久留米大学コンピュータジャーナル
VOL.37, 2022

［研 究 論 文］

Annual Trend of Correct Answer Rate by Exposure of Questions and Answers in Adaptive Online Testing

Hideo Hirose（Biostatistics Center, Kurume University）

Abstract:

This paper addresses learning effectiveness under conditions in which questions and answers are disclosed in adaptive online testing. Even when complete response matrices are available, research on the usefulness of item disclosure in large-scale testing is limited. In addition, despite the importance of the effectiveness of learning through adaptive online testing, research findings on the usefulness of item disclosure appear to be scarce. Since it seems difficult to analyze estimates for parameters in item response theory under question and answer exposure, this paper deals with the correct answer rate in order to discuss the effectiveness of learning. Using a large database of web-assisted online adaptive testing administered to undergraduate students, it is found that question and answer exposure in adaptive online tests contributes to student learning progress.

Keyword:

Effectiveness of Learning, Correct Answer Rate, Odds Ratio, Exposure of Questions and Answers, Adaptive Online Testing, Item Response Theory.

1. Introduction

Item response theory (IRT) (see [1, 2, 10, 19]) has been used to accurately and fairly assess examinee ability in a variety of settings, including the TOEFL. IRT is also capable of simultaneously estimating the difficulty of question items. These advantages make it possible to measure annual changes in examinee ability under conditions where the questions are not publicly available, e.g., [16, 17].

IRT can be used not only for assessment, but also for practice to enhance learning. Adaptive online tests, which attempt to automatically match the ability of the examinee ability with the difficulty of the questions, can be an efficient and effective method for such applications. Since the adaptive tests cannot create the complete response matrix consisting of user rows and item columns, only the student's ability is estimated and the item difficulties are always used as initial values; here user refers to the examinee and item refers to the question. Contrary to rigorous evaluation examinations that always keep question items hidden, adaptive testing often allows students to learn by disclosing question explanations and their answers.

In light of this situation, we have developed online testing systems for undergraduate students in a university [13] by creating a number of problem items for mathematics subjects [14]. Some of the questions are used for rigorous evaluation tests, while others are used for exercises. For this reason, we have separated the questions for evaluation and for practice.

Apart from the testing system operated by the university mentioned above, we have developed the web-assisted adaptive online testing systems [15] for undergraduate courses such as linear algebra, calculus, basic analysis, probability and statistics, ordinary differential equations, and basic physics as an additional tool for undergraduate textbooks. Similar to the adaptive testing for practice use in the university, explanations to the questions and their solutions are made available to the public.

In the case of rigorous assessment tests using complete item response matrices, many analytics research results have been reported such as [16, 17]. However, in the case of the above adaptive online tests, where questions and answer explanations are disclosed, it seems difficult to analyze examinees' abilities. Park et al. [20] pointed out that there has been limited research on the utility of item disclosure for large scale testing, and the issues requires ongoing and careful consideration. In response to this issue, we have investigated the effectiveness of learning in using the adaptive online testing under the condition that the questions and answers are disclosed.

Although the research results are limited, the following literature can be found. Bock et al. [3] claims that differential linear drift of item difficulty parameters over a ten-year period is demonstrated in data from the College Board Physics Achievement Test. However, Feinberg et al. [7] opposed to such the result that repeating the identical form does not create an unfair advantage in credentialing examinations. In addition, Wagner-Menghin et al. [28] mentions that no increase in individual scores when test items are reused, but information on change in item difficulty is lacking, and the observed decrease in mean item difficulty for reused items was insignificant. Also, Tang et al. [26] mentions that there are limited benefits from encountering the same items. In contrast, Raymond et al. [21] reports the validity of inferences based on scores from the second attempt. Moreover, Selvi et al. [25] mentions that item difficulty values obtained from initial item use were significantly lower than those obtained from repeated item use. Regarding the human memory characteristics, Ferreira et al. [8] describes that factor analyses indicates that visuospatial and verbal-numeric memory are distinct, but correlated variables. The difficulty for such treatment can be seen in Wood [29]; they mentions that how both memory advantages and disadvantages might manifest differently with item type. Similarly, Gilmer et al. [9] reports that effects of disclosure depend on the nature of the released items.

As the literature tells us, the reliability for ability parameters, the drifting (or shifting) phenomena to the estimates for difficulty parameters, and memory advantages are remain unclear under the item disclosure. In particular, we could not find relevant references in the adaptive online testing. Therefore, it would make a great deal of sense to investigate the effectiveness of learning in the adaptive online testing under the condition that the questions and answers are disclosed even now. Since it seems difficult to analyze the estimates for parameters in IRT such as the difficulty parameters, we have dealt with correct answer rate (CAR) for discussing the effectiveness of learning in this paper. The data we deal with is the case of the web-assisted adaptive online testing systems by [15].

2. Web-assisted Online Adaptive Tests

2.1 Common item response theory

In the common IRT, we assume probability P_{ij} that examinee i answered question j correctly is denoted as

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

where θ_i expresses the ability for examinee i, and a_j, b_j are constants in the logistic function for question j called the discrimination parameter and the difficulty parameter, respectively. Q_{ij} is the probability that examinee i answered question j incorrectly. This is the two-parameter mathematical model in IRT. 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,

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

where m and n are the number of examinees and the number of questions, respectively, and δ_{ij} is the indicator function such that $\delta_{ij} = 1$ if examinee i solved item j successfully and $\delta_{ij} = 0$ otherwise.

2.2 Online testing system

Since the tests we are dealing with here are adaptive, only the ability estimates are computed, and the values of item parameters such as the difficulty parameter and the discrimination parameter are given appropriately in advance.

Figure 1 shows the adaptive testing system. Assuming that a_j, b_j are given in advance, estimation of θ_i is straightforward. In the testing procedure, the initial question level is set to the skill level of an examinee, which is recorded if the examinee is a repeater. When the examinee visits the system for the first time, the very first question level is set to the intermediate level. Thereafter, the second and subsequent questions will be provided adaptively using the most recent estimated ability value of this examinee. In this system, the number of questions asked in a single adaptive test is set to either five or seven.

As mentioned earlier that students often requires the disclosures of explanations to the questions and solutions in the adaptive tests for their progresses, we have provided such materials. After examinees have finished the online tests, they will be able to learn how the solutions are obtained. Figure 2 shows a typical example for the explanation to a question and its solution.

2.3 Database size

Web-assisted online adaptive test systems are additional tools to accompany undergraduate text-books. The online testing system was first launched in January 2015 with linear algebra courses. Probability and statistics, calculus, physics, basic analysis, and ordinary differential equations (ODE) followed in sequence. Table 1 shows the start year, user size, item size, and access size for each subject through June 27, 2022, respectively. Here, access size means the total number of accesses to all question items.

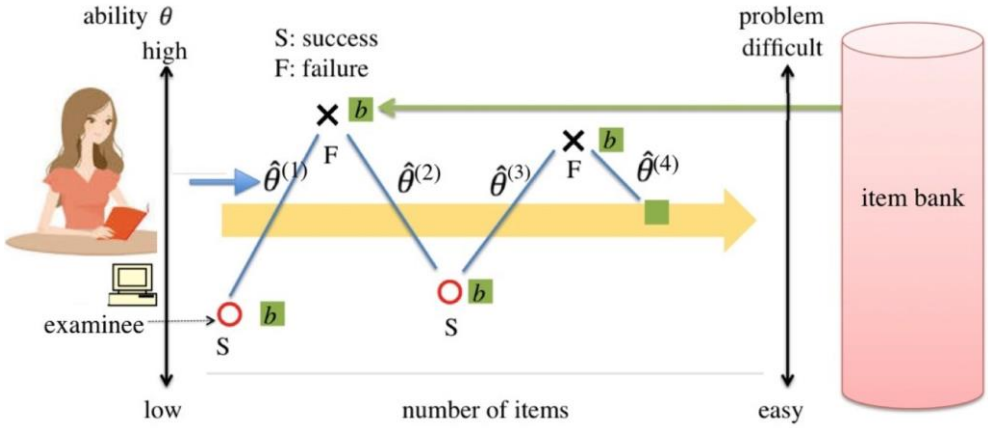

Figure 1: Online adaptive testing system [15].

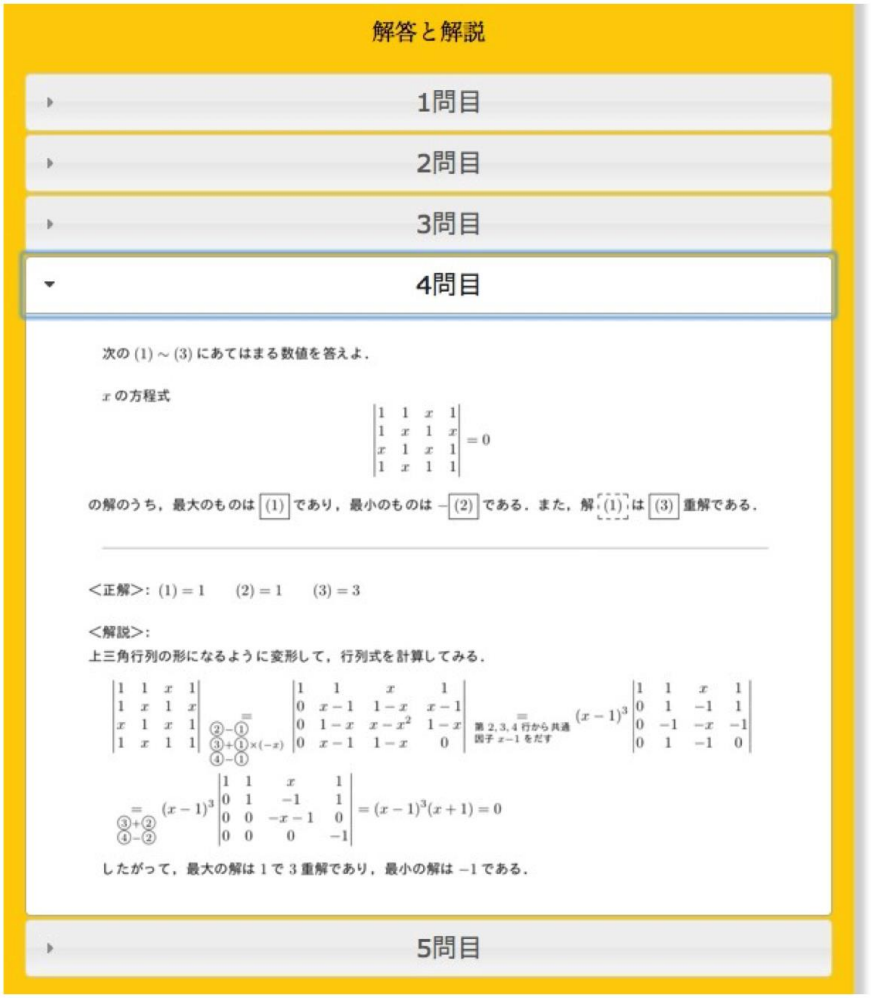

Figure 2: An explanation to a question and its solution [15].

Table 1: Web-assisted online adaptive test subjects and databases.

subject	launch	user size	item size	access size
Linear Algebra	Jan. 2015	2,201	860	71,418
Probability & Statistics	Jul. 2015	93	172	3,477
Calculus	Feb. 2017	853	636	36,618
Physics	Feb. 2017	72	133	2,310
Basic Analysis	Apr. 2020	223	760	1,315
ODE	Dec. 2021	9	151	53

2.4 Correct answer rate

Correct answer rate (CAR) is defined as the ratio of the total number of correct responses to the total number of attempts in the log files.

$$\mathrm{CAR} = \frac{\text{total number of correct responses}}{\text{total number of attempts}}. \tag{3}$$

In the log files, each response was recorded to each attempt. When the number of questions asked in a single adaptive test is k, the number of attempts is k in that single test.

CAR can also be defined to both user- and item-based, which is calculated for some user and for some item, respectively. To distinguish between the two, it is necessary to use different terms, e.g., $\mathrm{CAR}_{\mathrm{user}}$ and $\mathrm{CAR}_{\mathrm{item}}$. In this paper, the very first definition of (3) is primarily used, and $\mathrm{CAR}_{\mathrm{item}}$ will be dealt with in the discussion section. However, we will always use the term CAR for simplicity because there would be no confusion in distinguishing between these cases.

Looking at the log files that students worked on, three cases are seen: one in which they answered the question correctly, one in which they answered incorrectly, and one in which they finally gave up on solving the question. Whether this third case is counted as a case of incorrect solution (case 1) or not (case 2) can be determined by referring to Figures 3 and 4. The figures show the annual trends in CAR for linear algebra and calculus. There appears to be no clear differences between case 1 and case 2. Therefore, we consider case 2 as CAR here; that is, we disregard the unfinished responses. Thus, the numbers indicated as “sample size” in the figures mean the total numbers of correct answers and incorrect answers. The figures indicate that there are two databases (database 1 and database 2) due to the movement of system servers. However, both databases have the same problem items.

The two figures suggest that CAR values have improved over time, although small variations can be seen in the linear algebra subject. Therefore, to get a rough idea of these trends, we have compared CAR values between the two databases, i.e., data before fiscal year 2019 and after fiscal year 2020. Then, we have found the following.

2.5 Comparison of CAR values for data before 2019 and after 2020

Table 2 shows CAR value comparison between the two databases. We can see that CAR values before 2019 are smaller than 0.5, while those after 2020 are located near 0.5. Since the testing system adopts the adaptive style, it is efficient and effective that CAR value is 0.5. This means that the ability of an

examinee is matched to the difficulty of the questions, allowing the system to automatically select the right questions for each examinee's ability which is estimated at each question. In addition, CAR value improvement suggests that the examinees have learned a great deal from the explanations of the questions and solutions provided by the online testing system.

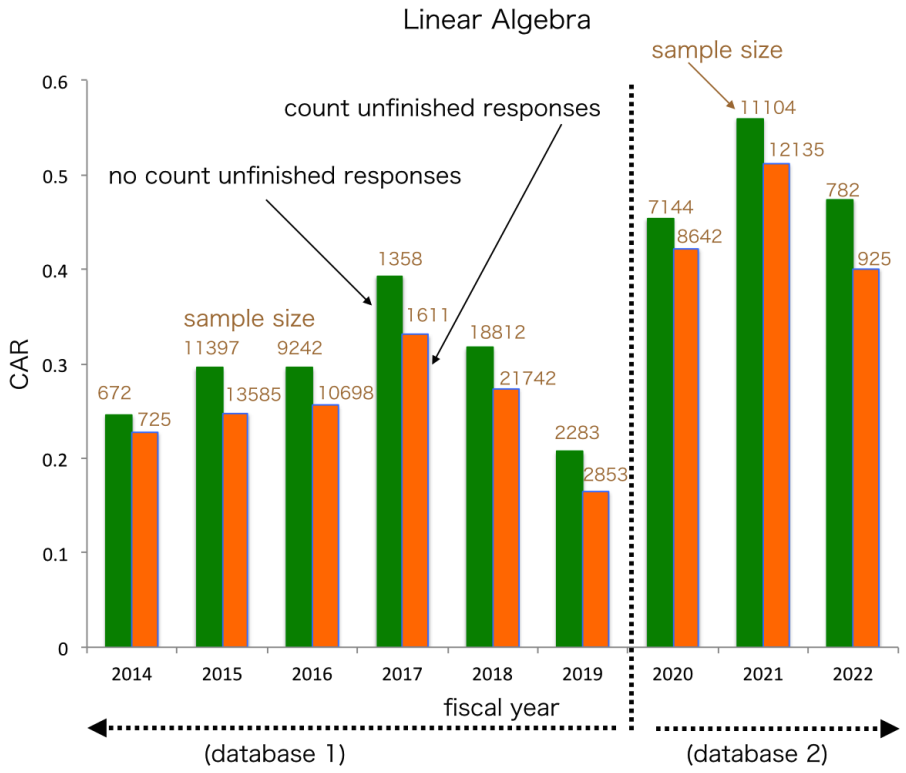

Figure 3: Difference of two CAR annual trend (linear algebra).

Table 2: Statistical test results for learning improvement.

	before 2019			after 2020		
subject	correct	incorrect	CAR	correct	incorrect	CAR
Linear Algebra	13244	30520	0.303	9600	8928	0.518
Probability & Statistics	156	734	0.175	38	56	0.404
Calculus	698	1141	0.380	14268	15194	0.484
Physics	203	499	0.289	37	51	0.420
Basic Analysis				644	527	0.550
ODE				20	27	0.426

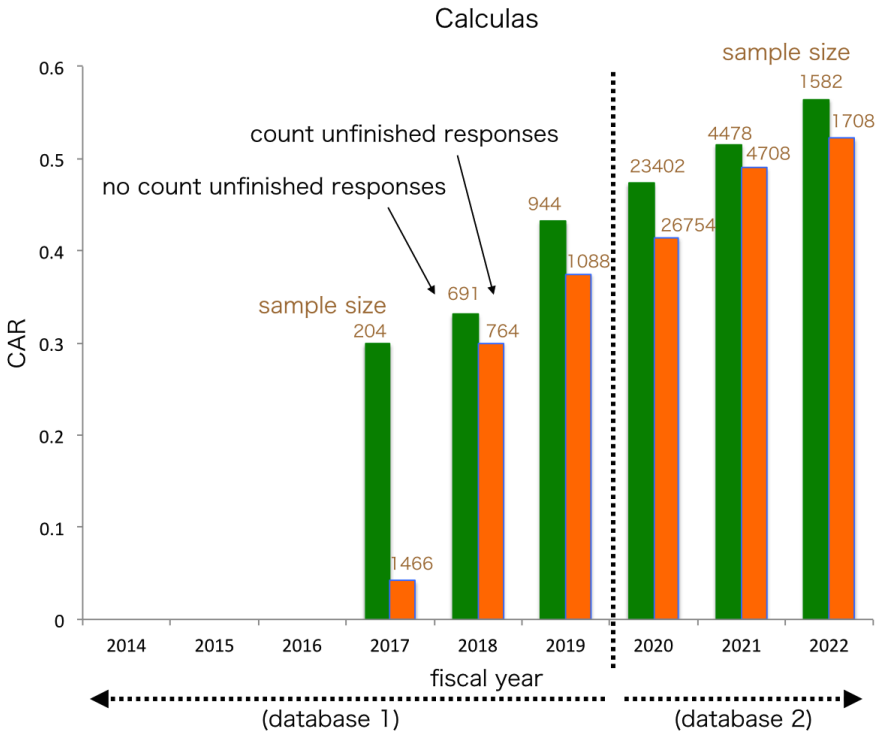

Figure 4: Difference of two CAR annual trend (calculus).

2.6 Statistical tests

Whether such item exposure effects are statistically significant or not, we have made statistical tests by using 2-way contingency table analysis. Dividing the CAR value for database 2 (2020 and beyond) by the CAR value for database 1 (pre-2019) yields the CAR ratio, which is analogous to the relative risk in a two-way contingency table analysis.

$$\text{CAR ratio} = \frac{\text{CAR value for database 2 (2020 and beyond)}}{\text{CAR value for database 1 (pre-2019)}}. \tag{4}$$

Moreover, like the common use of contingency table analysis, we will also apply for improvement inclination. The odds represents the ratio of the number of correct answers to incorrect answers,

$$\text{odds} = \frac{\text{number of correct answers}}{\text{number of incorrect answers}}, \tag{5}$$

and the odds ratio can be defined as

$$\text{odds ratio} = \frac{\text{odds for database 2 (2020 and beyond)}}{\text{odds for database 1 (pre-2019)}}. \tag{6}$$

Since there are no data in database 2 for the subjects of basic analysis and ODE, we deal with the other four subjects as shown in table 3.

We will use these ratios for finding improvement sign whether examinees have learned or not. Table 3

shows the CAR ratios and their 95% lower confidence limits as well as the odds ratios and their 95% lower confidence limits. Looking at the table, we see that all the 95% lower confidence limits are greater than 1, indicating that learning has increased with the use of the web-assisted online testing. These contingency table analysis results are in good agreement with the results using the bootstrap method [5, 6], and thus they are both reliable.

Table 3: Statistical test results for learning improvement.

	before 2019		after 2020	
subject	CAR ratio	95% lower CL	odds ratio	95% lower CL
Linear Algebra	1.712	1.679	2.473	2.392
Probability & Statistics	2.306	1.684	3.193	2.043
Calculus	1.276	1.203	1.535	1.393
Physics	1.454	1.071	1.783	1.133

3. Discussions

3.1 Distributions of CAR values and odds ratios

Based on the above analysis, it appears that, on average, learning can be improved by exposing the explanations of the problems and solutions. However, this item bank contains both difficult and easy questions together. Then, we may think that there could be some differences in responses between the difficult problems and easy problems.

Figures 5 compares the CAR ratios using database 1 and database 2 for each problem item, in the case of linear algebra. That is, CAR_{item}, in precise terms, are compared. The figure shows that for many problem items, the CAR values in database 1 are larger than the CAR values in database 2, but some of them have the opposite trend. On average, many of the CAR values lie near a straight line with a tangent of 1.7, which corresponds to the CAR ratio value of 1.7 in Table 3.

Figures 6 compares the log odds ratios for each problem item in the case of linear algebra. In the figure, vertical axis represents the odds ratio of database 1, and horizontal axis the odds ratio of database 2. The intersection of the vertical axis and the regression line deviates from 1 to the left, meaning that the odds ratio of database 2 is greater than that of database 1. These two plots show the rough distributions of CAR ratios and odds ratios, although the investigation is CAR_{item}-specific.

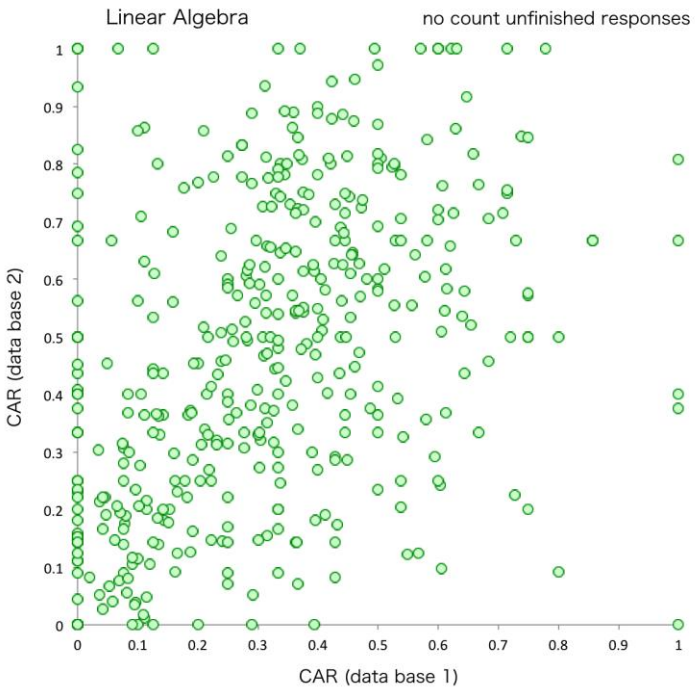

Figure 5: Comparison of CAR ratio between database 1 vs. database 2 (linear algebra).

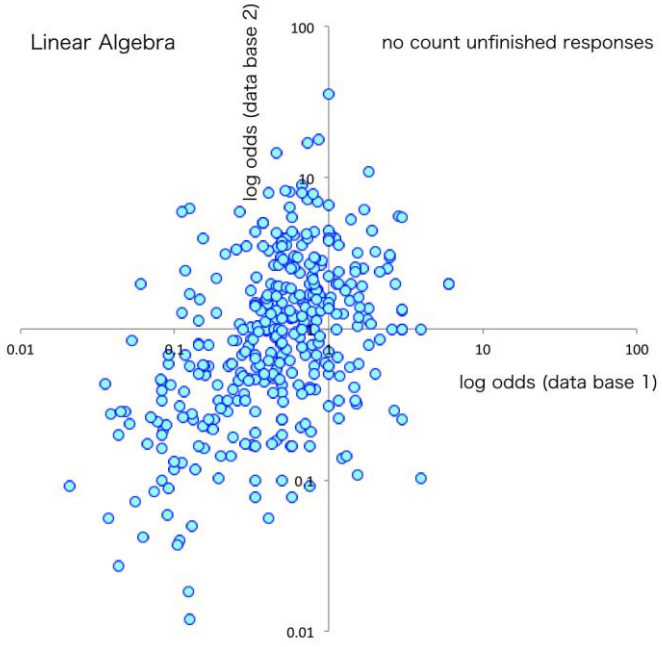

Figure 6: Comparison of odds ratio between database 1 vs. database 2 (linear algebra).

3.2 Relation between the CAR value improvement and effectiveness of learning

For subjects in the web-assisted online adaptive tests, we have seen that the CAR values in database 2 have been improved over the CAR values in database 1, as far as the two databases can be compared. This could be interpreted as indicating that learning with the online adaptive test is effective. However, it is not possible to determine to what extent the ability of each examinee has improved. At the very least, understanding the explanations of the problems and their solutions must have helped students learn. Further study is needed on the ability improvement by rigorous testing using general item response theory. Because of the time and effort required, the issues will be the subjects of future work.

4. Concluding Remarks

In rigorous assessment tests using item response theory, question items are usually hidden so that they can be used multiple times, which would seem to increase the reliability of the estimated results. However, in order to help students make progress, it is often required that the explanations to the questions and solutions are disclosed in the adaptive tests. Then, the reliability of the estimated ability and difficulty values is a concern.

Even when complete response matrices are available, research on the usefulness of item disclosure on large-scale tests is limited. In addition, whether learning with adaptive online tests is effective is important, research findings appear to be scarce.

This paper has investigated learning effectiveness in the adaptive online testing under the condition that the questions and answers are disclosed. Since it seems difficult to analyze estimates for parameters in item response theory, such as the difficulty parameters, this paper has dealt with the correct answer rate in order to discuss the effectiveness of learning. In the statistical tests using the two-way contingency table analysis, odds ratios have been also dealt with.

Using the database of large-scale web-assisted online adaptive tests administered to undergraduate students, it has been shown that exposure to questions and answers in adaptive online testing contributes to student progress.

5. Acknowledgment

The authors would like to thank professors, Yoshihisa Sato, Toshiharu Ikeda, Masayuki Hirokado, Yoshiaki Okazaki, Natsuo Saito, Tomoaki Okayama, Makoto Tagami, Tomokazu Fujino, Tetsuya Koyama, Yusuke Yamauchi, Masanori Takatou, Tooru Iwasa, Seiichi Kuwata, Chikoo Oosawa, Koichi Tanaka, for their contributions in preparing the problem items and in writing the explanations to solutions.

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] R. D. Bock, E. Muraki, W. Pfeifferberger, Item Pool Maintenance in the Presence of Item Parameter Drift, Journal of Educational Measurement, 25, 1988, 275-285.
[4] A.P.Dempster, N.M.Laird, andD.B.Rubin, Maximum Likelihood from Incomplete Data via the EM Algorithm, Journal of the Royal Statistical Society, Series B, 39, 1977, 1-38.

[5] B. Efron, Bootstrap Methods: Another Look at the Jackknife, The Annals of Statistics, 7, 1979, 1-26.
[6] B. Efron and R. J. Tibshirani, An introduction to the bootstrap, New York: Chapman & Hall, 1993.
[7] R. A. Feinberg, M. R. Raymond, S. A. Haist, Repeat Testing Effects on Credentialing Exams: Are Repeaters Misinformed or Uninformed?, Educational Measurement: Issues and Practice, 34, 2015, 34-39.
[8] A. I. Ferreira, L. S. Almeida, and G. Prieto, The Role of Processes and Contents in Human Memory: An Item Response Theory Approach, ISRN Education, 23, 2011, 873-885.
[9] J. S. Gilmer, The Effects of Test Disclosure on Equated Scores and Pass Rates, Applied Psychological Measurement, 13, 1989, 245-255.
[10] R. Hambleton, H. Swaminathan, and H. J. Rogers, Fundamentals of Item Response Theory, Sage Publications, 1991.
[11] H. Hirose and T. Sakumura, Test evaluation system via the web using the item response theory, in Computer and Advanced Technology in Education, 2010, 152-158.
[12] 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, 8-12.
[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, 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, 439-444.
[16] 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, 19-34.
[17] 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, 142-148.
[18] W. J. D. Linden and R. K. Hambleton, Handbook of Modern Item Response Theory, Springer, 1996.
[19] W. J. van der Linden, Handbook of Item Response Theory, Chapman and Hall/CRC, 2016.
[20] Y. S. Park and E. B. Yang, Three Controversies over Item Disclosure in Medical Licensure Examinations, Medical Education Online, 23, 2015, 1-5.
[21] M. R. Raymond, K. A. Swygert and N. Kahraman, Psychometric Equivalence of Ratings for Repeat Examinees on a Performance Assessment for Physician Licensure, Journal of Educational Measurement, 49, 2012, 339-361.
[22] L. A. Reyes, Considerations of Reusing Multiple Choice Items to Assess Medical Certification Repeat Examinees, Dissertation for the degree of Doctor of Philosophy in Educational Psychology University of Illinois at Chicago, 2018, 1-176.
[23] T. Sakumura, T. Kuwahata and H. Hirose, An Adaptive Online Ability Evaluation System Using the Item Response Theory, Education & e-Learning, 2011, 51-54.
[24] 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, 17-26.
[25] H. Selvi, Should Items and Answer Keys of Small-Scale Exams Be Published?, Higher Education Studies, 10, 2020, 107-113.
[26] X. Tang and M. Schultz, The Effect of Repeat Exposure to Simulation Based Items, Practical

Assessment, Research, and Evaluation, 25, 2020, 1-10.
[27] W. J. J. Veerkamp and C. A. W. Glas, Detection of Known Items in Adaptive Testing with a Statistical Quality Control Method, Journal of Educational and Behavioral Statistics, 25, 2000, 373-389.
[28] M. Wagner-Menghin, I. Preusche, and M. Schmidts, The Effects of Reusing Written Test Items: A Study Using the Rasch Model, ISRN Education, 2013, Article ID 585420, 1-7.
[29] T. J. Wood, The Effect of Reused Questions on Repeat Examinees, Advances in Health Sciences Education, 14, 2009, 465-473.