INFORMATION, Vol.17, No.2, February 2014

Score Allotment Optimization Method with Application to Comparison of Ability Evaluation in Testing between Classical Test Theory and Item Response Theory

Hideo Hirose, Takenori Sakumura and Takayuki Kuwahata

International Information Institute

INFORMATION
Volume 17, Number 2, pp.391-409
ISSN 1343-4500

Score Allotment Optimization Method with Application to Comparison of Ability Evaluation in Testing between Classical Test Theory and Item Response Theory

Hideo Hirose, Takenori Sakumura, Takayuki Kuwahata

Department of Systems Design and Informatics,
Kyushu Institute of Technology,
Fukuoka 820-8502, Japan

Abstract

Many researchers know the superiority of the item response theory (IRT) over the classical test theory (CTT) from a detailed test-evaluation viewpoint. However, teachers are still reluctant to use the IRT as a daily testing tool. The primary objective of this paper is to find the difference between the CTT and the IRT. In particular, we focus on the difference in ability evaluation. We compared the CTT and IRT evaluated abilities by using the hypothetically assumed abilities that are mimicked to a real case. By using a simulation study, we found that the IRT is superior to the CTT to some extent. The CTT uses pre-assigned allotments contrary to the IRT which has no allotment concept. However, if we regard the ability evaluation by the IRT as the standard, we can find the most appropriate allotments in the CTT so that the total scores of the CTT are adjusted as close as possible to the abilities obtained by the IRT. This is a kind of allotment optimization problem. We show the methodology in this paper. By applying our methodology to some simulation cases that mimic the real data case, we found an intriguing feature with respect to the pre-assigned allotments. If teachers want to raise the examination pass rate, we guess that they give higher scores than the actual scores achieved by students; we call this jacking-up. Using the allotment optimization, we have found that jacking-up causes higher allotments to easier problems in the CTT.

Key words: Ability evaluation; Classical test theory; Item response theory; Allotment optimization; Jack-up; Least square; Gradient descent.

1 Introduction

For effective evaluation of students' abilities, the item response theory (IRT) Hambleton and Swaminathan [1984], Hambleton et al. [1991], Linden and Hambleton [1996] is useful because it brings the difficulties of the test problems and the students' abilities together. The IRT may also enhance the students' skills and evaluate their abilities more accurately when several adaptive e-learning systems Mills et al. [2002] and test methods are appropriately used Hirose [2011], Hirose and Sakumura [2012], Sakumura and Hirose [2010b], Tsukihara et al. [2009], Sakumura and Hirose [2010a]. A student self-learning system that is embedded in the e-learning system by using an adaptive test method Sakumura et al. [2011] is also effective in performing the optimal test in terms of time and costs.

The IRT has been widely used in official test systems, such as the Test of English as a Foreign Language (TOEFL) and the Test of English for International Communication (TOEIC), and is a preferred method for the development of high-stakes tests, such as Graduate Record Examination (GRE) and Graduate Management Admission Test (GMAT). The superiority of the IRT over the classical test theory (CTT) is valid for many subjects in universities and high schools. One of the reasons why this method has often been used in tests for English is that the examinees can solve many items in a certain time period, for example, in two to three hours, and thus the method can effectively and easily be applied in such kinds of tests. However, even in tests for mathematics such as analysis, linear algebra, or probability, where students cannot solve as many problems in such a short time, the IRT is still effective to assess accurately the students' abilities as long as the preparation for test problems is appropriate Hirose and Sakumura [2010]. That is, inclusion of high- and low-level test items together makes assessment accurate and fair to a variety of students.

However, even now, teachers in universities and colleges do not use the IRT. One reason may be due to custom behaviors. Another reason may be that the difference in the ability evaluation between the CTT and the IRT is not correctly known. Therefore, in this paper, we highlight on this matter by comparing the ability evaluation in testing between these two methods and show the results by using typical simulation studies. To do this, we have newly developed a method to compare the CTT and the IRT.

The remainder of the paper is organized as follows. In the next section, the CTT and the IRT are briefly reviewed. In section 3, we discuss the relationship between the score and the ability. In section 4, we compare the CTT abilities and the IRT abilities by using the hypothetically assumed abilities. In section 5, we introduce the methodology of allotment optimization and show

the results obtained by using this methodology. In section 6, we discuss the outcomes and in section 7, we provide the conclusions of this study.

2 Classical Test Theory and Item Response Theory

2.1 Classical Test Theory

Classical test theory assumes that each person has a true score T. However, this cannot be observed directly on a test. We only observe the score value X which consists of T and the error component E Lord and Novick [1968]. That is, $X = T + E$. The reliability of the observed test scores is defined as the ratio of the true score variance to the observed score variance such that $\rho_{XT}^2 = \sigma_T^2/\sigma_X^2 = \sigma_T^2/(\sigma_T^2 + \sigma_E^2)$.

We usually measure students' abilities by summing up scores of problems where allotments to each problem are given in advance. If we assign many problem items in testing that have equally likely difficulties, the reliability of the true scores will increase according to the central limit theorem.

We set the student identifier as i (total number is n), and the problem identifier as j (total number is m). In the CTT, the total score to student i is given by

$$\psi_i = \sum_{j=1}^{m} t_{i,j} = \sum_{j=1}^{m} q_j \delta_{i,j}, \tag{1}$$

where $t_{i,j}$ is the score that the student i obtained for problem j, $\delta_{i,j}$ denotes the indicator function such that $\delta = 1$ for success and $\delta = 0$ for failure, and q_j is the weight assigned to each problem.

2.2 Item Response Theory

In the IRT, we assume a student i having ability θ_i takes a problem j. If the student is successful in giving the correct answer with probability P, such that

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

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

$$L = \prod_{i=1}^{n} \prod_{j=1}^{m} P_{i,j}(\theta_i; a_j, b_j)^{\delta_{i,j}} \times \{1 - P_{i,j}(\theta_i; a_j, b_j)\}^{1-\delta_{i,j}}, \tag{3}$$

where a_j and b_j are constants in the logistic function, and they are called the discrimination parameter and the difficulty parameter, respectively. The larger the value of a_j, the more discriminating the item is, and the larger the value of b_j, the more difficult the item is. In a statistical sense, $P_{i,j}$ in Equation (2) is a logistic probability distribution function with unknown parameters a_j and b_j; the random variable is θ_i. However, a_j, b_j, and θ_i are all unknown here. By maximizing L in Equation (3), the maximum likelihood estimates may be obtained. Figure 1 shows an item response theory estimation procedure. In the figure, the 0/1 response matrix for incorrect/correct answers is substituted into the likelihood function (3), and by solving the log-likelihood equations corresponding to Equation (3), a_j, b_j, and θ_i are numerically obtained.

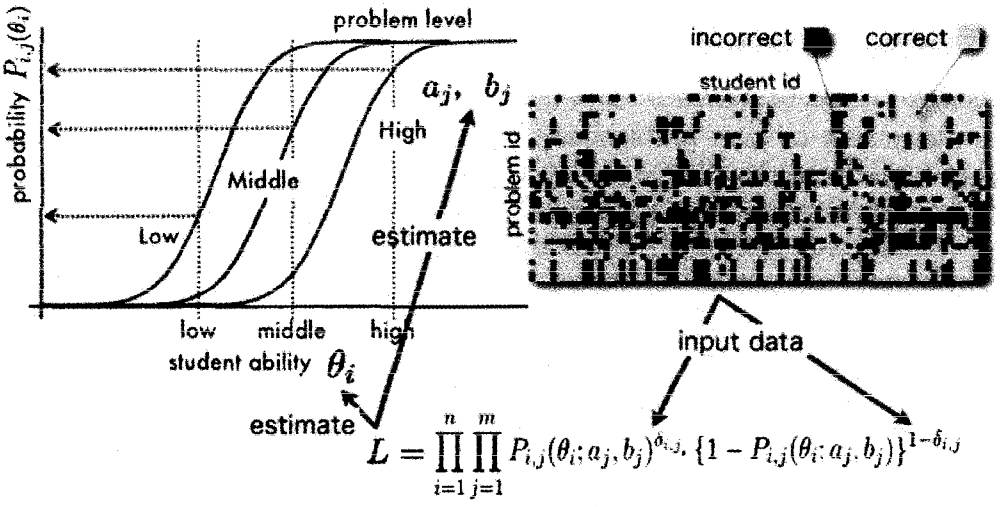

Fig. 1. Item response theory estimation procedure.

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

$$L(\boldsymbol{\delta}|a, b) = \prod_{i=1}^{n} [\int_{-\infty}^{\infty} g(\theta) \prod_{j=1}^{m} L(\delta_{i,j}|a_j, b_j) d\theta], \tag{4}$$

where $g(\theta)$ denotes the ability common to all the students (usually a standard normal distribution) and $\boldsymbol{\delta}$ denotes all the patterns of $\delta_{i,j}$, taking the value of 0 and 1. The EM algorithm Dempster et al. [1977] is often used in such a case [Baker and Kim, 2004]. 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. Some other method such as the Markov Chain Monte Carlo Method Patz and Junker [1999] is also useful in estimating the parameters. The errors for the estimates of students' abilities θ_i and item parameters a_j, b_j are obtained by using the Fisher information matrix corresponding to Equation (3).

3 Score and Ability

As Lazarsfeld and Henry say Lazarsfeld and Henry [1968], it is known that the person parameter represents the magnitude of latent trait of the individual, which is the human capacity or attribute measured by the test. It might be a cognitive ability, physical ability, skill, knowledge, attitude, personality characteristic, etc. We deal with this person parameter as student ability here.

Traditionally, the student's ability can be assessed by the total score on a test where many test problems have primal scores (allotments) assigned in advance by teachers. This is based on the CTT. In the IRT, the ability evaluation is interpreted in a very different manner as compared to traditional scores like number or percent correct. The individual's total score is not the actual score, but is rather computed on a likelihood principle using the points for each correct/incorrect response. Thus, the scoring methodology is totally different from each other in the CTT and the IRT, summing up the correct scores in the former, and in the latter assuming the individual's ability.

While scoring is much more sophisticated with the IRT, for most tests, the (linear) correlation between the θ estimate (which may be identified with the ability of a student) and a traditional score is very high. This is well known to the IRT researchers but unknown to the public. A graph of the IRT scores against traditional scores shows an ogive shape implying that the IRT estimates separate individuals at the borders of the range more than in the middle. This fact might be regarded as indicating that the ability of the IRT to discriminate the person's ability is much more than that of the CTT. We will show this in the next section.

Sometimes, teachers have to lower the baseline points so that the majority of students pass the examination when the problems were rather difficult. This will cause evaluation distortion and true evaluation may not be attained. We will show this effect in section 5 by using typical examples.

4 Comparison of Abilities between the CTT and the IRT Using the Hypothetically Assumed Abilities

Comparisons of abilities between the CTT and the IRT have been introduced using many testing cases up to now. Figure 2 shows such an example case where the number of students is 439, and the number of problems is 33. In the CTT, one point is assigned to each problem. Although abilities between the CTT and the IRT are strongly correlated as shown in the figure, we can see an ogive shape and some deviations to the CTT and the IRT abilities; when $\theta = 0.5$ in the IRT, the deviation is about 3 points, and when point $= 20$ in the CTT, the deviation is about 0.5. This means that the ability ranking in the CTT can be disturbed if we accept the IRT ranking, and *vice versa*. However, we do not know which one of the CTT and the IRT is close to the true ability. Thus, we next investigate this by a simulation study using hypothetically assumed abilities.

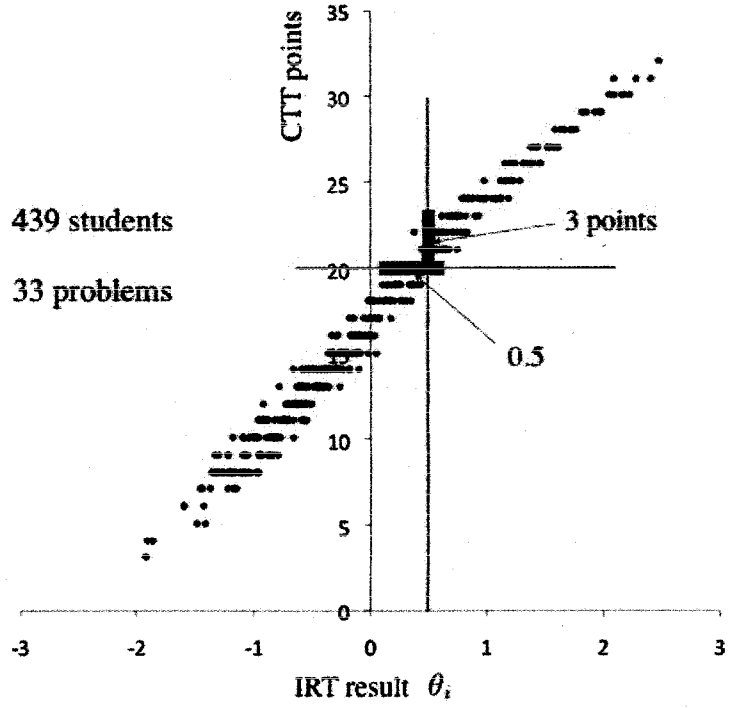

Fig. 2. Simple comparison of abilities between the CTT and the IRT.

We generate many response patterns according to the parameters obtained from a real case shown in Figure 2. Actually, we use Equation (2) with known parameters θ_i, a_j, and b_j, and determine that $\delta_{i,j} = 1$ if $P \geq 0.5$ and $\delta_{i,j} = 0$ if $P < 0.5$. Then, we obtain the total scores $\hat{\psi}_i$ in the CTT and abilities $\hat{\phi}_i$ in the IRT to each response pattern. We know now the true ability of each student by the seeds θ_i in the simulation. We can compare the abilities between the

CTT and the seeds and between the IRT and the seeds. Figure 3 shows the $\hat{\psi}_i$ and $\hat{\phi}_i$ in the box-plot style by using the 100 simulation cases. However, we cannot see the obvious difference between the two.

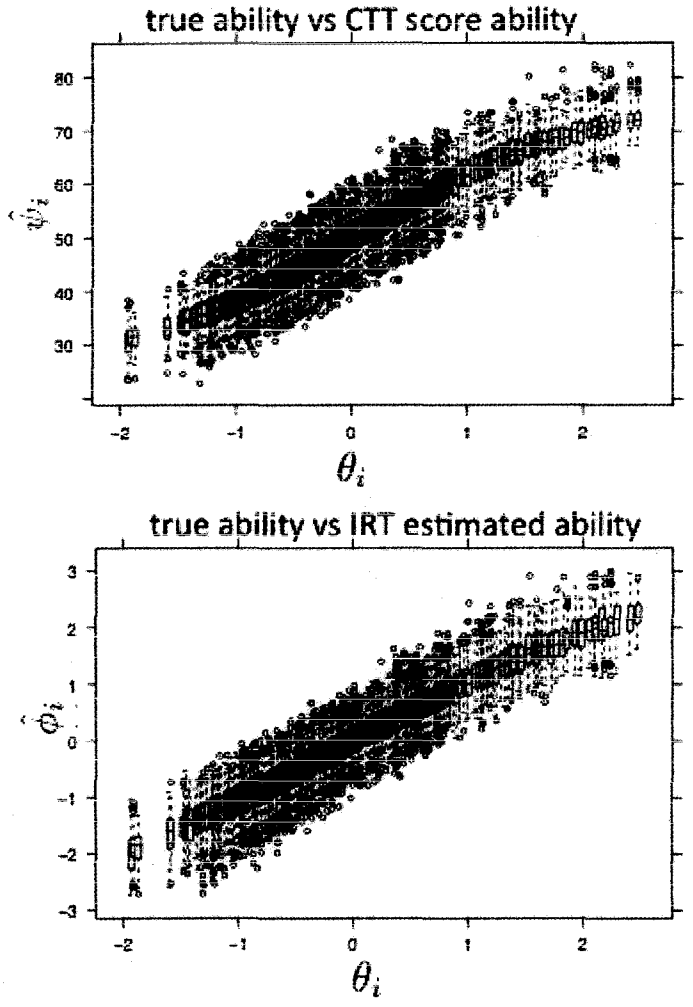

Fig. 3. Abilities $\hat{\phi}_i$ in the IRT and $\hat{\psi}_i$ in the CTT by using 100 simulation cases.

Thus, we next introduce three kinds of statistics to numerically evaluate the difference between the two methods. For comparison, we made a linear transformation from θ_i to ξ_i appropriately.

The three statistics to each response pattern are:

1) statistic S which stands for the square error for abilities:

$$S_{\text{CTT}} = \sum_{i=1}^{n}(\hat{\psi}_i - \xi_i)^2, \quad S_{\text{IRT}} = \sum_{i=1}^{n}(\hat{\phi}_i - \xi_i)^2, \tag{5}$$

2) statistic U which stands for the square error for disturbed ranks:

$$U = \sum_{i=1}^{n}([i] - (i))^2, \tag{6}$$

where (i) means the rank for $\hat{\phi}_i$, and $[i]$ means the corresponding rank for (i) in $\hat{\psi}_i$.

3) statistic V which stands for the number of identical orders:

$$V = \#([i] = (i)). \tag{7}$$

For S and V the lower the better, and for U the higher the better.

By using 100 simulation cases, we compared these three statistics as shown in Figure 4; each point in the figure expresses the values for $(S_{\mathrm{IRT}}, S_{\mathrm{CTT}})$, $(U_{\mathrm{IRT}}, U_{\mathrm{CTT}})$, $(V_{\mathrm{IRT}}, V_{\mathrm{CTT}})$. We have found that the IRT is superior to the CTT to some extent.

5 Comparison of Abilities between the CTT and the IRT Using the Score Allotment Optimization Method

In this section, we compare the ability evaluation between the CTT and the IRT in testing from another perspective. First, we describe the motivation for pursuing this theme. Next, we introduce a newly developed methodology to compare the ability evaluation between the CTT and the IRT when we regard the ability evaluation by the IRT as the standard, and then we show typical simulation study results that mimicked the real data case.

5.1 Motivation

By using the IRT, we can estimate the ability θ_i of each student. This estimated value $\hat{\theta}_i$ can be fluctuated by various testing conditions such as physical conditions or contents of the test. However, $\hat{\theta}_i$ is uniquely determined by Equation (3) to one testing.

On the contrary, in the CTT, teachers can assign the allotment distribution to the test in advance. This means that the total score $\hat{\psi}_i$ can be changed by the teacher's will. Sometimes, he wants to highly evaluate those who could solve the difficult problems. In some cases, he wants to lower the border in

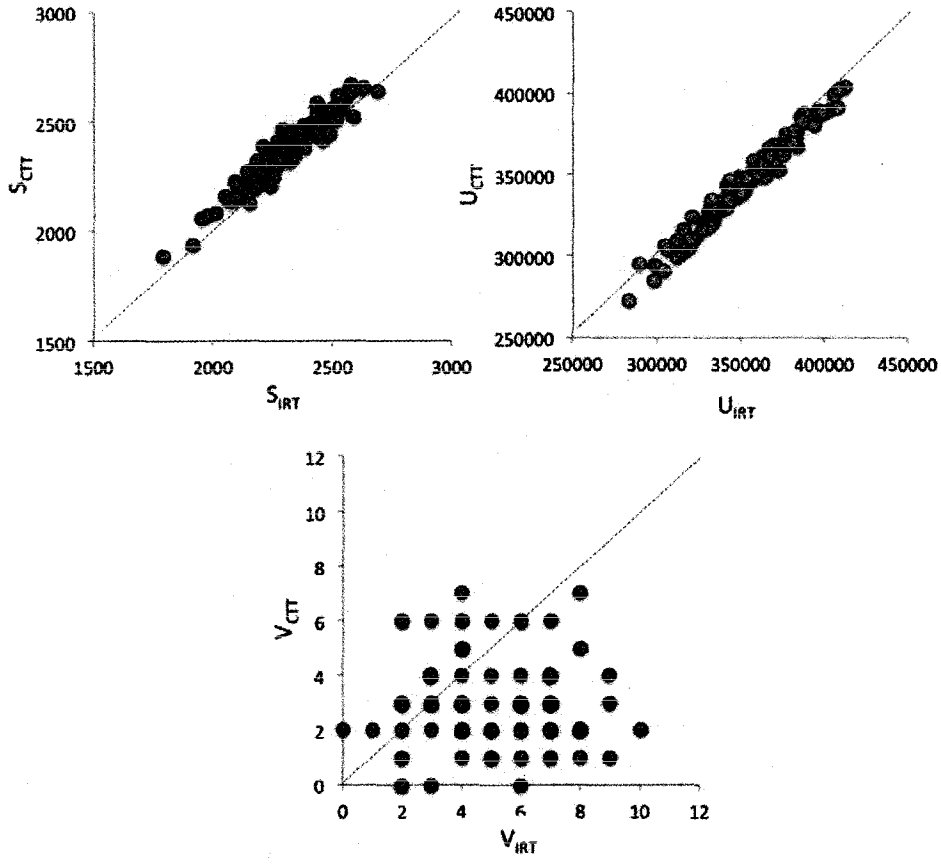

Fig. 4. Comparison of abilities between the CTT and the IRT by regarding the hypothetically assumed abilities as the standard.

order to raise the pass rate of the test. There may be unfairness with respect to the CTT. This is true, but not definitely so. Look at Figure 5; in the figure, in one case, the allotments are uniformly randomly (discretely from $U[1,5]$) given to 33 problems and 346 students, and in the other case, they are are all of the same value (3 points to each problem). We can see that they differ to some extent, but not definitely so. We next check if this tendency also holds for other cases.

To compare the IRT abilities with the CTT scores in which the teacher's will is incorporated, we developed an allotment optimization methodology. If we regard the ability evaluation by the IRT as the standard, we can find the most appropriate allotments in the CTT so that the total scores of the CTT are adjusted as close as possible to the abilities obtained by the IRT.

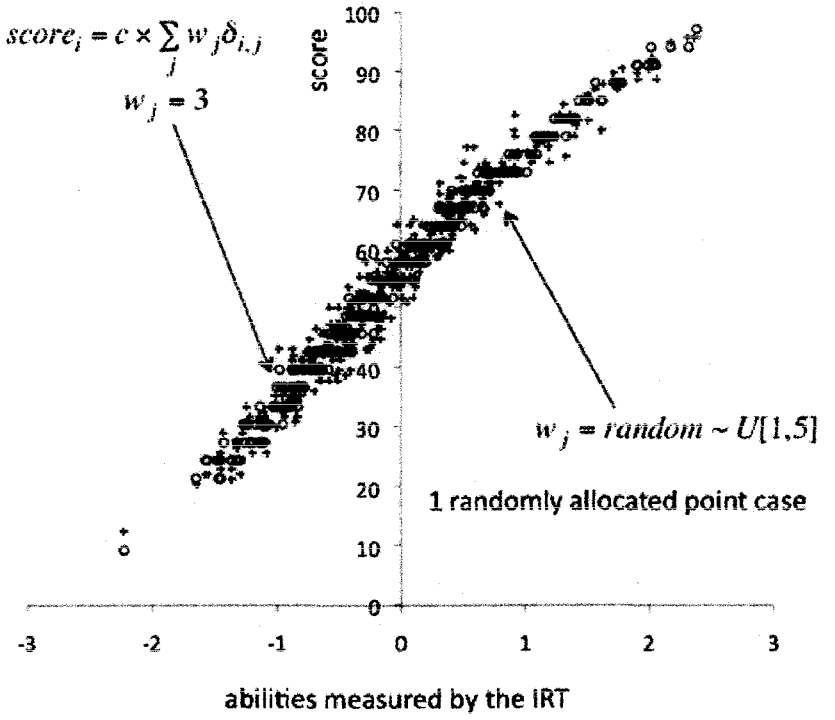

Fig. 5. CTT scores vs IRT abilities by the random allotments and the same value allotments to each problem.

5.2 Score Allotment Optimization Method

We assume that the IRT results provide accurate students' abilities. That is, we regard the IRT abilities as the standard. Our primary objective here is to fit the total scores $\hat{\psi}_i$ in the CTT to the IRT abilities $\hat{\theta}_i$. The problem is to minimize the following R.

$$R = \sum_{i=1}^{n} (\psi_i - \phi_i)^2. \tag{8}$$

Here, for comparison, we have made a linear transformation from $\hat{\theta}_i$ to $\hat{\phi}_i$ appropriately. To solve this problem, we used the gradient descent method. The number of unknown parameters is $m-1$ when the total score is restricted, and m when not restricted. The method requires the iteration such as,

$$q_j^{(k+1)} = q_j^{(k)} - \lambda (\partial R / \partial q_j)^{(k)}, \tag{9}$$

$$(\partial R / \partial q_j)^{(k)} = \sum_{i=1}^{n} 2(\psi_i - \phi_i)\delta_{i,j}, \tag{10}$$

where, λ is a tuning parameter.

For appropriate problem setting, we mimicked some test results that were officially performed. The problems are exactly the same as in Figure 2. To do so, as mentioned before, we first obtain the estimates for the IRT parameters, a_j, b_j, and θ_i. Then, using these values, we generated the response matrix $\delta_{i,j}$ for many cases, say 100 cases. Using these simulated response matrix, we estimate ϕ_i, ψ_i and q_j.

As mentioned earlier, teachers often raise scores superficially so that the majority of the students pass the examination, for example, 60 % of the scores are inflated. This means that teachers lower the border to increase the pass rate; see Figure 6. We assume three cases for that: (1) no jack-up, (2) 25-points jack-up, and (3) 50-points jack-up. Here, 25-points jack-up means that the lowest score goes by 25 points to the upper-side and the full score is maintained at 100 points. In addition to this, we computed the case of inversely jacking-up; that is, the lower the difficulty of b_j, the lower is the allotment as shown in Figure 7. Teachers may want to use this kind of allotment because this is intended to highlight smarter students. However, we often give the similar points to all the problems.

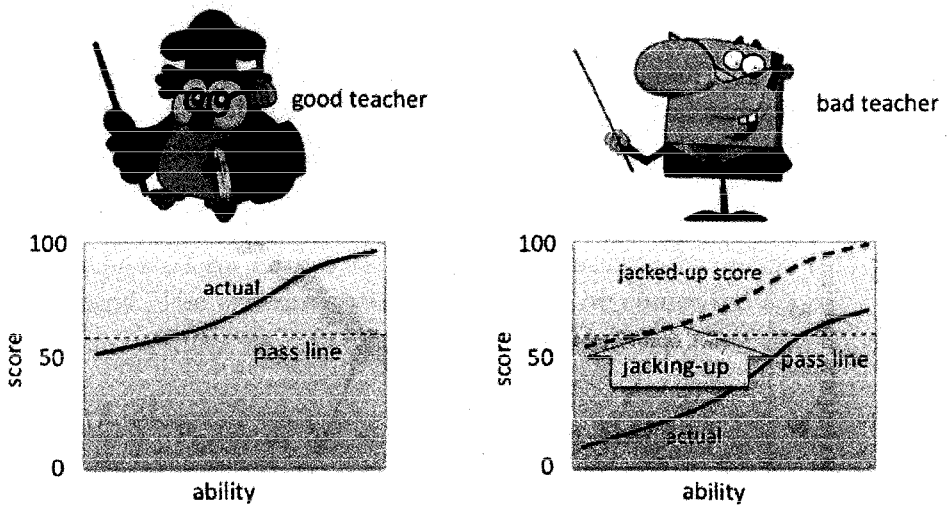

Fig. 6. Jacking-up the scores.

5.3 Comparison of the Results

First, we introduce a typical example case by using a simulation data case. Figure 8 shows a comparison between the abilities by using the IRT and the total scores by using the CTT. In the figure, square dots represent the relation between the abilities in the IRT (appropriately transformed from θ to scores of (0 - 100)) and the total scores by using the (evenly) pre-assigned allotments and circle dots represent the relation between the abilities in the IRT and the total scores by using the optimized allotments (to the abilities in the IRT).

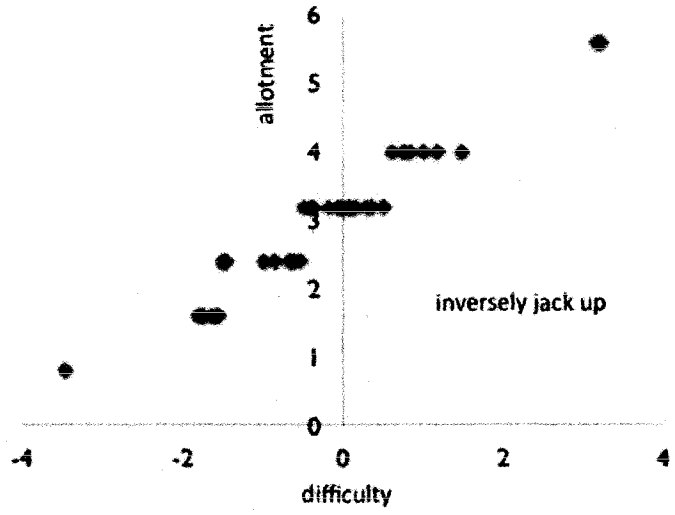

Fig. 7. Inversely jacked-up allotment.

The optimized total scores in the CTT show much more linearity to that of the IRT results. As mentioned earlier, we can observe the ogive shapes in the figure.

The corresponding allotments are shown in Figure 9. At a first glance, we cannot understand what has happened in the optimization for the allotments. However, the next example case reveals the information more clearly.

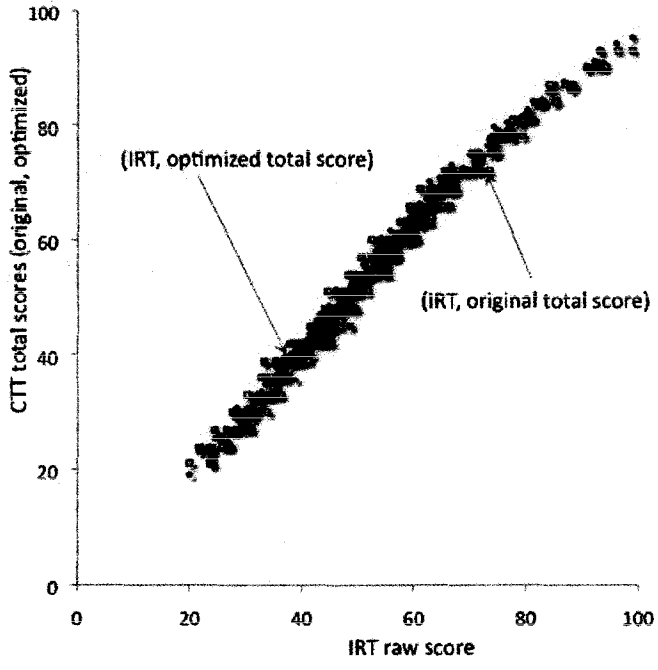

Fig. 8. Comparison of scores between the IRT and the CTT.

Figure 10 shows the relationship between the original problem difficulties b_j and the optimal allotments for three cases of (1) no jack-up, (2) 25-points

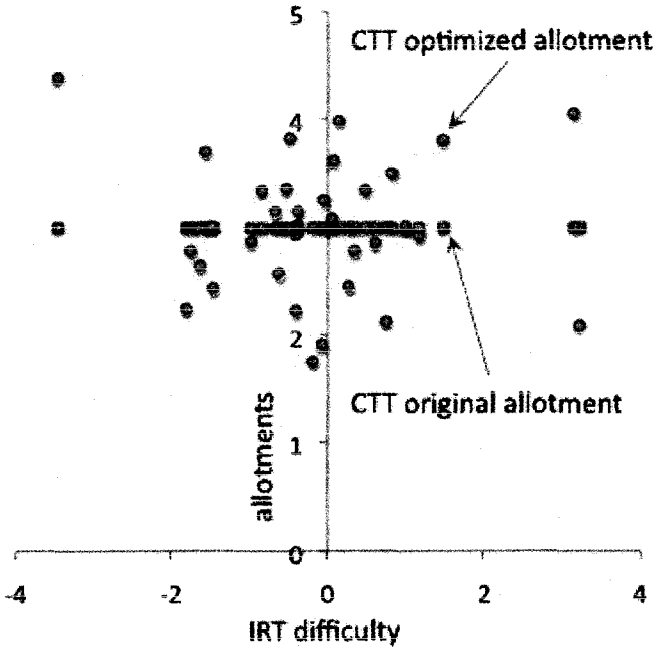

Fig. 9. Relationship between the original problem difficulties (b_j) in the IRT and the allotments in the CTT for the original and optimized cases.

jack-up, and (3) 50-points jack-up. We can see that the problem with lower difficulty results in higher allotments. We do not know now the exact reason as to why this tendency holds. This may be caused by the assignment of smaller number of problems that are extremely easy and are extremely difficult. This can be suggested by Figure 11 in which the uniformly distributed problems are allocated. The expected total score s_i can be computed by

$$s_i = \frac{1}{m}\sum_{j=1}^{m} P_{i,j}(\theta_i; a_j, b_j) \to 1 - \frac{1}{6}\int_{-3}^{3} P(\theta; a, b)d\theta, \tag{11}$$

and the results are illustrated in Figure 12 (when $a_j = 1$), where the ogive shape is seen. However, in solving the least square problem, we can confirm the validity of the estimates by looking at the convergence of R and q_j, which are shown in Figure 13 in the case of no jack-up. In the figure, the RMSE (root mean square error) provides the values scaled by $\sqrt{R/n}$.

Figure 14 shows the comparison of abilities between the IRT and the CTT in the four cases mentioned above in addition to the case of raw score (without adjustment by optimization of Equation (8)).

Next, we show the 100-simulation cases result. Figure 15 on the left shows the relationship between the original problem difficulties b_j and (1) no jack-up, (2) 25-points jack-up, and (3) 50-points jack-up. We can find an intriguing feature

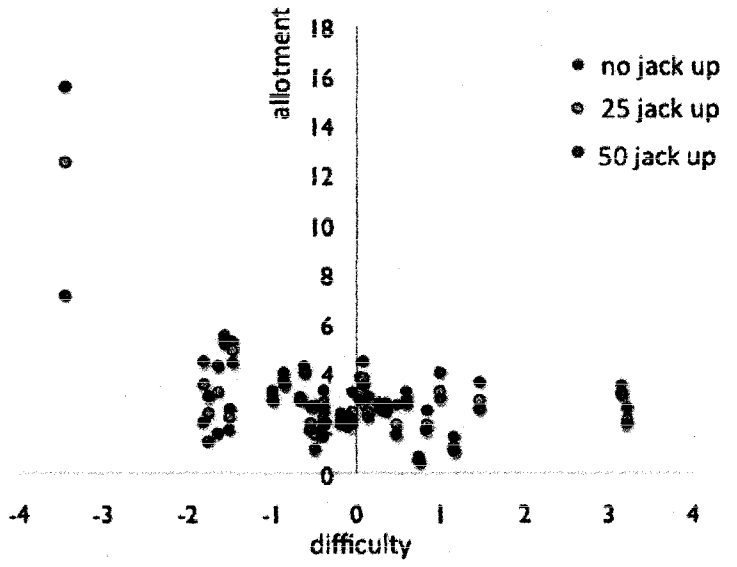

Fig. 10. Relationship between the original problem difficulties (b_j) and (1) no jack--up, (2) 25-points jack-up, and (3) 50-points jack-up.

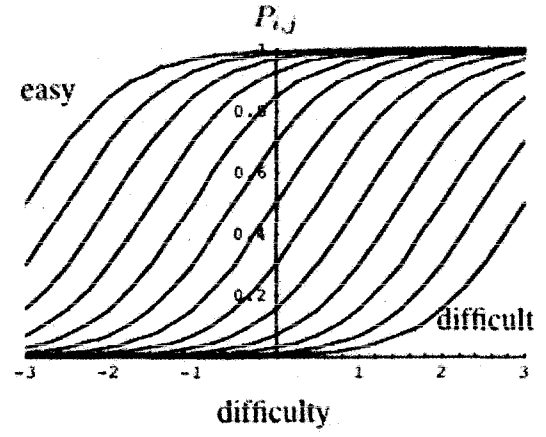

Fig. 11. A typical example case where uniformly distributed problems are allocated.

to these figures. To adjust the allotments to the IRT abilities which may be distorted by inflation (in the cases of positively jacked-up), we have to assign the higher points to the easier problems. This phenomenon is understood because the students with lower abilities can solve only the easier problems and the inflated scores (jacked-up scores) must benefit them.

Figure 15 on the right shows the comparison of abilities between the IRT and the CTT in the four cases mentioned above in addition to the case of raw score (without adjustment by optimization of Equation (8)). In the figure, we can observe that the adjusted scores by jacking-up are much more unreliable than those without the jack-up. This is probably the first time that this kind of a relationship is mathematically provided by using the optimized allotments.

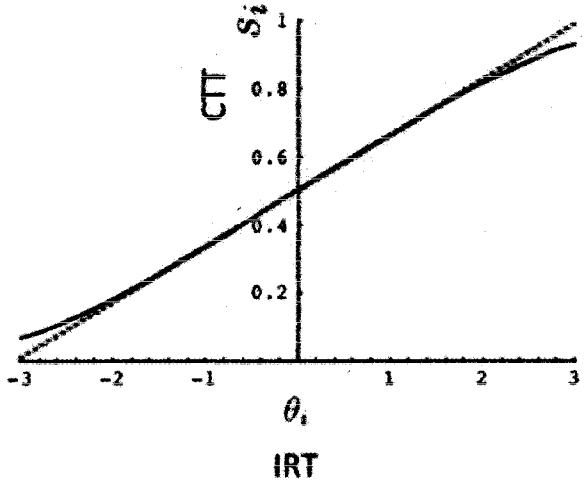

Fig. 12. Expected total score s_i with uniformly distributed problems allocation.

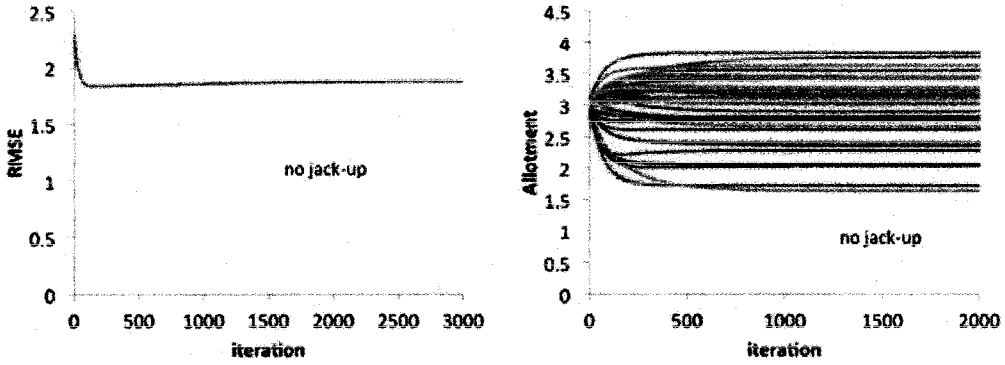

Fig. 13. Convergences of RMSE and allotments.

6 Discussions

We have developed a methodology to compare the ability evaluation between the CTT and the IRT by searching for the optimal distribution for allotments. Is there a possibility to persuade teachers to use the IRT based on our comparison of ability evaluation between the CTT and the IRT in a simulation study? The answer is yes.

First, the similarity between the total score by the CTT and the ability by the IRT will make it easy to change the evaluation method for teachers without obstacles. The superficial evaluation for ability by using the CTT is not so different from the IRT results as long as we do not use the jack-up process. In addition, the IRT provides us with the problem difficulties (by b_j) as well as the accurate estimates for students' abilities (by θ_i).

Second, we have revealed the relationship between the problem difficulties in

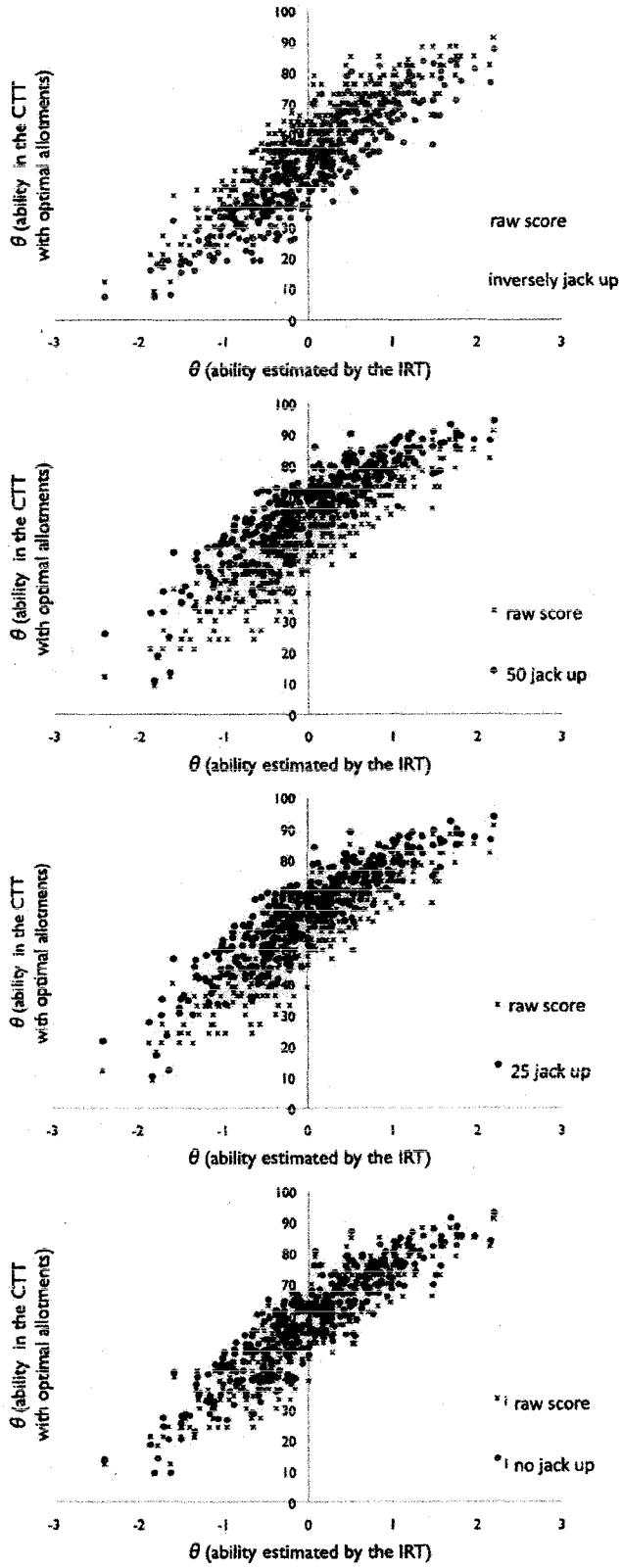

Fig. 14. Comparison of abilities between the IRT and the CTT with optimal allotments (1 case).

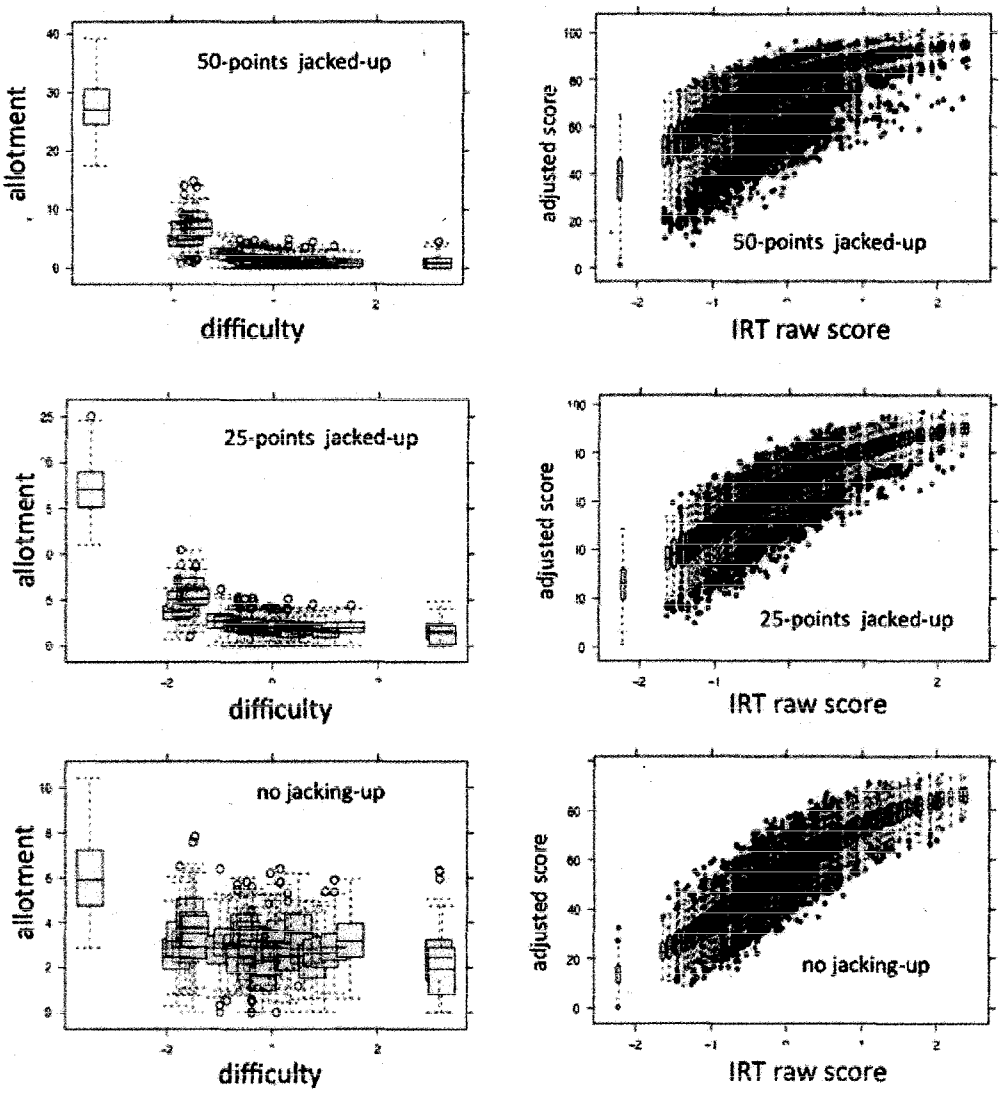

Fig. 15. Simulation results for (1) no jack-up, (2) 25-points jack-up, and (3) 50-points jack-up (100 cases).

the IRT and the optimal allotments in the CTT, when we regard the ability evaluation by the IRT as the standard. By applying our methodology to some simulation cases that mimic the real data case, we have found an intriguing feature with respect to the pre-assigned allotments in the CTT. To adjust the allotments in the CTT to the IRT abilities, the total scores may be distorted by inflation (in the cases of positively jacking-up), which may disturb the accurate evaluation of the students. Teachers can understand this phenomenon theoretically (for the first time, maybe) by the benefit of the proposed methodology.

7 Concluding Remarks

It is known that the item response theory (IRT) is superior to the classical test theory (CTT) with respect to the ability evaluation. However, teachers do not use the IRT as a regular testing tool in universities and colleges even now. Many teachers still use the CTT according to their custom behaviors. In this paper, we shed light on this matter by comparing the ability evaluation in testing between these two methods. However, the CTT uses pre-assigned allotments contrary to the IRT which has no allotment concept. Thus, we have newly developed a method to compare the CTT and the IRT by finding the most appropriate allotments in the CTT so that the total scores of the CTT are adjusted as close as possible to the abilities obtained by the IRT.

We first compared the CTT abilities and the IRT abilities by using the hypothetically assumed abilities that are mimicked to a real case. By using a simulation study, we have found that the IRT is superior to the CTT to some extent.

We next compared them by regarding the ability in the IRT as the standard. We have applied this to other simulation cases that mimic the real data case, and we found an intriguing feature with respect to the pre-assigned allotments. If teachers want to raise the examination pass rate, we guessed that they give higher scores (than the actual scores achieved) to students with lower abilities and called this process jacking-up. Using the proposed allotment optimization, we found that jacking-up causes higher allotments to easier problems in the CTT.

References

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

Dempster, A., Laird, N., Rubin, D., 1977. Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society: Series B (Methodological) 39, 1–38.

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., 2011. An optimal test design to evaluate the ability of an examinee by using the stress-strength model. Journal of Statistical Computation and Simulation 81, 79–87.

Hirose, H., Sakumura, T., 2010. Test evaluation system via the web using the

item response theory. In: Proceedings of the International Conference on Computer and Advanced Technology in Education (CATE 2010).

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

Lazarsfeld, P., Henry, N., 1968. Latent Structure Analysis. Boston: Houghton Mifflin.

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

Lord, F. M., Novick, M. R., 1968. Statistical theories of mental test scores. MA: Addison-Welsley Publishing Company.

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

Patz, R. J., Junker, B. W., 1999. Applications and extensions of mcmc in irt: Multiple item types, missing data, and rated responses. Journal of Educational and Behavioral Statistics 24, 342–366.

Sakumura, T., Hirose, H., 2010a. Student ability evaluation using the stress-strength model when ability is the random variable. In: Proceedings of the 2010 International Congress on Computer Applications and Computational Science (CACS 2010).

Sakumura, T., Hirose, H., 2010b. Test evaluation system via the web using the item response theory. Information 13, 647–656.

Sakumura, T., Kuwahata, T., Hirose, H., 2011. An adaptive online ability evaluation system using the item response theory. In: Education and e-Learning (EeL2011). pp.51-54.

Tsukihara, Y., Suzuki, K., Hirose, H., 2009. A small implementation case of the mathematics tests with the item response theory evaluation into an e-learning system. Computer and Education 24, 70–76.