

Test Evaluation System via the Web using the Item Response Theory

Takenori Sakumura, Hideo Hirose*

*Kyushu Institute of Technology, Department of Systems Design and Informatics,
Iizuka, Fukuoka, 820-8502 Japan
sakumura@ume98.ces.kyutech.ac.jp, *hirose@ces.kyutech.ac.jp*

Abstract

Although the superiority of the test method using the item response theory (IRT) over the classical test method is valid for regular examinations such as the midterm and final tests in universities and colleges, the IRT is not known to teachers in universities and colleges. To enhance the chance of use of this new method, we have developed a test evaluation system via the Web for university teachers. By simply dragging an EXCEL file in which 0/1 scores of the test result are stored to a program place of the Web page, teachers can obtain students' abilities and parameters for each problem attached in that EXCEL file. By using the IRT evaluation in tests, we have found that: 1) we can include high and low level test items together so that we can assess the student abilities more accurately and fairly; we do not worry about providing easier problems which will make the lecture level down; in other words, we do not care about finding the most appropriate problem levels to each student. 2) students do not raise claims about their scores; they are satisfied with this way of testing.

Key words: test evaluation; item response theory; e-learning; Web system.

1 Introduction

In Japan, the number of young people has been on the decline as compared to the growing number of universities and colleges. Almost half of the 18-year-old high school graduates can be enrolled in universities and colleges. Therefore, Japan's academia are supposed to accept a variety of students in the academic levels. For effective evaluation to such students, the item response theory (IRT) (Hambleton and Swaminathan (1984), Hambleton, Swaminathan and Rogers (1991), Linden and Hambleton (1996)) may enhance the students' skills and evaluate their abilities more accurately if several adaptive e-learning systems

(Mills, Potenza and Fremer (2002)) and test methods are appropriately used. A student self-learning system embedded in the e-learning system, Moodle (<http://moodle.org/>), is introduced (Tsukihara, Suzuki and Hirose (2009)), and a new adaptive test method is also proposed recently (Hirose (2009)) where the up-and-down method by Dixon and Mood (1948) and the stratified adaptive test method by Weiss (1973) are incorporated together to perform the optimal test.

Although the IRT has been widely used to test systems such as the Test of English as a Foreign Language (TOEFL), it is not well known to teachers in universities and colleges. The superiority of the IRT over the classical test method is also valid for many subjects in universities and high schools.

The IRT is new in the sense that it can assess the abilities of the examinees along with the difficulties of the problems (items). This method has often been used in many fields; TOEFL and TOEIC (Test of English for International Communication) are typical examples applying this theory. Examinees in tests for English can solve many items in a certain time period, e.g., in two to three hours. Thus, the method can effectively and easily be applied in such kinds of tests. However, those who take examinations in universities cannot solve as many problems in such a short time. It would be ideal if we can use fewer problems with higher evaluation accuracy in examinations. This paper aims to achieve this objective with much an easier method.

To enhance the chance of use of this new evaluation method in university teachers, we have developed a test evaluation system via the web. By simply dragging an EXCEL file in which 0/1 scores of the test result are stored to a program place of the Web page, teachers can obtain the students' abilities and parameters for each problem attached in that EXCEL file. By using the IRT evaluation in tests, we have found that teachers and students are both satisfied with this new system.

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

$$P_j(\theta_i; a_j, b_j) = \frac{1}{1 + \exp\{-1.7a_j(\theta_i - b_j)\}}, \quad (1)$$

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

$$L = \prod_{i=1}^N \prod_{j=1}^n P_j(\theta_i; a_j, b_j)^{\delta_{i,j}} \cdot (1 - P_j(\theta_i; a_j, b_j))^{1-\delta_{i,j}}, \quad (2)$$

where $\delta_{i,j}$ denotes the indicator function such that $\delta = 1$ for success and $\delta = 0$ for failure; 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 in common, P_j in Equation (1) 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 (2), the maximum likelihood estimates may be obtained. However, it is not easy to obtain the item parameters and the students' abilities together. There are $2 \times n + N$ 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

$$L(\delta|a, b) = \prod_{i=1}^N \left[\int_{-\infty}^{\infty} g(\theta) \prod_{j=1}^n L(\delta_{i,j}|a_j, b_j) d\theta \right], \quad (3)$$

where $g(\theta)$ denotes the ability common to all the students (usually a standard normal distribution) and δ denotes all the patterns of $\delta_{i,j}$, taking the value of 0 and 1; see Appendix. The EM algorithm (Dempster, Laird and Rubin (1977)) is usually used in such a case (Baker and Kim (2004)). Then, the students' abilities are obtained by maximizing the corresponding likelihood function; see Appendix. To circumvent the ill conditions so that all the items are correctly answered or incorrectly answered, the Bayes technique is applied (Baker and Kim (2004)). Some tuning parameters are seen in Appendix.

3 Test Evaluation System via the Web

Although the superiority of the test method using the IRT over the classical test method is valid for regular examinations such as the midterm and final tests in universities and colleges, the IRT is not known to teachers in universities and colleges. There may be some reasons: 1) despite the availability of test evaluation programs using the standard IRT such as Bilog (Bilog-MG, (2005)), there are still difficulties in using the programs, 2) due to the unknown tuning parameters in EM algorithm and Bayes method, the estimated values become slightly different, 3) many teachers adopt giving partial points to one test item such that 5 points to item one with full score of 10 points; they are not familiar with 0/1 evaluation.

To overcome these obstacles, we have developed a test evaluation system via the web. The specification is: 1) teachers provide the score matrix (column: student id, row: problem id) consisting of 0/1 element; whether an answer by a student is correct or not to each item should be decided in advance with 0/1 evaluation, 2) in response to the submission of the score matrix to the web, the system gives the estimated parameters; the student abilities are automatically converted to standard scores used in universities and colleges. This concept can be seen in Figure 1. Input and output images are also shown in Figures 2 and 3.

By using the IRT evaluation in tests of *statistics* and *numerical analysis* for several semesters, we have found the following: 1) we can include high and low level test items together so that we can assess the student abilities more accurately and fairly; we do not worry about providing easier problems which will make the lecture level down; in other words, we do not care about finding the most appropriate problem levels to each student. 2) students do not raise claims about their scores; they are satisfied with this way of testing. An example of the item characteristic curves is shown in Figure 4.

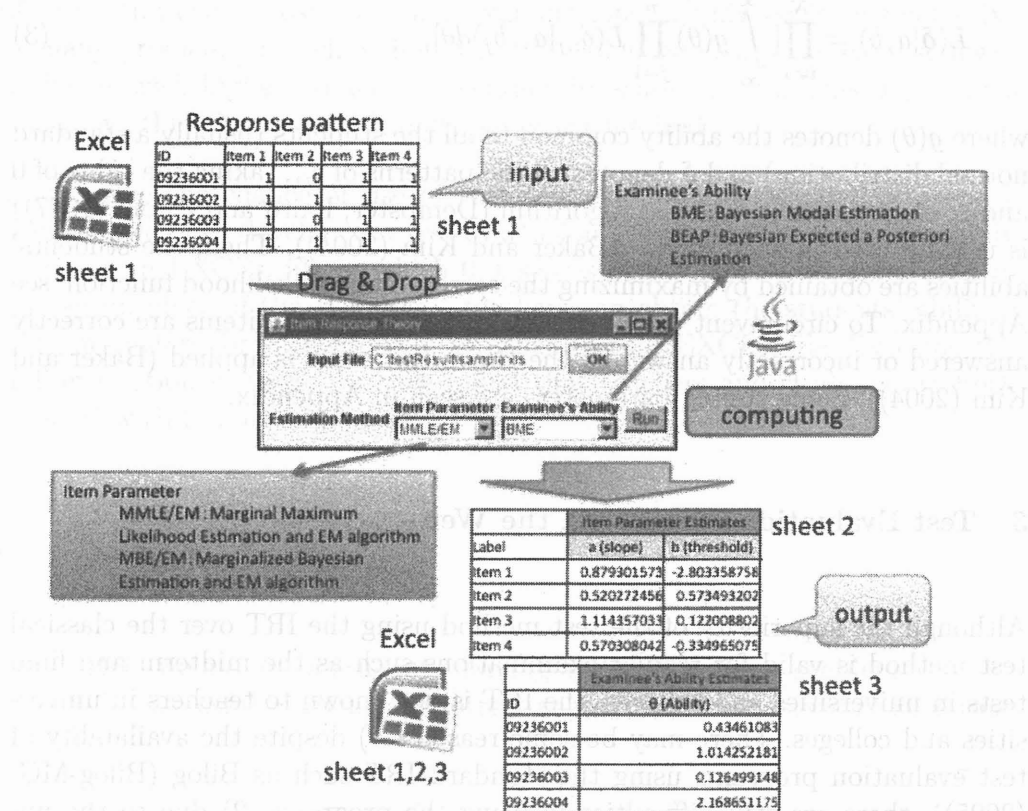
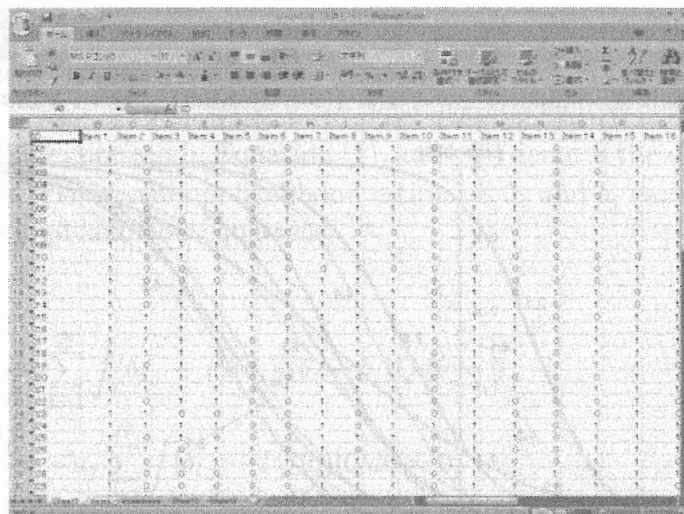
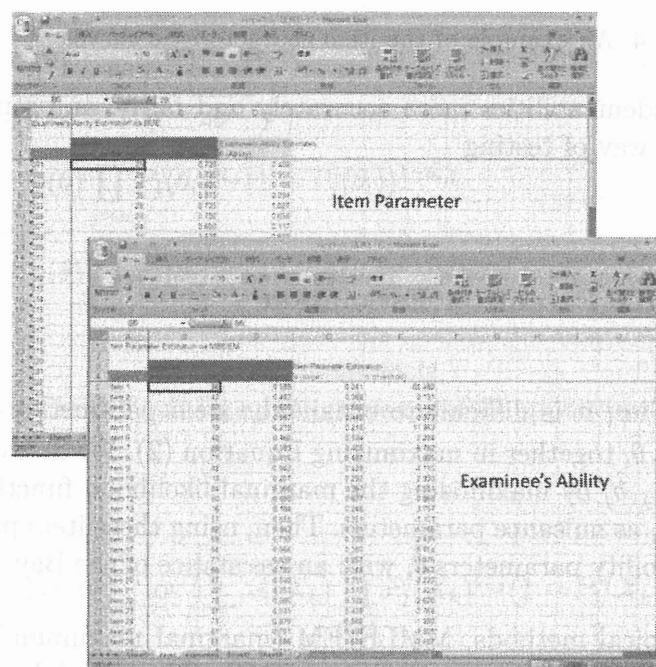


Fig. 1. Concept of the web system using the IRT method



	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10	Item 11	Item 12	Item 13	Item 14	Item 15	Item 16
1	0	1	1	0	0	1	1	1	0	1	1	0	1	1	1	1
2	1	0	1	0	0	1	1	1	0	1	1	0	1	1	1	1
3	1	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1
4	1	0	1	0	0	1	1	1	1	1	1	0	1	1	1	1
5	1	0	1	0	0	1	1	1	0	1	1	0	1	1	1	1
6	1	0	1	0	0	1	1	1	0	1	1	0	1	1	1	1
7	1	0	1	0	0	1	1	1	0	1	1	0	1	1	1	1
8	1	0	1	0	0	1	1	1	0	1	1	0	1	1	1	1
9	1	0	1	0	0	1	1	1	0	1	1	0	1	1	1	1
10	1	0	1	0	0	1	1	1	0	1	1	0	1	1	1	1
11	1	0	1	0	0	1	1	1	0	1	1	0	1	1	1	1
12	1	0	1	0	0	1	1	1	0	1	1	0	1	1	1	1
13	1	0	1	0	0	1	1	1	0	1	1	0	1	1	1	1
14	1	0	1	0	0	1	1	1	0	1	1	0	1	1	1	1
15	1	0	1	0	0	1	1	1	0	1	1	0	1	1	1	1
16	1	0	1	0	0	1	1	1	0	1	1	0	1	1	1	1
17	1	0	1	0	0	1	1	1	0	1	1	0	1	1	1	1
18	1	0	1	0	0	1	1	1	0	1	1	0	1	1	1	1
19	1	0	1	0	0	1	1	1	0	1	1	0	1	1	1	1
20	1	0	1	0	0	1	1	1	0	1	1	0	1	1	1	1

Fig. 2. Input image to the web system



Item Parameter

Item	a	b	c	d
Item 1	0.721	2.428	0.721	2.428
Item 2	0.721	2.428	0.721	2.428
Item 3	0.721	2.428	0.721	2.428
Item 4	0.721	2.428	0.721	2.428
Item 5	0.721	2.428	0.721	2.428
Item 6	0.721	2.428	0.721	2.428
Item 7	0.721	2.428	0.721	2.428
Item 8	0.721	2.428	0.721	2.428
Item 9	0.721	2.428	0.721	2.428
Item 10	0.721	2.428	0.721	2.428
Item 11	0.721	2.428	0.721	2.428
Item 12	0.721	2.428	0.721	2.428
Item 13	0.721	2.428	0.721	2.428
Item 14	0.721	2.428	0.721	2.428
Item 15	0.721	2.428	0.721	2.428
Item 16	0.721	2.428	0.721	2.428

Examinee's Ability

Examinee	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Exam 1	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185
Exam 2	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185
Exam 3	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185
Exam 4	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185
Exam 5	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185
Exam 6	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185
Exam 7	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185
Exam 8	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185
Exam 9	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185
Exam 10	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185
Exam 11	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185
Exam 12	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185
Exam 13	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185
Exam 14	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185
Exam 15	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185
Exam 16	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185
Exam 17	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185
Exam 18	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185
Exam 19	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185
Exam 20	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185	0.185

Fig. 3. Output image to the web system

4 Concluding Remarks

To enhance the chance of use of the IRT method, we have developed a test evaluation system via the web for university teachers. By simply dragging an EXCEL file in which 0/1 scores of the test result are stored to a program place of the Web page, teachers can obtain students' abilities and parameters for each problem attached in that EXCEL file. By using the IRT evaluation in tests, we can include high and low level test items together so that we

problem	1.1,	1.2,	2.1,	2.2,	2.3,	2.4,	3.1,	3.2,	3.3,	4.1,	4.2,	4.3,	4.4,
correct ans.	18,	9,	83,	23,	42,	21,	8,	6,	26,	62,	5,	56,	31
slope	0.7965	1.0239	1.6765	1.0533	1.2561	1.3568	0.95502	1.1708	0.5879	0.80878	0.95188	0.9102	1.4043
threshold	1.5233	2.1374	-0.7436	1.3997	0.5224	1.2382	2.2629	2.3597	1.2343	-0.2538	2.44004	-0.07849	0.8131

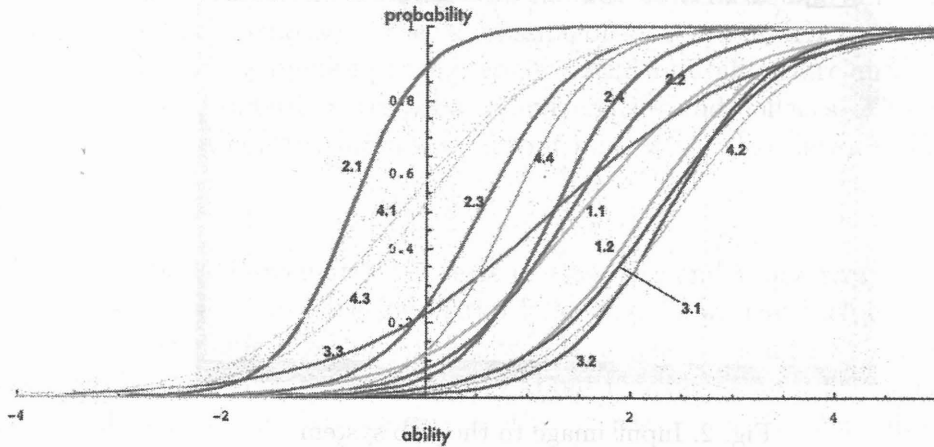


Fig. 4. An example of the item characteristic curves

can assess the student abilities more accurately and fairly, and students are satisfied with this way of testing.

5 Appendix: IRT Parameter Estimation

As mentioned earlier, it is difficult to obtain the item parameters \hat{a}_j, \hat{b}_j and ability parameters $\hat{\theta}_i$ together in maximizing Equation (2). We first obtain the item parameters \hat{a}_j, \hat{b}_j by maximizing the marginal likelihood function (3), in which we regard θ_i as nuisance parameters. Then, using these item parameters \hat{a}_j, \hat{b}_j , we obtain ability parameters $\hat{\theta}_i$ with an assistance of the Bayes method.

Here, we show typical methods, MMLE/EM (marginal maximum likelihood with EM algorithm) for obtaining \hat{a}_j, \hat{b}_j and BME (Bayes modal estimation) for obtaining $\hat{\theta}_i$. We will not describe BME/EM (Bayes modal estimation with EM algorithm) for obtaining \hat{a}_j, \hat{b}_j and BEAP (Bayes expected a posteriori estimation) for obtaining $\hat{\theta}_i$, although these methods are available in the proposed system.

5.1 Item Parameters

By taking logarithm to Equation (3), we obtain

$$\log L(a, b) = \sum_{i=1}^N \log \left[\int g(\theta) \prod_{j=1}^n P_j(\theta_i)^{\delta_{i,j}} (1 - P_j(\theta_i))^{1-\delta_{i,j}} d\theta \right], \quad (4)$$

where $P_j(\theta)$ is expressed in Equation (1), and $g(\theta)$ denotes the standard normal distribution. The maximum likelihood estimates \hat{a}_j and \hat{b}_j can be obtained by solving the log-likelihood equations,

$$\frac{\partial \log L}{\partial a_j} = \sum_{i=1}^N \int [\delta_{i,j} - P_j(\theta_i)] (\theta_i - b_j) Q d\theta = 0, \quad (5)$$

$$\frac{\partial \log L}{\partial b_j} = -a_j \sum_{i=1}^N \int [\delta_{i,j} - P_j(\theta_i)] Q d\theta = 0, \quad (6)$$

where,

$$g(\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left(-\frac{\theta - \mu}{2\sigma^2} \right), \quad (7)$$

$$Q = \frac{g(\theta) \prod_{j=1}^n P_j(\theta_i)^{\delta_{i,j}} (1 - P_j(\theta_i))^{1-\delta_{i,j}}}{\int g(\theta) \prod_{j=1}^n P_j(\theta_i)^{\delta_{i,j}} (1 - P_j(\theta_i))^{1-\delta_{i,j}} d\theta}. \quad (8)$$

To circumvent the difficult integration, we make an approximation to Equation (4) by discretization,

$$\log L(a_j, b_j) = \sum_{i=1}^N \log \left[\sum_k A(X_k) \prod_{j=1}^n P_j(X_k)^{\delta_{i,j}} (1 - P_j(X_k))^{1-\delta_{i,j}} \right], \quad (9)$$

with observed values $A(X_k)$ around point X_k . The approximate maximum likelihood estimates \hat{a}_j and \hat{b}_j can be obtained by using the EM algorithm by Dempster, Laird and Rubin (1977). The EM algorithm has two steps: E step and M step. These E and M steps are repeated until converged values of a_j and b_j are obtained.

E step: We compute the following.

$$L(X_k) = \prod_{j=1}^n P_j(X_k)^{\delta_{i,j}} (1 - P_j(X_k))^{1-\delta_{i,j}}, \quad (10)$$

$$\bar{f}_{jk} = \sum_{i=1}^N \left[\frac{L(X_k)A(X_k)}{\sum_k L(X_k)A(X_k)} \right], \quad \bar{r}_{jk} = \sum_{i=1}^N \left[\frac{\delta_{i,j} L(X_k)A(X_k)}{\sum_k L(X_k)A(X_k)} \right]. \quad (11)$$

M step: Using \bar{f}_{jk} and \bar{r}_{jk} , we solve the following equations using the Newton-Raphson method,

$$\frac{\partial \log L}{\partial a_j} = \sum_k (X_k - b_j) [\bar{r}_{jk} - \bar{f}_{jk} P_j(X_k)] = 0, \quad (12)$$

$$\frac{\partial \log L}{\partial b_j} = -a_j \sum_k [\bar{r}_{jk} - \bar{f}_{jk} P_j(X_k)] = 0, \quad (13)$$

where,

$$P_j(X_k) = \frac{1}{1 + \exp\{-1.7a_j(X_k - b_j)\}}. \quad (14)$$

5.2 Ability Parameters

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

$$H\{\theta \mid \delta, a, b\} \propto L\{\delta \mid \theta_i, a_j, b_j\} h(\theta), \quad (15)$$

where,

$$L\{\delta \mid \theta_i, a_j, b_j\} = \prod_{j=1}^n P_j(\theta_i)^{\delta_{i,j}} (1 - P_j(\theta_i))^{1-\delta_{i,j}}, \quad (16)$$

$$h(\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\theta - \mu_\theta}{2\sigma_\theta^2}\right), \quad (17)$$

$$\delta = (\delta_{i,j}). \quad (18)$$

We solve the likelihood equations,

$$Lh_{\theta_i} = \frac{\partial \log Lh}{\partial \theta_i} = \sum_{j=1}^n 1.7a_j(\delta_{i,j} - P_j(\theta_i)) - \left(\frac{\theta_i - \mu_\theta}{2\sigma_\theta^2}\right), \quad (19)$$

by using the Newton-Raphson method,

$$[\hat{\theta}_i]_{t+1} = [\hat{\theta}_i]_t - [\Lambda_{\theta\theta}]_t^{-1} [Lg_{\theta}]_t, \quad (20)$$

where,

$$Lh_{\theta_i} = L\{\delta \mid \theta_i, a_j, b_j\}h(\theta), \quad (21)$$

$$\Lambda_{\theta\theta} = \frac{\partial^2 \log Lh}{\partial \theta_i^2} = - \sum_{j=1}^n (1.7a_j)^2 P_j(\theta_i)(1 - P_j(\theta_i)) - \frac{1}{2\sigma_{\theta}^2}. \quad (22)$$

5.3 Tuning Parameters

In applyig the EM algorithm, we set the iteration times as,

iterative method	times
EM iteration	100
Newton-Raphson iteration in E step	20

The stopping conditions is that the relative error reaches 0.05.

References

- [1] Baker, F.B., and Kim, S-H., (2004). *Item Response Theory: Parameter Estimation Technique, 2nd edn.*, Marcel Dekker.
- [2] Bilog-MG, (2005). <http://www.ssicentral.com/irt/index.html>.
- [3] Dempster, A.P., Laird, N.M., and Rubin, D.B., (1977). Maximum Likelihood from Incomplete Data via the the EM Algorithm, *Journal of the Royal Statistical Society. Series B*, 39, pp.1-38.
- [4] Dixon, W.J., and Mood, A.M., (1948). A method for obtaining and analyzing sensitivity data, *Journal of the American Statistical Association*, 54, pp.109-126.
- [5] Hambleton, R.K., and Swaminathan, H., (1984). *Item Response Theory: Principles and Applications*, Springer.
- [6] Hambleton, R.K., Swaminathan, H., and Rogers, H. J., (1991). *Fundamentals of Item Response Theory*, Sage Publications
- [7] Hirose, H. (2009). An optimal test design to evaluate the ability of an examinee by using the stress-strength model, *Journal of Statistical Computation and Simulation*, to appear.
- [8] Linden, W.J.D., and Hambleton, R.K. (ed), (1996). *Handbook of Modern Item Response Theory*, Springer.
- [9] Mills, C.N., Potenza, M.T., and Fremer, J.J., (2002). *Computer-Based Testing: Building the Foundation for Future Assessments*, Lawrence Erlbaum.

- [10] <http://moodle.org/>
- [11] Tsukihara, Y., Suzuki, K., and 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*, Vol.24, pp.70-76. (in Japanese; Awarded the best paper in 2009)
- [12] Weiss, D.J., (1973). The stratified adaptive computerized ability test, *Research Report 73-3*, Psychometric Methods Program, Department of Psychology, University of Minnesota.