Practical Paper

INFORMATION AND SYSTEMS IN EDUCATION, Vol. 16, No. 1 2017
pp. 18–25

Application of the Graded Response Model of Item Response Theory to Computerized Dynamic Assessment in L2 English Education

Sei Sumi* and Yoshimitsu Miyazawa**

(Received 23 April 2017 and accepted in revised form 20 September 2017)

Abstract The purposes of this paper are to (a) develop an adaptive system with the help of the graded response model of IRT, (b) test the system with a small group of L2 learners, (c) examine the feasibility of computerized dynamic assessment, and (d) propose a new set of scoring mechanisms. We developed a computer-based adaptive system that provides students with appropriate test items and a set of graduated prompts from implicit to explicit under the conditions of mediation. We found that the set of graduated prompts successfully guided students to the correct answer and provided rich diagnostic information needed for L2 English education. With the help of the graded response model of IRT, computerized dynamic assessment will become a more advanced and practical educational assessment tool in the field of L2 studies.

Keywords: graded response model, IRT, computerized dynamic assessment, English education

1. Introduction

In April 2017, the Ministry of Education, Culture, Sports, Science & Technology in Japan released its latest survey results on third-year high school students' proficiency in the four English-language skills(1). According to the report, only 36.4% of third-year high school students from 3,390 state-run and other public high schools attained scores exceeding those of the pre-level 2 English Language Proficiency test, which represents an intermediate standard English proficiency level for Japanese high school students. To mitigate the situation, the Japanese government has introduced many initiatives. However, few of the anticipated improvements have been seen yet.

In addition, the sharp decline in the student population has worsened the situation. Japanese universities have started lowering their admission standards and requisite exam scores to enroll as many students as possible. Consequently, students often enter Japanese universities without a basic command of English. Some students have not even attained high school level English abilities. Many universities are addressing this challenge by offering small-group lessons and remedial education classes based on student needs. A key common concern for these universities is how to raise student English abilities in fundamental areas such as *vocabulary* and *grammar*.

With this context in mind, the concept of diagnostic language assessment (DLA) has gained much attention in the field of second and foreign language education(2). Lee, the chief editor of the special issue of *Language Testing*, described the two main parts of DLA: to (a) diagnose language learners' weaknesses and needs, and (b) provide diagnostic feedback and guidance for remedial learning and instruction(3). The concepts and approaches of DLA truly harmonize with the efforts made in Japanese universities.

Among the many articles related to DLA, the work of Poehner et al.(4) has attracted much interest. They insisted that diagnostic feedback should be incorporated from the outset into the design of achievement and proficiency testing to enhance language acquisition, with the aim of broadening the scope of DLA beyond simple diagnostic feedback to the integration of assessment into language curricula. In line with these ideas, Poehner et al. used the term *dynamic assessment*, rather than diagnostic assessment, and they developed a computerized dynamic assessment (C-DA) system.

We have been profoundly influenced by the work of Poehner et al., however, we identified some room for improvement that we address in this paper. The C-DA system designed by Poehner et al. only provides the same test items in the same order regardless of student

*School of Science and Technology, Kwansei Gakuin University, Japan

**Research Organization for Next-Generation Education, Tokyo Gakugei University, Japan

abilities and needs. To enhance the system, we used the graded response model (GRM) of the item response theory (IRT)[5] and developed an adaptive system. This practical paper reports on the use of our adaptive system, based on that of Poehner et al., with a small group of L2 learners. The new system made it possible for us to guide students to the correct answer under the conditions of mediation and to provide fine-grained diagnostic information.

2. Literature Review

2.1 Dynamic Assessment

Dynamic assessment has frequently been compared to its opposite, static assessment which includes IQ tests and standardized tests. Many critics have argued that static assessment has not been used enough to gain useful information for more effective instruction[6, 7]. In many cases with static assessment, learners who do not achieve the test target can easily be underestimated[8]. Even though dynamic assessment has a long history, it received little attention from the 1930s to the 1960s. Awareness of dynamic assessment surged in the late 1960s alongside the critique of standard psychometric tests[7].

Dynamic assessment is intended to intervene in and support the development of problem-solving skills by offering learners appropriate prompts or mediation. With the help of mediation from others who are more capable, learners are expected to demonstrate their abilities better than they could on their own. This idea is based on Russian psychologist L.S.Vygotsky's concept, the zone of proximal development (ZPD).

On the other hand, the zone of actual development (ZAD) is used as a concept that represents the current level of the learners' development, which they can attain independently without any mediation[9]. When new skills in the ZPD have been fully mastered and can be performed independently without assistance, it can be said that they have become part of the ZAD[10].

2.2 Computerized Dynamic Assessment

C-DA is an area that has been receiving increasing attention in recent years, and researchers are beginning to explore the possibility of electronically delivering mediation with its potential advantages, including the (a) capability of test administration for many learners, (b) repeatability of test administration as needed, and (c) automatic generation of reports on learner performance[11].

While there are several known advantages of C-DA, only a few applications have been reported so far in the field of L2 studies. Davoudi and Ataie-Tabar[12] investigated the effect of using C-DA for L2 English writing performance. In addition, Teo[13] developed a C-DA program for enhancing L2 English reading skills.

One of the large-scale and integrated C-DA projects is that of Poehner et al.[4], which covered Chinese, French, and Russian L2 education. The goal of the project was to develop assessment instruments and scoring mechanisms that allowed administrators to better capture and represent learner ZPDs. However, they have only reported on the system for L2 Chinese education.

2.3 C-DA System

The system developed by Poehner et al.[4] asked students to answer online multiple-choice questions on L2 Chinese reading and listening comprehension. The listening test consisted of 23 items while the reading test consisted of 24 items. The items were selected from the Chinese Proficiency Test (HSK), which is publicly available.

The HSK offers several proficiency levels. For the project, the lowest level was selected based on the needs of Chinese language programs at US universities. There were five choices for each multiple-choice question. Originally, there were only four choices available on the HSK; however, to allow students more attempts to answer each question and to offer prompts according to their answers, an additional distractor was added.

A set of prompts as mediation was embedded following questions, moving from implicit to explicit prompts to guide students to the correct answer. The maximum number of prompts students received was four. If a student answered an item correctly on the first attempt, an explanation of why that was the correct answer was offered as the first prompt. If a student answered incorrectly on the first attempt, the second mediating prompt was offered. The third prompt was offered similarly until the correct answer was displayed as the fourth prompt.

Although the C-DA system elaborated by Poehner et al. was both sophisticated and practical, we believed

it could be improved since the system could only provide the same test items in the same order regardless of student abilities and needs. To improve the system, we developed an adaptive system with the help of the GRM of IRT.

3. Method

3.1 Investigating Answering Patterns

To design graduated prompts according to student weaknesses and needs, we qualitatively investigated student answering patterns. We selected 14 Japanese students from a cooperating university in Kobe according to their TOEIC® Test Scores, which ranged from 250 to 640. Their English levels were considered to be low or low-intermediate. We also asked them to take a 100-question multiple-choice English test. The test was originally developed by the students' university and has been used as a yearly placement test. The test reflects the achievement standard of English education of the university in aspects of vocabulary and grammar.

All students took the test individually on separate days. While the students were taking the test, we asked them to think aloud and interactively gave them advice on a one-on-one basis when they seemed to be experiencing difficulty in answering. After they chose an answer, regardless of whether it was correct or incorrect, we asked them to articulate the reasons why they chose that answer. All data were video recorded.

The observation and interaction revealed one important factor to consider when designing prompts: insufficient knowledge of vocabulary and grammar. The many students who struggled with a lack of vocabulary and grammar could not precisely understand the meaning of the English sentences and choices of the test items. When unclear on sentence or answer item meanings, the students often guessed at the meanings with their limited knowledge. Their lack of knowledge led them to incorrect answers with a high probability. This phenomenon was observed frequently, especially among students with lower English proficiency. For example, male Student K, who was deemed to have low-intermediate English proficiency, scored 57 on the test. He answered Item 7 as follows:

Test B: Item 7
My sister is willing to give up the pleasure of eating to _____ weight.

a. lose b. spend c. give d. have

Student K: (started reading the sentence) *"My sister is willing to give up the pleasure of eating to…weight." Well…* (looking back and forth at the question, he chose "have" instead of "lose.")
Interviewer: What are you thinking?
Student K: *Well*, *weight means "fu to ru"* (gain weight in Japanese)
Interviewer: What do you think the whole meaning of the sentence is?
Student K: *"A ne ga"* (my sister), *there are two words I don't understand.*
Interviewer: What are they?
Student K: "Willing" and "pleasure." Then, I picked out the words I knew and connected them. I think I might have seen the word order, "have weight," so I chose it. Just a feeling.
Interviewer: I see. What do you think the whole meaning of the sentence is?
Student K: Well, my sister ate too much and gained weight?

3.2 Designing Graduated Prompts

Based on the data gained by the qualitative research and using the same 100 multiple-choice English test items, we designed graduated prompts from implicit to explicit under the conditions of mediation.

For the 1st degree, the system gives the student a multiple-choice test item without any prompts. If the student's answer is incorrect, the system moves on to the 2nd degree and gives the student the same test item with the first prompt. Since many students struggled with a lack of vocabulary and were carelessly prone to choose incorrect answers, the first prompt provides the Japanese translation of some important words from the sentence and choices. The important words were selected based on the student's error patterns and protocol data obtained in the qualitative research. If the student's answer for the 2nd degree is again incorrect, the system moves on to the 3rd degree and offers the same test item with the second prompt. The second prompt gives the student the Japanese translation of the blank along with grammatical tips. The former explicitly tells

the student what information he/she needs to fill in; the latter grammatically helps the student fill in the blank with the correct answer. The graduated prompts from the 2nd and 3rd degrees are arranged to guide students to the correct answer. If the student's answer is incorrect on the 3rd degree, the system gives the student the correct answer and the Japanese translation of the whole sentence. If the student cannot determine the correct answer even with the prompts, the system automatically estimates the student's ability and adaptively continues the test with less difficult test items.

3.3 Estimation of Item Difficulty

The estimation of item difficulty for each degree with graduated prompts was conducted with 47 students from the same cooperating university. They were different students from those who participated in the qualitative research. The same score range on TOEIC® Test scores (250–640) was applied to the selection of students.

The sample size was rather small for practical reasons. Considering this small sample size, we used Bayesian estimation. Bayesian approaches have been applied to IRT estimation in many ways[14, 15], and the potential advantage of its use has been reported over the use of marginal maximum likelihood in small samples and when the examinee ability distribution is not normal[16].

As for the default prior distributions, the item discrimination parameters were drawn from a log-normal distribution with a mean of 0 and a variance equal to 0.50. The item difficulty parameters were drawn from a normal distribution, also with a mean of 0 and a variance of 0.50. The posterior distribution was used as a reference for the next estimation procedure. The cycle was continued until the distribution of parameters converged consistently. We used *IRT Pro* ver. 2.1 for this procedure.

3.4 Development of the System

The GRM of IRT was applied in the development of the adaptive system. The GRM is an extension of the two-parameter logistic model and is one of the polytomous IRT models which include the partial credit model (PCM) and generalized partial credit model (GPCM). In a study comparing the PCM, GPCM, and GRM in the context of testlet scoring and a simulated data set, the calculated test information for the GRM was seen to be greater than either the PCM or the GPCM across most of the range of theta where theta is ability level[17]. Furthermore, the GRM is appropriate to use when item responses can be characterized as ordered categorical responses, as in our present study with its graduated prompts[18]. In addition, to the best of our knowledge, this is the first attempt at applying the GRM of IRT in the area of C-DA studies for foreign and second language education.

The GRM of IRT is mathematically expressed as:

$$P_{jc}(\theta) = \frac{1}{1+\exp\left(-1.7a_j\left(\theta - b_{jc}\right)\right)} - \frac{1}{1+\exp\left(-1.7a_j\left(\theta - b_{jc+1}\right)\right)} \quad (1)$$

where
$P_{jc}(\theta)$ is the probability of a student at ability level θ choosing category c of test item j correctly,
a_j is the slope parameter of item j, and
b_{jc} is the difficulty parameter of choice c of test item j.

The system is terminated when a student finishes answering the 25th multiple-choice test item. Figure 1 shows the sequential change of the system interface and the graduated prompts that are displayed according to a student's answering pattern. Each test item originally offered four choices, but we added three more to each test item to offer more selections.

4. Analysis

The group of 37 different students from the same cooperating university participated in a pilot test. In addition to student ability (θ), an *actual score*, a *mediated score*, and a *learning potential score* (LPS) were calculated following the model of Poehner et al.[4] The differences in answering patterns according to student abilities were also investigated.

4.1 Actual Score

The actual score represents independent and unassisted performance (ZAD). We assigned a maximum of three points to a correct answer and a minimum of zero

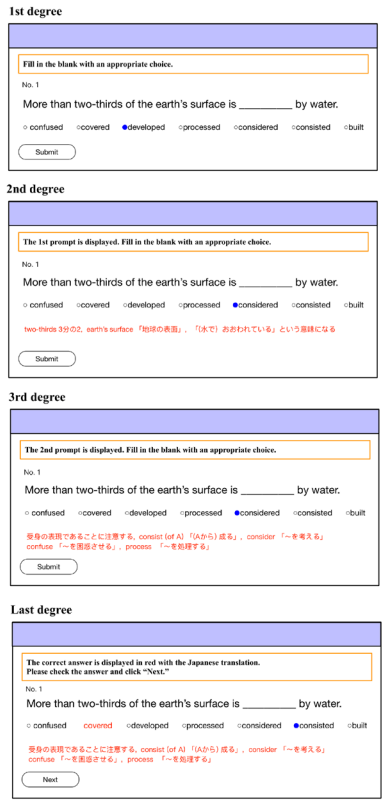

Figure 1. The System Interface.

points to an incorrect answer. The test had 25 test items, yielding a maximum score of 75.

Poehner et al. gave the students the same test items in the same order regardless of student abilities, making the value of each test item and the maximum score equivalent among students. Since our system adaptively provides students with different test items according to their abilities, test score values differ among individuals, but theoretically they reflect the maximum latent ability of each student.

4.2 Mediated Score

The mediated score represents the weighted score of each test item depending on the degree of graduated prompts provided (ZPD). The mediated score ranges from a maximum of three points to zero points. If students choose the correct answer for the 1st degree without any graduated prompts, they receive the maximum of three points. If students choose the correct answer for the 2nd degree, having been given the first prompt, they receive two points. Similarly, the actual score is reduced to one point for students choosing correctly for the 3rd degree, and zero points are given for students who cannot answer correctly even using all the prompts.

4.3 Learning Potential Score

The original idea of the LPS was elaborated by Kozulin and Garb[7], who gave reading strategy training between the pre-test and post-test. When they investigated the effects of this training, they found that students who performed similarly at the pre-test showed different outcomes at the post-test. To represent students' development between the pre- and post-tests via mediation, they conceptualized the LPS and proposed the mathematical formula below.

$$LPS = \frac{\left(Post\text{-}test\ Score - Pre\text{-}test\ Score\right)}{Maximum\ Score} + \frac{Post\text{-}test\ Score}{Maximum\ Score} = \frac{\left(2 \times Post\text{-}test\ Score - Pre\text{-}test\ Score\right)}{Maximum\ Score} \quad (2)$$

The (*Post-test Score*−*Pre-test Score*) indicates the gain score between the pre- and post-tests. This formula reflects how much the total of the gain and post-test score shares in the maximum score. If the total score exceeds the maximum score, the LPS should be above 1.00.

Based on this idea, Poehner et al.[4] developed the formula given in Eq. (3) that uses the term mediated score for the post-test score and the term actual score for the pre-test score. We also applied this formula to calculate LPS in this study.

$$LPS = \frac{\left(2 \times Mediated\ Score - Actual\ Score\right)}{Maximum\ Score} \quad (3)$$

5. Results and Discussion

5.1 Student Ability

The overall change in the students' estimated average ability (θ) was 0.82, and their ability remained con-

sistent at approximately θ=0.80 toward the end of the session set at the 25th test item. Figure 2 shows the overall change in the estimated English ability of four selected students at different levels. The estimated English ability for the 25th multiple-choice test item for these students was (clockwise from upper left) θ=2.68, θ=1.43, θ=0.54, and θ=−0.96. The color bar under the line graph indicates the degree of graduated prompts the students used. The orange indicates the 1st degree without any prompts, the yellow indicates the 2nd degree with the first prompt, the blue indicates the 3rd degree with the second prompt, and the white indicates the 4th degree with the correct answer displayed.

As Figure 2 shows, the student with the highest English ability (θ=2.68) could choose the correct answer for the 1st degree almost without any prompt, while the student with low-intermediate English ability (θ=0.54) more often used prompts provided for the 2nd and 3rd degrees. The student with the lowest English ability (θ=−0.96) rarely chose the correct answer for the 1st and 2nd degrees and was inclined to end up with incorrect answers. In addition, the line graphs of each student show that the students' English abilities gradually converged on a certain point and little change was observed after the 20th test item. This means that the system made it possible to efficiently estimate student English ability in a stable manner and to diagnose individual differences in a short time with a small number of test items.

Based on the results, we can say that the system made it possible to provide students with appropriate multiple-choice test items according to their abilities. With the help of the system, the students with higher English proficiency levels could work on more difficult multiple-choice test items, while the students with lower English proficiency levels were provided with an appropriate multiple-choice test. Comparing the degree of graduated prompts provided to the students' estimated abilities made it possible to understand more about individual students' needs.

5.2 Individual Differences

Table 1 shows the test results of six selected students. The categorical data of the LPS was used in a similar way as by Poehner et al.: high LPS ($\geq$ 0.91), mid LPS (0.71–0.90), and low LPS ($\leq$ 0.70). The ability data indicate the students' estimated ability (θ) on the final test item. Calculating a new set of scoring mechanisms yielded several interesting findings.

Students 5 and 6 had the same actual score of 3, which means that they could correctly answer only one out of 25 test items for the 1st degree without any graduated prompts. However, examining mediated scores and gain scores revealed differences between the students that were not apparent when considering only actual scores.

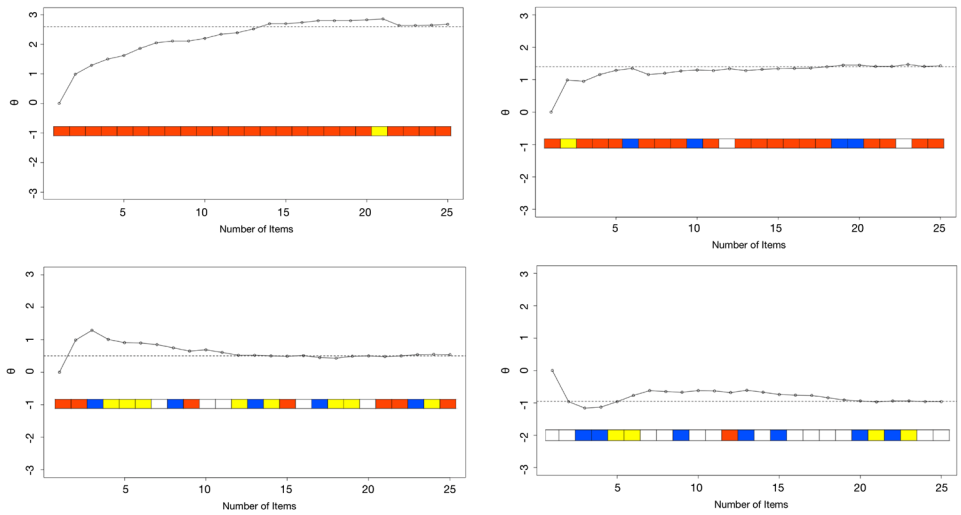

Figure 2. Estimated Abilities of Four Students.

Table 1. Individual Scores of Selected Students.

Student	Actual Score	Mediated Score	Gain Score	LPS	θ	1st Degree	2nd Degree	3rd Degree	Incorrect Answer
1	63	70	7	1.03	2.55	21	3	1	0
2	63	68	5	0.97	2.40	21	2	1	1
3	48	60	12	0.96	1.25	16	3	6	0
4	39	60	21	1.08	1.21	13	10	1	1
5	3	32	29	0.81	−0.29	1	6	17	1
6	3	18	15	0.44	−0.96	1	4	7	13

Student 5 had a mediated score of 32 while Student 6 had a score of 18; similarly, Student 5 had a gain score of 29 while Student 6 had one of 15. These results clearly reflect the differences in LPS. Student 5 was categorized as high with an LPS of 0.81, while Student 6 was categorized as low with an LPS of 0.44.

In addition, when we looked at the degree of graduated prompts that the two students used, we could observe the differences between them in detail. Student 5 chose the correct answer 17 times using the second prompt for the 3rd degree, which gave the student the Japanese translation of the blank and grammatical tips. This indicated the specific area of the student's needs in vocabulary and grammar. The result also told us his latent ability and demonstrated that the student could choose the correct answer at a high rate if he was provided with scaffolding in the form of the Japanese translation of the test items and grammatical tips. These kinds of rich diagnostic information can be useful assets when planning further instruction.

In contrast, Students 3 and 4 showed the same mediated score but different actual scores. Their LPS and estimated abilities were almost the same. However, a detailed look at their answering patterns revealed the difference between them. Student 4 used the second prompt 10 times for the 3rd degree but Student 3 used it only three times. This told us that that Student 4 had a limited vocabulary in comparison to Student 3.

The importance of examining the answering patterns could also be confirmed in the case of students who achieved high scores. Students 1 and 2 had actual scores of 63, and both had limited room for improvement when mediation was offered. As a result, their LPS scores and estimated ability were almost the same. Again, looking at the answering patterns in detail revealed the difference between them: Student 2 was slightly less knowledgeable about basic English vocabulary and grammar.

6. Conclusion

Based on the results, we confirmed that the use of the GRM of IRT and our developed adaptive system allowed C-DA to become a more advanced and practical educational assessment tool in the field of L2 studies. The actual score and estimated ability reflected the students' current maximum independent performance (ZAD). The mediated score, LPS, and answering pattern helped us identify each student's latent abilities and deficiencies (ZPD).

Certain limitations to our research did, however, become clear during the study. Since this study was a pilot practice for evaluating the feasibility of the system and scoring mechanism based on the work of Poehner et al.[4], the number of students who participated was quite small. To improve the accuracy of the system and the resolution of the scoring mechanism, more participants with different English proficiency levels are required. In addition, the study was only applied to a simple multiple-choice test. Future research should test other foreign or second language skills such as listening and reading.

In future studies, expanding the function of the GRM of IRT is a possible improvement. The GRM makes it technically possible to adaptively change the quantity and quality of graduated prompts provided according to student choices. With this technique, the system can come closer to being a real dynamic assessment procedure under the conditions of mediation.

Lee[3] described two main parts of DLA: to (a) diagnose language learners' weaknesses and needs, and (b) provide diagnostic feedback and guidance for remedial learning and instruction. The question becomes how to integrate these elements into language curricula with teaching and assessment in classrooms where more than 20 or 30 students are studying at the same time. As Poehner et al. mentioned, C-DA represents a solution to this concern. Our study also found that C-DA is compat-

ible with adaptive learning. While our findings were limited to one university with a small number of students, we sincerely hope that the present study will inspire further studies in the field of C-DA for foreign and second language education.

References

(1) MEXT: "The Investigation of Students' English Ability for the Improvement of English Education", http://www.mext.go.jp/a_menu/kokusai/gaikokugo/1384230.htm (accessed 2017.4.8) (in Japanese).

(2) Language Testing: "Special Issue on Future of Diagnostic Language Testing", Language Testing, Vol. 32, No. 3, pp. 295–416 (2015).

(3) Lee, Y. W.: "Diagnosing Diagnostic Language Assessment", Language Testing, Vol. 32, No. 3, pp. 299–316 (2015).

(4) Poehner, M. E., Zhang, J. and Lu, X.: "Computerized Dynamic Assessment (C-DA): Diagnosing L2 Development According to Learner Responsiveness to Mediation", Language Testing, Vol. 32, No. 3, pp. 337–357 (2015).

(5) Samejima, F.: "Estimation of Latent Ability Using a Response Pattern of Graded Scores", Psychometrika Monograph Supplement, Vol. 17, pp. 1–100 (1969).

(6) Feuerstein, R., Rand, Y. and Hoffman, M. B.: The Dynamic Testing of Retarded Performers: The Learning Potential Testing Device, Theory, Instruments, and Techniques, Baltimore, University Park Press, MD (1979).

(7) Kozulin, A. and Garb, E.: "Dynamic Assessment of EFL Text Comprehension", School Psychology International, Vol. 23, No. 1, pp. 112–127 (2002).

(8) Campione, J. C. and Brown, A. L.: "Linking Dynamic Assessment with School Achievement", in Dynamic Assessment, ed. Lidz, C. S., pp. 82–115. Guilford Press, NY (1987).

(9) Shabani, K., Khatib, M. and Ebadi, S.: "Vygotsky's Zone of Proximal Development: Instructional Implications and Teachers' Professional Development", English Language Teaching, Vol. 3, No. 4, pp. 237–248 (2010).

(10) McInerney, D. M., Dowson, M. and Van Etten, S. (eds.): Effective Schools., Information Age, Greenwich, CT (2006).

(11) Poehner, M. E.: Dynamic Assessment: A Vygotskian Approach to Understanding and Prompting L2 Development, Springer, NY (2008).

(12) Davoudi, M. and Ataie-Tabar, M.: "The Effect of Computerized Dynamic Assessment of L2 Writing on Iranian EFL Learners' Writing Development", International J. of Linguistics and Communication, Vol. 3, No. 2, pp. 1–11 (2015).

(13) Teo, A.: "Promoting EFL Students' Inferential Reading Skills through Computerized Dynamic Assessment", Language Learning & Technology, Vol. 16, No. 3, pp. 10–20 (2012).

(14) Béguin, A. A. and Glas, C. A. W.: "MCMC Estimation of Multidimensional IRT Models", Psychometrika, Vol. 66, pp. 541–562 (2001).

(15) Janssen, R., Tuerlinckx, F., Meulders, M. et al.: "A Hierarchical IRT Model for Criterion-Referenced Measurement", J. of Educational and Behavioral Statistics, Vol. 25, No. 3, pp. 285–306 (2000).

(16) Natesan, P., Nandakumar, R., Minka, T. et al.: "Bayesian Prior Choice in IRT Estimation Using MCMC and Variational Bayes", Frontiers in Psychology, Vol. 7, pp. 1–11 (2016).

(17) Cook, K. F. and Dodd, B. G.: "A Comparison of Three Polytomous Item Response Theory Models in the Context of Testlet Scoring", J. of Outcome Measurement, Vol. 3, No. 1, pp. 1–20 (1999).

(18) Hays, R. D., Morales, L. S. and Reise, S. P.: "Item Response Theory and Bealth Outcomes Measurement in the 21st Century", Medical Care, Vol. 38, No. 9, Suppl., II, pp. 28–42 (2000).

Sei Sumi received his Ph.D. from Kansai University. He specializes in CALL and has published many books and papers in the field. He is currently interested in applying adaptive technology to foreign language learning and teaching in response to the spread of ICT in education.

Yoshimitsu Miyazawa received his Ph.D. from the University of Electro-Communications and has published many papers in the field of the development and application of e-testing systems. He is currently interested in the application of technology to the dynamic assessment of learning.