# System for Online Detection of Aberrant Responses in e-Testing

Maomi Ueno and Toshio Okamoto
*Graduate School of Information Systems, University of Electro-Communications*
*E-mail: ueno@ai.is.uec.ac.jp*

## Abstract

*We have developed a method for online detection of examinees' aberrant responses. This method uses response time data in e-testing. Unique features of this method are: 1. It includes an outlier detection method using Bayesian predictive distribution. 2. It can be used with small-sample sets. 3. It provides a unified statistical test method of various statistical tests by changing hyper-parameters and provides more accurate test results than commonly used methods. 4. Outlier statistics are estimated by considering both examinee abilities and the difficulty level of items. We evaluated this system, and results of our evaluation show that it is effective.*

## 1. Introduction

Used for providing test items and handling examinee responses on the Internet, e-testing has recently become popular [1] [2]. A unique advantage of e-testing is that we can easily obtain various kinds of data not obtainable by paper-pencil tests, such as item response time data and response changing history. However, these data have not effectively been used, and finding a way to more effectively use them has become important.

Item response time data in e-testing reflects an examinee's cognitive processes rather than response (correct/incorrect) results. This data can therefore be used to assess an examinee's problem-solving processes in e-testing. In this paper, we propose a method for online detection of examinees' aberrant responses using item response time data in e-testing. Here, "aberrant responses" means responses acquired in such ways as cheating and guessing.

In psychometric society, several methods to detect aberrant response patterns in tests have been proposed based on test score data [3] [4]. These studies deal only with test results and do not involve cognitive processes.

Research has also been done on item response time data in e-testing. Ueno and Nagaoka [5] derived an expanded Gamma distribution as a test response time prediction function for e-testing. Furthermore, Ueno [2] developed a prediction tool for test response time that teachers can use when constructing a test. Thissen [6] proposed a log-normal distribution model for test response time distribution. As a response time distribution, Verhelst [7] used Gamma distribution, and Roskam [8] used Weibul distribution. Van der Linden [9] compared various models for a response time data model and reported that the log-normal model showed the highest performance with real data. However, all of this research was aimed at predicting response time distribution in e-testing, and was not aimed at assessing any cognitive processes in e-testing. We aimed to detect aberrant response processes by finding irregular response time data.

In knowledge, discovery, and data-mining (KDD) areas, many outlier detection techniques have been proposed [10-14]. However, applying these conventional techniques to the detection of aberrant responses in e-testing results in the following problems:

- If an examinee gives aberrant responses at the beginning of an examination, regular responses are regarded as irregular.
- Conventional techniques assume that all data in a time series corresponds to the same task. However, in e-testing situations, the tasks (items) in a time series are different depending on the level of difficulty of the items.
- In conventional techniques, the criteria specifying the outlier are not clear with regard to statistics.

We developed our method so that these problems could be avoided. Unique advantages of this method are:

1. It combines prior knowledge about content properties using the Bayesian approach; it disregards regular responses as aberrant responses even if an examinee gives aberrant responses at the beginning of an examination.
2. It uses a model in which outlier statistics are estimated by taking into account both the level of item difficulty and examinee abilities.
3. It is a unified statistical test method combining various statistical methods, and it has a clear mean and criterion with regard to statistical predictive distribution.

We also developed an online system that detects aberrant responses in e-testing. We evaluated our method and system, and evaluation results show that they are both effective.

978-0-7695-3167-0/08 $25.00 © 2008 IEEE
DOI 10.1109/ICALT.2008.171

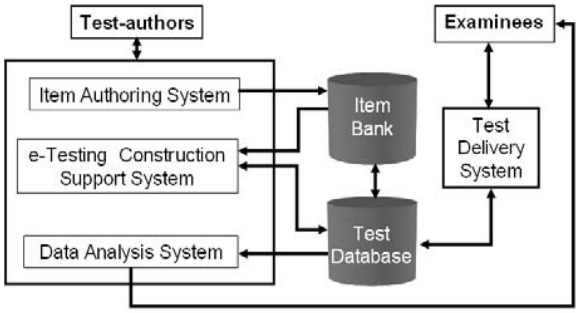


Figure 1. Framework of e-testing system

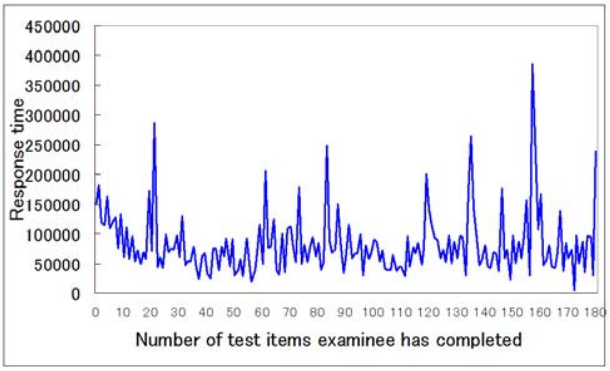


Figure 2. Response time data for items

## 2. e-Testing System

Ueno developed the e-testing system [2] to support test-authors when they create test items, produce and administer tests, and analyze test history data. The system has the following features: 1) an item authoring system, 2) an e-testing construction support system, 3) a data analysis system, 4) an item bank, 5) a test database, and 6) a test delivery system, as shown in Figure 1. An example of the e-testing items is shown in Figure 2. The e-testing system monitors examinee responses and saves them as log data in the test database. The saved data consists of A) test ID, B) examinee ID, C) the number of examinees that are taking the examination, D) test item ID, E) operation order ID, F) operation ID that indicates what operation was performed in the item answering process, G) date and time ID that indicates when the operation started, and H) time ID, which indicates the time taken to perform the operation.

## 3. On-Line Aberrant Response Detection

### 3.1 Data

The data used for aberrant response detection is response time data, as shown in Figure 2. The horizontal axis indicates the number of items on which an examinee has been tested, and the vertical axis indicates the response time for each item. Bayesian predictive distribution of new data $x_n + 1$, the examinee's response time data $x_1, \ldots, x_n$, and provision of a statistical test for outlier detection of the new data are the main ideas behind our method.

### 3.2 Model

In this section, we discuss Bayesian predictive distribution of new data $x_{n+1}$ based on the examinee's response time series data $x_1, \ldots, x_n$. Let $t_{ij}$ be examinee $j$'s response time for the $i$-th item in the following linear equation:

$$x_{ij} = \frac{t_{ij} - \bar{t}_i}{s_i} = \mu_j + e_j \qquad (1)$$

For Bayesian predictive distribution of new data $x_{n+1}$ based on the examinee's response time series data $x_1,\ldots, x_n$ is derived as:

$$p(x_{n+1} \mid X) = \int\int p(x_{n+1} \mid \mu, \sigma^2) p(\mu, \sigma^2 \mid x_1, \cdots, x_n) d\mu d\sigma^2 \qquad (2)$$

$$= \left( 1 + \left[ \frac{(x_{n+1} - \mu_*)}{\sqrt{\frac{n_0 + n + 1}{(n_0 + n)\nu} \lambda_*^{\,2}}} \right]^2 / \nu \right)^{-\frac{\nu+1}{2}}$$

where

$$t = \frac{(x_{n+1} - \mu_*)}{\sqrt{\frac{(n_0 + n + 1)}{(n_0 + n)\nu} \lambda_*^{\,2}}} \qquad (3)$$

and $t$ follows $t$ distribution with a degree of freedom ($\nu = n_0 + n - 1$).

Here, $\mu_*$ indicates the hyper parameter of the prior distribution, which is the mean parameter of the normal distribution. The value of $\mu_*$ is used to determine the mean of data $x$. In this case, data $x$ is standardized with a mean of zero and a standard deviation of one, then, $\mu_*$ is zero. This prior distribution combines prior knowledge of the item in relation with aberrant response detection. For example, even if an examinee gives aberrant responses at the beginning of a test, this method does not identify the response process as regular.

### 3.3. Aberrant Response Detection

From (3), aberrant responses can be detected as follows.

1. New data $x_{ij}$ is obtained.
2. The value of $t$ in (3) is calculated.
3. If $t$ is greater than the value of $t$ in $t$ distribution with $\alpha$, or $t$ is less than the value of minus $t$ in $t$ distribution with $\alpha$, then the new data is detected.

A unique feature of this test method is that it represents various conventional statistical test methods when the value of the hyper parameter $n_0$ changes, which is described as follows.

- When the value of the hyper parameter becomes high enough, the method is equivalent to a Z test.

Table 1. Comparisons of aberrant data detection methods with different hyper parameters

| N | Probabilities in which methods inaccurately detect regular data | | | | | | | Probabilities in which methods accurately detect aberrant data | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Z Test | $n_0$=0 | $n_0$=1 | $n_0$=5 | $n_0$=10 | $n_0$=15 | $n_0$=20 | Z test | $n_0$=0 | $n_0$=1 | $n_0$=5 | $n_0$=10 | $n_0$=15 | $n_0$=20 |
| 0–10 | .37 | .07 | .057 | .10 | .15 | .21 | .25 | .89 | .72 | .74 | .82 | .86 | .87 | .88 |
| 10–20 | .47 | .12 | .11 | .19 | .28 | .36 | .42 | .98 | .88 | .90 | .97 | .99 | .99 | 1.00 |
| 20–30 | .44 | .11 | .10 | .16 | .24 | .31 | .37 | .95 | .84 | .85 | .92 | .95 | .95 | .96 |
| 30–40 | .47 | .14 | .12 | .20 | .28 | .37 | .43 | .98 | .91 | .92 | .97 | .99 | .99 | .99 |
| 40–50 | .46 | .15 | .14 | .20 | .29 | .37 | .43 | .99 | .97 | .97 | .99 | .99 | .99 | 1.00 |
| 50–60 | .46 | .15 | .14 | .21 | .29 | .37 | .43 | .99 | .97 | .98 | .95 | 1.00 | 1.00 | 1.00 |
| 60–70 | .46 | .15 | .14 | .21 | .29 | .37 | .43 | .99 | .98 | .98 | .99 | 1.00 | 1.00 | 1.00 |
| 70–80 | .45 | .15 | .14 | .21 | .28 | .37 | .43 | .99 | .98 | .99 | .99 | 1.00 | 1.00 | 1.00 |
| 80–90 | .44 | .16 | .15 | .21 | .29 | .37 | .43 | .99 | .99 | .99 | .99 | 1.00 | 1.00 | 1.00 |
| 90–100 | .45 | .16 | .15 | .22 | .29 | .37 | .43 | .99 | .99 | .99 | .99 | 1.00 | 1.00 | 1.00 |

- When the value of the hyper parameter is equivalent to zero (called "non-information prior distribution"), the method is equivalent to the conventional *t* test.

The proposed method thus combines various conventional test methods.

## 4. Simulation Experiments

Although the proposed method represents various statistical test methods when the value of the hyper parameter $n_0$ changes, how to determine the value of the hyper parameter $n_0$ is still unknown. In this section, simulation experiments used to determine the optimum value of the hyper parameter are described. The flow of the experiment is as follows.

1. Fix learner *j* and generate random data from:
$$x_{ij} = \frac{t_{ij} - \bar{t}_i}{s_i} = \mu_j + e_j$$
2. Apply the method to the generated data.
3. The above two steps are repeated up to 1000 times.
4. Calculate the probabilities in which the methods correctly detect aberrant data and the probabilities in which the methods incorrectly detect regular data by changing the value of the hyper parameters.

Results are shown in Table 1. Column *n* in the table shows the numbers of random data sequences, which are used for aberrant data detection. For example, 0–10 indicates that aberrant data detection is completed using data from the first to tenth set of data in the data sequence. The calculated probabilities are obtained using the probability average of incorrectly detected regular data and the probability average of correctly detected aberrant data. In addition, this procedure is done by changing the value of the hyper parameter. The results show that the probabilities for correct detection increase along with the value of the hyper parameter. When the value of the hyper parameter is high, the probabilities corresponding to each value of this parameter are close to the probabilities in the Z test. On the other hand, results show that the probabilities of incorrect detection are high when the value of the hyper parameter is low. In such a case, the probabilities in $n_0 = 0$ are equivalent to those in the *t* test. We thus have to decide the value of the hyper parameter while considering the balance between two probabilities. A low probability of incorrect detection is important and so we use the value of the hyper parameter $n_0 = 1$, which minimizes the probability.

## 5. Detection Examples

Figure 3 shows the aberrant responses of examinee 4 with a detection curve corresponding to that shown in Figure 2. The four parallel lines in Figure 3 indicate the outlier detection line. For example, if the *t* value corresponding to a response time exceeds the top detection curve, the response time is irregularly too long. If the *t* value corresponding to a response time exceeds the bottom detection curve, the response time is irregularly too short.

In Figure 3, the aberrant response time for items 105–145 is shown. With regard to Figures 2 and 3, it should be noted that the responses, which seem very long or very short in Figure 2, are not always aberrant responses. The reason for this is the statistical value of *t* for the detection is estimated using both the examinee's response speed ability and the difficulty level of each item. For example, even if we find the responses that take longer to make than others, we cannot determine whether the responses are aberrant responses. This is because the item may require more response time than others. These features are quite different from conventional data mining methods using the outlier system. For example, discovering theft during a money deposit at a bank, or discovering an irregular invasion on a computer network that uses computer network processes. Figure 4 shows examples of the raw response time data and the detection curves in a case where there are few aberrant responses and in a case where there are many aberrant responses. Looking at the raw data for examinee 8, the response times for items 72–92 are irregularly long; however, the response times for items 77 and 78 are detected as outliers.

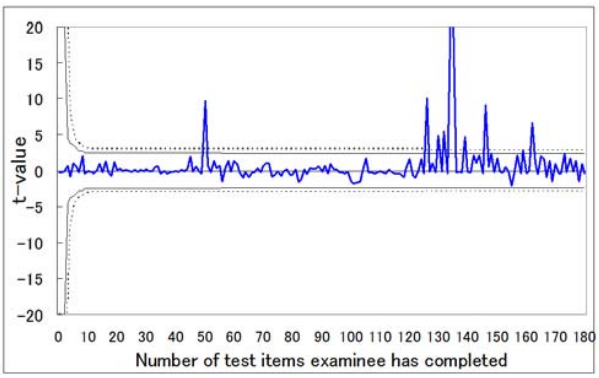


Figure 3. Detection curve corresponding to Figure 3

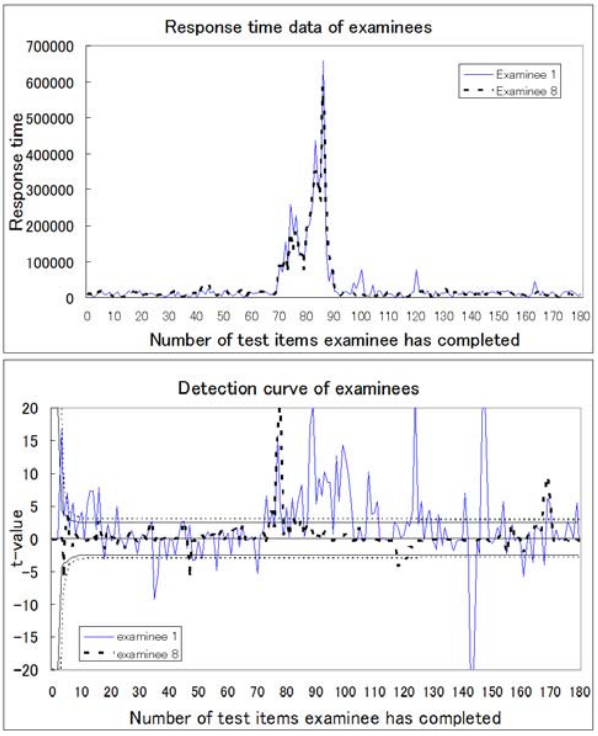


Figure 4. Detection curve examples for case of few aberrant responses and for case of many aberrant responses

It should be noted that the shape of raw data for examinee 1 is very similar; however, at the bottom of Figure 4, many aberrant responses in the data for examinee 1 are shown. The reason for this is that the proposed outlier statistics are estimated by considering both the level of item difficulty and examinee abilities; therefore, minor differences in response time data significantly affect the outlier statistics, especially when the average response time data is large and its variance is very small.

The proposed method can thus detect outlier data, which is barely noticed simply by analyzing raw response time data.

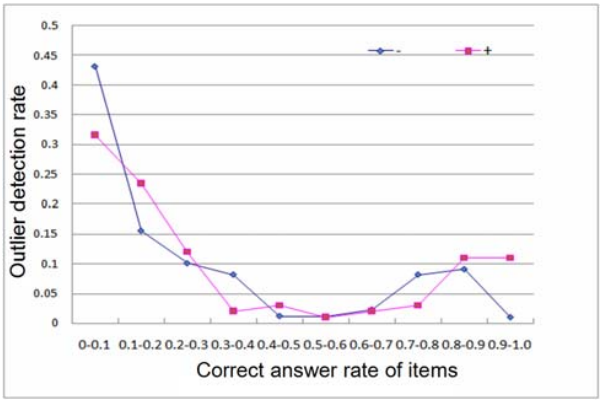


Figure 5. Relationship between item difficulty and aberrant response frequencies


Figure 6. Online aberrant response detection system

## 6. Relationship between item difficulty and aberrant response frequencies

The horizontal axis in Figure 5 indicates the correct answer rate for items, and the vertical axis indicates the detected aberrant response probabilities given the correct answer rate for items from a bank of 274 items. Here, "+" in the figure indicates "too late a response", and "-" in the figure indicates "too fast a response". The figure shows that the frequencies of aberrant responses, both too late a response and too fast a response, are very high, especially when the item is difficult. This means that examinees tend

to answer by guessing or deliberately take a long time to think about the items. If an examinee's response is aberrantly fast and a correct answer is made for a difficult item, then the correct answer might have been given as a result of guessing or cheating. If an examinee's response is aberrantly late and a wrong answer is given for an easy item, then the wrong answer might have been given as a result of a careless miss.

## 7. Online Aberrant Response Detection System

We developed an e-testing platform system that incorporates the aberrant response detection system. The outlier detection system is shown in Figure 6. The system shows: 1. the examinee's name, 2. the examinee's *t* value curve, and 3. the detected item; "+" in the figure indicates too late a response, and "-" indicates too fast a response. Users can therefore easily find aberrant responses from the test results in Figure 5.

## 8. Evaluation of System

In our evaluation experiment, a test with 40 items concerning computer science was given to 74 graduate students. We assigned one item that was randomly sampled to each examinee and showed each student the item and answer before the test. Next, we told the students that the item would be in the test and that they had to memorize its content and its corresponding answer. Their answer for this item is one type of aberrant response, and we expected the system to detect it. After the test, we estimated the probability of detecting the aberrant responses and the probability of incorrectly detecting regular responses. The results are shown in Table 2, which shows that the system effectively detects aberrant responses.

Table 2. Evaluation experiment results

| | Detection | No detection |
|---|---|---|
| Aberrant responses | 0.959 | 0.041 |
| Regular responses | 0.078 | 0.922 |

## 9. Conclusion

We created a method for online detection of aberrant responses using response time data in e-testing. The method is an outlier detection method that uses Bayesian predictive distribution.

Unique advantages of this method are:

1. It combines prior knowledge about content properties using the Bayesian approach; it disregards regular responses as aberrant ones even if an examinee gives aberrant responses at the beginning of an examination.
2. It uses a model in which outlier statistics are estimated by taking into account both the level of item difficulty and the examinee's abilities.
3. It is a unified statistical test method combining various statistical methods, and it has a clear mean and criterion with regard to statistical predictive distribution.

We performed simulation experiments to evaluate the proposed aberrant response detection method and showed that the method is effective. We also developed an online system that detects aberrant responses in e-testing. We evaluated this method and system, and results of our evaluation show that both are effective.

## 10. References


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

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

[3] K. K. Tatsuoka and R. L. Linn, "Indices for detecting unusual response patterns: Links between two general approaches and potential applications", *Applied Psychological Measurement, **7*(1)**, pp. 81–96 (1983)

[4] K. K. Tatsuoka and M. M. Tatsuoka, "Detection of aberrant response patterns", *Journal of Educational Statistics, **7*(3)**, pp. 215–231 (1982)

[5] M. Ueno and K. Nagaoka, "On-line Data-analysis of e-Learning Response Time using Gamma Distribution", Educational Technology Research, ***29*(1–2)**, pp. 41–56 (2006)

[6] D. Thissen, "Timed testing: An approach using item response theory," New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing, ed. D. J. Weiss, pp. 179–203, Academic Press, New York (1983)

[7] N. D. Verhelst, H. H. F. M. Verstralen and M. G. H. Jansen, "A logistic model for time limit tests (1997)" Handbook of modern item response theory, eds. W. J. van der Linden and R. K. Hambleton, pp. 169–185, Springer-Verlag, New York, (1997)

[8] E. E. Roskam, "Models for speed and time-limit test," Handbook of modern item response theory, eds. W.J. van der Linden and R. K. Hambleton, pp. 187–208, Springer-Verlag, New York, (1997)

[9] W. J. van der Linden, "A lognormal model for response times on test items," Journal of Educational and Behavioral Statistics, ***31***, pp. 181–204, (2006)

[10] V. Barnett and T. Lewis, "Outliers in Statistical Data", John Willy & Sons (1994)

[11] F. Bonchi, F. Giannotti, G. Mainetto, and D. Pedeschi, "A classification-based methodology for planning audit strategies in fraud detection", Proc. KDD-99, pp. 175–184 (1999)

[12] P. Burge and J. Shaw-Taylor, "Detecting cellular fraud using adaptive prototypes", Proc. AI Approaches to Fraud Detection and Risk Management, pp. 9–13, (1997)

[13] T. Fawsett and F. Provost, Combining data mining and machine learning for effective fraud detection, Proc. AI Approaches to Fraud Detection and Risk Management, pp. 14–19 (1997)

[14] W. Lee, S. J. Stolfo, and K. W. Mok, "Mining in data-flow environment experience in network intrusion detection", Proc. KDD99, pp. 114–124 (1999)