Proposal and Evaluation of a Method of Estimating the Difficulty of Items Based on Item Types and Similarity of Choices

Shinichi IKEDA[a], Teruhiko TAKAGI[b], Masanori TAKAGI[c], Yoshimi TESHIGAWARA[a]

[a] *Graduate School of Engineering, Soka University, Japan*

[b] *Graduate School of Information Systems, University of Electro-Communications, Japan*

[c] *Faculty of Software and Information Science, Iwate Prefectural University, Japan*

{e11m5202, teshiga}@soka.ca.jp

Abstract: In recent years, on the idea of supporting the composition of the tests by using statistical data, such as the difficulty level of the items that constitute the tests, has been investigated. In general, item response theory (IRT) is used to quantify the difficulty level of items. However, this approach requires that the items are answered by many learners in advance and it is difficult to ensure that all items in the bank are answered. We propose a method of estimating the difficulty level of unanswered items. In our method, the level of new items is estimated from the level of similar existing items based on the differences between item types and the similarity between choices. A simulation experiment shows that the difficulty levels calculated by IRT and by using the proposed method can have a reasonable correlation. However the results obtained using the new estimation method can be very different from the IRT results if incorrect answers to an item are similar to the correct answer.

Keywords: IRT, E-testing, Difficulty Level, Similar Item, Item Types

Introduction

In recent years, Web-based testing, commonly referred to as “e-testing”, has been attracting much attention [1][2]. In e-testing, more reliable tests can be conducted by preparing an item bank with managed statistical data [3] that includes information on the difficulty level of items and correct answer rate. In addition, a number of studies in the literature have shown support for composing tests through the use of such statistical data [4][5][6]. In these studies, item response theory (IRT) [7] is used to quantify the difficulty level of test items. In order to estimate the difficulty level, the items need to be answered by many test takers (subjects) in advance. Furthermore, new items are added periodically to replace items in the item bank. However, it is hard to ensure subjects answer all items in the item bank to gather complete data, and estimating the difficulty of new items when they are added to the item bank takes time and resources.

Therefore, the objective of this study is to estimate the difficulty level of unanswered items. The difficulty of items can change depending on how the question is asked (the item type) and the similarity of answer choices [8][9] and, in this paper, we focus on such differences. We also restrict our considerations to multiple-choice items. We propose a method of estimating the difficulty level of items by comparison with existing “similar” items. Similar items are defined as being those where the knowledge questioned and the

knowledge needed for the solution are similar. Items are then classified according to a measure of similarity [10].

1. Item Response Theory

This section examines a method for estimating the difficulty level of similar items using item response theory (IRT). A statistical model, called the IRT model, is used to reveal the statistical properties of test items. The properties of items are given by the item characteristic curve (ICC), where the vertical axis is the correct answer rate and the horizontal axis is latent ability (θ), representing the learning ability of the candidate, which does not depend on the candidate population. In this study, a commonly used two-parameter logistic model (2PLM) is applied. The probability of subject i with learning ability θ_i answering item j correctly is defined as

$$P_j(\theta_i \mid a,b) = \frac{1}{1+\exp\{-Da_j(\theta_i - b_j)\}} \tag{1}$$

where a_j is the discrimination level showing the degree to which item j discriminates between subjects, and b_j is the difficulty level of item j (typically, $-3 \le b_j \le +3$) [7]. Figure 1 shows three ICCs on the same graph. All have difficult combinations of discrimination level and difficulty level. When the curve moves to the right, the difficulty level of the item increases because the probability of a correct answer is low at the lowest ability level. When the curve becomes steep, the discrimination level of an item is high. In the 2PLM, the slope of the curve is maximized when the probability of a correct answer is 0.5, and this value of the slope is the discrimination level. In addition, when the answers of the n items of subject i are given by $\mathbf{u}_i=\{u_{i1}, u_{i2}, \cdots, u_{ij}, \cdots, u_{in}\}$, where $\mathbf{u}_i$ is 1 in the case of a correct answer and 0 in the case of an incorrect answer, the probability of the vector $\mathbf{u}_i$ is given by

$$P_j(u_i \mid \theta_i) = \prod_{j=1}^{n} p_j(\theta_i)^{u_{ij}} q_j(\theta_i)^{1-u_{ij}} \tag{2}$$

,

where $p_j(\theta_i)$ is the correct answer rate of subject i to item j, and $q_j(\theta_i) = 1 - p_j(\theta_i)$.

By using such a model, it is possible to estimate the learning ability θ of a subject, the discrimination level (a), and the difficulty level (b) from the test answers of all the subjects. However, the subjects must answer items in advance for these parameters to be estimated. In general, the number of answers required in order to estimate the difficulty level of items using IRT is 1,000 in 3PLM and 500 to 1,000 in 2PLM [11].Therefore, in this paper, a method for estimating the difficulty level of unanswered items is studied.

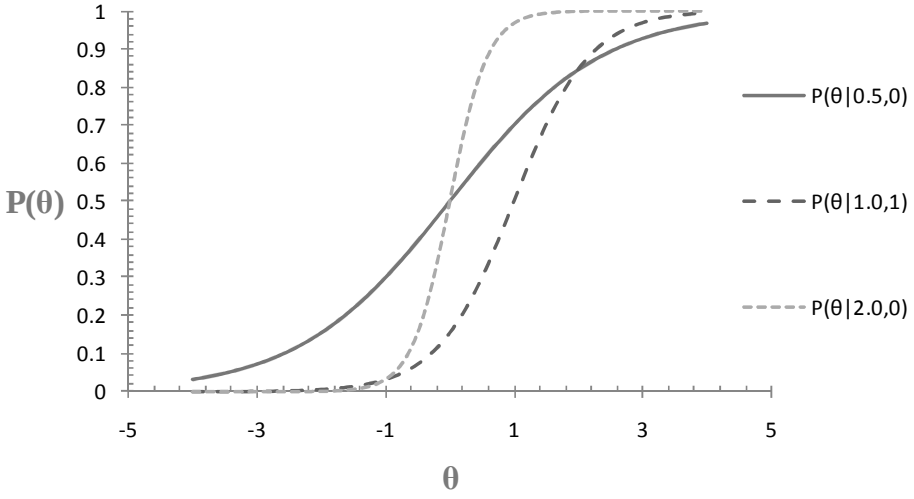

Figure 1 Item Characteristic Curve of 2PLM

Table 1 Item Types

Type ID	Item Type (Example)
Pa+	Select a correct example or explanation of a technical term. (Select the correct description of Morse code.)
Pa-	Select an incorrect example or explanation of a technical term. (Which of the following is not a Real-time Distributed System?)
Pb+	Select a technical term having the same type or attributes as a given technical term. (Which of the following is a type of visual communication?)
Pb-	Select a technical term having a different type or attribute from a given technical term. (Which of the following is not a type of visual communication?)
Pc+	Select a correct example or explanation of something relating to a technical term. (Which of the following is a problem affecting data management in a distributed environment?)
Pc-	Select an incorrect example or explanation of something relating to a technical term. (Which of the following does not have an impact on the structure of a computer network?)
Pd+	Select a correct technical word relating to a given technical term. (Which of the following devices is suitable for telephone communication?)
Pd-	Select an incorrect technical word relating to a given technical term. (Which of the following practical applications of a computer network does not appear in banks or convenience stores?)
Pe+	Select a correct combination of a technical term and an explanation of it. (Select a correct description of 4 layers in the OSI reference model.)
Pe-	Select an incorrect combination of a technical term and an explanation of it. (Which of the following is not a correct description of 4 layers in the OSI reference model?)
Pf	Select a correct technical term based on an example or explanation of it. (What is the host-centralized system which uses a single host computer and multiple terminals?)
	Others.

2. Method of Estimating the Difficulty Level

2.1 Item Type and Difficulty Level

In a preceding study, items were classified into 11 types according to how knowledge is tested [12]. We classified items based on the basis of their content and the answer choices. Also, when classifying items, we took into account whether the item requested the subject to select a correct or an incorrect answer. Table 1 shows the 11 item types and provides examples. In Table 1, the “Other” category includes computational items, fill-in-the-blank items, and flawed items.

The difficulty level of items can change depending on the phrasing used, such as whether the item seeks knowledge of a technical term (Pa and Pf in Table 1) or the item requires the subject to apply knowledge and use it for the answer (Pc and Pd in Table 1). Thus, for similar items, the difficulty level of unanswered items may be estimated by focusing on differences in item type.

In our proposed method, the difference of difficulty level between similar items *i* and *j* which arises from the differences in the item types is defined as

$$bp_{ij} = (D_{\max} - D_{\min}) \cdot sw_{ij} \quad (3)$$

where D_{max} is the maximum difficulty level in the Item bank, D_{min} is the corresponding minimum, and sw is the rate of changes for the range of difficulty levels (D_{max} - D_{min}). Thus, when the difficulty level of an item is known, the difficulty level of similar items can be estimated by adding the difference of difficulty level calculated by formula (3).

2.2 similarity of answer choices and difficulty level

In the case of multiple-choice items, the difficulty level may change according to the similarity of answer choices [9]. For example, the difficulty level of items is increased when the choices include a "confounding answer." On the other hand, the difficulty level of items is decreased when the choices contain an "obviously correct or incorrect answer." One possible measure of the difficulty level of items is the probability that each answer choice is selected (the selection probability). Thus, we propose a method of estimating the selection probability from the similarity of answer choices [13]. In this method, in order to estimate the selection probability, the similarity of each answer is calculated by representing the documents as a weighted collection of terms in a vector space. However, it may not be possible to calculate the similarity if there are few terms contained in the question and answer choices. So, terms that are related to the question or answer choices (related terms) are extracted from the item bank. Then, the similarity of answer choices is calculated using the related terms.

Therefore, in proposed method, the difference of the difficulty level which arises from the difference in the similarity of answer choices between similar items i and j is defined as

$$bc_{ij} = (D_{\max} - D_{\min})(cv_i - cv_j)bp_{\max} \quad (4)$$

where D_{max} and D_{min} are the same values as in formula (3), v_i and v_j are the unbiased variances of the selection probability for items i and j, c is the number of answer choices and bp_{max} is the maximum value of the difficulty level difference calculated using formula (3). Thus, when the difficulty level of an item is known, the difficulty level of similar items can be estimated by adding the difficulty level difference calculated using formula (4).

2.3 Calculation Procedure for Difficulty Level

In this study, the difficulty level of item i (the estimation item) is estimated using the formula

$$b_i = \frac{1}{n}\sum_{j=1}^{n}(bs_j + bp_{ij} + bc_{ij}). \quad (5)$$

Here bs_j is the difficulty level of one of n similar items (the comparison items), The changes in the difficulty level arising due the difference in item types bp_{ij} are estimated by IRT. The changes in the difficulty level between similar items bc_{ij} are based on the differences of selection probability for each answer choice. The difficulty level of estimation item i is calculated by adding bp_{ij} and bc_{ij} to the difficulty level of similar item bs_j. The average value over the n comparison items is used as the difficulty level of estimation item i.

Figure 2 shows the method for estimating the difficulty level proposed in this study. First, the difficulty level of items used in the test are estimated using IRT (estimated items - bs_j), as shown in Figure 2-(1). Then, from the same item bank, the target items for which the difficulty level must be checked are selected as the estimation items (i). Next, the estimation items are used to select similar items from the already estimated similar items (comparison items) as shown in Figure 2-(2). Then, the changes in difficulty level (bp_{ij}, bc_{ij}) are calculated based on the differences in the item type and the selection probability of each answer choice between estimation item i and comparison item j, as shown in Figures 2-(3)

and 2-(4). After that, Equation (5) is used to calculate the level of difficultly for the estimation items (b_i) using the results from steps 3 and 4. Finally, the calculated result (b_i) is registered as shown in Figure 2-(5).

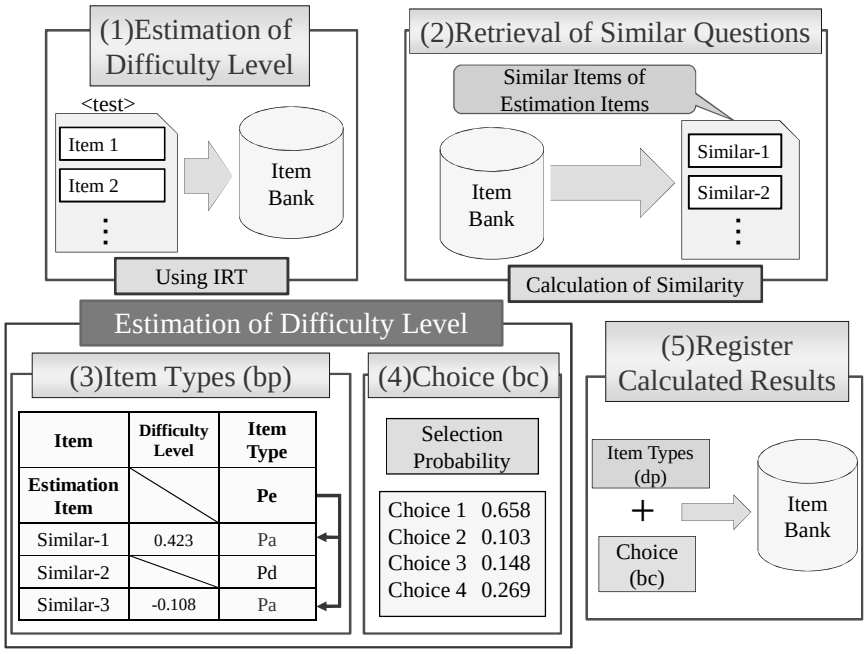

Figure 2 Procedure for calculating level of difficulty

3. Experiment

3.1 Experiment Outline

In this section we describe a comparative experiment that was conducted to verify the relevance of the difficulty level estimated by the proposed method. In this experiment, 1000 items given in previous "System Administrator" and the "Fundamental Information Technology Engineer" examinations are accumulated in the item bank. The differences and the correlation coefficient of difficulty level estimated by the proposed method and IRT are calculated. In the proposed method, the difficulty level is estimated in three ways: using only item types (Dp), using only similarity of answer choices (Dc), and using both item types and similarity of answer choices (Dp+Dc). The 1000 items are used to extract the related terms for estimating the selection probability of answer choices, and then bc_{ij} is calculated from the selection probability.

First, a test consisting of 20 items (Test 1) was conducted. Then, the difficulty levels of the items set in Test 1 were estimated using IRT with a 1PLM. These items were then used as comparison items. In this experiment, Test 1 was given to 82 students in three universities: Soka University, the University of Electro-Communications and Iwate Prefectural University. Second, 15 items similar to those in Test 1 were retrieved from the item bank. Furthermore, 5 items contained in Test 1 are used for items of Test 2 and those items are used for equating of Test1 and Test2. Third, a test consisting of these similar items (Test 2) was conducted and the difficulty levels of the items were estimated by IRT and using the proposed method in three ways (Dp, Dc, Dp+Dc). Finally, the differences and correlation coefficient of difficulty levels estimated by the proposed method and IRT were calculated.

Table 2 Estimation results for difficulty levels.

Item	1	2	3	4	5	6	7
Dp	0.50	0.78	-0.11	1.37	0.45	0.85	-1.95
Dc	-0.25	0.80	-0.60	1.26	0.78	-0.52	-1.88
Dp+Dc	0.48	1.29	-0.11	1.26	0.18	0.22	-1.85
IRT	-0.66	1.40	-0.26	-1.90	-2.36	4.35	-1.90

8	9	10	11	12	13	14	15
-0.57	-1.30	0.76	0.64	0.34	-0.73	-1.50	0.27
0.65	-0.72	0.65	0.86	0.26	-0.96	-1.60	0.30
0.05	-0.94	0.65	0.83	0.49	-1.18	-1.60	0.27
-0.26	-4.10	0.96	-2.36	-1.47	-1.90	-1.90	-1.06

Table 3 The differences in the difficulty levels for each method.

Item	1	2	3	4	5	6	7
Dp	1.16	0.61	0.15	3.27	2.81	3.50	0.05
Dc	0.41	0.60	0.33	3.16	3.14	4.87	0.02
Dp+Dc	1.15	0.11	0.15	3.16	2.54	4.13	0.05

8	9	10	11	12	13	14	15
0.31	2.80	0.20	3.00	1.82	1.17	0.41	1.33
0.91	3.39	0.31	3.22	1.73	0.95	0.30	1.36
0.31	3.16	0.31	3.19	1.96	0.72	0.30	1.33

3.2 Experimental Results

Table 2 shows the difficulty levels of items estimated by the proposed method and IRT. Table 3 shows the differences of difficulty levels between the proposed method and IRT. The correlation coefficients between the difficulty levels estimated by the proposed method and those found by IRT are 0.46 (Dp), 0.12 (Dc), and 0.37 (Dp+Dc). On the other hand, the correlation coefficients are 0.72 (Dp), 0.76 (Dc), and 0.80 (Dp+Dc) when items which have a large difference of difficulty level (items 4, 5, 6, 9, and 11) are removed. Therefore, the difficulty levels of items which have a small difference of, difficulty level estimated by IRT could be predicted quite well using the proposed method.

Since the correlation coefficient for method Dc is the lowest, the estimation of difficulty levels could be affected according to selection probability. Table 4 shows the selection probabilities of large difference items estimated using the proposed method (estimated selection probability) and calculated using answer data (calculated selection probability). The correct answer rates of items 4 and 5 of Test 1 and item 6 of Test 2 are 30% or less. In particular, the calculated selection probability of an incorrect answer choice is the highest in items 4 and 6. In the proposed method, the difficulty level of items is estimated from the difference of the variances of selection probability. However, the variances of selection probability may become equal even if the difficulty level of items is different. Thus, in the case of items for which selection probability of an incorrect answer choice is the highest, the difference between the difficulty level estimated by the proposed method and IRT becomes large because the difference of the difficulty levels is not calculated correctly from the selection probability. For item 5, the variance of selection probability is increased

Table 4 Selection probabilities of items with large differences.

Calculated Selection Probability		Choice 1	Choice 2	Choice 3	Choice 4
Item 4	Test 1	0.27	0.43	0.17	0.13
Item 5	Test 1	0.30	0.21	0.22	0.27
Item 6	Test 2	0.21	0.13	0.57	0.09
Item 9	Test 2	0.88	0.04	0.04	0.04
Item11	Test 1	0.21	0.17	0.24	0.38

Estimated Selection Probability		Choice 1	Choice 2	Choice 3	Choice 4
Item 4	Test 1	0.65	0.35	0	0
Item 5	Test 1	0.39	0.25	0.31	0.05
Item 6	Test 2	0	1	0	0
Item 9	Test 2	0.50	0.17	0	0.33
Item 11	Test 1	0	0.26	0	0.76

because the selection probability of answer choice 4 is estimated to be low. Item 9 of Test 2 is an easy item for which the correct answer rate is about 90%. However, if the selection probability of an incorrect answer rate is estimated to be high, the difference of difficulty levels becomes larger.

On the other hand, items for which the estimated selection probability is 0 are constrained because there are no related terms in the proposed method. In item 6 of test 2, the related terms for incorrect answer choices do not exist, and the selection probability of correct answer choice becomes 1. For item 11 of Test 1, for which the correct answer is "router", the selection probability of incorrect answers "gateway" and "repeater" are 0 because the related terms do not exist. However, these incorrect answers are similar to correct answer and the calculated selection probability is 0.2 for both incorrect answers. Thus, the difference of the difficulty levels becomes larger. The results of the experiment show that the estimation of the difficulty level is strongly influenced by the estimation of the selection probability. Therefore, it is necessary to develop a more accurate method for estimating the selection probability.

4. Conclusions

In order to estimate the difficulty level of unanswered items, the items were first classified into 11 types according to how knowledge is tested. We then estimated the difficulty level of items based on item types and similarity of answer choices. In our proposed method, the difficulty level of similar items is estimated by comparing the item types and the selection probability of answer choices with those of some similar items for which the difficulty levels have already estimated. In addition, a method for estimating the selection probability of answer choices based on representing the documents as weighted collections of terms in a vector space is proposed.

The results of an experiment show that the proposed method could provide estimates of difficulty levels which are close to those estimated by IRT. Therefore, the difficulty level of unanswered items could be estimated when new items are added to an item bank, thus reducing the costs and time when constructing an item bank. However, for some items the selection probabilities of answer choices are not estimated correctly, so the difference between the difficulty levels estimated by the proposed method and by IRT is large. In the future, we plan to develop an extended method to estimate the selection probability more

accurately, with a focus on the method of weighting related terms and the deletion of unnecessary terms.

References

[1] Ueno, M., (2005). Web based computerized testing system for distance education. *Educational Technology Research*, 28(1), 59–69.

[2] Ueno, M. & Okamoto, T., (2008). System for online detection of aberrant responses in e-testing. in *Proceedings of the2008 Eighth IEEE International Conference on Advanced Learning Technologies*, ser. ICALT '08. Washington, DC, USA: IEEE Computer Society, 824–828.

[3] Japan Society for Educational Technology, (2000). Educational Technology Dictionary: Jikkyo Press. (in Japanese)

[4] Songmuang, P., Ueno, M. (2008). Development of Prediction System of Score and Time in e-Testing. *Journal of IEICE* J91-D(9), 2225-2235. (in Japanese)

[5] Otomo, K. (2009). Principles and Selected Applications of Item Response Theory. *Journal of IEICE* 92(12), 1008-1012. (in Japanese)

[6] Takahashi, N., Nakamura, T. (2009). Development and Evaluation of the Adaptive Tests for Language Abilities (ATLAN). *The Japanese Journal of Educational Psychology* 57(2), 201-211. (in Japanese)

[7] Baker, F. B. (2001). *The Basics of Item Response Theory. 2nd ed*. (USA, ERIC Clearinghouse on Assessment and Evaluation).

[8] Ikeda, S., Takagi, T., Takagi, M. & Teshigawara, Y. (2011). A Study on a Method of Estimating the Difficulty of Quizzes Focused on Quiz Types, Proceeding from ICCE2011, *The 19th International Conference on Computers in Education*, Chiang Mai, Thailand, 312-316.

[9] Tsumori, S. Kaijiri, K. (2009). A method for automatic generation of multiple-choice questions adapted to student's understanding. *The Journal of Information and Systems in Education* 26(3), 240-251. (in Japanese)

[10] Takagi, T., Takagi, M. & Teshigawara, Y. (2009). A Proposal and Evaluation of a Method of Calculating Similarity between Quizzes Created by Students. *Journal of IPSJ*, 50(10), 2426-2439. (in Japanese)

[11] Ayala, R. J., (2009). *The Theory and Practice of Item Response Theory*. (New York, Guilford).

[12] Takagi, T., Takagi, M. & Teshigawara, Y. (2008). A Proposal and Evaluation of a Similar Level Calculation Method by Type of Quizzes Created by Students. *Proc. IPSJ Summer Symposium in Setouchi*, 95-102. (in Japanese)

[13] Ikeda, S., Takagi, T., Takagi, M. & Teshigawara, Y. (2011). A Study on a Method of Estimating the Difficulty of Items Focused on Item Types and Similarity of Choices. *Proc. IPSJ Summer Symposium in Setouchi*, 133-139. (in Japanese)