九州大学学術情報リポジトリ
Kyushu University Institutional Repository

DUALLY ADAPTIVE ONLINE IRT TESTING SYSTEM

Hirose, Hideo
Department of Environmental Design, Hiroshima Institute of Technology

https://doi.org/10.5109/2203024

出版情報：Bulletin of informatics and cybernetics. 48, pp.1-17, 2016-12. Research Association of Statistical Sciences
バージョン：
権利関係：

Bulletin of Informatics and Cybernetics, Vol. 48, 2016

DUALLY ADAPTIVE ONLINE IRT TESTING SYSTEM

By

Hideo HIROSE*

Abstract

The item response theory (IRT) provides us not only the abilities of examinees but also the difficulties of items (problems), and it is believed that the estimated abilities are fairer and more accurate than those obtained by the classical test methods. Using the estimated difficulties, we can construct an adaptive online testing system such that the system sequentially selects the most appropriate items to examinees automatically, resulting more accurate ability estimation and more efficient test procedures, where the term "adaptive" means the adequate item selection at each item selection step. However, as the number of examinees is growing in online testing, the difficulty values measured previously will possibly differ from those assessed by the new examinees. Then, calibration of the difficulty values may be required. For such conditions, we propose to use the dually adaptive online IRT testing system, where "dually adaptive" means that one is targeted to the adequate item selection and the other is targeted to the adjustment of the difficulty values for items. The key idea of this is to use the incomplete matrix completion. Using the proposed method, new items can be added and their difficulties are optimally adjusted without equating. We applied this method to mathematics testing cases, and we found that the system worked well.

Key Words and Phrases: item response theory; online adaptive testing; matrix completion; dually adaptive online IRT; item registration function.

1. Introduction

To evaluate the examinees' abilities accurately and fairly, the use of the item response theory (IRT) is considered to be one of the fundamental methods because it provides us the difficulties of the test items (problems) and the examinees' abilities together Ayala (2009), Hambleton et al. (1984), Hambleton et al. (1991), Linden et al. (1996). Once the difficulty values are obtained somehow, e.g., by using a monitor test, we can construct an adaptive online testing system which selects the most appropriate items to examinees automatically, resulting more accurate ability estimation and more efficient test procedures Barla et al. (2010), Chang et al. (2009), Kuo et al. (2013), Mills et al. (2002), Rajamani et al. (2013), Li et al. (2011). However, such an expectation is confirmed under the condition that the abilities of the adaptive online test examinees and those of the monitor test examinees are similar to each other because the difficulty values are static. Otherwise, calibrating the difficulty values should be required to those

* Department of Environmental Design, Hiroshima Institute of Technology 2-1-1 Miyake, Saeki, Hiroshima 731-5193 Japan. tel +81–82–921–9164 h.hirose.tk@hiroshima-it.ac.jp

who show different ability values from the monitor test examinees. Calibration of the difficulty values is often dealt with equating methods, and we have to find new monitor examinees in equating. However, it seems difficult to do that in the traditional equating because 400 samples are required even in Rasch model Kolen et al. (2004).

In actual cases, as the number of examinees is increasing, the difficulty values measured by the monitor test will possibly differ from those assessed by the new examinees. Then, it may be beneficial to use the item response results in adaptive testing to calibrate the difficulty values in dynamic. To do that, we have to obtain the difficulty parameters from the incomplete item response matrix. However, traditional IRT methods are incompetent to that problem.

Supposing another situation that we want to add new items to the item bank, and that we do not know the difficulty values of the items. This would be occurring when we accept a system that includes the item registration function which is available to item contributors. Using the temporary values for unknown difficulties, we may obtain the item responses to these new items. Since the item response matrix will become incomplete, it may beneficial to deal with the incomplete matrix.

Therefore, we propose the dually adaptive online IRT testing system, where "dually adaptive" means that one is targeted to the adequate item selection and the other is to the adjustment of the difficulty values for items Hirose, Aizawa (2014), Hirose et al. (2014), Hirose, Tokusada (2014). This system uses the matrix completion methods which can estimate the item difficulties and examinees' abilities altogether from incomplete item response matrices, which overcomes the problem of incomplete matrix in the traditional IRT methods. The relevant studies are seen in Eggen et al. (2011), Hirose et al. (2012), Little et al. (2002), Mazumder et al. (2010), Nydick et al. (2009).

The proposed system enables us to allow a novel aspect in using the item bank. Since the new items can be added at anytime and their difficulties are optimally adjusted without equating, we no longer require the additional monitor tests to new different examinees. This is new. In this paper, we illustrate that the system works by applying the method to mathematics testing cases.

2. A brief review to common item response theory and adaptive testing

2.1. Common Item Response Theory

We assume that an examinee (a student) i having ability θ_i takes a problem j. If the examinee is successful in giving the correct answer with probability P, such that

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

the likelihood for all the examinees, $i = 1, 2, \ldots, N$, and all the items, $j = 1, 2, \ldots, n$, will become

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

where $\delta_{i,j}$ denotes the indicator function such that $\delta = 1$ for success and $\delta = 0$ for failure; a_j, b_j , and c_j are constants in the logistic function, and they are called the

discrimination parameter, the difficulty parameter, and shift parameter, respectively. In a sense, $P_{i,j}$ in Equation (1) is a logistic probability distribution function with unknown parameters a_j, b_j, and c_j, and the random variable is θ_i. However, θ_i is also unknown here. With observed values of $\delta_{i,j}$, the maximum likelihood estimates for a_j, b_j, c_j and θ_i are obtained by maximizing L in Equation (2). If we assume $c_j > 0$, this is called the three-parameter IRT, and when $c_j = 0$, the two-parameter IRT. In addition, if we deal with the case of $a_j = 1$, this is called the one-parameter IRT, the Rasch model. We mainly deal with the case of the two-parameter IRT here. Figure 1 shows the concept of the common IRT parameter estimation in the two-parameter IRT case.

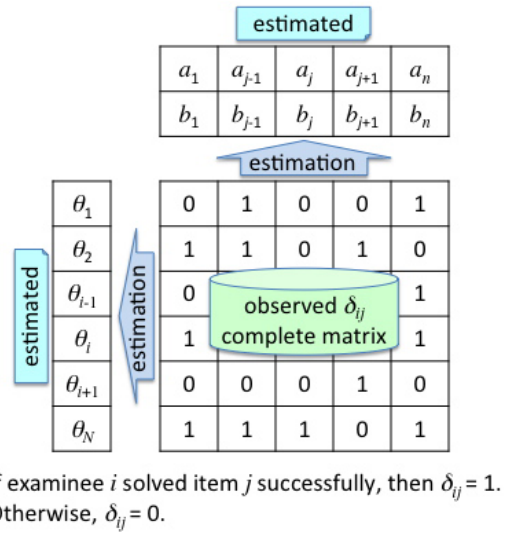

Figure 1: Common item response theory parameter estimation.

2.2. Adaptive Online Ability Evaluation Procedure

Assuming that item difficulty parameters are already obtained somehow in advance, e.g., by using a monitor test, then we can estimate the parameters θ_i by using these parameters. This is not a difficult task because the number of unknown parameters is only one. The conventional adaptive online systems use this kind of procedure.

A typical adaptive online testing is shown in Figure 2, where the most appropriate items are automatically chosen at each time according to the responses of the examinee. Usually, optimal selection of the items is executed by finding the maximum Fisher information to each item. In the figure, the very first problem is successfully solved, then the system provides the more difficult problem to the examinee. The level of the difficulty is set to around the estimated examinee's ability. This means that the probability to solve the given problem is nearly 0.5. We continue this procedure until the appropriate number of iterations.

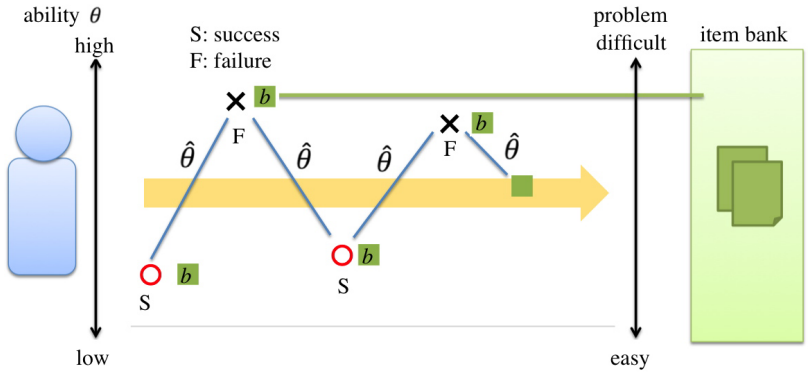

Figure 2: Typical conventional adaptive testing.

3. Dually adaptive online IRT testing

As explained before, as the number of examinees is growing, the difficulty values measured by the monitor test will possibly differ from those assessed by the new examinees (case 1). Then, calibrating the difficulty values may be required. In addition, when we want to add new items to the item bank, the difficulty values for them have to be adjusted appropriately (case 2). For such conditions, we have to deal with the incomplete matrix consisting of 0/1 observed responses and unobserved null responses. That is, in case 1, the examinees do not necessarily solve all the items in the adaptive testing, and in case 2, the parameter values for the item difficulties will not be revealed unless some monitor tests are performed. In such situations, the common IRT method is incompetent.

Then, we propose to use the matrix completion method for an incomplete matrix Hirose et al. (2012). The principle methodology for this will be explained in Appendix. See Figure 3. In the figure on the bottom, vacant elements mean that the corresponding items are not tackled; in the middle of the figure, numbers represented in real number larger than 0 and smaller than 1 mean the estimated response values by using the matrix completion method. As explained in Appendix, We allow the real number response values here.

Using this method in the adaptive online tests that allow the inclusion of the new items, we can construct the dually adaptive online IRT testing system, where "dually adaptive" means that one is targeted to the adequate item selection and the other is targeted to the adjustment of the difficulty values for items. The monitor tests are then not necessarily required for the newly added items in the proposed system.

4. Dually adaptive online IRT system configuration

Figure 4 shows the configuration of the dually adaptive online IRT testing system. The system is installed in a cloud system, and is connected to the internet. The system consists of the IRT computing core part, IRT user response database part, and IRT item bank part. There are four kinds of persons who play a role regarding the system: 1) examinees, 2) item contributors, 3) system manager, and 4) supervisor. They can access

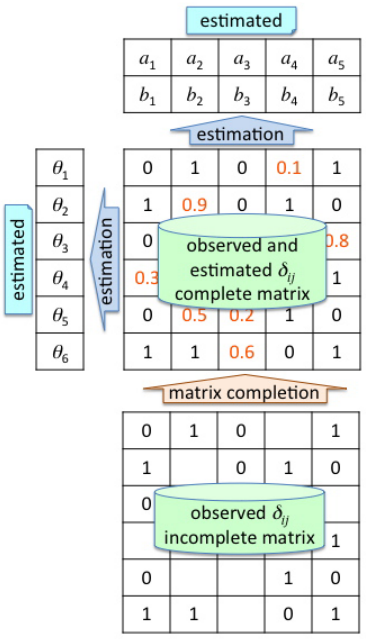

Figure 3: Dually adaptive online IRT testing using the matrix completion method.

to this system via the internet. Now, we define the functions of the system and the roles of the users of this system.

4.1. Function of the System

— IRT computing core part —

The common IRT computational method and the adaptive online IRT computing method targeted to the adequate item selection. In addition, the system has a function of the dually adaptive IRT computation, where the adequate item selection and the adjustment of the difficulty values of items are dually performed.

— IRT user response database part —

Examinees, responding the questions appropriately provided by the system, leave their marks on the database corresponding to the item-user matrix. This database grows as the examinees take new tests and the system accepts the new examinees.

— IRT item bank part —

This system allows item contributors to submit new items anytime. The IRT item bank usually has the non-variant item difficulty values in providing appropriate questions to examinees in the adaptive test. However, as the number of examinees grows and the new items are added to the system, item difficulty values should be updated (calibrated). The database grows as the system accepts new items.

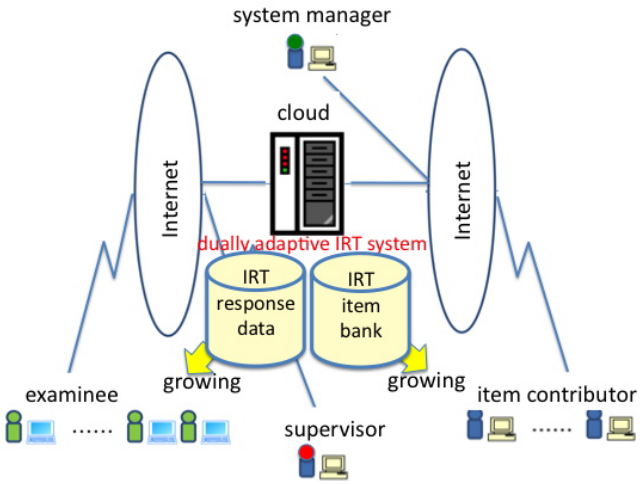

Figure 4: Dually adaptive online IRT system configuration.

4.2. User Roles in the System

— **Examinees** —

Examinees are persons who take examinations via the internet. They are registered in advance or as necessary. Those who want to be enrolled in the system can take examinations whenever a security check is passed.

— **Item contributors** —

This role is new. The system allows anyone who wants to submit new items to contribute to the item bank. Thus, the item bank is growing as a new item is registered to the system. The system stocks such items for some period (usually 24 hours), and the items are updated by using the incomplete matrix completion method. This update would be performed at midnight once a day.

— **System manager** —

The system manager usually works when the system is launching, when some fault happens, and when a major system update is needed. There is no duty, otherwise.

— **Supervisor** —

The supervisor will find possible failures and make update the system by watching the system carefully.

4.3. Processes in the System

There are mainly three processes in the system: 1) adaptive testing, 2) item registration, and 3) calibration.

— **Adaptive testing** —

Adaptive testing is performed in a conventional manner. That is, the most appropriate question is selected by analyzing the past history of responses in a sequence of testing. The system will open the gate whenever the applicant wants to enter the

system. This means the system allows the parallel executions.

— **Item registration** —

Teachers who want to contribute new items to the system can submit items when they are permitted to access the system. The item registration in detail is shown later.

— **Calibration** —

Periodically, the system makes a calibration for the item difficulties and user abilities, resulting the update of the item difficulty values. Next day's testing is performed using the latest values.

5. Typical procedure of the dually adaptive online IRT testing system

We introduce here a typical example of the dually adaptive online IRT testing system. We assume four kinds of persons as stated before. However, we only show the two roles, the examinees and the item contributors.

5.1. An Examinee Taking a Test

By clicking the examinee button, they can enter the adaptive test course. The system asks the fields to be tested. By clicking one of the button, they can take the adaptive test in their field.

The system provides a question as shown in Figure 5. In this case, the examinee can select one button which seems to be a correct answer. On the top of the figure, the icon indicating the previous successful result is shown.

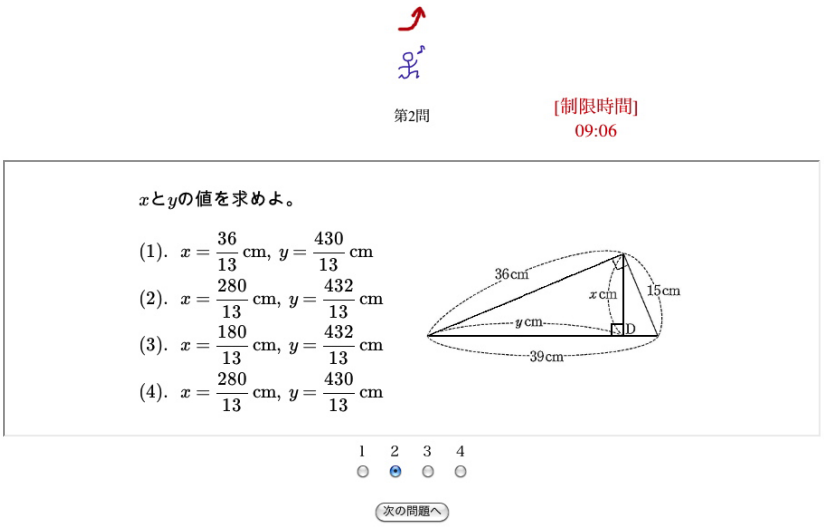

Figure 5: An example of the question.

Finally, the system gives the final result to the examinee as shown in Figure 6. In this page, he can also learn the correct answer in detail.

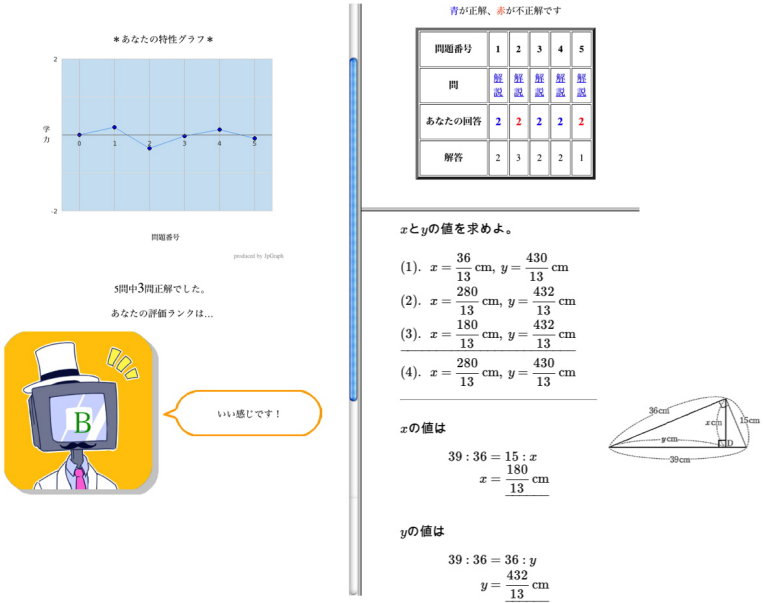

Figure 6: The system gives the result.

5.2. An Item Contributor Submits an Item

From the welcome page, the item contributor can go into the item registration page as shown in Figure 7. He provides the pdf file which includes the problem and the answer in separate pages. He can select the question style either 1) selecting the button or 2) choosing the number, character, or sign. In the illustrative case in the figure, he chose two numbers (one is the ten's place, and the other is the one's place). On the bottom of the page, an explanation to the answer is shown.

The registration system has functions of 1) initial registration, 2) modification, 3) adding, 4) deleting, and 5) confirming.

6. Applications

We developed the dually online adaptive system, and applied it to high-school mathematics testing. Here, we show the effects of the proposed system, 1) by comparing the evaluations using the monitor test and those using the actual tests, and 2) by comparing the evaluations using the monitor test and those using the actual tests with monitor test. There were three fields in the tests: a) algebra, calculus, trigonometry (these may be in memory subjects), b) geometry (these may require some inspiration), and c) logic, probability (these may be in logical thinking).

6.1. Monitor Test Evaluation and Actual Calibration Test Evaluation

We had two high-school student test cases in addition to the monitor test. One test case was taken in 2013, and the other in 2014. In 2013, we prepared the monitor

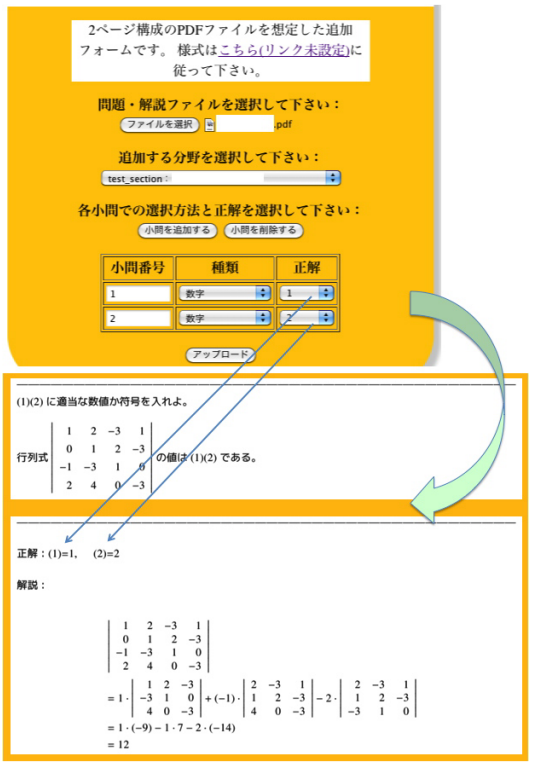

Figure 7: Item registration.

test as common adaptive tests do. The number of items provided are 37, 32, and 30 to each a), b), c) field, respectively. The monitor test was executed with a help of university students (more than 30 students from undergraduate to graduate). As usual, we obtained the difficulty values estimated by using the equating method. Here, we call this "monitor".

We have 84, 32, and 22 examinees (called "2013 examinees") to each field in the 2013 high-school student test. The ability evaluation in the adaptive testing can be accomplished by using the difficulty values via the monitor. After the test was over, we estimated the calibrated item difficulties and the examinee abilities together for the obtained incomplete response matrix. That is, we had the different difficulty values from the monitor. This is called the 2013 calibration. The ability evaluation can be accomplished by using the difficulty values 1) via the monitor, and 2) via the 2013 calibration. In this computation, we note that only the 2013 examinees were used.

In 2014, we have added the new items to each field such as 44 to a), 13 to b), and 33 to c), and the total number of items becomes 81 to a), 45 to b), and 63 to c), respectively. We did not prepare the monitor test for these newly added items, and we managed to estimate the difficulty values. In 2014, we had two test days. We used the initial difficulty values of $a = 1, b = 0$ for all the new items in testing. Then, we try to update the difficulty values using the obtained response pattern via the dually adaptive system. The numbers of examinees to a), b), c) fields on the very first day were 15, 10, 5, respectively. Using this day's examinees and 2013 examinees, we estimated the difficulty values for the newly added items; for items without any traces of tackling, the difficulty values remain the same values. In two days in 2014, we had 47, 43, and 19 examinees (called "2014 examinees") to each field in 2014 high-school student test, and the total number of examinees in 2003 and 2014 is 131 to a), 75 to b), and 41 to c), respectively. This is called the 2014 calibration. That is, the 2014 calibration consists of 2014 examinees and 2013 and 2014 items. The ability evaluation can be accomplished by using the difficulty values 1) via the monitor, and 2) via the 2014 calibration. In this computation, we note that only the 2014 examinees were used.

See Table 1 for the number of items and examinees, and see Figure 8 for the matrices configured by various items and examinees.

Table 1: Number of items and examinees.

year	field	category	# items	# examinees
2013	total		99	138
	a) calculus etc.	memory	37	84
	b) geometry	inspiration	32	32
	c) logic etc.	thinking	30	22
2014	total		90	109
	a) calculus etc.	memory	44	47
	b) geometry	inspiration	13	43
	c) logic etc.	thinking	33	19
2013	total		189	247
&	a) calculus etc.	memory	81	131
2014	b) geometry	inspiration	45	75
	c) logic etc.	thinking	63	41

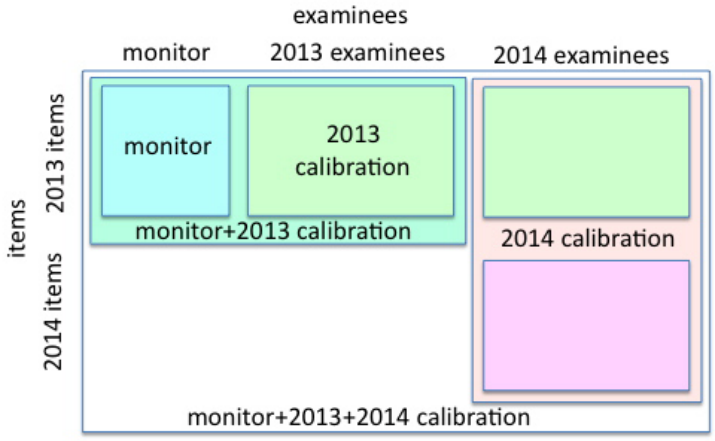

Figure 8: Matrix configurations by various items and examinees.

By using the incomplete response matrix obtained in 2013 and 2014, we can estimate the calibrated difficulty values of a and b by the matrix completion. If the abilities of the examinees in the monitor test are located at the same (or similar) level to the high-school examinees, the monitor difficulties are similar to the difficulties obtained by using the high-school examinees. Figure 9 shows each b value in the algebra, calculus, trigonometry field in 2013 and 2014. We can see that some items of b values in 2014 were updated and others were not updated, which means that the calibration worked. Figure 10 shows the comparison of the difficulties of b in the algebra, calculus, trigonometry field. The horizontal axis expresses the value obtained by the monitor test (monitor equating method). The vertical axis expresses the value by the calibration. The figure indicates that there seems no strong relationship between the monitor results and calibration results. This is because the examinees of the monitor test and those of the actual tests in 2013 and 2014 were different from each other.

This may recall us that the estimated abilities differ from each other if the difficulty values are different because the estimates for abilities are affected by the difficulty values to some extent. Figure 11 shows the comparisons of the abilities between the use of the monitor and that of the 2013 and 2014 calibrations. The figure indicates, however, that the monitor result and the calibration result are strongly correlated in both years. This may suggest that the difficulty values will not affect the ability values much. However, also by looking at the figure, we see that the ability orders can easily be changed by the difficulty values. The accurate estimates for the difficulty values are still required for accurate ability estimation.

6.2. Monitor Test Evaluation and Actual Calibration Test Evaluation with Monitor Test

In the previous subsection, we did not use the monitor data in the calibration, resulting that the difficulty values from the monitor and those from the calibration were different from each other. In a sense, this can be understood because there was no room that the monitor test and the high-school student tests interfered.

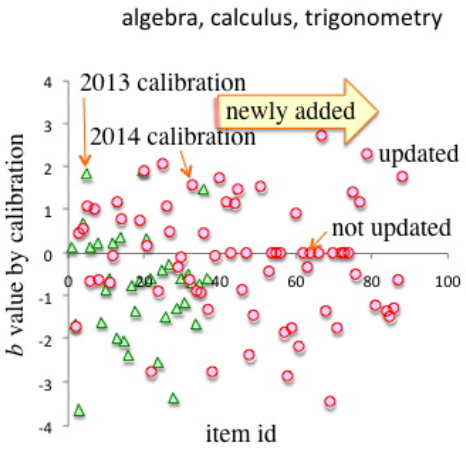

Figure 9: Comparisons of the difficulties of b to each item.

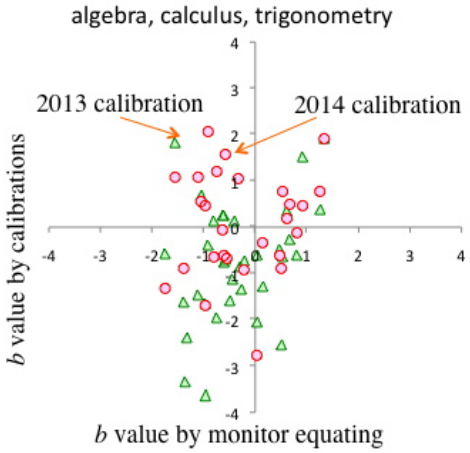

Figure 10: Comparisons of the difficulties of b: monitor to calibrated.

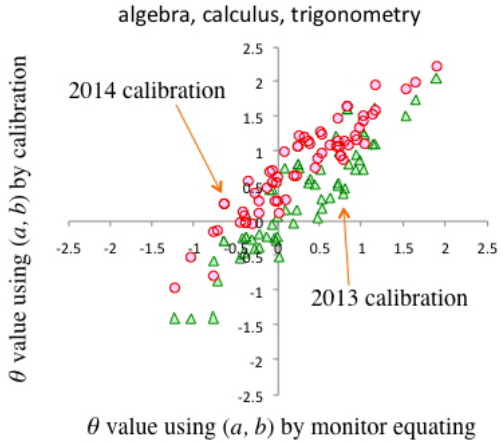

Figure 11: Comparisons of the abilities between the use of the monitor and the calibration.

In the actual application to the dually optimal adaptive testing, we can easily imagine that the new items will be added and the new examinees will join the system. The difficulty parameters should be updated seamlessly. Thus, we performed the dually adaptive procedure, including the monitor test result.

Figure 12 shows the comparison of the b value. Horizontal axis means the difficulty b obtained by the monitor equating. Vertical axis means the difficulty b obtained by the calibration with the monitor (monitor+2013 calibration and monitor+2013+2014 calibration in Figure 8). Figure 13 shows the comparison of the ability θ value. Horizontal axis means the ability θ obtained by using the monitor equating. Vertical axis means the ability θ obtained by the calibrations with the monitor (monitor+2013 calibration and monitor+2013+2014 calibration in Figure 8). From both the figures, we can see a much more natural relationship between the monitor and the calibrations with the monitor. This means that the calibration using the dually adaptive method works.

7. Discussion

When examinees use the adaptive online IRT system, we need not care about the computational cost because it depends only on the number of trials. However, as the numbers of items and examinees increase, i.e., the size of item response matrix increases, the computational cost for the IRT computation increases. This is related to the calibration procedure, and it may occur periodically, e.g., daily, weekly, or monthly. In the calibration procedure, the computatinal cost is proportional to the number of items and the number of examinees since we adopt the marginal likelihood in estimating the item parameters. In the future, however, we have to tackle with the huge size of response matrix such as $100,000 \times 10,000$. Here, we investigated the actual computational times to the small size of response matrices. The sizes of examinees and items are shown in Table 2. The computational times with no calibrations are also shown in Table 2; the machine is ThinkPad X240 with 7MB memory size. We may roughly estimate the computational

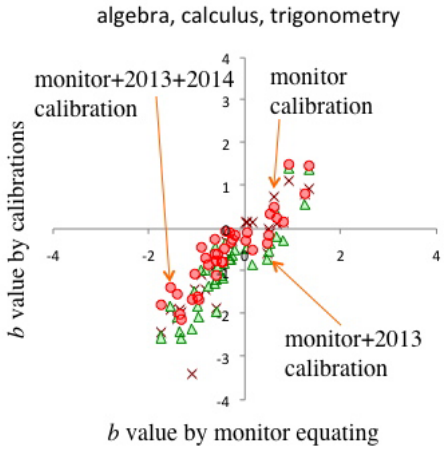

Figure 12: Comparisons of the difficulties of b: the monitor to calibration with the monitor.

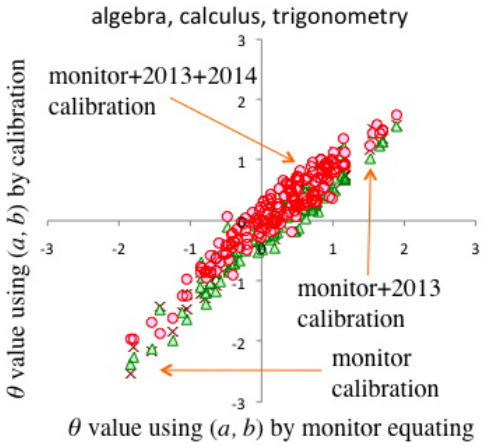

Figure 13: Comparisons of the abilities between the use of the monitor and the calibration with the monitor.

time for the matrix $100,000 \times 10,000$ is 280×10^3 using case id=5. This approximately corresponds to three days computation. Considering that calibration procedure requires 10 to 100 times to ordinary IRT full matrix computation. The next challenge will be the computational cost reduction.

Table 2: Computational cost.

case id.	# items	# examinees	# elements	time*
1	100	100	10000	3.9
2	1000	100	100000	38.0
3	1000	200	200000	60.1
4	1000	500	500000	145.5
5	1000	1000	1000000	282.3
6	1000	2000	2000000	567.0

*: second

8. Concluding remarks

We proposed the dually adaptive online IRT testing system, where "dually adaptive" means that one is targeted to the adequate item selection and the other is targeted to the adjustment of the difficulty values for items. The proposed system can estimate the difficulty values for the newly added items as well as the adjustment of item difficulty values for the items already existing appropriately. The key idea of this is to use the incomplete matrix completion. This is a dynamic system in a sense that the item bank grows automatically unlike the traditional adaptive online testing. Obviously, the monitor tests are not necessarily required for the newly added items in the proposed system, and we no longer require the equating. We applied this method to mathematics testing cases, and we found that the system worked well.

9. Appendix

9.1. Matrix completion by using the EM-type IRT

We briefly introduce parameter estimation procedure here. In using Equation (2), it is common that the value of δ should be an indicator such that $\delta = 1$ when successfully solved and $\delta = 0$ otherwise. However, we can admit the value of δ as a rational number, if we regard the value of $\delta = l/m$ such that an examinee solved l times successfully out of m given different problems which have the same difficulty. As m goes to infinity, we can admit the value of $\delta \in [0, 1]$ as close as to any real number.

First, we fill $\delta^0_{i,j} \in [0, 1]$ with some initial values to the vacant elements; observed values of $\delta_{i,j} = 0, 1$ remain the same. These values are any values as long as $0 \leq \delta^0_{i,j} \leq 1$ holds. For example, μ_j, mean value for successful ratio to problem j, μ_i, mean value for successful ratio to examinee i, or μ, mean value for successful ratio to all the observed cell values can be used. Thereby, the initial log-likelihood value L^0 is obtained from Equation (2) with appropriate initial values for problems and abilities, say, a^0_j, b^0_j, and θ^0_i.

Using this complete matrix of $\{\delta^0_{i,j}\}$, we can estimate the parameters of abilities

and difficulties together by maximizing the likelihood L in Equation (2), resulting L^1, a_j^1, b_j^1, and θ_i^1. This procedure can be a maximization procedure, either by the two-step algorithm or the MCMC method, i.e., the common IRT method.

Once, these temporal parameters are computed, we then compute the expected scores using the probability $P_{i,j} \in [0,1]$ to the target places using Equation (1). This $\hat{\delta}_{i,j} = P_{i,j}$ becomes temporal $\delta_{i,j}^1$, and this procedure can be considered as an expectation procedure.

We continue such a procedure, and obtain L^k, $\delta_{i,j}^k$, a_j^k, b_j^k, and θ_i^k $(k = 0, \ldots)$. Then, for any values of $\delta_{i,j}^0 \in [0,1]$, we could expect the converged values, (L^∞, $\delta_{i,j}^\infty$, a_j^∞, b_j^∞, and θ_i^∞). We cannot always guarantee that the converged values are uniquely determined Suen et al. (1994), Yen et al. (1991). However, we have experienced that the converged values are uniquely determined in many cases. These two-step procedure is called here the EM-type IRT method. We have investigated that the simulation study was consistent Sakumura et al. (2014).

References

Barla, M., Bielikova, M., Ezzeddinne, A., Kramar, T., Simko, M., Vozar, O. (2010). On the impact of adaptive test question selection for learning efficiency, *Computers and Education, 55, 846-857.*

Chang, H., Ying, Z. (2009). Nonlinear sequential designs for logistic item response theory models with applications to computerized adaptive tests, *The Annals of Statistics, 37, 1466-1488.*

de Ayala, R. (2009). The Theory and Practice of Item Response Theory, *Guilford Press.*

Eggen, T. J. H. M., Verhelst, N. D. (2011). Item calibration in incomplete testing designs, *Psicologica, 32, 107-132.*

Hambleton, R., Swaminathan, H., Rogers, H. J. (1991). Fundamentals of Item Response Theory, *Sage Publications.*

Hambleton, R. K., Swaminathan, H. (1984). Item Response Theory: Principles and Applications, *Springer.*

Hirose, H., Sakumura, T. (2012). An adaptive online ability evaluation system using the item response theory, *Proceedings of the IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE 2012), pp.8-12.*

Hirose, H., Aizawa, Y. (2014). Automatically Growing Dually Adaptive Online IRT Testing System, *Proceedings of the IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE 2014), pp.528-533.*

Hirose, H., Tokusada, Y., Noguchi, K. (2014). Dually Adaptive Online IRT Testing System with Application to High-School Mathematics Testing Case, *Proceedings of the IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE 2014), pp.447-452.*

Hirose, H., Tokusada, Y. (2014). A Simulation Study to the Dually Adaptive Online IRT Testing System, *Proceedings of the IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE 2014), pp.97-102.*

Kolen, M., Brennan, R. (2004). Statistical Analysis with Missing Data, 2nd Ed, *Springer.*

Kuo, C., Wu, H. (2013). Toward an integrated model for designing assessment systems: An analysis of the current status of computer-based assessments in science, *Computer and Education, 68, 388-403.*

Li, X., Wang, Z., Wu, X., Li, Y. (2011). The design of adaptive test paper composition algorithm based on the item response theory, *Proceedings of the 6th IEEE Joint International Information Technology and Artificial Intelligence Conference (ITAIC 2011).*

Linden, W. J. D., Hambleton, R. K. (1996). Handbook of Modern Item Response Theory, *Springer.*

Little, R. J. A., Rubin, D. B. (2002). Statistical Analysis with Missing Data, 2nd Ed, *Wiley.*

Mazumder, R., Hastie, T., Tibshirani, R. (2010). Spectral regularization algorithms for learning large incomplete matrices, *The Journal of Machine Learning Research, 11, 2287-2322.*

Mills, C. N., Potenza, M. T., Fremer, J. J. (2002). Computer-Based Testing: Building the Foundation for Future Assessments, *Lawrence Erlbaum.*

Nydick, S. W., Weiss, D. J. (2009). A hybrid simulation procedure for the development of cats, *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing, pp.1-23.*

Rajamani, K., K. V. (2013). An adaptive assessment system to compose serial test sheets using item response theory, *Proceedings of the International Conference on Pattern Recognition, Informatics and Mobile Engineering (ITAIC 2011).*

Sakumura, T., Tokunaga, M., Hirose, H. (2014). Making up the complete matrix from the incomplete matrix using the em-type irt and its application, *Transactions of Information Processing Society of Japan, TOM, 7, 17-26*

Suen, H. K., Lee, P. S. C. (1994). Constraint Optimization: A Perspective if IRT Parameter Estimation, *Objective Measurement: Theory Into Practice, 2nd Ed., pp.289-300.*

Yen, W., Burket, G., Sykes, R. (1991). Nonunique solutions to the likelihood equation for the three-parameter logisitic model, *Psycometrika, 56, 39-54.*

Received June 3, 2015
Revised January 14, 2016