Collaborative E-Test Construction

Using Predicted Response-Time and Score Distributions to Improve Reliability

Pokpong Songmuang and Maomi Ueno
The University of Electro-Communications, Japan

Abstract: Analysis of collaborative e-test construction identified the number of test-authors as the most important factor in test validity, while test reliability depends more on participation of an expert. Based on these findings, a collaborative e-test construction system was developed that uses predicted response-time and score distributions to improve the reliability of tests constructed by novice test-authors. A gamma distribution is used as the predicted response-time distribution, and a mixed model of binomial distributions is used as the predicted score distribution. An experiment in which a novice and an expert test-author each constructed tests by using and not using these predicted distributions showed that those constructed using them were more reliable, although those constructed by the expert had even higher reliability.

Keywords: Collaborative e-test construction, reliability, predicted response-time distribution, predicted score distribution.

1. INTRODUCTION

Test administration on computers has become more common over the past decade. This “computer-based testing (CBT)” is done using either tests developed specifically for the computer or tests converted into a computer-based format. More recently, along with the diffusion of e-learning, CBT has been extended to web-based testing, or “e-testing.” This e-testing has become a common method of evaluation for e-learning, and much attention has been paid to the use of e-testing to deliver an on-line test function to various places. Moreover, it enables collaborative test construction by several test-authors in different places. There are many advantages to such collaboration.

- It provides validation-checking mechanisms as part of the criticism process, something a machine does not provide (Miyake 1986).

- The distributed cognition provides an opportunity to distribute work activities, thereby improving the complex information analysis (as described, for example, by Hutchins and Klausen 1996).
- It enables effective and efficient solving of ill-structured problems (Simon 1973), which need complex expert knowledge to solve.

By applying collaboration to e-test construction, we can obtain several benefits, including stimulation of test-authors' reflections and thus improved test validation, increased test reliability due to distributed cognition; and more sophisticated test construction due to the sharing of expert knowledge, particularly tacit knowledge and ill-structured knowledge.

In a previous paper, Ueno (2005) proposed a web-based computerized testing system for assisting test-authors in sharing the used item database and in collaborative test construction. However, this paper did not focus on a collaborative e-test construction system and did not provide any analysis of collaborative test construction.

Our interest here is improving the effectiveness of collaborative test construction. 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 describe a collaborative e-test construction system that provides a predicted response-time distribution and a predicted score distribution that can be used to improve the reliability of tests constructed by novice test-authors. We use a gamma distribution (Ueno and Nagaoka 2005) as the predicted response-time distribution and a mixed model of several binomial distributions as the predicted score distribution. Both distributions help a test-author better understand the status of a constructed test. An experiment was performed to compare the reliability of tests constructed by an expert and by a novice test-author with and without the distributions. The reliability of those constructed using them was better although those constructed by the expert had even higher reliability.

2. TEST THEORY

Extensive research related to test construction can been summed up as a test theory (as described by Lord and Novick 1968). Traditional test theory describes two concepts related to test construction criteria.

2.1 Validity

Validity can be defined in a number of ways. 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 the 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.

2.2 Reliability

The central concept of classical test theory using statistics exists in the concept of "reliability." Test theory assumes that the square root of the reliability is the correlation between the true and observed scores (Lord and Novick 1968). Consequently, Cronbach's α can be used as a measure of test reliability. 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 test reliability (Lord and Novick 1968).

According to this theory, the validity and reliability of a test should both be maximized for it to be a good test.

3. COLLABORATIVE E-TEST 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, and data on the construction process was collected and stored.

The validity of each test was measured using a test item database we constructed including some incorrect items. The number of incorrect items included in the test was used as the measure of its validity. To evaluate test reliabilities using IRT, we used a three-parameter logistic model:

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

where θ is the person ability parameter, and a_i, b_i, and c_i are item parameters. The b_i represents the item location, which, in the case of attainment testing, is referred to as item difficulty. The a_i represents the discrimination of the item, that is, the degree to which the item discriminates between persons in different regions on the latent continuum. This parameter characterizes the slopes of the item response curves. For items such as multiple-choice, parameter c_i is used in an attempt to account for the effects of guessing on the probability of a correct response. 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.

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

We used the Pearson correlation coefficient and t-test value to calculate the correlation between test construction parameters as shown in the table 1.

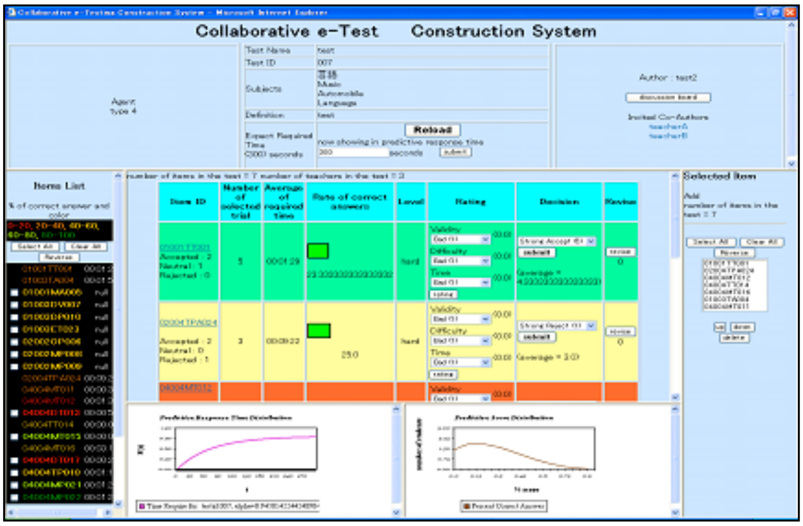
Table 1: Pearson correlation coefficients and t-test values between test construction parameters

	a	b	c	d	e	f	g	h
a								
b								
c	0.94 (0.010)	0.10 (0.010)						
d	-0.76 (0.076)	0.13 (0.578)	-0.77 (0.011)					
e	-0.33 (0.042)	0.86 (0.000)	-0.13 (0.013)	0.23 (0.001)				
f	0.97 (0.008)	-0.21 (0.037)	0.70 (0.071)	-0.63 (0.009)	-0.49 (0.015)			
g	0.88 (0.040)	-0.22 (0.037)	0.90 (0.007)	-0.86 (0.042)	-0.41 (0.099)	0.89 (0.006)		
h	0.95 (0.272)	-0.47 (0.206)	0.24 (0.010)	-0.49 (0.323)	-0.18 (0.014)	0.43 (0.008)	0.51 (0.038)	

a. number of test-authors
b. participation of an expert
c. average test construction time
d. average number of incorrect items
e. average test information
f. average number of times an item was added
g. average number of times an item was deleted
h. average number of times an item was created

- 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.

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


Collaborative e-Test Construction System

Test Name: test
Test ID: 007
Subjects: Music, Automobile, Language
Definition: test

Author : test2

Figure 1: Collaborative e-test construction system

4. COLLABORATIVE E-TEST CONSTRUCTION SYSTEM

As shown by the results above, test reliability depends on the participant of an expert, so we investigated ways to improve the reliability of tests constructed by novice test-authors. We developed a collaborative e-test construction system, as illustrated in Figure 1. Its basic function is to enable test-authors in distant places to share 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. The system also provides a discussion board to enable the test-authors to share opinions and ideas during their collaboration.

The main focus of this description is on the use of the predicted score and response-time distributions, which are used to support the authors' decision making, as shown at the bottom of Figure 1.

4.1 Predicted score distribution

The system presents the predicted score distribution of the test being constructed to enable the authors to visualize its current status. The set of the probabilities of correct answers for m items $\Theta = \{\theta_i\}, \{i = 1,...,m\}$ is estimated using historical data. A Bayesian estimation based on a binomial distribution is estimated using:

$$\theta_i = \frac{n_i + a'}{n + a'}, (i = 1,...,m), \qquad (1)$$

where m is the number of items on the test, n_i is the number of examinees who provided correct answer, n is the number of examinees, and a' is the value of a hyper parameter. We set $a' = 1$, reflecting our assumption that the prior distribution is uniform. Let $x, (0,...,m)$ be a score random variable for a test with m items. The mixed model of several binomial distributions is defined by

$$p(x|\Theta) = \sum_{i=1}^{m} \left[p(m_i) p(x|m_i, \theta_i) \right], \qquad (2)$$

where $m_i(1,...,m)$ means the *i*-th model.

4.2 Predicted response-time distribution

Ueno and Nagaoka (2005) analyzed e-learning 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 study provides a predicted response-time distribution. We use the gamma distribution described by Ueno and Nagaoka as the predicted response-time distribution along with item historical data. We assume that any testing process consists of α repetitions of simple cognitive processes.

Moreover, the response time for a simple problem-solving process is assumed to follow a distribution so as to maximize, and, given minimum response time t_0 and average response time *E,* what is given by:

$$H\left[f_s(t)\right] = \int_{t_0}^{\infty} f_s(t) \log f_s(t) dt . \qquad (3)$$

The required time for a simple problem-solving process is given by an exponential distribution:

$$f_s(t) = \frac{1}{\tau} e^{-(1/\tau)t} . \qquad (4)$$

While the testing process is generally viewed as consisting of α layers of the process given by (4) and is thus obtained by a convolution integral of (4), we introduce β, the time required for solving a simple problem, and calculate α convolution integrals under the restriction that $\alpha\beta = E$. (5)

Thus, the gamma distribution obtained as the distribution model for the required learning time is

$$f(t) = \frac{t^{\alpha-1} \exp\left(-\frac{t}{\beta}\right)}{\beta^{\alpha} (\alpha-1)!} . \qquad (6)$$

The predicted response-time distribution is then given by

$$F(t) = \begin{cases} 0 & t < t_0 \\ \int_{t_0}^{t} f(t) dt & t \geq t_0 \end{cases} . \qquad (7)$$

5. EXPERIMENT

We compared the reliability of tests constructed by a novice test-author and by an expert test-author with and without the predicted distributions. The constructed tests measured Japanese language proficiency and were equivalent to the Level 4 Japanese Proficiency Test given by the Japanese government. The tests were constructed based on the same item database, and data on the construction process was collected and stored. Each constructed test had about 30 items.

Table 2. Average information of constructed tests

Test-author	Novice w/o distributions	Expert* w/o distributions	Novice with distributions	Expert* with distributions
Test information	3.279	3.952	4.805	8.735

**Had Japanese language proficiency equal to or better than Level 2 on Japanese Language Proficiency Test.*

As shown in Table 2, the average information of the tests constructed using the predicted distributions were higher. This means that the test reliability was improved by using the predicted distributions.

However, the average information of the tests constructed by the expert with and without the distributions was higher than that of the novice. This means that expert knowledge is still an important factor in test construction.

6. CONCLUSION

We have developed a collaborative e-test construction system that provides a predicted response-time distribution and a predicted score distribution that can be used to improve test reliability. A gamma distribution is used as the predicted response-time distribution, and a mixed model of binomial distributions is used as the predicted response-time distribution.

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 predicted distributions. The tests constructed using them had higher reliability although those constructed by the expert had even higher reliability. We plan to develop an agent system that plays the role of a domain expert in order to increase the reliability of tests constructed by novice test-authors.

7. REFERENCES

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, no. 2, pp. 151–177.

Simon, H. A. (1973), The Structure of Ill-Structured Problems, *Artificial Intelligence*, vol. 4, pp. 181–201.

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 e-Learning Time based on Gamma Distributions, *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2005,* pp. 3629–3637. Norfolk, VA: AACE.