Development of Japanese-Computerized Adaptive Test

Shingo Imai
International Student Center, Yamaguchi University
Yoshida 1677-1, Yamaguchi 753-8511, Japan
E-mail: shingoimai@hotmail.com

Abstract: The Japanese-Computerized Adaptive Test (J-CAT) [1] evaluates proficiency in Japanese as a second or foreign language via web. J-CAT consists of a item pool, test algorithm, and data base for recording the response patterns and results of test takers. The items are assigned discrimination parameters, difficulty parameters, and guessing parameters based on the Item Response Theory (IRT). The CAT system delivers suitable items for each test taker according to his or her ability. Unnecessary items for evaluation such as items that are too difficult or too easy for the test taker are not given. This technique reduces testing time while keeping high reliability. Furthermore, IRT guarantees that the test is sample-independent and test-independent. Therefore, the scores have an absolute value even if test takers are given different items.

1. Introduction: Item Response Theory

The testing theory used for J-CAT is Item Response Theory (IRT) [1], which is a logistic model based on the probability of correct/incorrect response patterns of a test taker's ability. Person's parameters and item parameters are calculated from the response patterns. Items parameters include the discrimination parameter, namely how well the item discriminate the ability of test takers, difficulty parameter, namely how difficult the items are, and guessing parameter, namely the possibility of correct response by guessing. In (1) $p_i(\theta)$ indicates the probability of a correct response on item *i* by a test taker whose ability is θ.

(1) $$p_i(\theta) = c_i + \frac{(1-c_i)}{1+e^{-Da_i(\theta-b_i)}}$$

In this formula, *e* is an exponential constant. *D* is a scaling factor, whose value 1.7. *D* is used to change this logistic model equivalent into the normal ogive model, from which the IRT was originally developed from.

The discrimination parameter on item *i* is given as a_i, The b_i is the difficulty parameter on item *i*, and c_i is the guessing parameter on item *i*. Since this model includes three item parameters, it is called the Three Parameter Model. Figure 1 shows the logistic curve of the Three Parameter Model. This curve is also called the Item Characteristic Curve. It indicate the probability of a person with a given ability level getting the correct answer on item *i*. Although the value of theta or ability may infinitively extend to both negative and positive sides theoretically, most examinees fall in between -3 (lower ability) and +3 (higher ability) in a real situation.

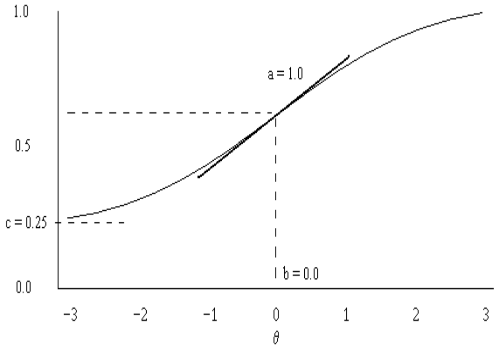

Fig.1. Item characteristic curve (Wikipedia [2])

If c_i is removed as in (2), it becomes a Two Parameter model,

(2) $$P_i(\theta) = \frac{1}{1+e^{-Da_i(\theta-b_i)}}.$$

A One Parameter Model as in (3), the value of a_i is 1.

(3) $$P_i(\theta) = \frac{1}{1+e^{-D(\theta-b_i)}},$$

The curve of ICC shifts from left to right as the *b*-parameter gets higher, or the item becomes more difficult. If the *c*-parameter is set to 0 as in a Two Parameter Model, the central point, or inflection point of the curve is at *p=0.5*. It means that a person whose ability is θ=0 has 0.5 chances of answering the item correctly.

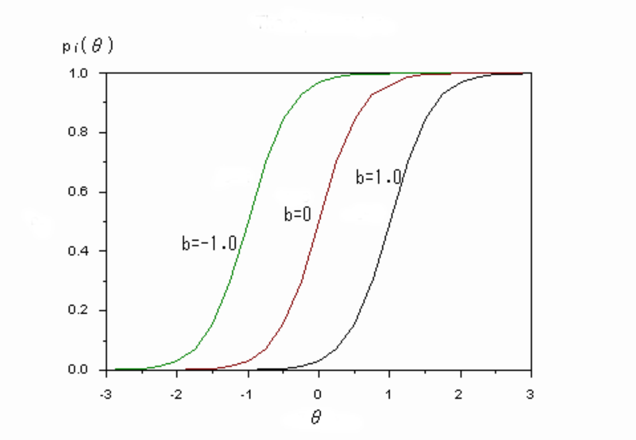

Fig.2. Shift of ICC according to difficulty parameters

The discrimination power is graphically represented as the steepness of the curve of ICC. The curve becomes steeper as the *a*-parameter or discrimination parameter increases. The curve of an item with lower value of the discrimination parameter becomes flat. Such items have less power of discriminating the ability of test takers and are useless in the test.

[1] J-CAT is a registered trademark of Yamaguchi University in Japan.

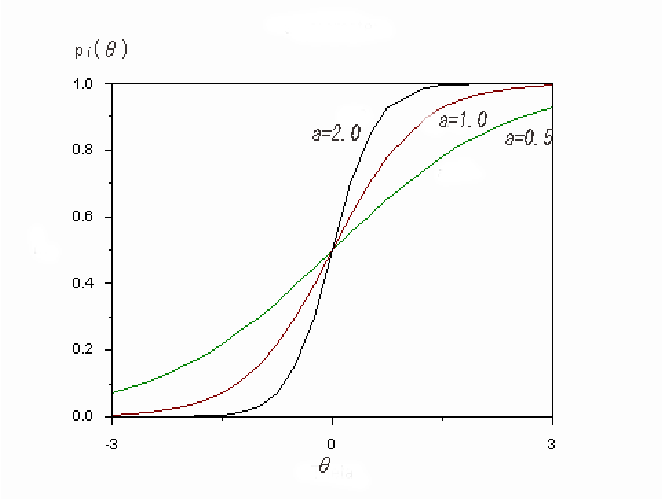

Fig.3. Shift of ICC according to discrimination parameters

Unlike the parameters in the Classical Testing Theory, the parameters of IRT are not sample-dependent nor test-dependent. Thus IRT provides significantly greater flexibility in situations where different samples or test forms are used. These IRT properties are foundational for computerized adaptive test.

2. Mechanism of J-CAT: Testlet version

In this section, the structure and mechanism of J-CAT are introduced. .Several versions of J-CAT have been produced in the course of the development of J-CAT for the last five years. The latest version incorporates Baysian inference, while earlier versions utilize item sets or testlets to determine the level of test takers. In this paper, we will introduce the testlet adaptive system.

J-CAT (testlet version) system consists of several units; the Input unit, Output unit, Time meter unit, Data Unit, Setting unit, Answer units, Level movement unit, Level computation unit, and Result generation unit.

Figure 4 shows the Input unit, Output unit, Time unit, Data unit, and Setting unit. The Input unit and Output units are interface units between the J-CAT system and clients' computer terminals.

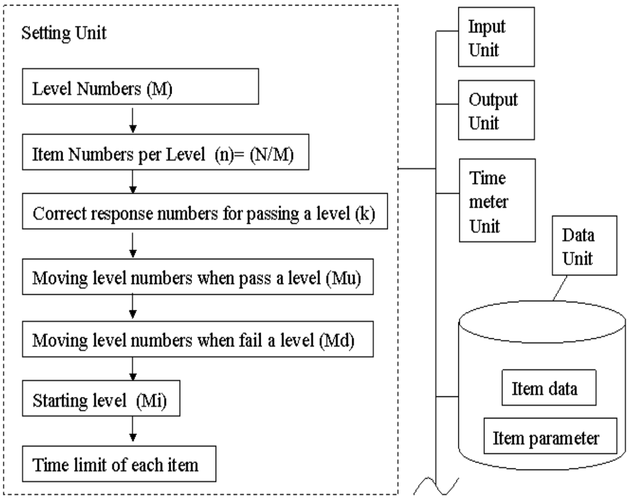
Setting Unit
Level Numbers (M)
Item Numbers per Level (n)= (N/M)
Correct response numbers for passing a level (k)
Moving level numbers when pass a level (Mu)
Moving level numbers when fail a level (Md)
Starting level (Mi)
Time limit of each item
Input Unit
Output Unit
Time meter Unit
Data Unit
Item data
Item parameter

Fig.4. Setting unit

The Data unit stores several kinds of data. The Item data include item IDs, items and multiple choices. J-CAT has four sections; vocabulary, grammar, reading, and listening. All items are in tandem with four multiple choices. One of those four choices is the correct answer and the other three are rated incorrect without any partial points. In a reading test item, a short passage is shown on the screen. In a listening test item, a recorded sound of a short dialog or monolog is delivered via streaming audio. Some listening items are given with picture(s) along with the recording sound. Each item consists of only one question, which is in accordance with the requirement of local independence in IRT. Each item is tagged with a difficulty parameter and a discrimination parameter in the testlet version and with an additionally guessing parameter in the latest version of J-CAT.

In the Setting unit, items are put into groups according their difficulty parameters. Discrimination parameters are referred to when more than required items in a group exist. An item with a higher discrimination parameter is preferred to one with lower discrimination parameter. Item Numbers per Level (*n*) refers to numbers of items in each set or testlet. Each testlet may consist of only one item or more. The accuracy of the test increases in proportion to the number of items in each testlet. However, if the number of items in each testlet increases, the length of the test also increases. There is a trade-off between the accuracy and the length of the test. The adaptive test is designed to reduce test length with the minimum reduction of accuracy.

In J-CAT 2006 testlet version, there are 60 items in each section. There are 20 levels. Therefore, the number of items per level (*n*) is three. In J-CAT 2007 version, three out of four items at each level are randomly chosen and presented. In J-CAT 2006 version and 2007 version, *k* is set to two items, *Mu* to three levels, *Md* to two levels, and *Mi* is to level 10. These numbers can be set according to the test developer's requirements.

The time limit can be set to be 30 seconds to five minutes for each item. The time indicator is shown on the screen of a terminal, and if a test taker does not reply within it, the response will be considered as incorrect.

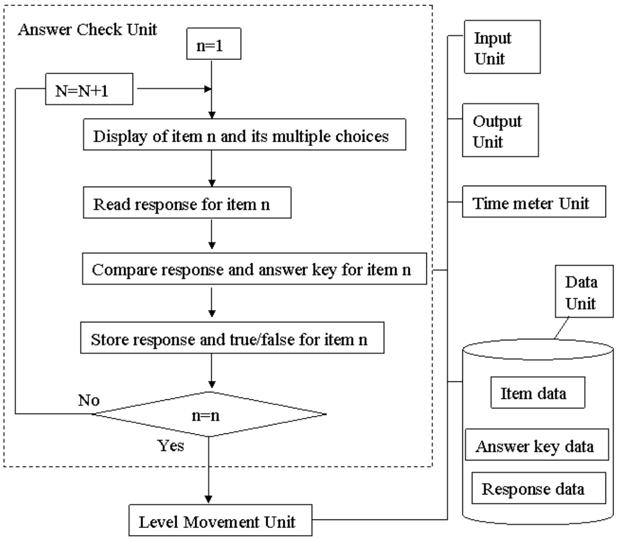

Fig.5. Answer check unit

Responses or answers given by test takers are evaluated in the Answer check unit in Figure 5. A response of each item is compared to an answer key. If both match, the response is counted as correct, otherwise incorrect.

Number of correct responses *n1* is compared to *k* in the Level movement unit as shown in Figure 6.

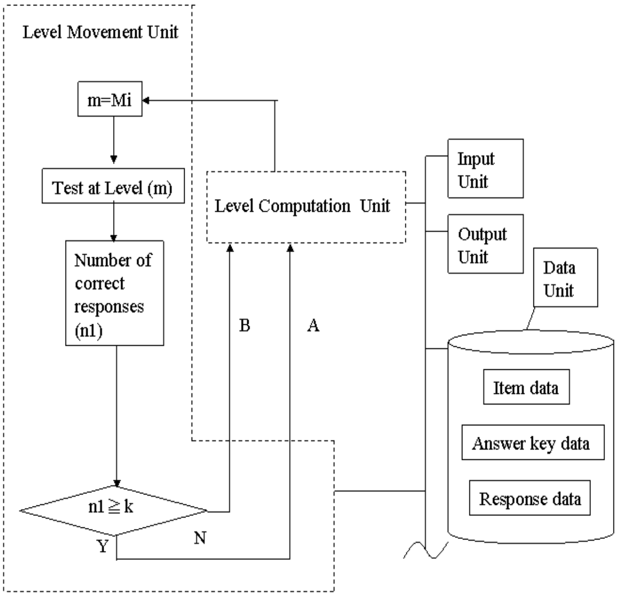

Fig.6. Level movement unit

In the Level movement unit, *n1* is compared with *k* and its result is sent to the Level computation unit shown in Figure 7. In the 2006 and 2007 versions, if two correct/incorrect responses are given consecutively, the level moves up/down after two items, otherwise the third item at the same level is given. Because of the movement of levels according to the response patterns of a test taker, items given in the test are "adapted" to the test taker's ability.

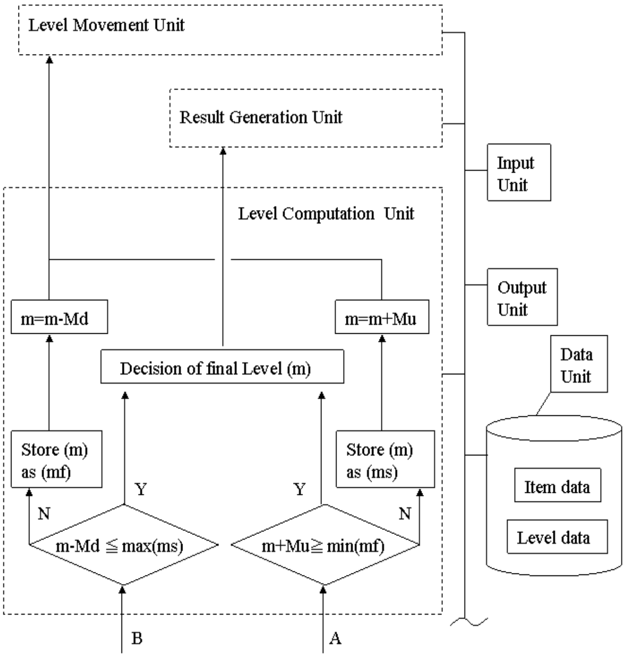

Fig.7. Level computation unit

When *n1* is smaller than *k*, *m-Md* is compared with *max(ms)*. *Max(ms)* denotes the highest level that the test taker passed or succeeded in a previous cycle. In the first cycle, that is at the beginning of the test, the value of *max(ms)* and *min(mf)* is null. If *m-Md≦max(ms)* is met, the level of the test taker is finalized at the level of *max(ms)*, otherwise, the level of the items given to the test taker in the following cycle moves down by *Md* (e.g. one level for 2006 and 2007 versions). The value of *m* is stored as *mf* in the Date unit.

When *n1* is greater than or equal to *k*, *m+Mu* is compared with *min(mf)*. *Min(mf)* denotes the lowest level that the test taker failed in a previous cycle. If *m+Mu≧min(mf)* is met, the level of the test taker is finalized at the level of *min(mf)*. Otherwise, the level of the items given to the test taker in the following cycle moves up by *Mu* (e.g. three levels for 2006 and 2007 versions). The value of *m* is stored as *ms* in the Date unit.

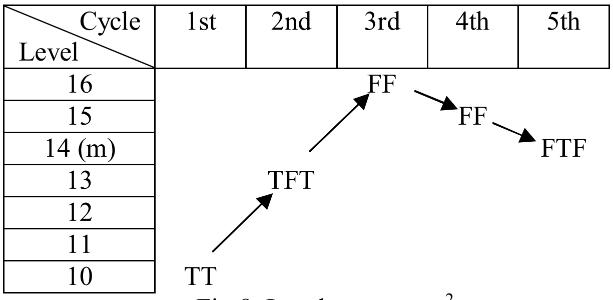

Fig.8. Level movement[2]

Cycles are repeated until either *m-Md≦max(ms)* or *m+Mu≧min(mf)* is met and the value of *m* is finalized. The value of the finalized *m* indicates the level of the ability of the test taker. The result of each section and the average of four sections are shown on the screen at the end of the test.

3. Features of J-CAT

We faced several challenges in the course of development of J-CAT, which bore fruits as the characteristics of the system. How to display special Japanese fonts, namely Kyokasho-tai, with "furigana" characters or ruby texts on a screen is a unique problem for a Japanese language test. Prevention of copying data and prevention of downloading data and/or storing data in a cache of a client's terminal are critical issues for protecting items of the test when the test is delivered on the internet.

J-CAT is intended to be used not only in Japan but also in other countries. Windows operating systems[3] prior to Windows XP do not support multilingual fonts by default. Even though XP and Vista supports Japanese fonts and other multilingual fonts, they are required to be installed by the users. Furthermore, Kyokasho-tai fonts used in J-CAT are not supported in English version of XP, though Kyokasho-tai (i.e.. text book style) is the fonts used in most

[2] T: correct response, F: incorrect response

[3] Windows is a registered trademark of Microsoft Corporation in the United states and other countries.

of basic Japanese language text books. In addition, ruby texts (furigana) are added on top of kanji. Ruby texts (furigana) can be displayed on the screen by using <ruby> tag in HTML, however, adding <ruby> tags to many of kanjis is not easy task for test developers. To make it possible to show all of such special fonts and furiganas, all the texts were converted into image files and displayed on the screen by using Adobe Flash[4]. Images of J-CAT are free from operating system dependency. Utilization of Adobe flash enables the synchronization of image and sound in a listening section.

Prevention of copying, downloading, and caching are achieved by combination of Adobe Flash and a PHP program with an HTML header incorporated in it[5]. The image of this system is shown in Figure 9. The outputter.php which is called for by Flash files opens files in the Data folder and returns data to Flash files in turn. HTTP header in *outputter.php* prevents the data from being kept in the cache. In this system, using Flash itself avoids downloading and copying data.

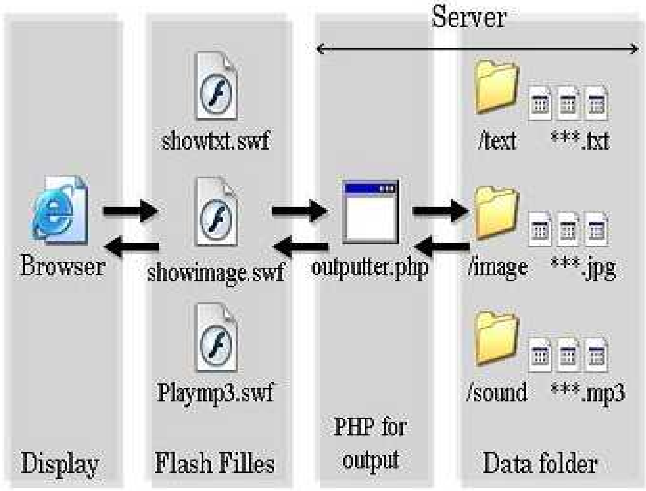

Fig.9 Copy protection

Prevention of data leakage is the top priority in operational testing systems. J-CAT as well as all computer systems, which are under potential threat of abuse as the advance of technology, are required to have a constant update and upgrade in its security system. J-CAT utilizes the above mentioned device in combination with other security systems to protect items in its item pool against illegal access and leakage.

4. Present and Future of J-CAT

J-CAT has been used as a placement test for international students at Yamaguchi University for the last three years. The latest version of J-CAT 2008 has been recently released. It uses a Baysian inference system instead of a testlet system used for the 2006 and 2007 versions and described in this paper. The Baysian inference system calculates a test taker's ability at each response and chooses the most suitable item from the item pool. Both versions will be evaluated in comparison to determine the best system for operation.

The development of system is just one step of three main processes of developing J-CAT as a whole. The first step is to create items. Trained item writers compose the passages and scenarios of items. according to the specification or guideline of the test. These items are printed and recorded for pretests. In the second step, items are tested by volunteers. Response patterns are analyzed by an IRT software [3] to obtain parameters of each item. Three parameter model of IRT requires approximately one thousand persons for each item in a pretest session. Only after the parameters are obtained, as the third step, items are put into an item pool to be used in a test. The item pool must be continuously expanded and improved to achieve the maximum reliability and validity of a test.

5. Remarks

We are grateful to project members, item writes, and participants in pretests, from more than 30 institutions around the globe, who have contributed much for the development of J-CAT. The system and its programming discussed in this paper has been and will be developed by the system group members of J-CAT project, T. Hiramura, K.Kikuchi, Y. Nakamura, T. Kitamoto, and S. Imai.

The latest operational version of J-CAT is accessible at http://www.j-cat.org.

This project is supported by the Grant-in-Aid for Scientific Research (A) (18202012) from. MEXT (The Ministry of Education, Culture, Sports, Science and Technology).

References

[1] R.K. Hambleton and H. Swaminathan, *Item Response Theory: Principles and Applications*, Kluwer-Nijhoff, 1985.

[2] Wikipedia, “Item response theory,” http://en.wikipedia.org/wiki/Item_response_theory, May 2008.

[3] M. Toit (ed.), *IRT from SSI: BILOG-MG, MUTILOG, PASCALE, TESTFACT*, Scientific Software International, 2003.

[4] Adobe and Flash are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States and/or other countires.

[5] This device has been developed by Takekatsu Hiramura, a member of the J-CAT project.