems, we can measure the time durations that students spend in taking tests, and thus we may find precise relationships between the student skills and the time durations spent in taking tests. By looking at the relationships between the numbers of correct answers and the time durations that students spend in taking tests, we have found that there are typical three patterns.

The patterns of the time durations spent in taking the test to each number of correct answers depends on the difficulties of the questions. To easy problems to solve, smart students use less time to solve the problems and students with low academic ability need much time to solve. To moderate problems to solve, every student requires the similar time duration to solve the problems. To difficult problems to solve, many students tend to use full time to the pre-specified time duration, but some students with low ability are reluctant to tackle the problem.

Acknowledgment

The author would like to thank mathematical staffs at Hiroshima Institute of Technology.

Appendix: Relationships between the time durations spent in taking tests and the numbers of correct answers

We show the relationships between the time durations spent in taking tests and the numbers of correct answers in the cases of analysis basic in the first semester, linear algebra in the first semester, analysis basic in the second semester, and linear algebra in the second semester in all the LCT.

References

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

[2] R.P. Beaulieu and B. Frost, Another look at the time-score relationship, Perceptual and Motor Skills, 78, 1994, pp.40-42.

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

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

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

[6] W.E. Herman, The relationship between time to completion and achievement on multiple choice exams, Journal of Research and Development in Education, 30, 1997, pp.113-117.

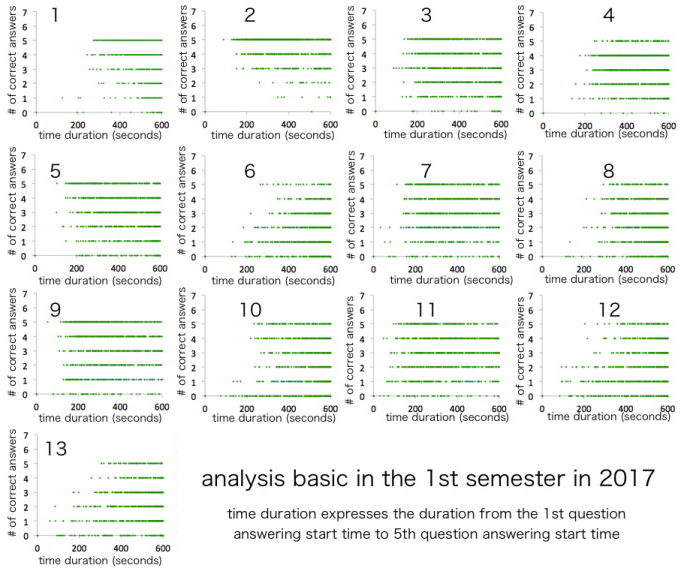

Figure 17: The relationships between the time durations spent in taking tests and the numbers of correct answers in the cases of analysis basic in the first semester.

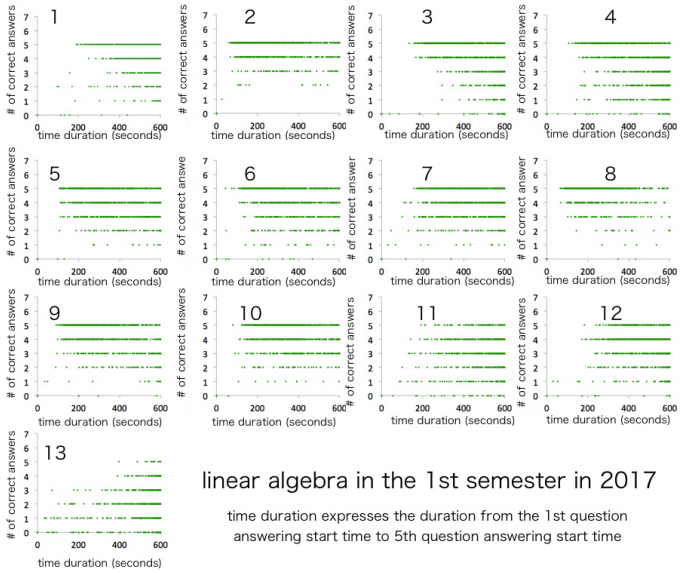

Figure 18: The relationships between the time durations spent in taking tests and the numbers of correct answers in the cases of linear algebra in the first semester.

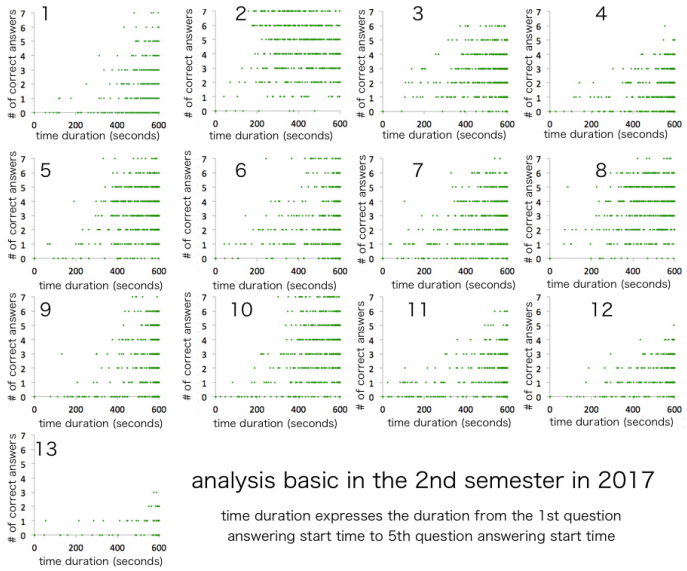

Figure 19: The relationships between the time durations spent in taking tests and the numbers of correct answers in the cases of analysis basic in the second semester.

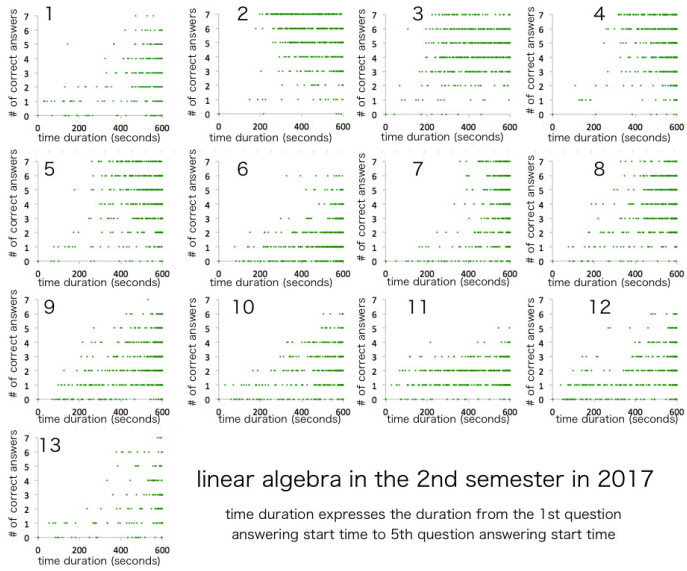

Figure 20: The relationships between the time durations spent in taking tests and the numbers of correct answers in the cases of linear algebra in the second semester.

[7] 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.

[8] 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 Con-ference on Learning Technologies and Learning Environments, 2016, pp.433-438.

[9] H. Hirose, aLearning 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.

[10] 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.

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

[12] H. Hirose, Difference Between Successful and Failed Students Learned from Analytics of Weekly Learning Check Testing, Information Engineering Express, Vol 4, No 1, 2018, pp.11-21.

[13] H. Hirose, A Large Scale Testing System for Learning Assistance and Its Learning Analytics, Proceedings of the Institute of Statistical Mathematics, Vol.66, No.1, 2018, pp.79-96.

[14] 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.

[15] H. Hirose, T. Sakumura, T. Kuwahata, Score allotment optimization method with application to comparison of ability evaluation in testing between classical test theory and item response theory, Information, 17, 2014, pp.391-410.

[16] R.E. Landrum, H. Carlson, W. Manwaring, The relationship between time to complete a test and test performance, Psychology Learning and Teaching, 8, 2009, pp.53-56.

[17] K. Noguchi, H. Hirose, A relationship between the adaptive online IRT evaluation and the response time, National Conference of IEEJ 2013, 11-2P-07, 2013.

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

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

[20] C. Terranova, Relationship between Test Scores and Test Time, The Journal of Experimental Education, 40, 2015, pp.81-83.

[21] 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.

[22] 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.

[23] 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.