US-China Education Review B 5 (2011) 627-635
Earlier title: US-China Education Review, ISSN 1548-6613

Validities of the Signed and Unsigned Lecture Questionnaires Using the Item Response Theory*

Hideo Hirose
Kyushu Institute of Technology, Fukuoka, Japan

Teachers often raise a question that whether the lecture questionnaires are necessary or not. In this paper, we first show the recent statistical analysis for the official unsigned questionnaire evaluation results took in our faculty. We have found that: (1) the evaluation scores of lectures by students have been rising up year by year, which means that lectures have been improved; and (2) to take a look at the distribution of the evaluation scores as well as the mean value is crucial, which indicates that taking the questionnaires enhances the teaching skills of teachers. In addition to the official questionnaires, the author has been taking the Web-based signed lecture questionnaires to three mathematics subjects for more than five years. Using these stocked data, we have, next, analyzed the relationship between the signed and unsigned lecture questionnaires and found that: (1) although few unsynchronized relationships between the signed and unsigned evaluation scores are observed, the trends between them are roughly the same; and (2) detailed analysis for the questionnaires is also important to grasp the student lecture comprehension and satisfaction, via the IRT (item response theory) to investigate whether the official lecture questionnaires in the department and the Web-based signed lecture questionnaires are reliable or not, the questionnaires are reliable.

Keywords: lecture questionnaire, evaluation, IRT (item response theory), signed form (registered form), unsigned form (bearer form), Web-based questionnaire

Introduction

Although the official FD (faculty development) systems are mandatory since 2008 in Japan by the suggestion of MEXT (Ministry of Education, Culture, Sports, Science and Technology), our faculty has been continuing our own FD activities since 2002. The main subject is the lecture questionnaires. Until 2004, the results of the questionnaires are sent to each teacher and the whole statistics are closed. From the second semester in 2004, we have changed the system and each result became open to teachers, students and staffs. At the beginning, we tackled the FD activities aggressively. However, as the day passes, teachers often raise a question that the lecture questionnaires are still necessary or not. The author also began to use the minute paper (Davis, Robert, & Wilson, 1983) and have been using the Web-based lecture questionnaires since 2002. The author feels vaguely that continuing such an activity is still important (Berk, 2006). Thus, the author investigate will here what the continuing FD activities brought us.

First, we show the recent statistical analysis for the official unsigned (bearer) lecture questionnaires took

*Acknowledgement:** The author appreciates Mr. Sakumura for his cooperation to this study.
Hideo Hirose, Ph.D., professor, Department of Systems Design and Informatics, Kyushu Institute of Technology.

in our faculty. All the statistics regarding the questionnaires are being processed in open. The basic statistical tools are the correlation to know the relationship between the unsigned official lecture questionnaires and Web-based signed (registered) lecture questionnaires.

Next, we analyze the relationship between the signed and unsigned lecture questionnaires in detail using the IRT (item response theory). That is, we investigate whether the official lecture questionnaires in the department and the Web-based signed lecture questionnaires are reliable or not. This analysis will show that it is important to grasp the students' lecture comprehension and satisfaction.

Trend of Faculty Official Unsigned Lecture Questionnaires

In our faculty, we have two semesters and more than 350 lectures are opened in a semester. The number of students is about 1,500 from freshman to senior. All the teachers are mandatory to carry out the lecture questionnaires at the end of the lectures. Students mark 2, 4, 6, 8 and 10 scores as the evaluation points; the higher, the better. The class sizes are from 10 students to 100 students.

Figure 1 shows the trend of the mean values of the evaluation points for all the lectures since 2004. The results for the first and second semesters are separately dealt with because the types of classes are a bit different from each other. We can see that the evaluation scores of lectures by students have been proportionally rising up year by year, which means that lectures have been improved. The author believes that the decision of score opening to the public made this progress as well as the teacher's effort.

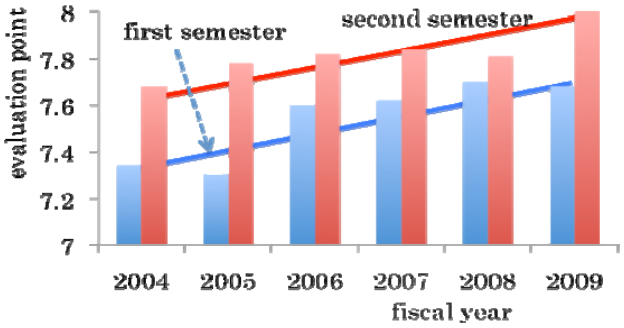

Figure 1. Trend of the mean evaluation points for all the official faculty unsigned lecture questionnaires.

Trend of Web-Based Lecture Questionnaires

Since 2003, the author has been using the Web-based questionnaires at each lecture time. Thus, 15 times answers are obtained to each student. The questions are that: (1) to urge review of the lecture, the author asked "what the points were"; (2) to know how attractive the lecture, the author asked "what the discoveries were"; (3) to know to what extent students understand the lecture, the author asked "what the questions were"; and (4) to find the technical skill for lectures, the author asked "what the improvements were". In addition, "the comprehension points" and "the overall satisfactory points" are marked by 1-10 scores; the higher, the better.

Figure 2 shows a trend of the evaluation points to subject A (Statistics and Data Analysis) that scores are the understanding points and the overall points. We can see that: (1) the overall points are a bit larger than the understanding points; (2) there is a high correlation (correlation coefficient value is 0.83) between the overall points and the understanding points; and (3) the points have been rising up year by year similarly to the official unsigned results. This also indicates that continuing the questionnaires may enhance the improvement of the lecture skills.

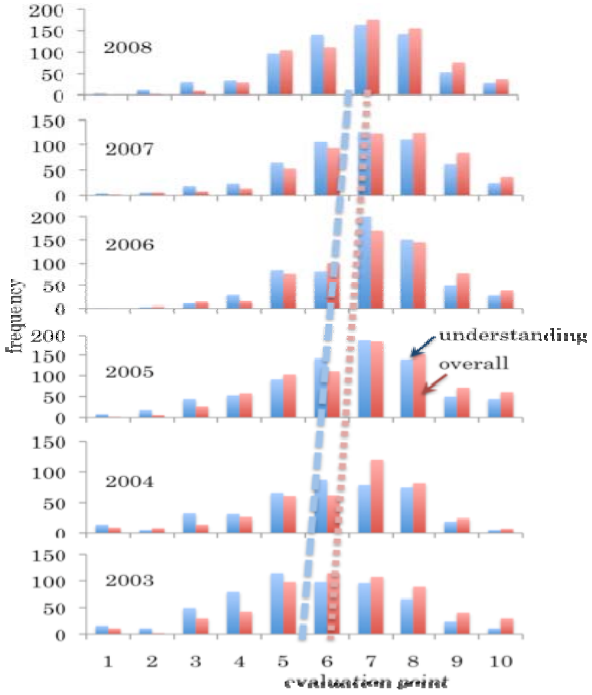

Figure 2. Trend of the mean evaluation points of a signed lecture questionnaire result.

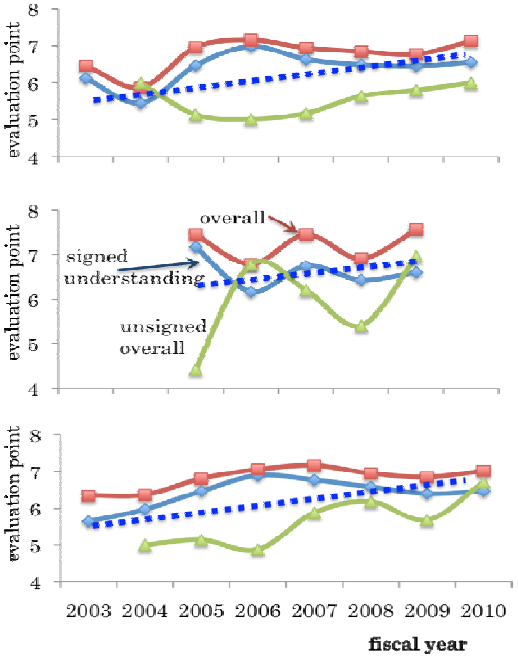

Figure 3. Comparison between the signed and unsigned questionnaires.

Signed and Unsigned Lecture Questionnaires

The official questionnaires are carried out in unsigned; on the contrary, the Web-based questionnaires are carried out in signed. Are there any differences between the two answers? Although we cannot know the differences to each student, the mean values can be compared with each other. Figure 3 shows the differences between the two evaluation scores to subjects A, B (mathematical computation), and C (probability theory). If

there is a positive correlation between the two evaluation scores, we can use one of either signed or unsigned questionnaires. Although few unsynchronized relationships between the signed and unsigned evaluation scores are observed, the trends between them are roughly the same. These trends are similar to those in Figure 1. The unsigned questionnaires are carried out only one time at the final lecture time and the Web-based signed questionnaires are carried out at every lecture time. Did this cause the difference? Figure 4 indicates a denial to this conjecture. The figure shows the evaluation scores to each lecture time. Fluctuations due to the lecture time are not observed. Thus, we may rely on the use of one of either signed or unsigned questionnaires. We will later discuss the detailed analysis for the unsigned Web-based lecture questionnaires using the IRT.

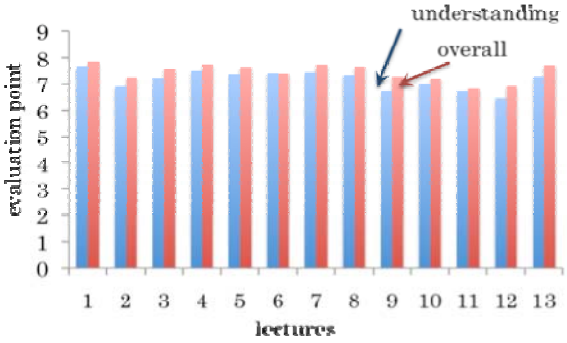

Figure 4. Evaluation scores to each lecture time.

Distribution of the Evaluation Scores

We can roughly grasp the trend of evaluation by looking at the mean values of scores for questionnaires. However, we should be cautious and mean value of 5 is obtained when all the students give point 5 at every lecture time, and also obtained when half the students give point 1 at every lecture time and the rest of half gives point 9 at every lecture time. The evaluation distribution reveals the details for the evaluation. Figure 5 shows typical cases for evaluation distribution.

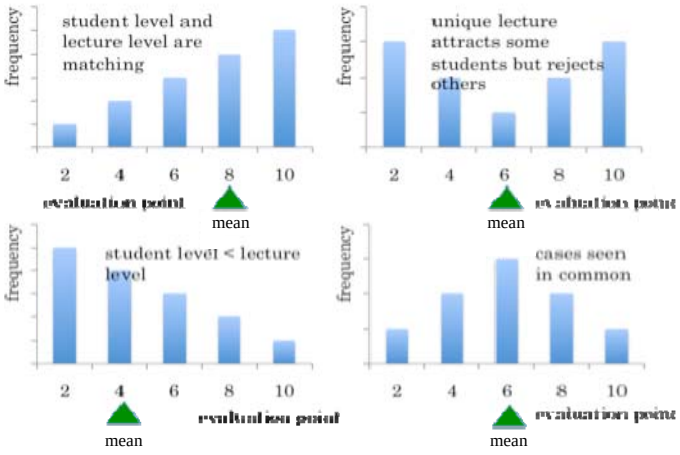

Figure 5. Typical cases for evaluation distribution.

Figure 6 shows an actual distribution of subject B. This subject starts in 2005. First, the teacher delivered the high level to students, in 2006, he changed the lecture level much easier and in 2007, he could find the

appropriate lecture level. To know the distribution of the evaluation, scores is important.

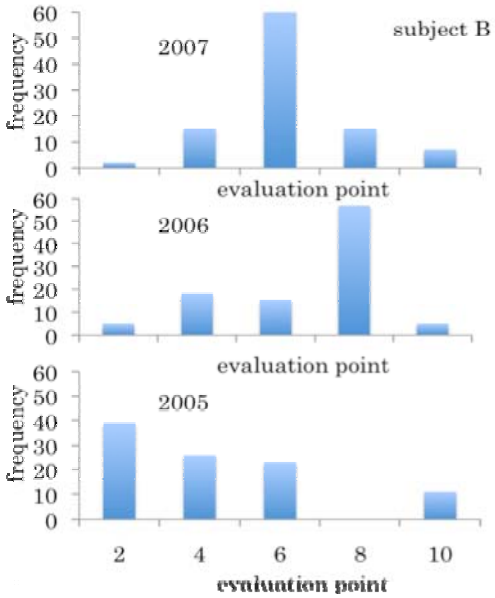

Figure 6. An actual evaluation distribution of subject B.

Analysis of Lecture Questionnaires Using the IRT

So far, we have investigated the results of the evaluation scores for questionnaires using the basic statistical methods believing that the evaluation scores are true. However, some students may give scores 5 every time, and some may carefully give appropriate scores. We may not believe in the scores for questionnaire as they are. Thus, we have tried to analyze the scores as accurate as possible using the IRT which will be shown in appendix.

Figure 7 shows the score data matrix of Subject A in 2006; column corresponds to lecturing date (lecture time) and line corresponds to student id; dark purple regions mean the absences. We can see that some students give high points at every lecture and some low. With a small number of exceptions, we do not observe the drastic fluctuation during the series of lectures. On the right in the figure, the results of IRT analysis are given. After obtaining the IRT parameters of problem difficulties and students' abilities, we have rebuilt the score matrix using the success (evaluation) probability equation (1) to every cell in the matrix.

Looking at the two figures, we can see that: (1) there is a similarity between the two matrixes; (2) the unreliable evaluation fluctuations seem to be relaxed; (3) each student's evaluation stands out sharply; and (4) there seem little fluctuations among lecture times. Therefore, we may believe in the true scores from the Web-based questionnaires. This assures that the unsigned questionnaires are also reliable due to the similarity between the signed and unsigned questionnaires shown before.

The parameters of problem difficulties are seen in Figure 8 where the curves of item characteristics to each lecture day are shown and the discrimination parameters are all small and similar to each other. Then, we can proceed to use this result with confidence; that is, the results of the lecture questionnaires are reliable.

Figure 9 shows the re-evaluated trend for overall scores using the IRT from 2003 to 2008 fiscal years. Comparing this figure with Figure 2, we can find the sharpness to Figure 9. That is, it is recommended to use the IRT re-evaluation together rather than to use the basic statistics alone.

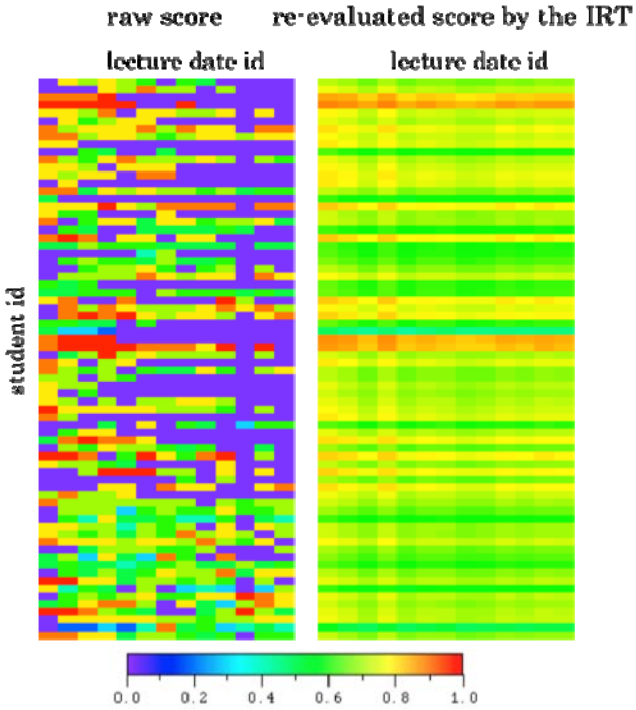

Figure 7. An example of the evaluation distribution matrix (on the left the raw data are given, on the right the IRT re-evaluated scores are given).

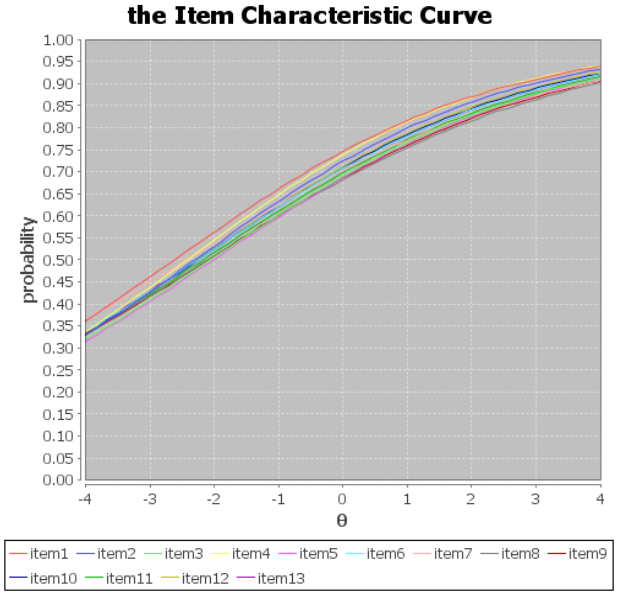

Figure 8. An example of the item characteristic curve.

Concluding Remarks

In this paper, firstly, we have shown the recent statistical analysis results for the official unsigned lecture questionnaire evaluation results took in our faculty, where we have found that the evaluation scores of lectures by students have been rising up year by year, which means that lectures have been improved. The author believes that the decision of score opening to the public made this progress as well as the teacher's effort. We have, next, investigated the features for the signed and unsigned questionnaires. We have found the following:

(1) continuing the questionnaires will improve the lecture skills; (2) although few unsynchronized relationships between the signed and unsigned evaluation scores are observed, the trends between them are roughly the same; and (3) it is important to know the distribution for the evaluation scores. Lastly, we have analyzed the questionnaire evaluation results using the IRT, where the results of the lecture questionnaires are found to be reliable.

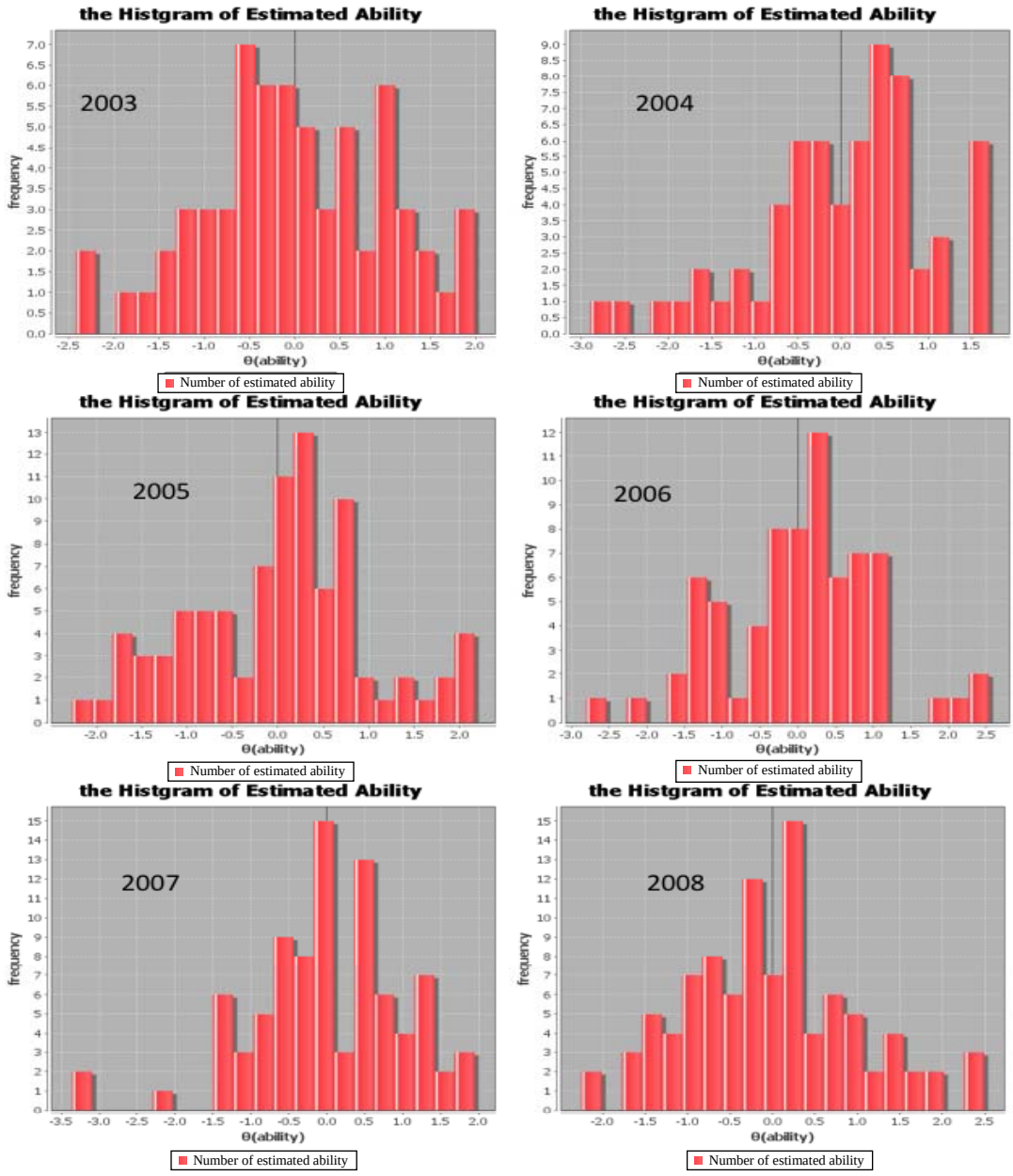

Figure 9. An example of the re-evaluated trend using the IRT from 2003 to 2008 fiscal years.

References

Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation technique* (2nd ed.). Marcel Dekker.

Berk, R. A. (2006). *Thirteen strategies to measure college teaching: A consumer's guide to rating scale construction, assessment and decision making for faculty, administrators and clinicians*. Stylus Pub, LLC.

Bilog, M. G., (2005). *Bilog for Windows, program.* Retrieved from http://www.ssicentral.com/irt/index.html

Davis, B. G., Robert, L. W., & Wilson, C. (1983). *A Berkeley compendium of suggestions for teaching with excellence*. University of California.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society (Series B), 39*(1), 1-38.

Dougiamas, M., & Taylor, P. (2003). Moodle: Using learning communities to create an open source course management system (pp. 171-178). Proceedings of *World Conference on Educational Multimedia, Hypermedia and Telecommunications*.

Hambleton, R. K., & Swaminathan, H. (1984). *Item response theory: Principles and applications*. Springer.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications.

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*(1), 79-87.

Hirose, H., & Sakumura, T. (2010). An accurate ability evaluation method for every student with small problem items using the item response theory (pp. 152-158). Proceedings of *the International Conference on Computer and Advanced Technology in Education*, Hawaii.

Linden, W. J. D., & Hambleton, R. K. (1996). *Handbook of modern item response theory*. Springer.

Mills, C. N., Potenza, M. T., & Fremer, J. J. (2002). *Computer-based testing: Building the foundation for future assessments*. Lawrence Erlbaum.

Sakumura, T., & Hirose, H. (2010a). Test evaluation system via the Web using the item response theory. *Information, 13*(3), 647-656.

Sakumura, T., & Hirose, H. (2010b). Student ability evaluation using the stress-strength model when ability is the random variable (pp. 865-868). *The 2010 International Congress on Computer Applications and Computational Science*, Singapore.

Tsukihara, Y., Suzuki, K., & Hirose, H. (2008). A small implementation case of the mathematics tests with the item response theory evaluation into an e-learning system. *Computer and Education, 24*(1), 70-76.

Appendix: Item Response Theory

For effective evaluation of students' abilities, the IRT (Hambleton & Swaminathan, 1984; Hambleton, Swaminathan, & Rogers, 1991; Linden & Hambleton, 1996; Baker & Kim, 2004) can be used, and this gives the students' abilities accurately in addition to the problem difficulty. Adaptive e-learning systems (Mills, Potenza, & Fremer, 2002) and test methods appropriately used may enhance this feature. We have introduced a student self-learning system (Tsukihara, Suzuki, & Hirose, 2008) embedded in the e-learning system, via Moodle (Website, http://moodle.org/), and a new adaptive test method is also proposed recently (Hirose, 2011) to perform the optimal test. Moreover, we have introduced a Web-based students' evaluation system (Sakumura & Hirose, 2010; Hirose & Sakumura, 2010) and a stress-strength model based ability evaluation system (Sakumura & Hirose, 2010).

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

$$P_j(\theta_i, a_j, b_j) = 1/[1+\exp\{-1.7a_j(\theta_i - b_j)\}] \quad \text{..... (1)}$$

where denotes the indicator function such that 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, 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 (see Figure 10).

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

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:

$$L(U \mid a,b) = \prod_{i=1}^{N} [\int_{-\infty}^{\infty} g(\theta) \prod_{j=1}^{n} L(u_{ij} \mid a_j, b_j) d\theta] \quad \text{..... (3)}$$

Where $g(\theta)$ denotes the ability common to all the students (usually a standard normal distribution) and U denotes all the patterns of u_{ij}, taking the value of 0 and 1. The EM (estimation maximization) algorithm (Dempster, Laird, & Rubin, 1977) is usually used in such a case (Baker & 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 (Baker & Kim, 2004).

To the scores of the lecture questionnaires, we cannot use Equation (2) as it is; that is, we have assumed that $\delta_{ij} = 0.1$, the discrete value. Thus, we have modified to allow the continuous value to δ_{ij}, such as $0 \le \delta_{ij} \le 1$. For convenience, the vacant cells are occupied in advance with the mean observed values to each student.

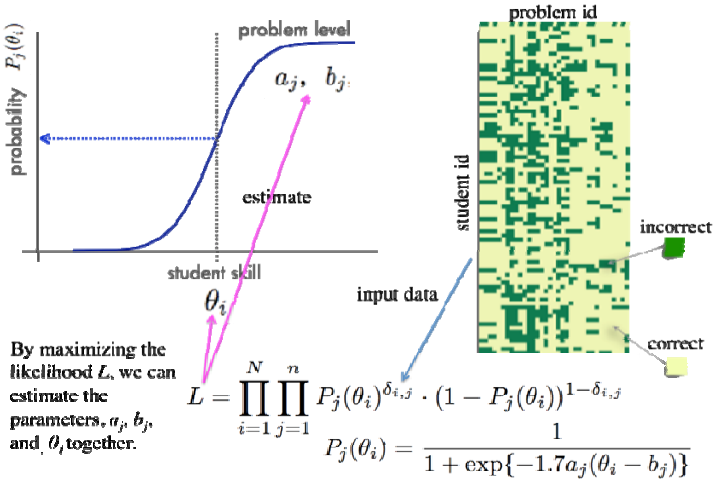

Figure 10. IRT estimation procedure.