International Journal of Learning Technologies and Environments
International Institute of Applied Informatics
2021, Vol. 4, No. 1, 1 – 20

Accurate Student Ability Estimation by Removing Teacher Evaluation Bias via Full Computer Based Testing

Hideo Hirose [*] [†]

Abstract

A test score does not represent the exact ability of an examinee. It only shows just one aspect of the examinee, even if the coverage of the test is restricted. Due to this, for example, we may not see obvious relationships between entrance examination scores and academic records in universities, even in mathematics subjects. Thus, in order to make clear such a relationship in a statistical sense, we have investigated three testing records of the placement test, the learning check test, and term examinations.

Then, we have shown mainly three consequences from the investigation: 1) by using the full computer based testing results of the placement test, we have become aware of the magnitude of irreducible probabilistic fluctuations; 2) in using the description type testing, it would be inevitable to accept biased evaluations by teachers; 3) by adopting full computer based testing in the placement test, the learning check test, and term examinations, we can remove the teacher's evaluation bias occurred in the description type testing, and can obtain the more accurate student's ability.

In addition, we have proposed a fundamental ability equation on student's ability including irreducible probabilistic fluctuations.

Keywords: ability estimation, item response theory, computer based testing, irreducible probabilistic fluctuation, evaluation bias, description type testing, multiple choice type testing, academic growth, ability equation, learning analytics.

1 Introduction

Many people often regard examinees' apparent performances in testings as their true potential abilities. However, a test score does not represent the exact true ability of an examinee. It only shows just one aspect of the examinee, even if the coverage of the test is restricted. It may be a result of probabilistically fluctuated outcome due to the examinee's condition, selected problems, teacher's evaluations, and etc.

Here is an another aspect for the fluctuation phenomenon. It is well-known that we cannot see obvious relationships between entrance examination scores and academic records in universities in mathematics subjects (see [10, 42], e.g.). There may be many reasons for

[*] Kurume University, Fukuoka, Japan

[†] This work was supported by JSPS KAKENHI Grant Number 17H01842.

this. One is that examinees are classified into a successful group and a failed group by the entrance examination and that academic records in the successful group will become much more similar to each other; this is so-called the regression fallacy (see [38, 44, 48]). The second is that the contents taught in universities become much more difficult to understand than those in high schools, resulting in larger differences of scores in universities. The third is that assessment results in universities are versatile and academic records deeply depend on evaluation ways by teachers. That is, teachers' evaluations to students' abilities may be biased as Meissel et al. [34] mentions such as the psychological belief and the narrow focus of standardized tests. Regarding such biases, they suggest that there may be possible causes that we have not yet considered, and future research should investigate alternative explanations for these results to develop a better understanding.

If we can remove some of such fluctuation factors, we may obtain students' true abilities more accurately. In such a sense, the purpose of the paper is to reveal such factors and to eliminate them. To accomplish these, firstly, we pay attention to the academic score itself. We show that the score is fluctuated with some probability. Next, after we have seen such fluctuations, we will focus on evaluation bias elimination due to teachers' evaluation methods and attitudes. To do this, we have changed the testing style from description type examinations to multiple choice type examinations in the end-term examinations. (see Appendix). Thirdly, we try to see academic growths of students by education ways of teachers. From these investigations, we have found new interesting insights. They are the following: 1) by using the full computer based testing results of the placement test, we have become aware of the magnitude of irreducible probabilistic fluctuations; 2) in using the description type testing, it would be inevitable to accept biased evaluations by teachers; 3) by adopting full computer based testing in the placement test, the learning check test, and term examinations, we can remove the teacher's evaluation bias occurred in the description type testing, and can obtain the more accurate student's ability. In addition, under such situations, we have proposed a fundamental ability equation on student's ability including irreducible probabilistic fluctuations [27].

Here, we remark that the term "teacher evaluation" in this paper does not mean so-called "student evaluations of teaching (SETs)" in a narrow sense. SETs are used for evaluating a teacher's teaching effectiveness, often by using questionnaires asking students to rate their perception of the course teacher. Regarding such discussions, there are many references (see, e.g., [1, 46, 47]), while other references are also seen ([3, 4, 6, 9, 11, 32, 33, 35–37]).

2 Irreducible Probabilistic Fluctuation

We take into account of various type of testing results. They are placement tests took just after the enrollment, the learning check testing (LCT) for every class to check if students comprehend the contents of lectures or not (see [17–20, 22–25, 29, 30, 41, 45]), the midterm examination results, and the final (end-term) examination results. By comparing results among these testings, we have tried to extract irreducible probabilistic fluctuations.

2.1 Placement Test

First, in order to grasp the magnitude of irreducible probabilistic fluctuations, we deal with the placement test analysis. All the questions in the test consist of multiple choice type questions. Figure 1 shows the histogram of the placement test results applied to 1130 freshman students in 2017. The horizontal axis expresses the correct answer rates (CAR). A

CAR score for student i is computed from $\frac{1}{T}\sum_j \delta_{i,j}$, where $\delta_{i,j}$ denotes the indicator function ($\delta_{i,j} = 1$, if the student answered the question correctly, and $\delta_{i,j} = 0$, incorrectly). T is the total number of questions, and here, it was 77. We assume that all of the questions are independent from each other. The mean value for all the CAR is 0.624.

If we assume that a student has the same ability to solve each question, the probability distribution of CAR taken from the even number of questions and that from the odd number of questions may be similar. Figure 2 shows the scatter plot for this, superimposed two histograms corresponding to these two CAR. The mean value to even number of questions is 0.634, and 0.615 to the odd number of questions. The coefficient of correlation is 0.948. Looking at Figure 2, we could assume such a hypothesis to the student's ability.

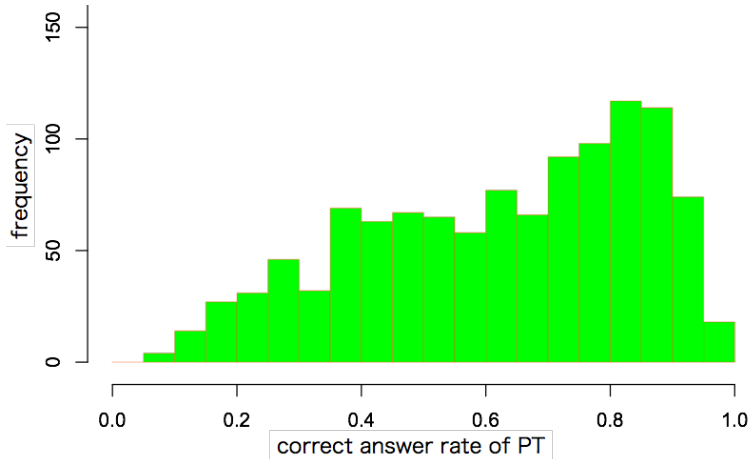

Figure 1: Histogram of the correct answer rates (CAR) of the placement test results to 1130 freshman students in 2017. The total number of questions was 77.

Figure 2 reveals us that the ability of a student does not vary to a large extent and that the magnitude of the probabilistic fluctuation to the corresponding ability can be measured to some extent. That is,

$$\phi_i = \mu_i + \varepsilon_i, \tag{1}$$

where, ϕ_i, μ_i, and ε_i are student's observed ability, student's true ability, and probabilistic fluctuation, respectively. This is a basic idea for the ability equation.

In this situation, since the number of questions using even id question numbers and using odd id question numbers are both almost half of the total questions, the magnitude of the fluctuations in Figure 2 is larger than the true magnitude of the fluctuation. Actually, they are $\sqrt{2}$ times larger of the true value. This is because we assumed that all the questions are independent from each other and a student has the same ability of μ_i to solve each question; the standard deviation ε_i to student i is approximately computed from $\sqrt{\frac{2}{T}\mu_i(1-\mu_i)}$, assuming a binomial distribution. When $\mu_i = 0.5$, then ε_i is 0.08, and the approximate 95% confidence interval for ϕ_i becomes $[0.34, 0.66]$. This result is consistent to Figure 2. Such a probabilistic fluctuation is considered to be irreducible.

To find the magnitude of the probabilistic fluctuations, we have divided the observed CAR result into two groups. This treatment is considered to be the two-fold cross-validation method. The bootstrap method is another method to find such a fluctuation shown below.

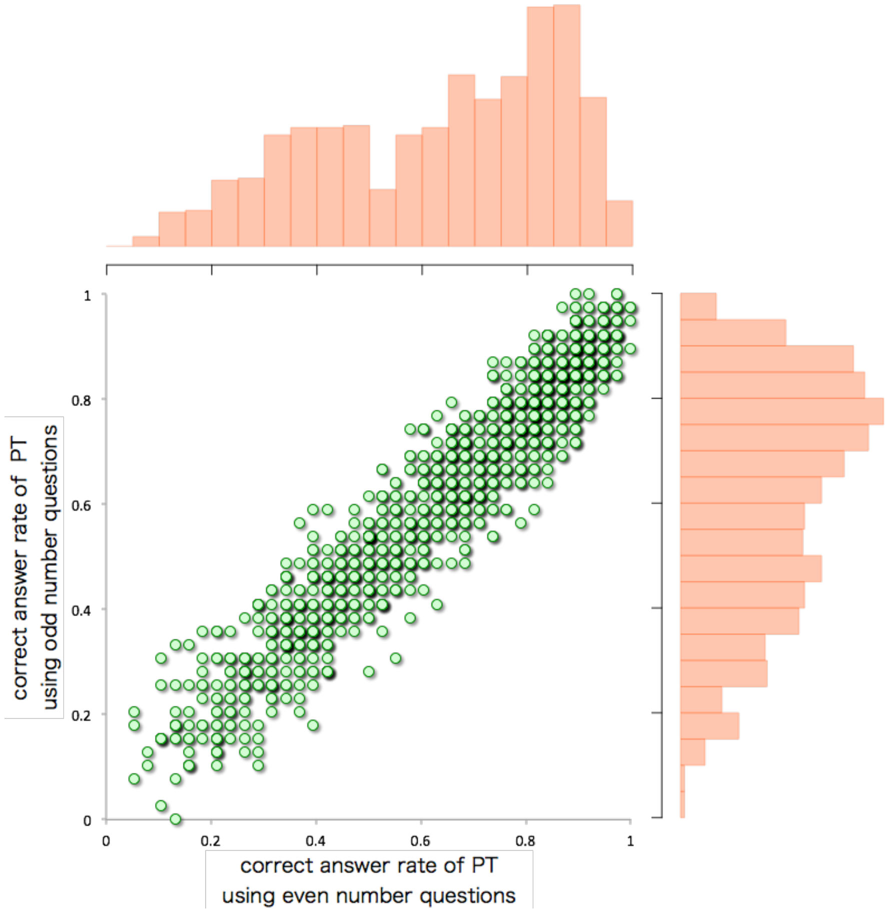

Figure 2: Scatter plot of the CAR taken from the even id question numbers and that from the odd id question numbers from the same data as in Figure 1 with two histograms corresponding to these two CAR. The number of questions having even id question number is 33, and the number of questions having odd id question number is 34. A point with longer shadow indicates that the frequency at the point is larger.

2.2 Midterm and End-term Examinations

Figure 3 shows a part of the response matrix from the midterm and end-term examination results performed in the first semester in 2019 to analysis basic subject. The number of questions to the midterm examination is 31 and the number of questions to the end-term examination is 36, and the total number of questions is 67; they are all consisting of multiple choice type questions.

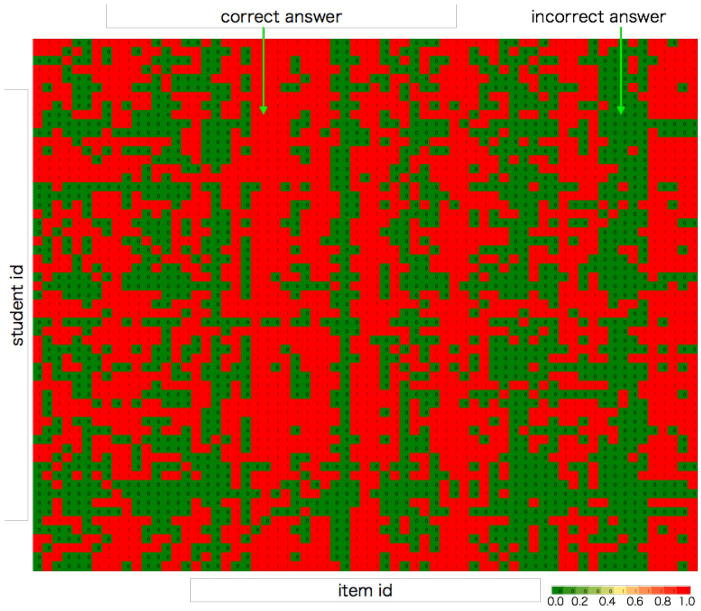

Figure 3: A part of the response matrix from the midterm and end-term examination results performed in the first semester in 2019 to analysis basic subject. Only 55 rows in the response matrix is shown from more than 209 students, and 67 columns in the response matrix represent full results. Red color suggests that the answer was correct, and green color suggests that the answer was incorrect.

From this response matrix, we can estimate students' abilities and difficulty and discrimination parameters of questions altogether using the item response theory (IRT); for general explanations, see [2, 8, 31], and for specific explanations to online testing applications, see [12–16, 40]. We deal with the cases of the standard IRT evaluation using the two-parameter logistic function $P_{i,j}(\theta_i; a_j, b_j)$ shown below.

$$\begin{aligned} P_{i,j}(\theta_i; a_j, b_j) &= \frac{1}{1+\exp\{-1.7a_j(\theta_i - b_j)\}}, \\ &= 1 - Q_{i,j}(\theta_i; a_j, b_j), \end{aligned} \tag{2}$$

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. 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,

$$L = \prod_{i=1}^{N}\prod_{j=1}^{n}\left(P_{i,j}^{\delta_{i,j}} \times Q_{i,j}^{1-\delta_{i,j}}\right), \tag{3}$$

where $\delta_{i,j}$ denotes the indicator function such that $\delta = 1$ for success and $\delta = 0$ for failure in answering a question. When student i miss a question j in the LCT, we regard $\delta_{i,j} = 0$ in that LCT. N is the number of students, and n is the number of questions.

After we obtained the estimated parameters $\hat{\theta}_i$ and $\hat{a}_j, \hat{b}_j$, we can perform the bootstrap simulation to generate a random response matrix. There are many methods to gen erate such a matrix. For example, we could choose N whole column results from column questions randomly. However, this may cause the difficulty in numerical computation due to the matrix singularity (results of two columns could be the same with probability $1/e$). Thus, we have generated $\hat{\delta}_{i,j}$ by using (2). Figure 4 shows one example of the generated $\hat{\delta}_{i,j}$ response matrix by such a bootstrap method. To each element of the matrix, warm colors correspond to high probability to solve question j by student i, and cool colors low probability.

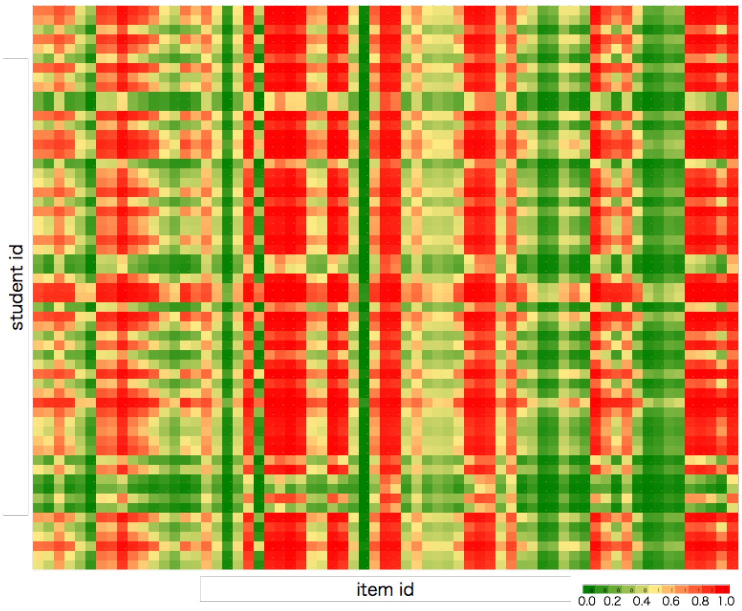

Figure 4: A generated $\hat{\delta}_{i,j}$ response matrix by the bootstrap corresponding to Figure 3. We have generated $\hat{\delta}_{i,j}$ by using (2). Warm colors corresponds to high probability to solve question j by student i, and cool colors lower probability.

After that, we regenerate the response matrix. In regenerating the response matrix, we first generate a uniform random number $p \in [0,1]$ and obtain $\hat{P}_{i,j}(\theta_i; a_j, b_j)$, then we set $\hat{\delta}_{i,j} = 1$ if $\hat{P}_{i,j}(\theta_i; a_j, b_j) \geq p$, and $\hat{\delta}_{i,j} = 0$ otherwise. Figure 5 shows the regenerated $\hat{\delta}_{i,j}$ response matrix. Figure 5 looks similar to Figure 3, but slightly different from each other.

We have performed this procedure for B times and have obtained the B bootstrapped estimates. In this example, we set $B = 100$. We have picked up typical five students' cases from 209 students, and Figure 6 shows these five student cases. In the figure on the right, histograms of abilities estimated by the IRT to each student are superimposed. The mean values and the standard deviations of the 100 bootstrapped estimated $\hat{\theta}_i$ to five students are

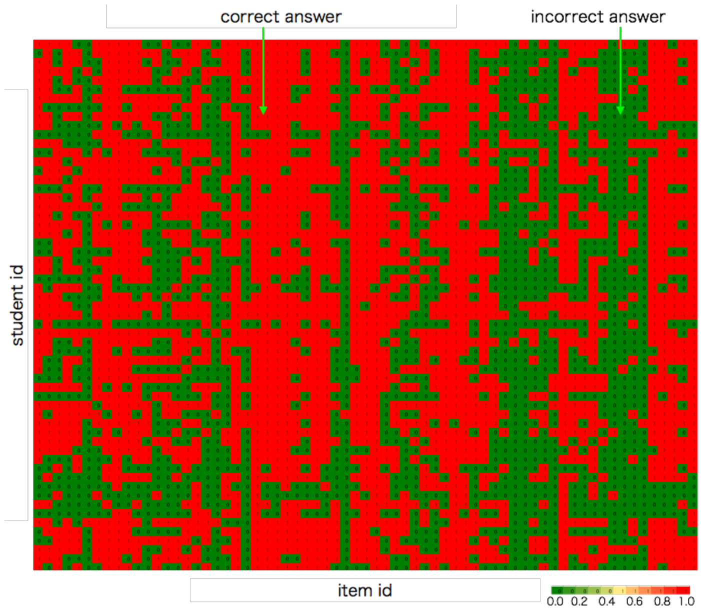

Figure 5: A bootstrapped regenerated response matrix corresponding to Figure 3. We have generated $\hat{\delta}_{i,j}$ by using (2) such that if $\hat{P}_{i,j} \geq 0.5$ then we set $\hat{\delta}_{i,j} = 1$, and $\hat{\delta}_{i,j} = 0$ otherwise. Red color suggests that the answer could be correct, and green color suggests that the answer could be incorrect.

shown in Table 1. These standard deviations are similar to the mean value of 0.224 by the computed standard deviations to each examinee by using the IRT. We can roughly grasp the magnitude of probabilistic fluctuations. We understand again the ambiguity of the academic scores. They are not deterministic, of course.

Table 1: The mean values and the standard deviations of the 100 bootstrapped estimated $\hat{\theta}_i$ to typical five students

student id	mean	standard deviation
1	-1.401	0.218
2	-0.912	0.218
3	-0.223	0.225
4	0.682	0.245
5	1.319	0.261

Figure 7 shows the scatterplots of the abilities among originally computed abilities from observed $\delta_{i,j}$ values and the estimated abilities using the bootstrapped simulation results. The correlation coefficients between the original case and the bootstrap cases are located around 0.95, and that among bootstrap cases are located between 0.91 and 0.93, which are a little bit smaller than 0.95. In both the cases, they are highly correlated with each other.

From these two trials, we can recognize the ambiguity of classifying classes by using a threshold score in a test performed just only once such as the placement test, the end-term examination, and the entrance examination.

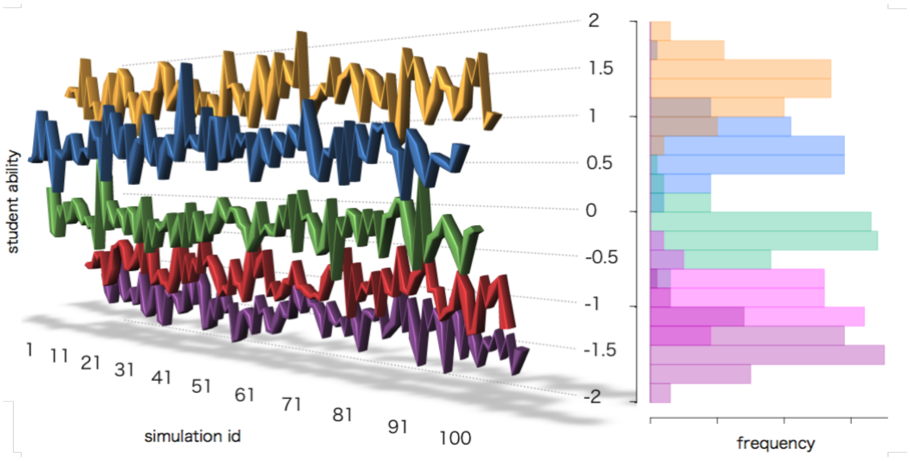

Figure 6: Typical five students' cases for 100 bootstrapped simulation times. We see each student's 100 simulated ability values; each simulated ability is consisting of his/her own true ability with his/her own probabilistic fluctuation. On the right, histograms of 100 simulated abilities can be seen to each student.

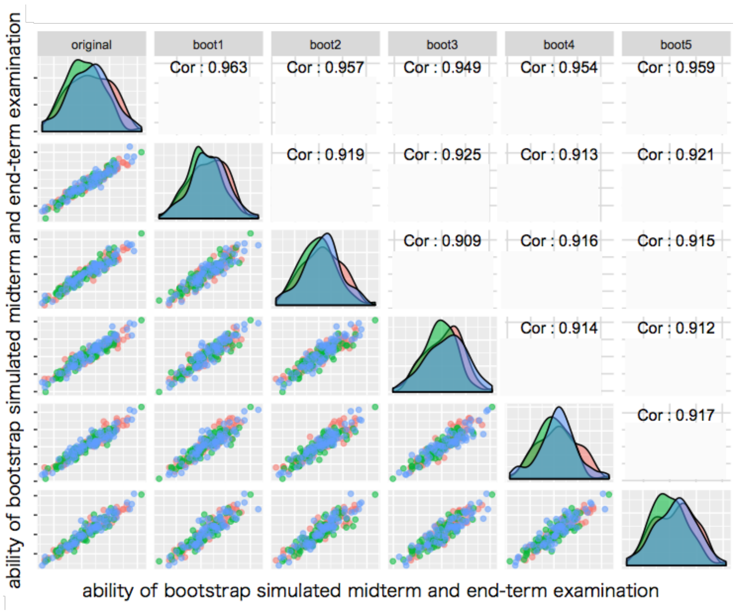

Figure 7: Scatterplots of the estimated abilities among the bootstrapped simulation results and the original testing result.

3 Evaluation Bias in End-term Examination

In many classes, teachers charged themselves with the responsibility of teaching students in their classes. They teach students in their own ways; some teacher focuses on important points, some teacher teaches everything thoroughly; some teacher's problems in testing are very easy; some teacher's problems are tough; some teacher is too generous in evaluation; some teacher evaluates students' academic records very rigorously. However, in some cases, e.g., students are required to be learned in small classes to one department, then the teaching material inevitably must be the same to all the teachers and the problems in the final examination shall be the same. In such a condition, in some cases, students in a department are classified into classes where the placement test score distributions are almost equivalent among classes. Then, we may expect that end-term examination evaluation distributions are also equivalent among classes. However, this expectation becomes to be negative in reality.

3.1 An Example that Evaluation Bias Appears

Figure 8 shows the three testing cases (placement test, LCT, end-term examination) with two teachers. This case is corresponding to the results of linear algebra in the first semester in 2017. At the beginning of the classes, students were split into two classes equally; that is, odd numbered students in descending order scores are classified into class A (teacher a), and even numbered students are into class B (teacher b). On the left and top in the figure, we see that placement test scores are totally the same in these two classes because the distributions are the same. Two classes are considered to be equal to each other. In the middle of the figure, we see the distributions of the LCT ability results, where two classes also show almost the same aspect. This means that students' abilities in two classes seem to be almost the same both at the beginning and at the end of the semester.

However, we find very different patterns of evaluation distributions between the two classes in the end-term examination. One class teacher evaluated "C" to many students, and the other teacher evaluated "A" to most of the students. This could be regarded as the teacher's evaluation bias due to the teacher's way of evaluations because as mentioned above students' latent abilities in two classes appear to be almost the same.

We have investigated whether differences exist or not between two classes by using the hypothesis test of Wilcoxon signed rank test [43]. Table 2 shows that although there is no difference of mean values between two classes in the placement test, we recognize a very clear difference of mean values between the two classes in the end-term examination.

Table 2: Wilcoxon signed rank test for linear algebra in the first semester in 2017

hypothesis	H_0 : a = b
placement test	0.703
LCT	0.048
end-term exam.	**0.00000761**

numbers mean p-values

H_0 : a = b means that mean value in teacher a class equals to that in teacher b class

This may be merely an unusual case showing a clear difference of end-term examina-

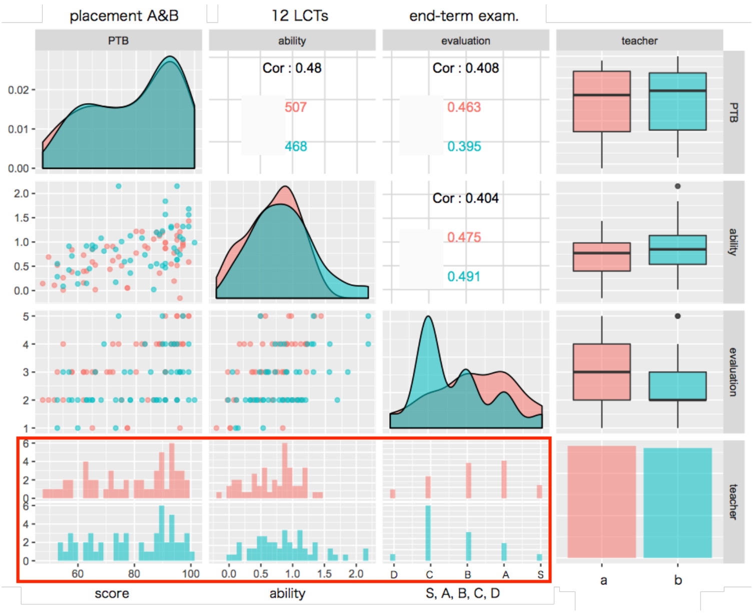

Figure 8: Comparison of records among three testing cases (placement test, LCT, end-term examination) with two teachers performed to linear algebra in the first semester in 2017. At the beginning of the classes, placement test scores are totally the same; the LCT ability results show almost the same aspect; at the end-term examination, we could see the teacher's evaluation biases due to teacher's way of evaluations.

This may be merely an unusual case showing a clear difference of end-term examination evaluations between some two classes. However, such a phenomenon can be seen in common; see discussion section 5.1.

3.2 Removal of Teacher's Evaluation Bias

In the above case, the term examinations were performed in description type testing style. In such a condition, teacher's evaluation bias can be occurred unless teachers set previous arrangements for evaluation because we can assume that the ability distributions in two classes at the end-term examination are equal to each other by the fact that the LCT results were observed to be almost the same in these two classes. Setting this kind of arrangement seems to be difficult in general, and eventually the bias could appear in many cases. This is considered to be unfair in evaluation.

To avoid such an inconvenience, we have changed the testing style of end-term examination from description type to multiple choice type. Computers mark the examination automatically using the IRT, and there is no room for teacher's own manner of evaluation.

Figure 9 shows a similar comparison among three testing cases (extended placement test, LCT, midterm and end-term examinations) performed to analysis basic (calculus) in the first semester in 2019. Similar to the previous case, students in three classes (teachers e, f, g) were split into the three classes equally. The LCT results also show the similar aspect to the previous case. However, the abilities from the evaluation of the combination of midterm and end-term examinations via the IRT seem to be almost equally distributed unlike the previous case. The problems in the examinations are totally the same, and scores are automatically computed without human working.

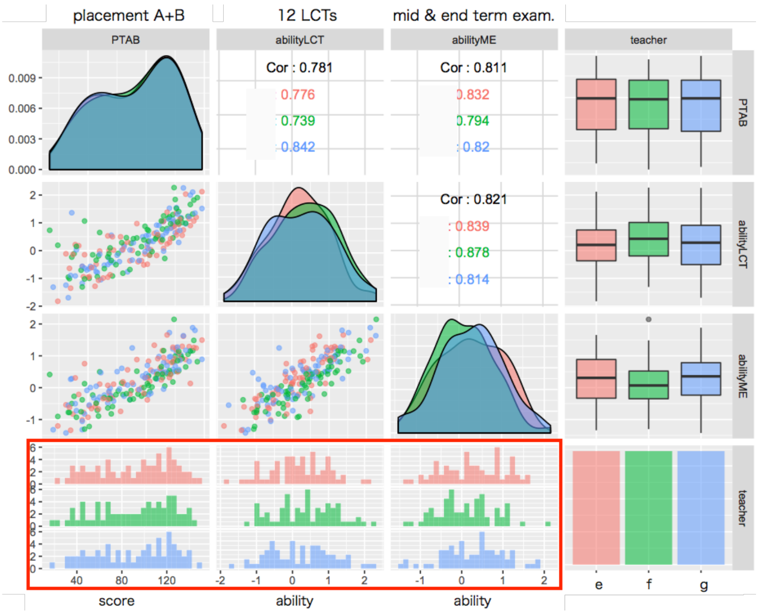

Figure 9: Comparison of records among three testing cases (extended placement test, LCT, midterm and end-term examinations) with three teachers performed to analysis basic (calculus) in the first semester in 2019. Students were split into the three classes equally; the LCT results show the similar aspect; in addition, the abilities from the combined evaluation of midterm and end-term examinations seem to be almost equally distributed.

We have investigated whether differences exist or not among three teachers by using the hypothesis test of Wilcoxon signed rank test again. The hypothesis test cases of H_0 : e = f, H_0 : f = g, and H_0 : g = e are all rejected in the extended placement test case (extended placement means placement test A and placement test B), LCT case, and combined evaluation of midterm and end-term examinations. As a result, there are no differences among all three classes of extended placement test case, LCT case, and combined evaluation case by midterm and end-term examinations; see Table 3. In these three classes, we may think that teaching manners are not so different in the three classes, and thus the fully automated examination style using multiple choice type testing admits no room for existence of teacher's evaluation bias.

Table 3: Wilcoxon signed rank test for analysis basic in the first semester in 2019

hypothesis	$H_0 : \mathrm{e} = \mathrm{f}$	$H_0 : \mathrm{f} = \mathrm{g}$	$H_0 : \mathrm{g} = \mathrm{e}$
placement test	0.91	0.99	0.91
LCT	0.17	0.36	0.71
end-term exam.	0.15	0.12	0.95

numbers mean p-values

H_0 : e = f means that median value in teacher e class equals to that in f

4 Student’s Growth Owing to Teaching Skill

We have found a much fairer evaluation method using the multiple choice type testing than using the description type testing. This enables us to look at further aspects for education. Education works when students are educated; that is, students show the growth by education. There may be a large extent growth or a small extent growth due to teaching skills by teachers.

Figure 10 shows also a similar comparison among three testing cases (placement test, LCT, end-term examination) performed to linear algebra in the second semester in 2019; the placement test and the LCT took in the first semester. Similar to the previous two cases, the classes were split into three equal classes using the placement test results. Obviously, in the figure, there are no differences among these three classes in latent academic skills. The LCT results show a similar aspect to the two cases mentioned before; there seems no differences among the classes.

However, on the contrary to the case mentioned just above, it seems that there could appear the differences among three classes. As indicated in the figure, the median in the class of teacher x is located lower than those in the classes of teachers y and z. To check if the academic records in these three classes are the same or not, we again used Wilcoxon signed rank test. The results are shown in Table 4. The hypothesis that the median in class of x equals to that in y is rejected; z equals to y is also rejected. This suggests that students in classes of teachers y and z are grown much than students in class of teacher x.

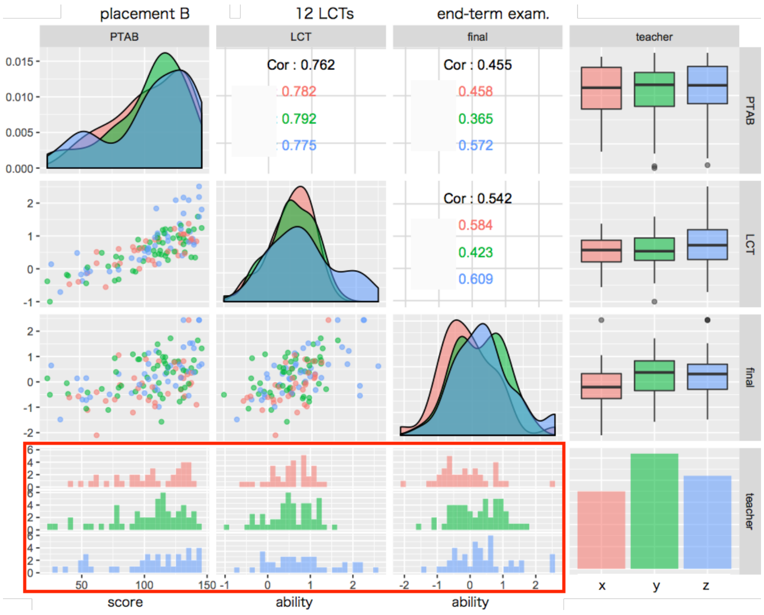

Figure 10: Comparison of records among three testing cases (placement test, LCT, end-term examination) with three teachers performed to linear algebra in the second semester in 2019. Students were split into the three classes equally; the LCT results show the similar aspect; however, the abilities from midterm and end-term examinations show teacher dependent learning growth.

Table 4: Wilcoxon signed rank test for linear algebra in the second semester in 2019

hypothesis	H_0 : x = y	H_0 : y = z	H_0 : z = x
placement test	0.92	0.53	0.63
LCT	0.96	0.15	0.20
end-term exam.	**0.02**	0.73	**0.01**

numbers mean p-values

H_0 : x = y means that median value in teacher x class equals to that in y

5 Discussions

5.1 Existence of Evaluation Biases

As mentioned in 3.1, we can commonly observe cases showing a clear difference of end-term examination evaluations among equally distributed classes at the beginning of the semester. Figure 11 shows such examples. In Figure 11, looking at on the left top two comparison figures of placement score distribution and linear algebra (LAA) end-term examination grade in classes A&B, we see that placement scores in two classes colored by pink and blue are equally distributed and that LAA end-term examination grade seems to be a little bit different from each other. However, we can find that there is no difference between the two classes by using Wilcoxon signed rank test. Then, there are no marks of difference in red color. Next, looking at on the left second top two comparison figures in classes C&D, we see a mark of difference in red color between green colored distributions and blue colored distributions.

We see 18 LAA classes on the left and 18 analysis basic classes (ABA) on the right in Figure 12, and the possible numbers of comparison between two classes is 11 in LAA and 13 in ABA; in total, 24 cases can be tested. The number of cases we observed the difference between two classes in the end-term examination evaluation is 12. Thus, almost half of the two classes comparison shows the evaluation difference between the two classes. The probability that such a case just occurs is $\binom{24}{12}/2^{24} = 0.16$. The probability that more than 12 cases show the difference between two classes is $\sum_{k=12}^{24}\binom{24}{k}/2^{24} = 0.58$. Thus, it is proved that the example shown in 3.1 is not a rare case. We can see that it is important to remove such an evaluation bias.

5.2 Extension of the Ability Equation

In section 2, we have mentioned that irreducible probabilistic fluctuations may occur even if an examinee happens to take very similar examinations twice assuming that the examinee's ability is not changed. That is, ε_i in (2) could be measured. This can be estimated by using a binomial distribution assumption. For example, if someone took an examination with 36 questions and 18 answers are correct, then the standard deviation for this score (1 point for one question) is computed to be $\varepsilon = \sqrt{36 \times 0.5 \times (1-0.5)} = 3$, and the 95% confidence interval could be $[12, 24]$. We should understand the observed score in such a way. In splitting a class into several classes using some threshold in pre-tests, e.g., placement test, we should pay attention to this effect.

Looking at Figures 9, the values of coefficient of correlation among the placement test, the LCT, and the term examinations are approximately 0.8, and this value is smaller than

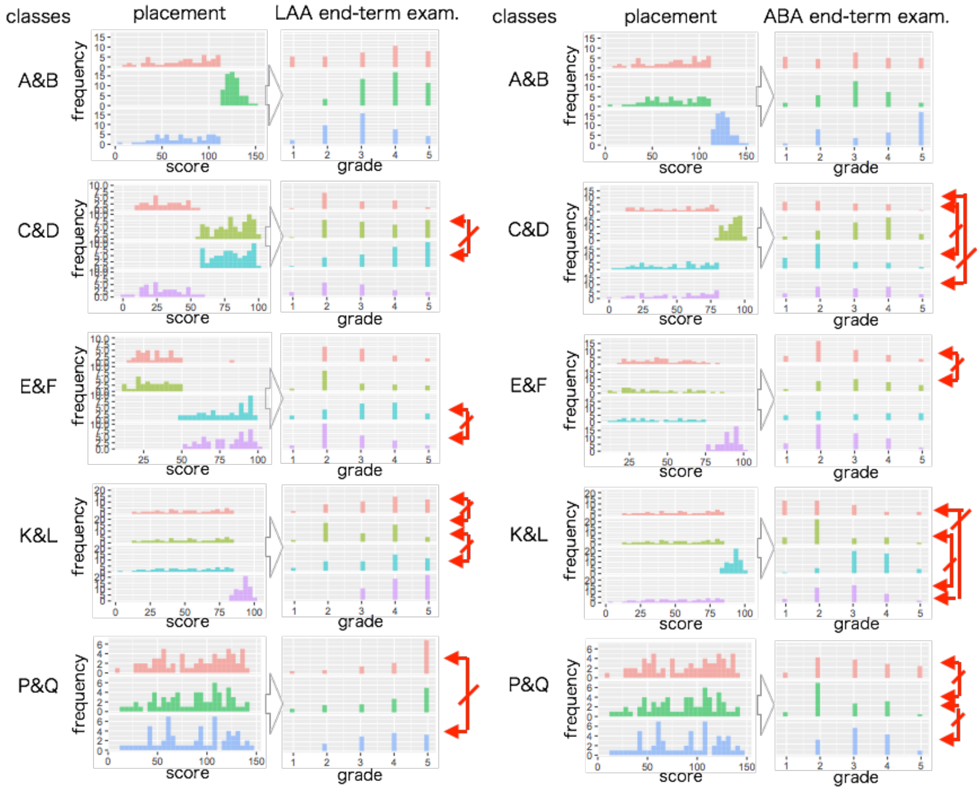

Figure 11: Comparison of records among three testing cases (placement test, LCT, end-term examination) with two teachers performed to analysis basic (calculus) in the first semester in 2017. At the beginning of the classes, placement test scores are totally the same; the LCT ability results show almost the same aspect; at the end-term examination, we see the teacher's evaluation biases due to teacher's way of evaluations.

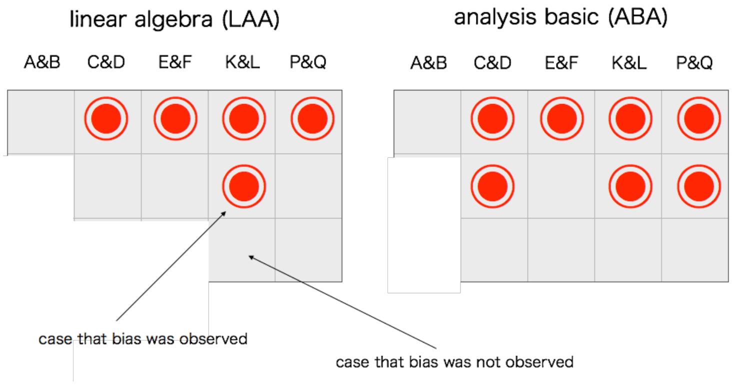

Figure 12: Comparison of records among three testing cases (placement test, LCT, end-term examination) with two teachers performed to analysis basic (calculus) in the first semester in 2017. At the beginning of the classes, placement test scores are totally the same; the LCT ability results show almost the same aspect; at the end-term examination, we see the teacher's evaluation biases due to teacher's way of evaluations.

0.9 (or more concretely 0.91 to 0.95 seen in Figure 7) mentioned in the irreducible fluctuations. Such discrepancies are considered to be caused by variations of the students' effort or laziness. That is, we may assume that μ_i in (2) was not a constant value to each student; μ_i can be described as $\mu_i(t)$ depending on time. For example, some student showed $\mu_i(0) = 0.4$ and $\mu_i(1) = 0.7$ (he/she made effort), and some student showed $\mu_j(0) = 0.6$ and $\mu_j(1) = 0.3$ (he/she was reluctant to study), where $t = 0$ means the beginning of the class and $t = 1$ means the end of the class.

Looking at Figure 10, the value of coefficient of correlation between the placement test and the end-term examination is approximately 0.5, which is smaller than the value of 0.8 mentioned above. Since the value of coefficient of correlation between the LCT and the end-term examination is approximately 0.76, we may suppose that there may be teaching differences among teachers.

On the other hand, by looking at Figure 8, the values of coefficient of correlation among the placement test, the LCT, and the end-term examinations seem to be badly disturbed by some effects. This may be teacher's evaluation bias.

Therefore, we can extend equation (1) to

$$\phi_i(t) = \mu_i(t) + \beta_j(t) + \varepsilon_i(m), \tag{4}$$

where, t means time ($t = 0$ when a class begins, and $t = 1$ when the class ends), m means the number of questions; i corresponds to student i and j corresponds to teacher i. β expresses the teacher's evaluation bias. We may assume that $\varepsilon_i(m)$ is normally distributed. By taking into account the multiple choice type testing, we could eliminate the teacher's evaluation bias $\beta_j(t)$. Then, we could measure the student's academic growth by the difference between $\mu_i(0)$ and $\mu_i(1)$. This is the proposed equation for ability equation. The concrete methodology to estimate each term is planned to show in the future. This is beyond this paper's scope.

5.3 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 [7]). 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 [5]). 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 [39]).

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.

6 Concluding Remarks

A test score does not represent the exact ability of an examinee. It only shows just one aspect of the examinee, even if the coverage of the test is restricted. To answer a typical question why we cannot see obvious relationships between entrance examination scores and academic records in universities, we have investigated three testing results; one is the placement test, the second is the learning check test, and finally the term examinations. In this paper, firstly, we have analyzed the irreducible probabilistic fluctuations of academic scores by using the placement test and term examinations. Then, we could catch the magnitude of irreducible probabilistic fluctuations. Next, we have compared the distributions of the three testing scores in equally split classes using the placement scores. In testings, unlike the common testing style of the description type in mathematics subjects, the multiple choice type testing was applied. Then, we have found two crucial points. One point is that we can remove the teacher's evaluation biases. The other is that student's academic growth could be measured more clearly. We have also proposed a fundamental equation on student's ability including irreducible probabilistic fluctuations.

7 Appendix

We illustrate a multiple choice type question and a description type question below.

7.1 Multiple Choice Type Problem

Select numbers from $\{0,1,\cdots,9\}$ and fill them into (1), (2), (3), and (4) boxes, so that function$f(x)$ becomes a differentiable function.

$$f(x)=\begin{cases} x^2 & (x>0) \\ \boxed{(1)}\,x+\boxed{(2)} & (x\leqq 0) \end{cases}$$

$$f(x)=\begin{cases} x^2+2x+2 & (x>0) \\ \boxed{(3)}\,x+\boxed{(4)} & (x\leqq 0) \end{cases}$$

7.2 Description Type Problem

Answer the following questions.

1. Show the definition to obtain the differential function of $f(x)$.

2. Show the differential function of $f(x)=\sin x$ by following the definition 1).

3. In Figure A, if the length of arc PQ is r, and the length of line segment HP is s, then obtain $\lim_{d\theta \to 0} \frac{s}{r}$ using θ.

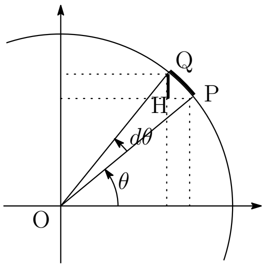

Figure A

References

[1] S. L. Annan, S. Tratnack, C. Rubenstein, E. Metzler-Sawin, L. Hulton, An Integrative Review of Student Evaluations of Teaching: Implications for Evaluation of Nursing Faculty, Journal of Professional Nursing, 29, 2013, pp.10-24.

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

[3] M. Bamber, The impact on stakeholder confidence of increased transparency in the examination assessment process, Assessment & Evaluation in Higher Education, 40, 2014, pp. 471-487.

[4] A. Campbell, Application of ICT and rubrics to the assessment process where professional judgement is involved: the features of an e marking tool, Assessment & Evaluation in Higher Education, 30, 2005, pp.529-537.

[5] 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.

[6] D. E. Clayson, Student evaluation of teaching and matters of reliability, Assessment & Evaluation in Higher Education, 43, 2018, pp.666-681.

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

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

[9] E. H. Haertel, Reliability and Validity of Inferences about Teachers Based on Student Test Scores, Mathematics Education Trends and Research, The 14th William H. Angoff Memorial Lecture, 2013, Educational Testing Service.

[10] S. He, K. Kempe, Y. Tomoki, M. Nishizuka, T. Suzuki, T. Dambara, T. Okada, Correlations between Entrance Examination Scores and Academic Performance Following Admission, Juntendo Medical Journal, 2015, pp. 1-7.

[11] K. K. Hensley, Examining the effects of paper-based and computer-based modes of assessment on mathematics curriculum-based measurement, Mathematics Education Trends and Research, Ph.D Thesis in Teaching and Learning, 2015, The University of Iowa.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

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

[24] 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.

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

[26] 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.

[27] 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.

[28] 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.

[29] H. Hirose, Key Factor Not to Drop Out is to Attend Lectures, Information Engineering Express, 5, 2019, pp.59-72.

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

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

[32] H. W Marsh, Students' Evaluations of University Teaching: Dimensionality, Reliability, Validity, The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective, Springer, 2020, pp.319-383.

[33] H. W Marsh, Student Evaluations of Teaching Encourages Poor Teaching and Contributes to Grade Inflation: A Theoretical and Empirical Analysis, Basic and Applied Social Psychology, 42, 2020, pp.276-294.

[34] K. Meissel, F. Meyer, E. S. Yao, C. M. Rubie-Davies, Subjectivity of teacher judgments: Exploring student characteristics that influence teacher judgments of student ability, Teaching and Teacher Education, 65 (2017) 48-60.

[35] B. O'Connell, P. De Lange, M. Freeman, P. Hancock, A. Abraham, B. Howieson, K. Watty, Does Calibration Reduce Variability in the Assessment of Accounting Learning Outcomes?, Assessment & Evaluation in Higher Education, 41, 2016, pp. 331-349.

[36] K. Otani, B. J. Kim, J.-IL Cho, Student Evaluation of Teaching (SET) in Higher Education: How to Use SET More Effectively and Efficiently in Public Affairs Education, Journal of Public Affairs Education, 18, 2012, pp.531-544.

[37] F. Ostad-Ali, M.H. Behzadi, A. Shahvarani, Descriptive Qualitative Method of Evaluation from the Viewpoint of Math Teachers and Its Comparison with the Quantitative Evaluation (Giving scores) Method (A Case Study on the Primary Schools for Girls in Zone 1 of Tehran City), Mathematics Education Trends and Research, 20, 2015, pp.50-56.

[38] D. Quah, Galton's Fallacy and Tests of the Convergence Hypothesis. The Scandinavian Journal of Economics. 95, 427-433, 1993.

[39] 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.

[40] 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.

[41] 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.

[42] H.M. Salahdeen, B.A. Murtaba, Relationship between Admission Grades and Performances of Students in the First Professional Examination in a New Medical School, African Journal of Biomedical Research, 8, 2005, pp.51-57.

[43] S. Siegel, Non-parametric statistics for the behavioral sciences. New York: McGraw-Hill, 1956.

[44] S.M. Stigler, Regression towards the mean, historically considered, 6, 1997, pp.103-114.

[45] 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.

[46] P. Spooren, B. Brockx, D. Mortelmans, On the Validity of Student Evaluation of Teaching: The State of the Art, Review of Educational Research, 83, 2013, pp.598-642.

[47] W. Stroebe, Student Evaluations of Teaching Encourages Poor Teaching and Contributes to Grade Inflation: A Theoretical and Empirical Analysis, Basic and Applied Social Psychology, 42, 2020, pp.276-294.

[48] `https://www.slideshare.net/HideoHirose/regression-fallacies`