We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,300

Open access books available

172,000

International authors and editors

190M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

E-Testing Construction Support System with some Prediction Tools

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

1. Introduction

In actual test construction, predictions of test score and response-time distributions have become important (Ueno, 2006, Ueno, 2007). The following situations exemplify this importance:

1. In the case of absolute grading in qualifying examinations, the qualifying score is selected in advance. The test score and response-time distributions must be approximately constant so that among tests, the degree of difficulty does not differ.
2. In selection tests such as entrance examinations at national centres, examinees can select subject choices such as mathematics A or B and science A, B or C. The distributions in both subjects must be approximately similar in both score and response-time distributions so that the degree of difficulty of the test is not different between the subject choices.
3. In the selection tests, examinees can choose a large problem during tests such as entrance examinations. In this case, the score and response-time distributions must be approximately similar across each large problem so that the degree of difficulty is relatively equal among them.

As mentioned above, test-authors should construct tests while predicting the score and response-time distributions to control the equality of difficulty. Nevertheless, there are many cases in which test-authors construct tests referring to past test constructions. However, it was noticed that the national center for entrance examinations persisted in the aim of showing an average score, but test reliability was seriously ruined in the actual tests (Ueno, 2006, Ueno, 2007).

Recently, with the increasing use of e-testing, test construction utilizing an item bank has become the standard method, allowing predictions of score and response-time distributions during actual test construction. For example, Nagaoka (2000) has developed a test score prediction system using a truncated exponential distribution. Ueno (2005b) has proposed a web-based computerized testing system with a test construction support system employing the mixture binominal distribution as a prediction of score distribution. Meanwhile, the only research which focuses on score distribution is that by Keats and Lord (1962). In this study, they applied the beta-binominal distribution as a score distribution.

On the other hand, a lot of research has been done on test response-time distributions. In the educational engineering field, the response curve has been investigated for analyzing response-time data obtained from response analyzers and computer-based testing systems (for example, Nagaoka, 2003, Nagaoka and Wu, 1989, Ueno, 2005a). Moreover, in the field of psychological measurement, Thissen (1983) applied a lognormal distribution to test response-time data, Verhelst et al. (1997) used a gamma distribution for a speed test, and Roskam (1997) applied the Weibull distribution for test response-time.

As mentioned above, the related studies have rarely showed comparisons among several models from the viewpoint of prediction accuracy in score and response-time distributions. Therefore, we first did some comparison experiments of the prediction models of the test score and response-time distributions using actual data. This paper proposes 1) a mixture beta-binomial distribution, which combines the beta-binomial and mixture binomial distributions; 2) a score distribution based on item response theory (IRTs: the Rasch model (Rasch, 1966) and a two-parameter logistic model (Lord and Novick, 1968)), since few test score prediction models have been proposed. In particular, we compared the proposed models with the binomial distribution (Ueno, 2005b), the mixture binomial distribution (Ueno, 2005b), the beta-binomial distribution, and the truncated exponential distribution (Nagaoka, 2000). In addition, comparison experiments among response-time prediction models were performed. We compared the prediction precision of normal distribution, lognormal distribution, the extended gamma distribution (Ueno, 2005a) and the Weibull distribution for test response-time.

The results show that the score distribution based on IRT (two-parameter logistic model) and the extended gamma distribution produced the best prediction precision for score and response-time distributions, respectively. Furthermore, the prediction tool using the score distribution based on IRT (two-parameter logistic model) and the extended gamma distribution was installed in the e-testing construction support system. Some evaluation experiments done using this system are described here. The results show the effectiveness of the proposed system. In addition, evaluation obtained from questionnaires also confirms the effectiveness of this system.

2. E-Testing System

An e-testing system (Songmuang and Ueno, 2005) was developed to support test-authors in creating test items, constructing tests, delivering tests, and to analyzing test history data. The system consists of the following features: 1) an item authoring system, 2) e-testing construction support system, 3) data analysis system, 4) item bank, 5) test database, and 6) test delivery system, as shown in Figure 1. This paper focuses on the e-testing construction support system and the development of a tool to predict test score and response-time distributions.

Fig. 1. Framework of e-testing system

	Test Content	Number of items	Number of examinees	Rate of correct answers	Standard deviation	Average response-time (seconds)	Standard deviation
A	Japanese Language Test	69	91	0.445	0.198	1421.05	1366.60
B	Introduction to Computers	54	62	0.270	0.217	659.10	569.54
C	Information Theory	20	72	0.739	0.903	204.11	284.33
D	Network Technology	37	66	0.241	0.142	622.91	614.07

Table 1. Summary information of sample data

3. Comparison Experiment of the Prediction Models

3.1 Method

In this section, to develop a prediction tool for the e-testing construction support system, we evaluated several score and response-time prediction models using actual data. Table 1. lists the sample data for experiments obtained from ‘N’ University. Test A concerned the Japanese Language Proficiency Test. Tests B, C and D focused on information science. The sample data was taken by students who participated in e-learning classes.

The experiment was divided into two parts. Initially, we evaluated the fit of the estimated distributions using the models to the sample data by comparing the difference between the estimated distributions and the sample data. Subsequently, we evaluated the prediction precision of the predicted distributions using the models. In this part, the sample data were randomly divided into halves, which were used to predict the distributions (the training data) and verify the predicted distributions (the validation data). This procedure was repeated 1000 times for each set of test content data.

However, in both parts of the experiment, the evaluation of the score distribution transformed the estimated/predicted score distributions and actual data into discrete

distributions in ten stages. The evaluation of the response-time distribution was performed by transforming both the estimated/predicted response-time distributions and actual data into cumulative distributions. The difference between the estimated/predicted distributions and the actual data, which indicates the prediction precision of the models, was calculated using the root mean square error (RMSE).

Fit	Binomial Distribution	Mixture Binomial Distribution	Beta-binomial Distribution	Mixture Beta-binomial Distribution	Rasch Model	2-Parameter Logistic Model	Truncated Exponential Distribution
A	0.00356	0.00042	0.00052	0.00041	0.00059	<u>0.00041</u>	0.00052
B	0.00570	0.00235	0.00319	0.00120	0.00160	<u>0.00118</u>	0.00159
C	0.00766	0.00332	0.00321	0.00262	0.00267	<u>0.00192</u>	0.00277
D	0.00464	0.00098	0.00178	<u>0.00060</u>	0.00093	0.00075	0.00079
Predic tion	Binomial Distribution	Mixture Binomial Distribution	Beta-binomial Distribution	Mixture Beta-binomial Distribution	Rasch Model	2-Parameter Logistic Model	Truncated Exponential Distribution
A	0.00346	0.00055	0.00062	0.00056	0.00067	<u>0.00053</u>	0.00062
B	0.00591	0.00239	0.00324	0.00134	0.00168	<u>0.00132</u>	0.00173
C	0.00737	0.00339	0.00331	0.00269	0.00279	<u>0.00179</u>	0.00286
D	0.00422	0.00105	0.00182	0.00073	0.00095	<u>0.00068</u>	0.00089

Table 2. RMSEs between estimated score distributions and actual data, and RMSEs between predicted score distributions and actual data

3.2 Experiment Results of Score Distribution

In this section, we introduce new prediction models for test score distribution. The mixture beta-binomial distribution, which combined the mixture binomial distribution (Ueno, 2005b) and beta-binomial distribution (Keats and Lord, 1962), is defined as follows:

$$p(x|\alpha, \beta) = \sum_{i=1}^m \left[\frac{1}{m} \binom{m}{x} \frac{B(\alpha_i + x, \beta_i + n - x)}{B(\alpha_i, \beta_i)} \right]$$

(1)

where $i,(1,...,m)$ indicates the i -th item, n indicates the total number of examinees taking the test, $B(\alpha, \beta)$ is the beta function, and α and β are the estimated parameters of each item. Various IRT models exist, but our approach uses the Rasch model (Rasch, 1966) and the two-parameter logistic model (Lord and Novick, 1968). For the IRT, u_{ij} indicates the response of examinee $j,(1,...,n)$ on item $i,(1,...,m)$ as follows:

$$u_{ij} = \begin{cases} 1 : \text{when examinee } j \text{ answers item } i \text{ correctly} \\ 0 : \text{other cases} \end{cases}$$

The probability of a correct answer in the two-parameter logistic model is:

$$p(u_{ij} = 1 | \theta_j) = \frac{1}{1 + \exp\{-1.7a_i(\theta_j - b_i)\}} \quad (2)$$

where θ_j is person j 's ability parameter, a_i is i -th item's discrimination parameter, and b_i is the i -th item's difficulty parameter, which represents the degree to which the item discriminates between persons in different regions on a latent ability scale. The Rasch model is a special case of the two-parameter logistic model by setting $a_i = 1$. Here, the ability distribution is assumed to be the standard normal distribution. In this paper, we propose a test score distribution based on the IRTs as follows:

$$p(x) = \int_{-\infty}^{\infty} \sum_{i=1}^m \frac{1}{m} \binom{m}{x} p(u_{ij} | \theta_j) d\theta_j \quad (3)$$

Next, the comparison experiments and results are described. The score distribution models used in this experiment were the binomial distribution (Ueno, 2005b), the mixture binomial distribution (Ueno, 2005b), beta-binomial distribution, the mixture beta-binomial distribution, the truncated exponential distribution (Nagaoka, 2000), and the score distribution based on IRTs (Rasch model and two-parameter logistic model). The parameters of the models were estimated using the sample data.

The fits of the estimated score distribution to the actual data were estimated using RMSE. The results are shown in the top of Table 2. The smallest RMSEs across all models of each test are underlined. According to the results, the score distribution based on the IRT (two-parameter logistic model) gave the best results to fit the score distribution of tests A, B and C. The best model to fit the score distribution of test D data was the mixture beta-binomial distribution, which is also proposed in this paper.

The prediction precisions of the predicted score distribution models were evaluated using the above-mentioned method. Based on the results in the bottom of Table 2, the score distribution based on IRT (two-parameter logistic model) gave the best results in predicting score distributions of data from tests A, B, C and D.

From the above results, the score distribution based on IRT (two-parameter logistic model), which is proposed in this paper, had the best performance results in both evaluations.

3.3 Experiment Results of Response-Time Distribution

The response-time distribution models in this experiment were normal distribution, lognormal distribution, the extended gamma distribution (Ueno, 2005a), and the Weibull distribution. The parameters of these models were estimated using the sample data.

The fits of the estimated response-time distributions to the actual data were estimated using RMSE. The results are shown in the top of Table 3. The smallest RMSEs across all models of each test are underlined. According to these results, the extended gamma distribution gave the best results to fit the response-time distribution of data from tests A and D. The best model to fit the response-time distribution of test B data was the normal distribution, while that to fit the response-time distribution of test C data was the lognormal distribution. Thus, we could not determine which model offered the best performance results in this experiment because the results were sparse.

Fit	Normal	Lognormal	Gamma	Weibull
A	0.758	1.195	<u>0.157</u>	0.92
B	<u>1.043</u>	3.167	2.812	4.266
C	5.232	<u>1.226</u>	3.199	1.271
D	0.979	1.294	<u>0.534</u>	1.330
Prediction	Normal	Lognormal	Gamma	Weibull
A	4.481	1.961	<u>1.032</u>	1.802
B	4.261	4.455	<u>1.578</u>	4.271
C	4.826	2.954	<u>1.493</u>	1.630
D	1.007	2.080	<u>1.002</u>	1.220

Table 3. RMSEs between estimated response-time distributions and actual data, and RMSEs between predicted response-time distributions and actual data

Next, the prediction precisions of the response-time distribution models were evaluated using the above-mentioned method. As can be seen in the bottom of Table 3, the extended gamma distribution produced the best results in predicting the response-time distributions of data from tests A, B, C and D. From the above results, the extended gamma distribution gave the best results in both evaluations.

4. E-Testing Construction Support System with Score and Time Prediction Tool

From the results of the previous section, we employed the score distribution based on IRT (two-parameter logistic model) and the extended gamma distribution as the prediction models for the score and response-time distributions in the prediction tool, respectively. The tool is installed into the e-testing construction support system. Figure 2 shows the interface of the developed system. The top frame of the interface shows the test attributes (test name, test ID, etc.). The left frame of the interface displays a list of item IDs and classifies them by difficulty using a colour scheme which changes from green, indicating lower difficulty, to red, indicating a higher degree. When the mouse is clicked on an item ID, the content of the item is presented in a pop-up window, as shown in the Figure 2. Test-authors select items to the test by clicking the check boxes at the left of the item IDs. Then, the selected items are registered in the test memory and shown in the central frame of the interface. The average response-time and the degree of difficulty are shown. Test-authors can change the order of items with the $\uparrow \downarrow$ buttons or delete them by clicking the “Delete” button. The prediction tool for score and response-time distributions utilizes statistical history data of registered items in the test memory. The predicted distributions (the score distribution based on IRT (two-parameter logistic model) and extended gamma distribution) are presented in the bottom frame of the interface. In addition, the predicted average score, the predicted standard deviation in score, the predicted average response-time, and the predicted standard deviation in response-time are also presented under each graph. Thus, test-authors can decide whether to discontinue test construction or to replace or to add an item using this prediction tool.

Fig. 2. e-Testing construction support system

5. System Evaluation

This section evaluates the effectiveness of the proposed e-testing construction support system. For this experiment, we previously determined the average score, score standard deviation, average response-time, and the response-time standard deviation, which were the goal attributes of the test, and compared differences between the test goal and actual test attributes constructed using the following systems: the proposed system (system ①), system ① without the prediction tool (system ②), and system ② without showing the difficulty and average response-time of items in the central frame (system ③). Thirty Japanese postgraduates of university ‘N’ were assigned to use systems ①, ② and ③ at random in groups of ten, and tests were constructed with each system. Seventy items from Level 1 of the Japanese Proficiency Test were used as response data and obtained from 197 examinees. Concretely, a random sampling of historical data of 97 examinees from the 197 examinees was used for the prediction tool. The RMSEs were calculated using the remaining 100 examinees’ data as the validation data of the test constructed using systems ①, ② and ③. The results show that the average RMSE of the test constructed by system ① was lowest of all of the goal attributes. Moreover, Table 4 shows the results of the multiple comparisons (WSD method) between the tests’ RMSE. *d* is the difference of two compared average pairs, WSD shows the test statistic, * indicates there is a significant difference in the significant level within 5%. According to the results in Table 4, the proposed system and the comparison systems show a 5% significant difference from the results in average score,

standard deviation score, average response-time, and standard deviation response-time, indicating the prediction tool of this system was highly effective.

Average RMSE of average score	<i>d</i>	WSD	
Ⓐ-Ⓑ	2.267	2.088	*
Ⓑ-Ⓒ	1.368	2.564	
Ⓒ-Ⓐ	3.635	2.811	*
Average RMSE of standard deviation score	<i>d</i>	WSD	
Ⓐ-Ⓑ	2.946	2.651	*
Ⓑ-Ⓒ	4.930	3.256	*
Ⓒ-Ⓐ	7.876	3.570	*
Average RMSE of average response-time	<i>d</i>	WSD	
Ⓐ-Ⓑ	9.387	14.962	*
Ⓑ-Ⓒ	11.054	16.360	
Ⓒ-Ⓐ	20.441	18.159	*
Average RMSE of standard deviation response-time	<i>d</i>	WSD	
Ⓐ-Ⓑ	31.693	21.748	*
Ⓑ-Ⓒ	6.363	19.890	
Ⓒ-Ⓐ	38.056	24.167	*

* means there is a significant difference in the significance level within 5%.

Table 4. Multiple comparisons (WSD method) results

6. Questionnaire Analysis

Furthermore, we compared the questionnaire results between test-authors who used the prediction tool and those who did not. The contents of the questionnaires and results are listed in Table 5. The questionnaires consisted of multiple-choice questions with five answers: 1 definitely no, 2 no, 3 not sure, 4 yes, and 5 definitely yes. The results in Table 5 indicate that the average score of questions related to the standard-deviation of the score (with and without the prediction tool) are higher than that of the average of the score. This means that the prediction of the standard deviations in score of the constructed tests was more difficult than predicting the average of the score wherever the system was used or not. This is also similar in the case of the response-time. However, the average scores of all questions answered by test-authors supported by the prediction tool are higher than those answered by test-authors who were not. Moreover, there was a significant difference 1% in all questions between cases where the system was used or not used. This difference shows that it was easier to predict each goal attribute utilizing the proposed system.

Question	Without the prediction tool: average score (standard deviation)	With the prediction tool: average score (standard deviation)	t	p-value
Was it difficult to predict the average response-time of the constructed test?	2.844 (1.322)	1.968 (1.092)	2.886	0.005
Was it difficult to predict the standard deviation of response-time of the constructed test?	3.969 (0.782)	2.313 (1.030)	7.245	0.001
Was it difficult to predict the average score of the constructed test?	3.250 (1.244)	1.813 (0.931)	5.232	0.001
Was it difficult to predict the standard deviation of score of the constructed test?	4.125 (0.793)	2.281 (0.991)	8.215	0.001

Table 5. Results of questionnaire

7. Conclusion

We presented an e-testing construction support system with the prediction tool for predicting the test score and response-time distributions. Moreover, 1) the mixture beta-binomial distribution, which combines the beta-binomial distribution and the mixture binomial distribution, and 2) the test score distribution based on item response theory (Rasch model and the two-parameter logistic model) were proposed in this paper. The proposed models were compared with the traditional models (the binomial, the binomial mixture, the beta-binomial, and the truncated exponential distributions). Furthermore, we compared the traditional response-time distributions (the normal, the lognormal, the extended gamma, and the Weibull distributions).

The results indicated that the score distribution based on item response theory (two-parameter logistic model) for test score distribution and the extended gamma distribution for response-time distribution showed the best prediction performance. Using these results, we developed the e-testing construction support system to visualize the test score and response-time distributions for the constructed test. We showed the effectiveness of the system from evaluation experiments.

We compared the difference of the errors between goal attributes and actual values of tests constructed with and without the system using the actual item bank. In addition, the average score, the standard deviation of score, the average response-time, and the standard deviation of response-time as the test goal attributes were decided prior to the evaluation. The results show that errors in goal attributes can be significantly decreased utilizing the proposed system.

Moreover, evaluations concerning the test construction process of test-authors who used and did not use this system were compared from questionnaires. The results reveal the effectiveness of this system. A future task is to increase type and number of data consider when to evaluating the prediction models.

8. References

- Keats, J. A. & Lord, F. M. (1962). A theoretical distribution for mental test scores. *Psychometrika*, Vol. 27, 59-72, ISSN 0033-3123.
- Lord, F.M. & Novick, M.R. (1968). *Statistical Theories of Mental Test Scores*. Addison-Wesley, ISBN 0394347714, Massachusetts.
- Nagaoka, K. (2000). On Development of Computerized Testing System for Practical Use with Estimation Function. *Japan Journal of Educational Technology*, Vol.24, No.1, 63-72, ISSN 0385-5236. (In Japanese)
- Nagaoka, K. (2003). Response Curve of Group Learning. In: *Educational technology dictionary*, 265-266, Jikkyo Shuppan. ISBN 4407051108. (In Japanese)
- Nagaoka, K. & Wu, Y. (1989). Analysis of response time data from computer testing. *Japanese Journal of Educational Technology*, Vol.12, No.4, 129-137, ISSN 0385-5236.
- Rasch, G. (1966). An item analysis which takes individual differences into account. *British journal of mathematical and statistical psychology*, Vol.19, 49-57, ISSN 0007-1102.
- Roskam, E.E. (1997). Models for speed and time-limit test. In: *Handbook of modern item response theory*, W.J. van der Linden and R.K. Hambleton (Eds.), 187-208, Springer, ISBN 0387946616, New York.
- Songmuang, P. & Ueno, M. (2005). e-Testing Management System. *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2005*, pp. 3139-3148.
- Thissen, D. (1983). Timed testing: An approach using item response theory. In: *New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing*, D. Weiss (Ed.), 179-203, Academic Press, ISBN 0127427805, New York.
- Ueno, M. (2005a). On-line data-analysis of e-learning response time using gamma distribution. *Japan Journal of Educational Technology*, Vol.29, No.2, 107-117, ISSN 0385-5236.
- Ueno, M. (2005b). Web-based computerized testing system for distance education. *Educational Technology Research*, Vol.28, 59-69, ISSN 0387-7434.
- Ueno, M. (2006). Introduction to test theory. *Trend of University Entrance Examination Research*, Vol.23, 63-70. (In Japanese)
- Ueno, M. (2007). Statistical analysis of national center for university entrance examinations (basic information relations). *Research Bulletin of National Center for University Entrance Examinations*, Vol.36, 71-99. (In Japanese)
- Verhelst, N. D., Verstralen, H. H. F. M. & Jansen, M. G. H. (1997). A logistic model for time-limit tests. In: *Handbook of Modern Item Response Theory*, W. J. van der Linden and R. K. Hambleton (Eds.), 169-185, Springer, ISBN 0387946616, New York.

Advances in Learning Processes

Edited by Mary Beth Rosson

ISBN 978-953-7619-56-5

Hard cover, 284 pages

Publisher InTech

Published online 01, January, 2010

Published in print edition January, 2010

Readers will find several papers that address high-level issues in the use of technology in education, for example architecture and design frameworks for building online education materials or tools. Several other chapters report novel approaches to intelligent tutors or adaptive systems in educational settings. A number of chapters consider many roles for social computing in education, from simple computer-mediated communication support to more extensive community-building frameworks and tools. Finally, several chapters report state-of-the-art results in tools that can be used to assist educators in critical tasks such as content presentation and grading.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Pokpong Songmuang and Maomi Ueno (2010). E-Testing Construction Support System with some Prediction Tools, *Advances in Learning Processes*, Mary Beth Rosson (Ed.), ISBN: 978-953-7619-56-5, InTech, Available from: <http://www.intechopen.com/books/advances-in-learning-processes/e-testing-construction-support-system-with-some-prediction-tools>

INTECH
open science | open minds

InTech Europe

University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China

Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

© 2010 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the [Creative Commons Attribution-NonCommercial-ShareAlike-3.0 License](https://creativecommons.org/licenses/by-nc-sa/3.0/), which permits use, distribution and reproduction for non-commercial purposes, provided the original is properly cited and derivative works building on this content are distributed under the same license.

IntechOpen

IntechOpen