Collaborative e-Testing Construction Using Prediction Tools To Improve Test Reliability

Pokpong Songmuang, Maomi Ueno, and Toshio Okamoto

Graduate School of Information Systems
The University of Electro-Communications
Japan
pokpong@ai.is.uec.ac.jp

Abstract

This paper focuses on one of the features of e-testing, which has not been discussed until now, called collaborative test construction by several test-authors in distant places. It is well known that collaborative work has many advantages. The main idea of this paper is to improve test qualities by applying the unique advantages of collaborative work to e-testing construction and using prediction tools while constructing test. The analysis of collaborative e-testing construction identified the number of test-authors as the most important factor in test validity, while test reliability depends more on the participation of an expert. Based on these findings, a collaborative e-testing construction system was developed that using several prediction tools to improve the reliability of tests constructed by novice test-authors. An experiment in which a novice and an expert test-author each constructed tests by using these prediction tools in one case and not using in the other showed that those constructed using them were more reliable, although those constructed by the expert had even higher reliability.

1. Introduction

Test administration on computers has become more common over the past decade. This is often called Computer Based Testing (CBT) (See, for example, Cynthia, G. et al, 2002). More recently, along with the diffusion of eLearning, CBT has been extended to web-based testing, or e-testing. This e-testing has become an ordinary method of evaluation for eLearning.

Many researches propose e-testing construction system, item authoring tools, item sharing opportunity for test-author, and basic analysis of examinee's test history. For example, Ueno (2005) proposed a design of Web base Computerized Testing System (WCTS) which is designed for assists the authors to shared database item and interactive construction of a test. However, Ueno (2005) did not focus on how to improve the test quality.

In order to identify the important factors in constructing high-quality tests, we compared the effectiveness of test construction by one, three, and five test-authors. The effectiveness was measured in terms of reliability and validity based on test theory (as described by Lord and Novick, 1968, for example,). The results showed that the reliability of a test constructed by an expert or a group of test-authors including an expert was better than that of one constructed by novice test-authors alone. They also showed that test validity increased with the number of test-authors.

The main idea of this paper is to propose e-testing construction system which provides

a collaborative environment and prediction tools in order to improve the test quality. We apply Item Response Theory (IRT) using test history as a predictive test information curve and predictive test characteristic curve instead of knowledge and experience of expert test-authors. Furthermore, the system implements Gamma distribution as a predictive response-time distribution and the mixed model of several binomial distributions as a predictive score distribution.

Finally, we demonstrate the efficiency of the prediction tools by compare the reliability of test constructed by novice and expert test-authors with and without prediction tools. The results showed that the reliability of test construction with tools by novice increased and it became close to the reliability of the expert case.

2. Test Theory

The extensive research related to test construction has been summed up as a test theory (as described, for example, by Lord and Novick, 1968). Traditional test theory describes two concepts related to test construction criteria.

Validity

There are various definitions of validity in the area of test theories (For example, Lord and Novick, 1968). This paper employs one of the most popular definitions. In this definition, the validity means that the ability actually measured by test item represents the ability which should be measured. In other words, the validity indicates that the test item content exactly reflects the test domain. Content validity checking is required intuitive judgment of test-author which machine is unable to provide it.

Reliability

The central concept of the classical test theory using statistics exists in the concept "reliability". The test theory assumes that the square root of the reliability is the correlation between the true and observed scores (For example, Lord and Novick, 1968). Consequently, Cronbach's α can be used as a test reliability measure. Recently, a more sophisticated model, item response theory (IRT), has replaced classical test theory. Here we use the test information function of IRT as the measure of the reliability of a test (Ueno 2005).

According to the test theory, the validity and reliability of a test should be maximized for it to construct an effective test.

3. Collaborative work

It is well known that collaborative work has many advantages. For example, as Miyake (1986) observed, "Because each participant works from different starting schema, what is obvious and natural to one may not be so to the other. This leads to criticism. In this sense, criticisms are the expression of validation checks, and studying them should reveal something of this validation checking mechanism". This validation-checking mechanism is an advantage of collaborative work that no machine can provide. And, as Hutchins and Klausen (1996) pointed out, the distributed cognition provides redundant error checking. In short, these studies demonstrate that applying the concept of collaborative work to test construction should improve test validity.

4. Collaborative e-testing construction analysis

To analyze the effectiveness of collaborative test construction, we compared the validities and reliabilities of tests constructed by different numbers of test-authors (one, three, and five) with and without the participation of an expert in the test domain. The constructed tests measured

Japanese language proficiency and were equivalent to the Level 4 Japanese Proficiency Test given by the Japanese government. (Level 1 is the highest, and level 4 is the lowest.) The tests were constructed based on the same item database. We preferred Japanese knowledge domain in this analysis because it has specialize standard to separate test-authors into expert and novice.

The validity of each test was measured using a test item database we constructed that included some incorrect items. The number of incorrect items included in a constructed test was used as the measure of its validity.

To evaluate the test reliabilities using IRT, we used a three-parameter logistic model:

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

where θ is the person ability parameter, and a_i, b_i, and c_i are item parameters. Using a Bayesian method, we estimated the values of these parameters from the data for the constructed tests. The following function was used to calculate the test information which we used as an index of test reliability.

$$I(\theta) = \sum_{i=1}^{m} a_i^2 p_i(\theta)\left[1 - p_i(\theta)\right] \quad (2)$$

We used the Pearson correlation coefficient and t-test value to calculate the correlation between test construction parameters. The results shown that;

- The test information and the participation of an expert were highly correlated.
- The average number of incorrect items was correlated with the test construction time and the average number of times an item was added.
- The test construction time was correlated with the number of test-authors and the average number of times an item was added.

That is, test validity increased with the number of test-authors and the test construction time, while test reliability depended on the participation of an expert

5 Collaborative e-testing construction system

We previously developed a collaborative e-test construction system (Songmuang and Ueno, 2005). The basic function of the system is to enable test-authors in distant places to share the items in a used item database and to create new items. The test-authors are able to add items to and delete items from the constructed test.

This system provides an opportunity to colleagues of test-authors who live in distance places to participate in the test construction by using web technology. Moreover, the system also provides synchronize discussion board to support communication among test-authors. The sample of the collaborative e-testing construction system is illustrated in Fig 1.

5.1 Prediction tools

The prediction tools interactively monitor the status of the constructed test by using history data from response database, as shown in bottom of Fig1. This system provides four types of prediction tools; 1) predictive test characteristic curve,

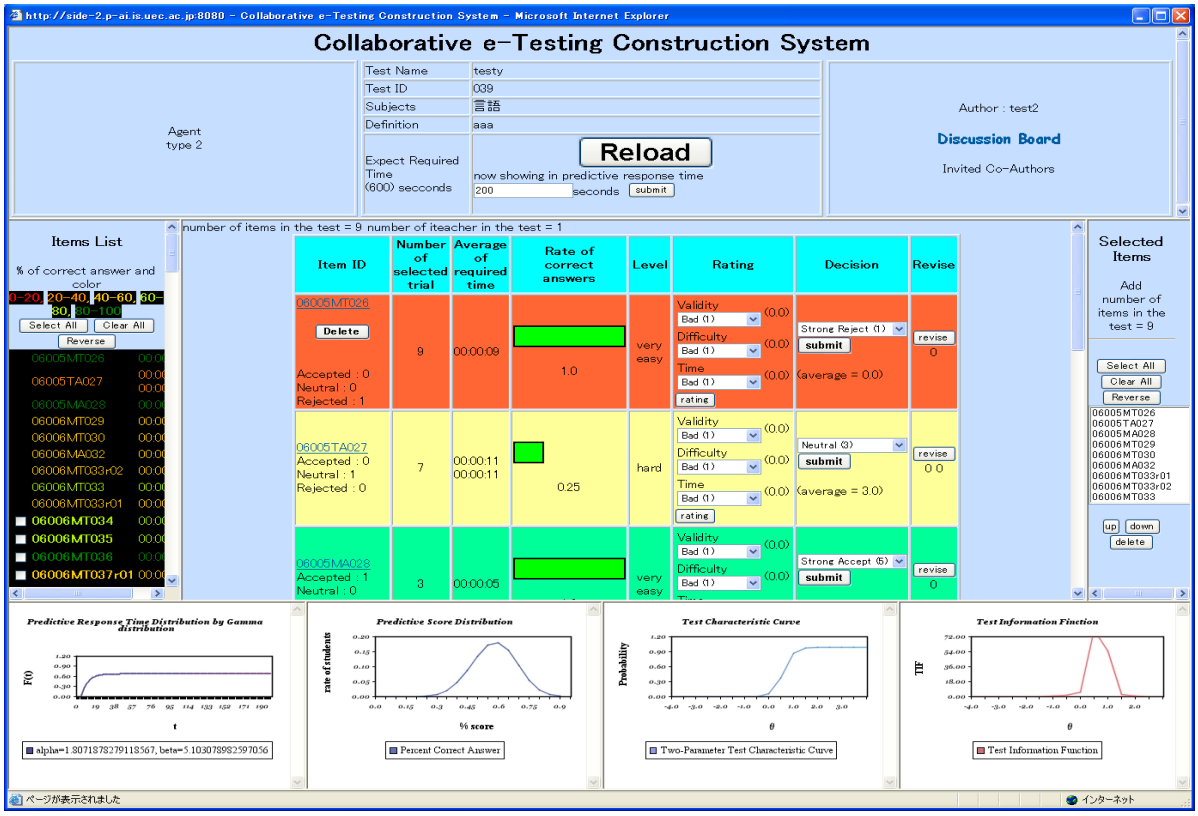

Figure 1 Collaborative e-Testing Construction System Interface

2) predictive test information curve, 3) predictive score distribution, and 4) predictive response-time distribution. Details of the prediction tools are describe as follows:

5.1.1 Predictive test characteristic curve

We employ the Two-parameter Logistic Ogive Model as Item Characteristic Curve.

The curve function is define by

$$P_i(\theta) = \frac{1}{1+e^{-z_i}}. \qquad (3)$$

We apply average of all Item characteristic curves of constructed test as the predictive test characteristic curve.

5.1.2 Predictive test information curve

The expression for the Item Information Function (IIF) of two-parameter logistic are:

$$I_i(\theta) = \alpha_i^2 P_i(\theta)(1-P_i(\theta)). \qquad (4)$$

Test information function is the sum of all IIFs in the constructed test. This function is employed as the predictive test information function.

5.1.3 Predictive response-time distribution

Ueno and Nagaoka (2005) analyzed eLearning time based on a gamma distribution with parameters α and β representing the complexity of the learned content and the expected time of a simple cognitive process. To visualize the current status of the constructing test required time, the proposed paper provides a predictive response-time distribution. We use the gamma distribution described by Ueno and Nagaoka (2005) as the predictive response-time distribution along with item historical data.

5.1.4 Predictive score distribution

The system presents the predictive score distribution of the test being constructed to enable the authors to visualize its current status. Ueno (2005) applied the mixture model of several binomial distributions as a predictive score distribution, this paper also employs it as a predictive score distribution.

6. Experiment

The experiment of this paper was conducted in order to evaluate the effectiveness of the prediction tools by compare the reliability of tests constructed by group of three expert and novice test-authors with and without prediction tools. The constructed tests measured Japanese language proficiency and were equivalent to the Level 4 Japanese Proficiency Test.

Table 1. Test construction conditions

Test-authors	Expert* (groups)	Novice (groups)
Without tools	2	3
With tools	2	3

*The Japanese language proficiency of the experts was equal to or exceeded Level 2 on the Japanese Language Proficiency Test.

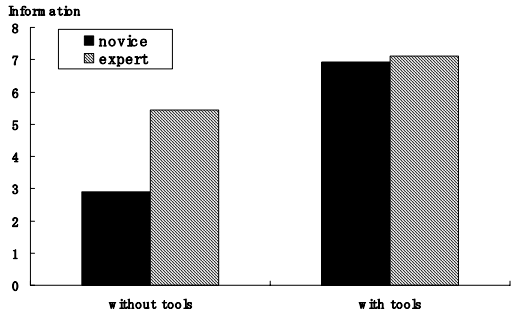

Figure 2. Average test information of constructed tests

As shown in Fig2, the information of test constructed by novice with the prediction tools are improved and become close to one of expert case. This means that the reliability of tests constructed by novice is improve by using the prediction tools.

Moreover, the required time of test construction with prediction tools decrease to half of the case without tools.

7. Conclusions

We have developed a collaborative e-test construction system that provides a predictive test characteristic curve, predictive test information function, predictive response-time distribution and predictive score distribution that can be used to improve test reliability.

To evaluate the effectiveness of this approach, we compared the reliability of Japanese language proficiency tests constructed by a novice test-author and by an expert test-author with and without the prediction tools. The tests constructed using them had higher reliability and we can decrease difference between the information of test constructed by novice and one by expert. Furthermore, the test construction with the prediction tools required time less than the case without tools, and it shows that these tools make the efficiency collaborative test construction improve.

8. References

Cynthia, G. et al (2002), *Practical Considerations in Computer-Based Testing*. Springer-Verlag New York, Inc.

Hutchins, E. and Klausen, T. (1996) "Distributed Cognition in an Airline Cockpit", In Middleton, D. and Engeström, Y. (eds.), *Communication and Cognition at Work*. Cambridge University Press, Cambridge. pp. 15–54

Lord, F. M. and Novick, M. R. (1968) *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley

Miyake, N. (1986) "Constructive Interaction and the Iterative Process of Understanding", *Cognitive Science*, Vol. 10(2), pp. 151–177

Songmuang, P. and Ueno, M. (2005) **"**e-Testing Management System", *Proc. of e-learn2005*, pp. 3139–3148

Ueno, M. (2005) "Web based computerized testing system for distance education", *Educational Technology Research*, Vol. 28, pp. 59–69

Ueno, M. and Nagaoka, K. (2005) "On-Line Analysis of eLearning Time based on Gamma Distributions", In *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2005,* pp. 3629–3637. Norfolk, VA: AACE