International Journal of Learning Technologies and Learning Environments
International Institute of Applied Informatics
2023, Vol. 6, No. 1, IJLTLE741

Fluctuations of Ability Estimates in Testing in Item Response Theory

Hideo Hirose *

Abstract

By analyzing the fluctuations of ability estimates in testing, we first obtain the purely probabilistic fluctuations of ability estimates in a one-time testing under the condition that the students' abilities can be estimated by using the item response theory, and next, by taking into account such the probabilistic fluctuations, we find students who reveal the discrepancies of observed abilities between two separated testings. When such discrepancies of abilities are observed, test results are considered to be affected by some factors such as the physical conditions of the examinees, the teacher's teaching skills, and students' study skill developments. To describe such a phenomenon, we proposed a basic formula. The accuracies are obtained under the situation that the observed data follows the item response theory. To investigate whether we can assume such a condition or not, we have introduced the matrix decomposition perspective, and confirmed that the item response theory were used properly. Using an example case took in a university mathematics testing, we have shown how we have extracted the purely probabilistic fluctuations and segregated fluctuations due to other factors.

Keywords: item response theory, purely probabilistic fluctuation, basic formula representing fluctuations, bootstrap method, Fisher information matrix, learning analytics, matrix decomposition.

1 Introduction

Since testing is an indirect method to measure the ability of an examinee in a specific field, the result from a test does not always represent the true ability of the examinee. Sometimes, the examinee is not good at questions in a test even if he/she often can do well to other tests. Therefore, it is commonly seen that the results of two consecutive tests are often not highly correlated, even though the test interval between the two tests is short and the difficulties of the tests are very similar.

For example, TOEIC tests are believed, in general, to be very reliable such that a test score by an examinee discloses the absolute potential value of the examinee's English skill. However, as shown in Figure 1, the scores of 296 examinees fluctuate very much person to person although the mean value of test A (horizontal axis) is almost the same as that of

* Kurume University, Fukuoka, Japan

test B (vertical axis); the interval between the two tests (A and B) is about one year. If we assume that English skill is admitted to be acceptable when the score of TOEIC test is larger than 350, the number of rejected examinees using test A is 118 out of 296. Out of 118 examinees, 69 examinees again failed in test B. Thus. if we predict the second rejection using the first result, the hit rate to that becomes only 0.58. In order to improve such a hit rate, we investigate the score fluctuation characteristics in this paper.

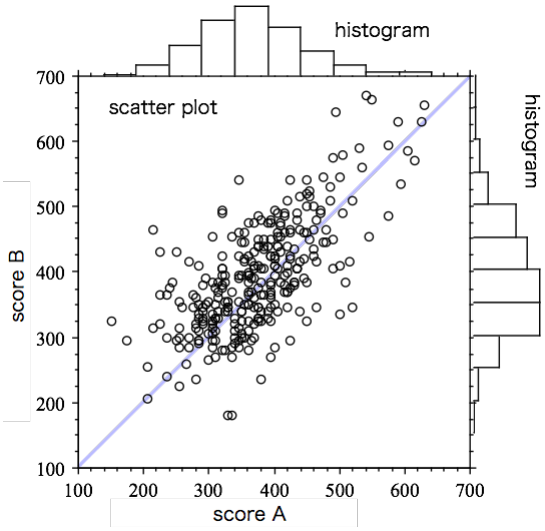

Figure 1: Two consecutive TOEIC test results. The coefficient of correlation between the two tests is 0.63.

Such a large fluctuation phenomenon is not considered to be caused only by probabilistic errors; there may exist many factors such as personal physical conditions, advantageous/disadvantageous problems given, personal studying results, or assistances by good teachings. In this paper, we consider whether we can deal with the fluctuation phenomena in testings. More concretely, our primary matter of concern is a possibility that we can segregate the pure probabilistic errors from the total fluctuations.

2 Basic Formulation of Fluctuations

In the previous paper [23, 24], we showed the existence of the irreducible probabilistic fluctuations in testing along with the basic formula representing such the fluctuations as

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

where, ϕ_i, μ_i, and ε_i are student's observed ability, student's true ability, and purely probabilistic fluctuation, respectively. However, as mentioned above, student's observed ability may include other factors affected by testing chances. Thus, we extend equation (1) more precisely to include such factors $\alpha_i(t)$ such that

$$\phi_i(t) = \mu_i + \alpha_i(t) + \varepsilon_i, \tag{2}$$

where, $\phi_i(t)$ and $\alpha_i(t)$ are time-dependent, and μ_i and ε_i are time-independent. In [23, 24], we also considered the evaluation bias effect model by teacher k such as

$$\phi_i(t) = \mu_i + \beta_k(t) + \varepsilon_i. \tag{3}$$

This bias effect $\beta_k(t)$ can be included in equation (2) if we interpret that $\alpha_i(t)$ absorbs $\beta_k(t)$ affected by teacher k, or class k.

Thus, we think that the observed value of student's ability fluctuates due to two elements of $\alpha_i(t)$ and ε_i. For just a one-time testing, μ_i in equation (1) is the same as the ability parameter in the item response theory (IRT) which will be denoted by θ_i in equation (4) in the next section. Here, we show two methods to evaluate the magnitude for ε_i; one is to use the bootstrap method and the other is to use the Fisher information matrix in likelihood principle.

We try to find the purely probabilistic fluctuation ε_i under the IRT with two-parameter logistic function. For general explanations to the IRT, see [1], [7], [28], for example. For specific explanations to online testing applications, see [8–12, 32]. However, we, here, introduce the IRT very briefly for clarity.

3 Item Response Theory

We assume the probability P_{ij} that an examinee i answer to a question j correctly is denoted as

$$\begin{aligned} P_{ij}(\theta_i; a_j, b_j) &= \frac{1}{1+\exp\{-1.7a_j(\theta_i - b_j)\}}, \\ &= 1 - Q_{ij}(\theta_i; a_j, b_j), \end{aligned} \tag{4}$$

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. Then, 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,

$$L = \prod_{i=1}^{m} \prod_{j=1}^{n} \left(P_{ij}^{\delta_{ij}} \times Q_{ij}^{1-\delta_{ij}} \right), \tag{5}$$

where m is the number of students, and n is the number of questions.

However, it is not easy to obtain the item parameters and the students' abilities together. There are $2 \times n + m$ unknown parameters to be estimated. Therefore, the item parameters are first estimated by using the marginal likelihood function by removing the students' abilities such as

$$L(\delta|a,b) = \prod_{i=1}^{m} [\int_{-\infty}^{\infty} g(\theta) \prod_{j=1}^{n} L(\delta_{ij}|a_j, b_j) d\theta], \tag{6}$$

where $g(\theta)$ denotes the ability common to all the students and δ denotes all the patterns of δ_{ij}, taking the value of 0 and 1. We often apply $g(\theta)$ to a standard normal distribution. The EM algorithm [3] is usually used in such a case [2]. Then, the students' abilities are obtained by maximizing the corresponding likelihood function. To circumvent the ill conditions so that all the items are correctly answered or incorrectly answered, the Bayes technique is applied [2]. However, we sometimes meet other cases such as the uniform distribution to $g(\theta)$. In such a situation, other formulations should be established.

The indicator function δ_{ij} takes values such that $\delta = 1$ for success and $\delta = 0$ for failure in answering a question. However, we may extend δ_{ij} characteristic from discrete value of $\delta = 0, 1$ to continuous value of $\delta \in [0, 1]$ to accept the partial correctness to answer a question.

4 Evaluation of Purely Probabilistic Fluctuations

4.1 The bootstrap method

After a test is taken, we obtain the estimated parameters $\hat{\theta}_i$ and $\hat{a}_j, \hat{b}_j$ by maximizing the likelihood function (6). From the test results taken just a one-time, we attempt to obtain the magnitude of the probabilistic fluctuation. To do so, we perform bootstrap simulations to generate random response matrices.

Example testing cases we use here are the midterm examination and the end-term examination for the undergraduate freshman mathematics subject, analysis basic (or calculus). They were performed in the first semester in 2019. See relevant references such as [13–22,25–27,33,34]. 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. The number of examinees took the midterm examination is 216, and the end-term examination 215. For the sake of comparison, we chose the examinees for analysis who took both the midterm examination and end-term examination together. The numbers of such examinees was 207. In dealing with the grade evaluation, we have used all the question results together in the IRT analysis.

We have considered two bootstrap cases although there are many methods to generate such a matrix. One is that we draw n whole column response results with replacement from n question columns randomly; this is a kind of the nonparametric bootstrap method. We call this random item selection; see Figure 2 on the left. The second is that we generate new $\hat{\delta}_{ij}$ by using the cumulative probability distribution (4) with the estimated $\hat{\theta}_i$ and $\hat{a}_j, \hat{b}_j$; this is a kind of the parametric bootstrap method. In regenerating the response matrix, we first generate a uniform random number $p \in [0,1]$ and obtain $\hat{P}_{ij} = P_{ij}(\theta_i; a_j, b_j)$, then we set $\hat{\delta}_{ij} = 1$ if $\hat{P}_{ij} \geq p$, and $\hat{\delta}_{ij} = 0$ otherwise. We call this random delta determination; see Figure 2 on the right.

In the figure on the bottom, two scatter plots are shown. In both the scatter plots, the horizontal axis represents the estimated ability using the original (observed) response matrix; the vertical axis represents the estimated ability using the bootstrapped response matrix. We see there are no obvious differences between the two scatter plots. We may use any of these two results. However, the former case may cause the difficulty in numerical computation in maximizing the likelihood function due to the matrix singularity (results of two columns could be the same with probability $1/e$). The example in Figure 2 happens to avoid such an inconvenient case. From now on, we use the second type bootstrap method, i.e., the parametric bootstrap method.

From these scatter plots, we can see that the estimated ability points seem not to spread in an elliptic shape (this means the points are approximately normally distributed), but in an band-wise shape. This spreading characteristic is totally different from the feature in Figure 1 where spreading shows elliptic shape. Such a difference can be understood that Figure 1 case includes the variable $\alpha_i(t)$ and Figure 2 cases do not.

We have generated 100 response matrices using the random delta determination. Then, to each student i, 100 estimates of $\{\theta_i^l\}, (l = 1, \ldots, 100)$ are obtained. Using these estimates, we can draw the box-plot of the estimated abilities to each student i. Figure 3 shows the box-plots for all the students in ascending order of the originally estimated ability $\hat{\theta}_i$ for clarity. In the figure, the thick line in the center of the box-plot indicates the median of $\{\theta_i^l\}, (l = 1, \ldots, 100)$, $\mathrm{Q}_{2,i}$. The top edge of the box represents the third quartile, $\mathrm{Q}_{3,i}$, and the bottom edge of the box represents the first quartile, $\mathrm{Q}_{1,i}$. In addition, the upper and lower

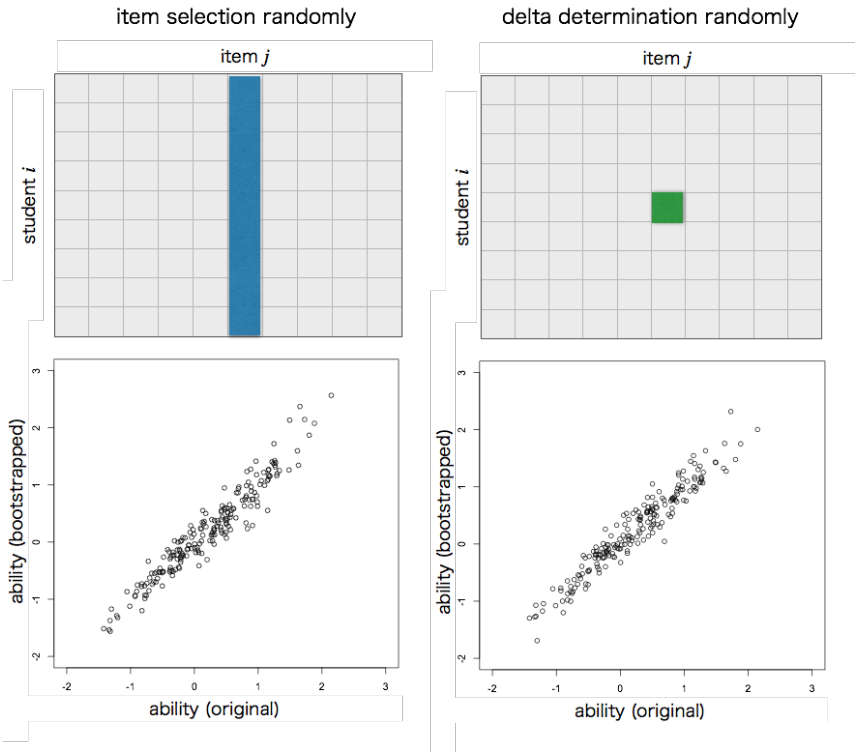

Figure 2: Two kinds of the bootstrap methods. On the left hand side, the nonparametric bootstrap method is illustrated where full questions in column are selected randomly. On the right hand side, the parametric bootstrap method is illustrated where $\hat{\delta}_{ij}$ is regenerated randomly, using equation (4).

whiskers are, $Q_{1,i} - 1.5 \times IQR_i$, and $Q_{3,i} + 1.5 \times IQR_i$, where IQR_i is $Q_{3,i} - Q_{1,i}$. Circles represent data points larger or smaller than the whiskers, that is, outliers. We can roughly grasp the purely probabilistic fluctuations from this figure.

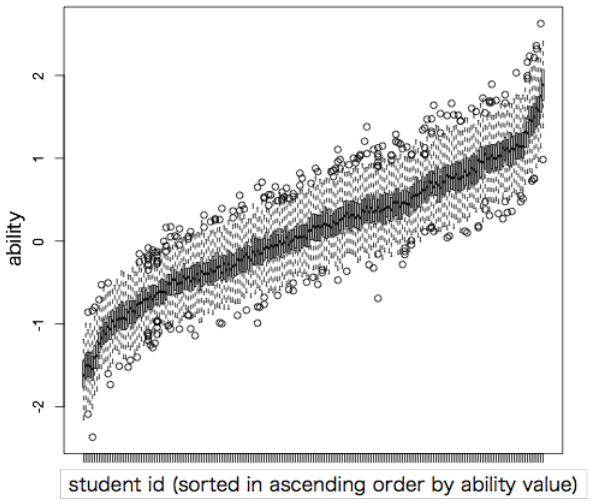

Figure 3: The box-plots of the bootstrapped ability estimates for all the students in ascending order of the originally estimated ability $\hat{\theta}_i$.

To capture the fluctuations more intuitively, we illustrated the standard deviations of the bootstrapped ability estimates for all the students in ascending order of the originally estimated ability $\hat{\theta}_i$ as shown in Figure 4. Looking at the figure, we see that the standard deviations do not vary much. This may indicate that the magnitudes of ε_i in equation (2) are similar to all the students.

4.2 Use of the Fisher information matrix

Assuming that the ability parameters θ_i follow the normal distribution, then we can use a Bayes method to obtain the ability parameters by maximizing,

$$H\{\theta \mid \delta, a, b\} \propto L\{\delta \mid \theta_i, a_j, b_j\} h(\theta), \quad \delta = (\delta_{ij}), \tag{7}$$

where,

$$L\{\delta \mid \theta_i, a_j, b_j\} = \prod_{j=1}^{n} P_{ij}(\theta_i)^{\delta_{ij}} (1 - P_{ij}(\theta_i))^{1-\delta_{ij}}, \tag{8}$$

$$h(\theta) \quad = \quad \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\theta - \mu_\theta}{2\sigma_\theta^2} \right). \tag{9}$$

Figure 4: The standard deviations of the bootstrapped ability estimates for all the students in ascending order of the originally estimated ability $\hat{\theta}_i$.

The Fisher information matrix (in this case, it is scalar) can be obtained as follows.

$$\begin{aligned} I(\theta) &= E[(\frac{\partial \log Lh}{\partial \theta_i})^2] \\ &= -E[\frac{\partial^2 \log Lh}{\partial {\theta_i}^2}] \\ &\approx \sum_{j=1}^{n} (1.7a_j)^2 P_{ij}(\theta_i)(1-P_{ij}(\theta_i)) \\ &+ \frac{1}{2{\sigma_\theta}^2}. \end{aligned} \tag{10}$$

Using the same example in the subsection above, we can obtain the probabilistic fluctuations regarding the Fisher information matrix. Figure 5 shows the standard errors for the ability estimates using the Fisher information matrix for all the students in ascending order of the originally estimated ability $\hat{\theta}_i$. The figure corresponds to Figure 4, and we see that they provide approximately the same features.

We have compared the probabilistic errors using the bootstrap method and using the Fisher information matrix against the examinee's estimated ability values. Figure 6 shows such a comparison. We can see that both the probabilistic errors are close to each other.

In addition, to compare these two fluctuations, we provide Figure 7 where the standard deviations using the bootstrap method located in horizontal axis and the standard errors using the Fisher information matrix located in vertical axis. These two fluctuations are well correspond to each other. Thus, we can grasp the purely probabilistic fluctuation magnitude to each examinee using the parametric bootstrap method (and also using the nonparametric bootstrap method) and using the Fisher information matrix from just a one-time testing result.

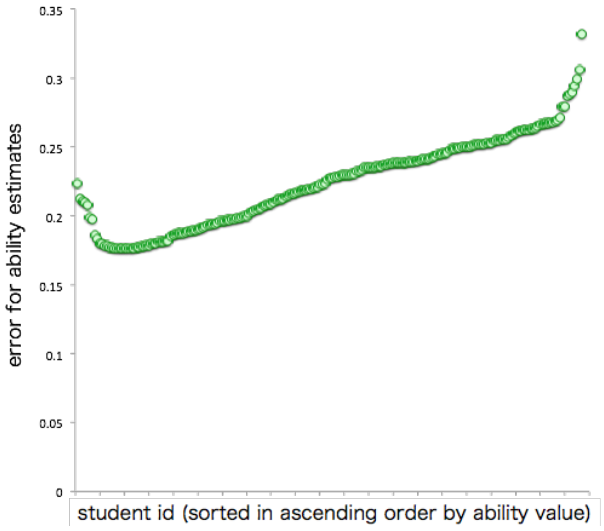

Figure 5: The standard errors for the ability estimates using the Fisher information matrix for all the students in ascending order of the originally estimated ability $\hat{\theta}_i$.

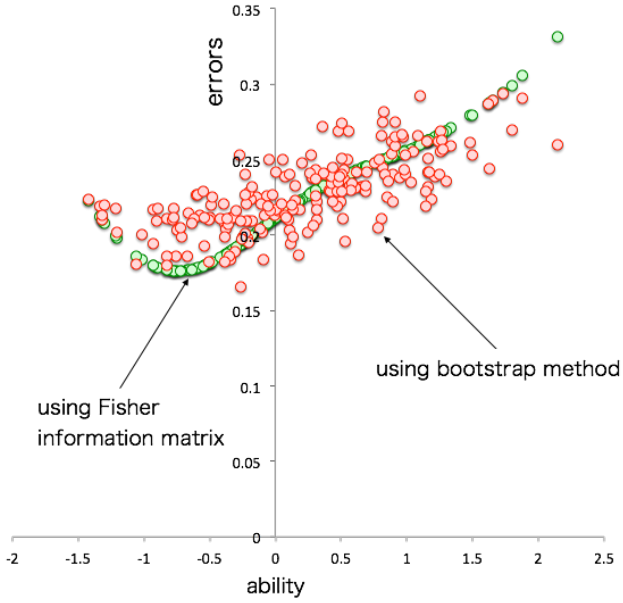

Figure 6: Comparison the probabilistic errors between using the bootstrap method and using the Fisher information matrix against the examinee's estimated ability values.

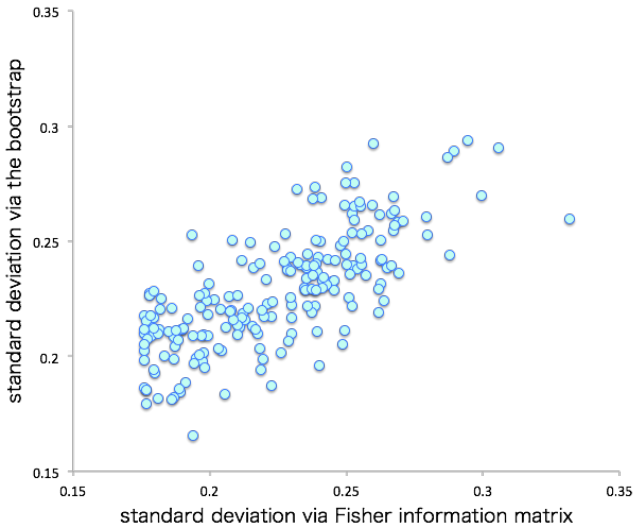

Figure 7: Comparison of the fluctuations using the bootstrap method and those using the Fisher information matrix.

5 Comparison between the Purely Probabilistic Fluctuations and the Other Factor Add-on Fluctuations

We have shown that the purely probabilistic fluctuations can be grasped using either from the bootstrap method or the Fisher information matrix for just a one-time testing. As mentioned before, there may be some other factors affecting the ability results between the two test intervals; to each test, we can obtain the magnitude of the purely probabilistic fluctuations. We next attempt to extract other factor fluctuations. We assume that such the fluctuations are independent from the purely probabilistic fluctuations in just a one-time testing.

We again use the test results as introduced before, that is, the midterm examination results and the end-term examination results. First, we plot these two test results via the estimated ability values in Figure 8. We may consider that the coefficient of correlation is high such as a value close to 1; however, actually, it is 0.72, and this is not so small, but also not so large enough.

We again performed the bootstrap method to the end-term examination. Figure 9 shows an ability plot of the original end-term estimates (horizontally located) and the bootstrapped estimates (vertically located). Contrary to Figure 8, the coefficient of correlation is 0.95 which is considered to be high. Thus, we may assume that other factors different from the purely probabilistic fluctuation in a one-time testing are existing.

To confirm such an assumption, we provided additional two figures. Figure 10 shows the ability estimates of the end-term examination and their 95% upper and lower confidence limits computed by using the Fisher information matrix along with the bootstrapped ability estimates to the end-term examination. Looking at the figure, we can see that almost all estimates obtained by the bootstrap method are included in the 95% confidence band. Figure 11 shows the ability estimates of the end-term examination and the 95% upper and lower confidence limits computed by using the Fisher information matrix along with the ability estimates to the midterm examination. From the figure, we can find that many points

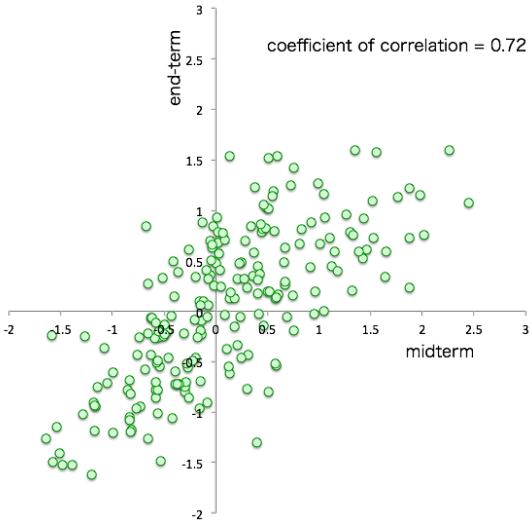

Figure 8: The relation between the midterm examination results and the end-term examination results.

located outside the 95% confidence band. We think that such points may include the factors $\alpha_i(t)$ different from the purely probabilistic fluctuations in a one-time testing. A student (student id is 78) took 1.85 in ability value in the midterm examination, but his ability value in the end-term examination dropped steeply to -0.25. We can imagine that his testing condition was not so good or something wrong happened to him. On the contrary, another student (student id is 83) took -1.67 in ability value in the midterm examination, but his ability value in the end-term examination soared to -0.21. In such cases, we can utilize the fluctuation analysis in a statistical manner.

6 Another Look at the Response Matrix

In analyzing the fluctuations for ability estimates, either parametric or nonparametric, we have been using the IRT. This means that the results are those restricted under the assumption that the IRT can be fitted to the data. However, the IRT requires a strict assumption that δ_{ij} and δ_{ik} is independent from each other, where $j \neq k$. That is, if question j is related to question k to some extent, we may assume that an examinee who solved question j could also solve question k with higher probability than the examinee's latent ability. This is so-called the local independency [4]. In order to investigate whether the results obtained by the IRT are affected by such a dependency or not, we have to look at the response matrix from another perspective. We introduce the matrix decomposition (MD) [29] perspective here. The MD does not require such an independency among questions.

6.1 Matrix decomposition perspective

The matrix decomposition is similar to the singular value decomposition, SVD [6]. The SVD deals with complete matrix, and the MD deals with incomplete matrix. When the matrix is complete, both the SVD and the MD can be used, and the decomposed matrices are almost the same.

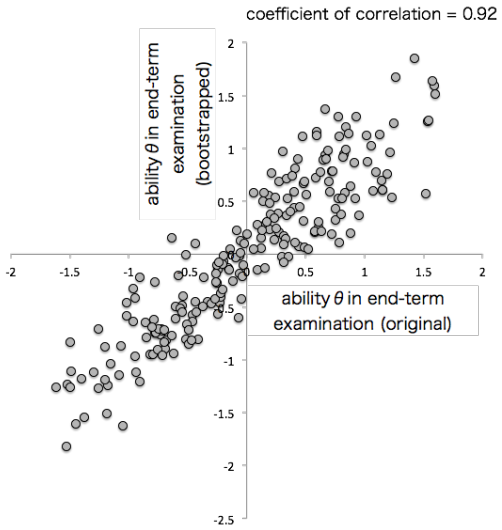

Figure 9: The relation between the original ability estimates of the end-term examination and the bootstrapped ability estimates to the end-term examination.

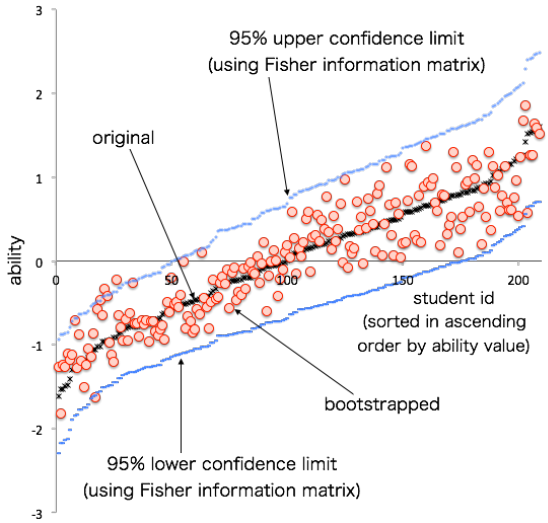

Figure 10: The ability estimates of the end-term examination and the 95% upper and lower confidence limits computed by using the Fisher information matrix along with the bootstrapped ability estimates to the end-term examination in ascending order of the originally estimated ability $\hat{\theta}_i$.

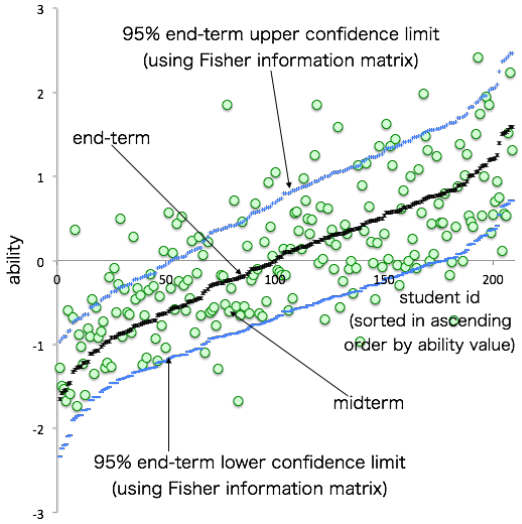

Figure 11: The ability estimates of the end-term examination and the 95% upper and lower confidence limits computed by using the Fisher information matrix along with the ability estimates to the midterm examination in ascending order of the originally estimated ability $\hat{\theta}_i$

We can construct a target matrix $R \in \mathbb{R}^{m\times n}$ with two matrices $U \in \mathbb{R}^{m\times k}$ and $V \in \mathbb{R}^{n\times k}$. Matrix decomposition is described as

$$R = UV^{\mathsf{T}}. \tag{11}$$

Using the non-null element values in $A = (a_{ij})$, we find U and V so that

$$E = \sum_{i=1}^{m}\sum_{j=1}^{n} I(i,j)(a_{ij} - r_{ij})^2 \tag{12}$$

becomes small, where $r_{ij} = \sum_{l=1}^{k} u_{il}v_{jl}$, and $I(i,j)$ is the indicator function such that $I(i,j) = 1$, if a_{ij} is non-null, $I(i,j) = 0$, if a_{ij} is null. For stable computation, we use another function such that

$$W = \sum_{i=1}^{m}\sum_{j=1}^{n} I(i,j)(a_{ij} - r_{ij})^2 + k_u \sum_{i=1}^{m}\sum_{l=1}^{k} u_{il}^2 + k_v \sum_{j=1}^{n}\sum_{l=1}^{k} v_{jl}^2, \tag{13}$$

where, k_u and k_v are regularization coefficients to prevent overfitting. To find the optimum values, we use the descent method [5, 30]. From appropriately set initial values of $u_{il}^{(0)}$ and $v_{jl}^{(0)}$, we proceed the following iteration until $|u_{il}^{(t+1)} - u_{il}^{(t)}|$ and $|v_{jl}^{(t+1)} - v_{jl}^{(t)}|$ are both small enough.

$$\begin{aligned} u_{il}^{(t+1)} &\leftarrow u_{il}^{(t)} - \mu \frac{\partial W}{\partial u_{il}}\Big|^{(t)} \\ v_{jl}^{(t+1)} &\leftarrow v_{jl}^{(t)} - \mu \frac{\partial W}{\partial v_{jl}}\Big|^{(t)}, \end{aligned} \tag{14}$$

where, μ is a learning coefficient.

We use the data obtained from the midterm examination result mentioned above. The number of questions to the midterm examination is 31. However, the number of examinees is 216. Using the estimates of the IRT parameters, we can reconstruct the response matrix by using equation (4). Similarly, by using the MD method, we can build the response matrix. Figure 12 shows these two reconstructed response matrices along with the original observed response matrix. To find a relationship between the estimate $\hat{\theta}_i$ for ability parameter in the IRT and the estimated mean value $\hat{\delta}_i$ of $\hat{\delta}_{ij}$ for examinee i such that

$$\hat{\delta}_i = \frac{1}{n}\sum_{j=1}^{n} \hat{\delta}_{ij}, \tag{15}$$

we have shown Figure 13. We can see that there is a clear relationship between them, and thus we may proceed that we evaluate the estimation accuracy by using the reconstructed response matrix.

Regarding Figure 12, the two constructed matrices seem to be similar at a first glance. To compare the accuracy of the constructed matrices to the originally observed response matrix numerically, we have computed the root mean squared error, RMSE. Here, the RMSE between a matrix $A = (a_{ij})$ and a matrix $B = (b_{ij})$ is defined as below

$$\mathrm{RMSE} = \Big(\frac{1}{\sum_{i=1}^{m}\sum_{j=1}^{n} I(i,j)} \sum_{i=1}^{m}\sum_{j=1}^{n} I(i,j)(a_{ij}-b_{ij})^2\Big)^{1/2}, \tag{16}$$

where a_{ij} and b_{ij} are real numbers in $\mathbb{R}$, and $I(i,j) = 0$ if element (i,j) is null, $I(i,j) = 1$ if element (i,j) is not null. The RMSE using the IRT is 0.3915, and that using the MD is 0.3854 when $k = 2$. We have selected k such that the number of free parameters are to be comparable in these two mathematical models. From these RMSE, we may regard that the independency assumption among questions will not affect the response matrix reconstruction much. In other words, although there are dependencies among questions, we may ignore such an effect as long as we measure the accuracy via the RMSE for the response matrix.

6.2 Fluctuations using training data and test data

Usually, the IRT uses the full observed data for estimation, which may cause overfitting. In machine learning, to avoid such overfitting effect, we first select training data appropriately from the observed data, and then build a mathematical model for prediction. After that, we see if the model is properly assumed or not by using the test data set which is a complement set to the training data. However, the common IRT cannot perform such a procedure. Thus, we extend the IRT to overcome this flaw such that we can use an incomplete matrix [31,32].

Figure 14 shows one case for reconstructed response matrices using the incomplete training data matrix by the IRT and the MD along with the training data matrix. The size of training data is set to 9/10 of the total number of elements in the observed matrix, and the test data is set to 1/10. The training data are selected at random. We have performed such trials for ten times. Table 1 shows the RMSE values for training data sets and test data sets.

Looking at the table, there seems to be no large discrepancies among the RMSE using the 10 training data and among the RMSE using the 10 test data. Also, the RMSE between the IRT and the MD show similar values in training data and in test data, respectively. We can regard the IRT results are almost the same as those by the MD. That is, the results obtained from the common IRT are reliable in reality.

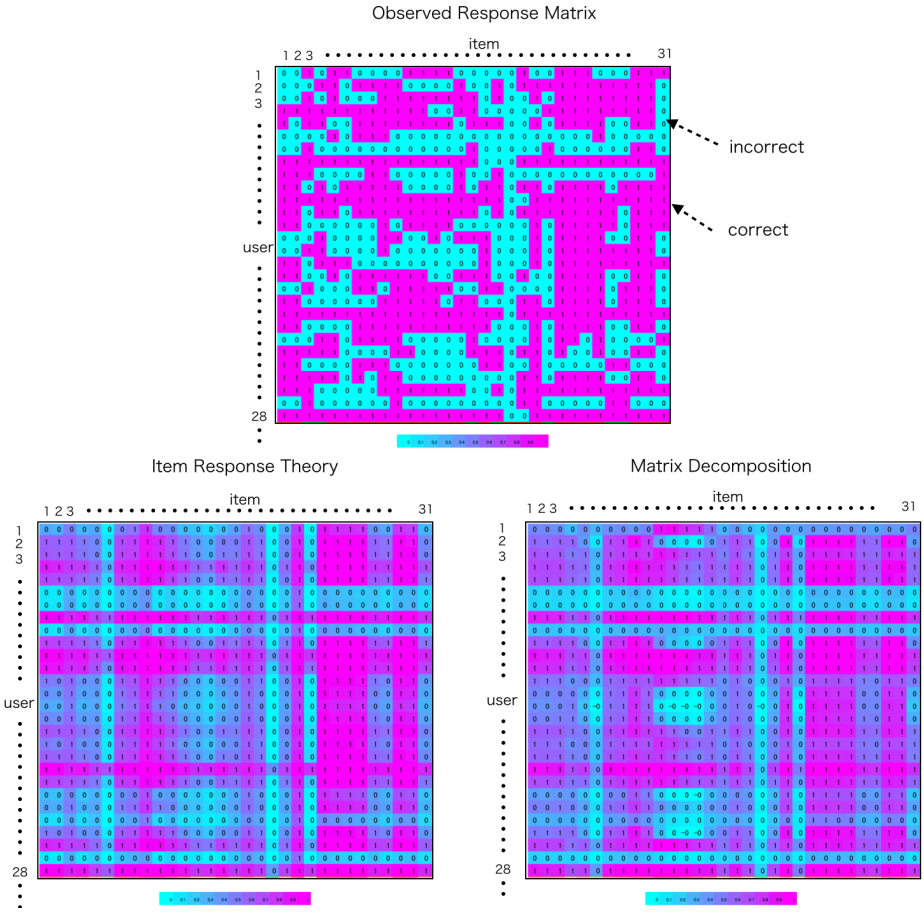

Figure 12: Observed response matrix and the reconstructed response matrix using the IRT and the reconstructed response matrix using the MD.

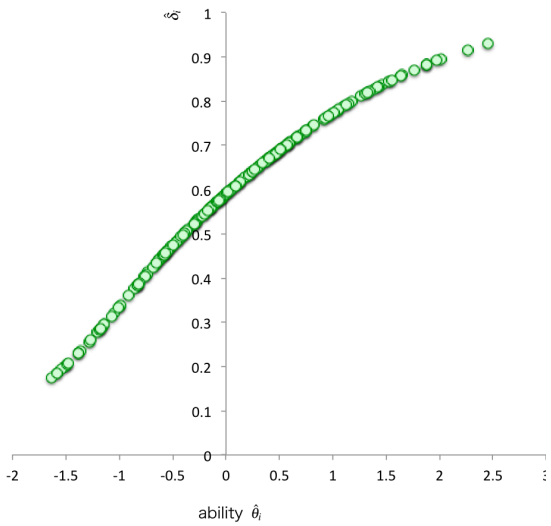

Figure 13: Relation between the estimated ability parameter $\hat{\theta}_i$ and the mean value of the corresponding $\hat{\delta}_i$ computed by equation (15) for student i.

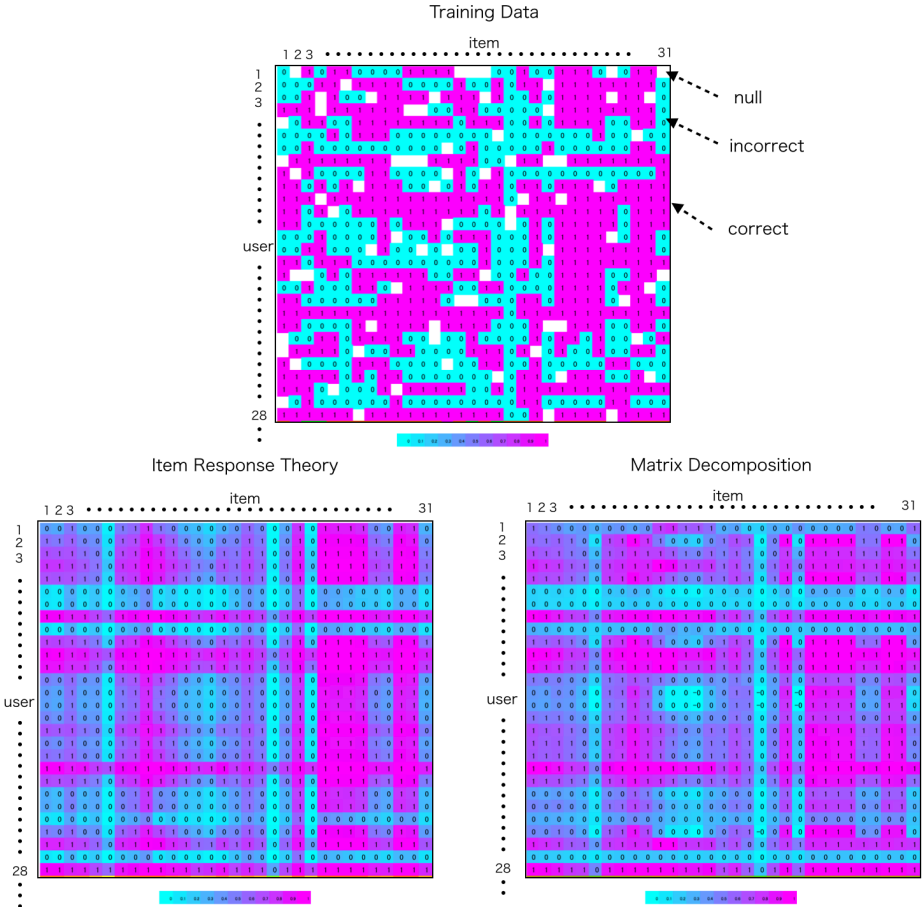

Figure 14: Reconstructed response matrices using the incomplete training data matrix by the IRT and the MD along with the training data matrix. On the top, elements where 0 and 1 numbers appear are used in the training data, and elements where 0 and 1 numbers are hiden are used in the test data.

Table 1: RMSE values for training data set and test data set using the IRT and the MD.

trial	MD training	MD test	IRT training	IRT test
1	0.3844	0.4137	0.3909	0.4093
2	0.3842	0.4106	0.3905	0.4097
3	0.3848	0.4158	0.3907	0.4120
4	0.3843	0.4136	0.3898	0.4174
5	0.3849	0.4162	0.3906	0.4059
6	0.3849	0.4057	0.3912	0.4046
7	0.3848	0.4088	0.3915	0.4044
8	0.3848	0.4076	0.3909	0.4048
9	0.3865	0.3918	0.3922	0.3923
10	0.3845	0.4085	0.3909	0.4080
mean	0.3848	0.4092	0.3909	0.4068

7 Concluding Remarks

Since testing is an indirect method to measure the ability of an examinee in a specific field, the result from a test does not always represent the true ability of the examinee. We often observe the fluctuations affected by many factors. In this paper, we have analyzed such fluctuations of ability estimates in testing. By analyzing the fluctuations, we can first obtain the purely probabilistic fluctuations of ability estimates in a one-time testing under the condition that the students' abilities can be estimated by using the item response theory. Next, by taking into account such the probabilistic fluctuations, we can find students who reveal the discrepancies of observed abilities between two separated testings. When such discrepancies of abilities are observed, test results are considered to be affected by some factors such as the physical conditions of the examinees, the teacher's teaching skills, and students' study skill developments. To describe such a phenomenon, we proposed a basic formula. The accuracies are obtained under the situation that the observed data follows the item response theory. To investigate whether we can assume such a condition or not, we have introduced the matrix decomposition perspective, and confirmed that the item response theory were used properly. Using an example case took in a university mathematics testings, we have shown how we extracted the purely probabilistic fluctuations and segregated them from fluctuations due to other factors in a statistical manner.

References

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

[2] F.B. Baker and S-H. Kim, Item Response Theory: Parameter Estimation Technique, 2nd edn., Marcel Dekker, 2004.

[3] A.P. Dempster, N.M. Laird, and D.B. Rubin, Maximum Likelihood from Incomplete Data via the EM Algorithm, Journal of the Royal Statistical Society. Series B, 39, 1977, pp.1-38.

[4] M.C. Edwards, C.R. Houts, and L. Cai, A diagnostic procedure to detect departures from local independence in item response theory models, Psychol Methods, 23, 2018, pp.138-149.

[5] R. Fletcher, Practical Methods of Optimization, Wiley, 2000.

[6] G.H. Golub, C.F. Van Loan, Matrix Computations, Johns Hopkins Univ. Press, 2012.

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.

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

[22] 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.

[23] 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.

[24] H. Hirose, Analysis of Fluctuations of Ability Estimates in Testing, 10th International Conference on Learning Technologies and Learning Environments, 2021, pp.148-153

[25] 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.

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

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

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

[29] Y. Koren, R.M. Bell and C. Volinsky, Matrix Faxctorization Techniques for Recommender Systems, Computer, 42, 2009, pp.30-37.

[30] E. Polak, Optimization : Algorithms and Consistent Approximations, Springer, 1997.

[31] T. Sakumura, T. Kuwahata and H. Hirose, An Adaptive Online Ability Evaluation System Using the Item Response Theory, Education & e-Learning, 2011, pp.51-54.

[32] 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.

[33] 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.

[34] 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.