PAPER *Special Section on Advanced Technologies in Knowledge Media and Intelligent Learning Environment*

Latent Conditional Independence Test Using Bayesian Network Item Response Theory

Takamitsu HASHIMOTO[†,††], *Nonmember and* Maomi UENO[†], *Member*

SUMMARY Item response theory (IRT) is widely used for test analyses. Most models of IRT assume that a subject's responses to different items in a test are statistically independent. However, actual situations often violate this assumption. Thus, conditional independence (CI) tests among items given a latent ability variable are needed, but traditional CI tests suffer from biases. This study investigated a latent conditional independence (LCI) test given a latent variable. Results show that the LCI test can detect CI given a latent variable correctly, whereas traditional CI tests often fail to detect CI. Application of the LCI test to mathematics test data revealed that items that share common alternatives might be conditionally dependent.

key words: *item response theory, Bayesian network model, local independence, conditional independence test, latent variable*

1. Introduction

Lord and Novick [1] used a modern mathematical statistical approach to formulate the basic constructs of the item response theory (IRT). Since then, a great deal of research effort has been spent in developing their idea from different perspectives (e.g., statistical theory and parameter estimation algorithms). There are many possible IRT models, which differ in the mathematical form of the item characteristic function and/or the number of parameters specified in the model, for example, the Rasch model [2], the normal ogive model [1], the two parameters logistic model [3], and the three parameters logistic model [4]. There are more general and well-known IRT models such as the graded response model [5], the free response model [6], the partial credit model [7], and the nominal response model [8]. All IRT models incorporate one or more parameters describing the subject. IRT rests on three basic postulates: (1) The performance of a subject for a test item can be predicted (or explained) by a set of factors called traits, latent traits, or abilities. (2) The relation between the subject's item performance and the set of traits underlying item performance can be described using a monotonically increasing function called an item characteristic function or item characteristic curve. (3) When the ability variables influencing the test performance are held constant, the subject's responses to any pair of items are statistically independent, which is often called local independence. It should be particularly noted that assumption (3) states that a subject's responses to different items in a test are statistically independent (Fig. 1). For this assumption to be true, a subject's performance for one item must not affect, either positively or negatively, the response to any other item in the test. Regarding this conditional independence (CI) assumption, Yen [9] previously pointed out that actual situations often violate this assumption. Furthermore, many previous studies [10]–[13] have shown that the parameter estimation often fails when the CI assumption is violated.

Consequently, CI tests among items given a latent variable are necessary for the application of IRT to test data. However, it is difficult to realize the CI test given a latent variable. Considering this problem, several CI tests given a latent variable have been proposed. For example, Yen's Q_3 statistic [9] is defined as the correlation coefficient of two items' fitting errors between the expected and actual responses. Chen and Thissen's G^2 statistic [10] is the log-likelihood ratio of an observed frequency to an expected frequency derived from an IRT model. These statistics marginalize out the latent variable and provide CI tests of only the two target items. Namely, the traditional methods implicitly assume that the CI tests of two variables are not affected by any other two items' dependencies when the latent variable is marginalized out. However, in the present study, we found through some simulation experiments that this assumption does not hold. That is, we show that the other two variables dependencies affect the CI test of the two target variables even when the latent variable is marginalized out.

To solve this problem, we propose a new CI test between two items given a latent variable. The unique feature of our method is a CI test of the target items given all the other items, which are assumed to be mutually dependent (complete structured variables). In addition, we use Bayesian network IRT [14], which alleviates the CI assumption given the latent variable, to calculate our CI test. We

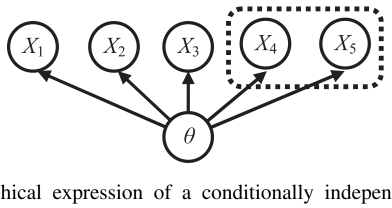

Fig. 1 Graphical expression of a conditionally independent structure given a latent ability variable.

Manuscript received June 29, 2010.
Manuscript revised December 18, 2010.

[†]The authors are with the University of Electro-Communications, Chofu-shi, 182–8585 Japan.
[††]The author is with the National Center for University Entrance Examinations, Tokyo, 153–8501 Japan.

DOI: 10.1587/transinf.E94.D.743

also prove that our CI test correctly detects the dependency of the target variables given the latent variable even when two other items are dependent. The effectiveness of our test was confirmed in some simulation experiments.

This paper is organized as follows. Section 2 briefly reviews IRT. Section 3 explains the traditional CI tests. Section 4 describes the problems of the traditional CI tests. Section 5 presents our proposal for solving the problem. Section 6 describes some experiments in comparison with traditional CI tests and Sect. 7 presents examples of application to real data. Section 8 discusses the results and presents our conclusions.

2. IRT

IRT represents the probability of an examinee answering an item correctly as a function of the examinee's latent ability variable θ. In the usual dichotomous response formulation of IRT, the correctness of the i-th item in a test is indicated by a random response variable X_i taking the value 1 for a correct response and 0 for an incorrect response. For example, a two-parameter logistic (2PL) model [4], which is the most popular IRT model, is expressed by

$$P(X_i = 1|\theta, a_i, b_i) = \frac{1}{1 + \exp\{-1.7a_i(\theta - b_i)\}}, \tag{1}$$

where item parameters a_i and b_i are called the "discrimination parameter" and "difficulty parameter", respectively.

When the discrimination parameters of all items are equivalent, this model is called the one-parameter logistic model or Rasch model [2]. When one more item parameter, called the "guessing parameter", is added to the 2PL model, the model is called the three-parameter logistic (3PL) model [4].

The likelihood function of the 2PL model is given by

$$\begin{aligned} L(\mathbf{X}|\theta, \xi) = \prod_{j=1}^{N}\prod_{i=1}^{m}&\left[\frac{1}{1 + \exp\{-1.7a_i(\theta_j - b_i)\}}\right]^{x_{ji}} \cdot \\ &\left[1 - \frac{1}{1 + \exp\{-1.7a_i(\theta_j - b_i)\}}\right]^{1-x_{ji}}, \end{aligned} \tag{2}$$

where

$\mathbf{X} = \{\, x_{ji} \,\} \quad (j = 1, \cdots, N;\ i = 1, \cdots, m)$

$$x_{ji} = \begin{cases} 0 & \text{for the } j\text{−th examinee's incorrect} \\ & \text{response to the } i\text{−th item} \\ 1 & \text{for the } j\text{−th examinee's correct} \\ & \text{response to the } i\text{−th item} \end{cases}$$

$\theta = \{\, \theta_j \,\} \quad (j = 1, \cdots, N)$

θ_j : the j−th examinee's latent ability variable

$\xi = \{\, \xi_i \,\} \quad (i = 1, \cdots, m)$

$\xi_i = (a_i,\ b_i)^t \quad (i = 1, \cdots, m)$

N : number of examinees

m : number of items

Equation (2) assumes that a subject's responses to different items are conditionally independent given the variable θ. This assumption is called the "local independence" of items. However, as described in Sect. 1, in actual educational assessment, many factors violate the local independence assumption. For example, Yen [9] pointed out the following causes of local dependence of items: external assistance or interference, speed, fatigue, practice, item or response format, passage dependence, item chaining, explanation of a previous answer, and scoring rubrics or raters.

If we apply the IRT model to data that do not satisfy the local independence assumption, it will suffer seriously biased estimates. Therefore, it is important to detect items that violate the local independence assumption and remove them in order to estimate parameters reliably.

3. Previous Work on CI Detection Given a Latent Variable

Several methods for testing CI between a pair of items given a latent variable have been proposed.

Yen [9] proposed the use of the Q_3 statistic as a score for identifying pairs of items that display CI given a latent variable. The Q_3 statistic of the i-th and i'-th items is the correlation coefficient of the following d_{hi} and $d_{hi'}$:

$$d_{hi} = x_{hi} - \hat{P}_i(\hat{\theta}_h)\ ,\ d_{hi'} = x_{hi'} - \hat{P}_{i'}(\hat{\theta}_h), \tag{3}$$

where x_{hi} denotes the score of the h-th examinee for the i-th item, $\hat{\theta}_h$ denotes the estimate of the h-th examinee's estimated latent variable, and $\hat{P}_i(\hat{\theta}_h)$ represents the probability of the correct answer given the estimated parameters. Actually, Q_3 requires estimates of latent variables and item parameters. These estimates are obtained assuming local independence of all items, even when some items are locally dependent. Thus, the estimates are biased and cause error in local independence detection. Moreover, Q_3 has a high computational cost because it requires estimation of the latent variable.

Another statistic suggested for identifying CI given a latent variable is the G^2 statistic developed by Chen and Thissen [10]. Unlike Q_3, G^2 does not require any estimation of latent variables. The G^2 statistic is calculated through the following procedure.

Let $N_{x_i x_{i'}}$ be the number of examinees whose responses to the i-th item X_i and i'-th item $X_{i'}$ are x_i and $x_{i'}$ ($x_i, x_{i'} = 0, 1$), respectively; let N be the total number of examinees. Under a null hypothesis, the expected number of examinees with x_i and $x_{i'}$ is given by

$$\begin{aligned} E_{x_i x_{i'}} = N \int & \hat{P}_i(\theta)^{x_i} \hat{P}_{i'}(\theta)^{x_{i'}} \cdot \\ & \left[1 - \hat{P}_i(\theta)\right]^{1-x_i} \left[1 - \hat{P}_{i'}(\theta)\right]^{1-x_{i'}} p(\theta)d\theta. \end{aligned} \tag{4}$$

Here, $\hat{P}_i(\theta)$ is the item characteristic curve in which item parameter estimates are substituted. The G^2 statistic is computed as

$$G^2 = 2 \sum_{x_i=0}^{1} \sum_{x_{i'}=0}^{1} N_{x_i x_{i'}} \log_e \frac{N_{x_i x_{i'}}}{E_{x_i x_{i'}}}. \tag{5}$$

Although G^2 integrates out the latent variable, it still requires estimates of item parameters. Since these estimates are obtained under the local independence assumption, the same problem as that of Q_3 occurs: detection bias caused by parameter estimation bias.

The following simulation experiments demonstrated this problem.

4. Problems of Previous Work

Yen's Q_3 statistic [9] and Chen and Thissen's G^2 statistic [10] implicitly assume that the CI tests of two items are not affected by the dependencies of any other two items when the latent variable is marginalized out. In this section, we describe how we applied these tests to two locally independent items when some other items were locally dependent.

4.1 Method

We considered the following six structures in our simulation experiments.

Three structures each consisted of 7 items. The following three cases were assumed: (a) completely independent case, (b) one pair dependent case, and (c) two pairs dependent case. In case (a), all items are locally independent (Fig. 2). This is the case assumed by traditional methods. In case (b), one pair of items were locally dependent, and the other pairs of the items were locally independent (Fig. 3). In case (c), two pairs of items were locally dependent, and the other pairs of the items were locally independent (Fig. 4).

The other three structures each consisted of 20 items. The following three cases were assumed: (d) completely independent case (Fig. 5), (e) one pair dependent case (Fig. 6), and (f) nine pairs dependent case (Fig. 7). Only case (d) met the assumption of the traditional methods.

For these cases, item parameters were determined as follows.

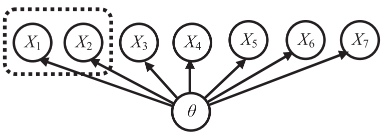

Fig. 2 Case (a) (completely independent case).

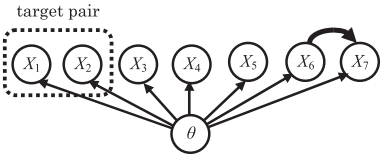

Fig. 3 Case (b) (one pair dependent case).

Parameters of locally independent items In all cases, some items were locally independent of the other items. In Figs. 2–7, such items are not linked to the other items by arcs. Parameters of such items were generated randomly from the following distributions:

$$\log_2 a_i \sim N(0,\ 1)$$
$$b_i \sim N(0,\ 1).$$

Parameters of locally dependent items In cases (b) and (c), some items were assumed to be locally dependent. The responses to some items changed the difficulty of other items. Items that changed difficulty parameters of other items were designated "parent" items, and items with difficulty parameters that were changed by parent items were designated "child" items. Parameters of parent items were generated in the same way as independent items. When an examinee answered the parent item correctly, the difficulty parameter of the child item was assumed to be smaller. Otherwise, that parameter was assumed to be larger. Parameters of child items were generated as

$$\log_2 a_i \sim N(0,\ 1)$$
$$b_i \sim N(0,\ 0.25)$$
$$d \sim N(1.85,\ 0.25)$$

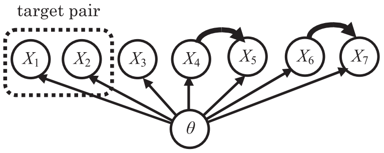

Fig. 4 Case (c) (two pairs independent case).

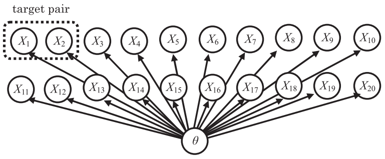

Fig. 5 Case (d) (completely independent case).

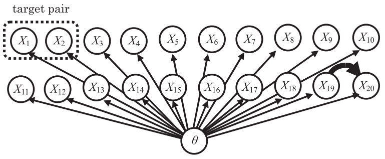

Fig. 6 Case (e) (one pair dependent case).

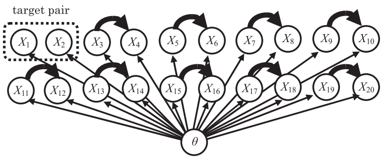

Fig. 7 Case (f) (nine pairs independent case).

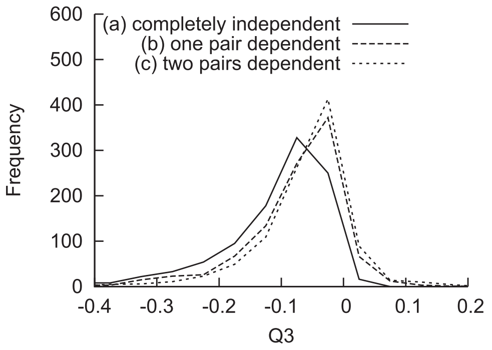

Fig. 8 Frequency distributions of Q_3 of locally independent pairs (number of items: 7).

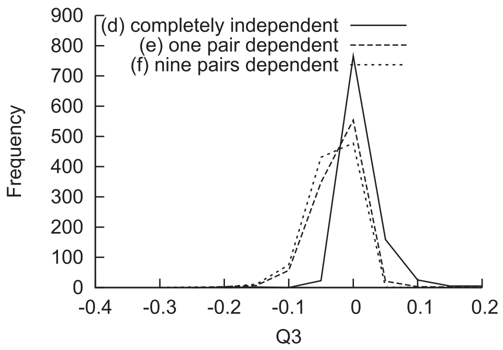

Fig. 9 Frequency distributions of Q_3 of locally independent pairs (number of items: 20).

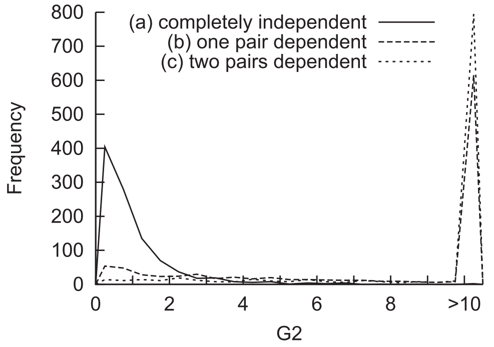

Fig. 10 Frequency distributions of G^2 of locally independent pairs (number of items: 7).

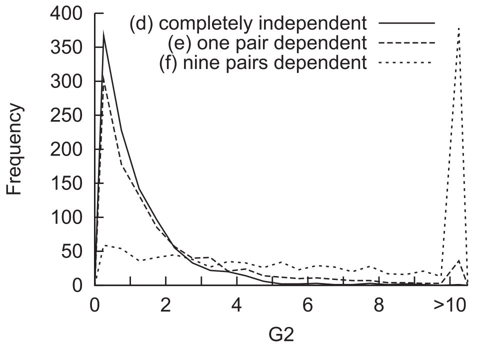

Fig. 11 Frequency distributions of G^2 of locally independent pairs (number of items: 20).

$$
\begin{aligned}
b_{i0} &= b_i + d \\
b_{i1} &= b_i - d,
\end{aligned}
$$

where b_{i0} denotes the difficulty parameter of the i-th child item when the answer to the parent item was incorrect, and b_{i1} denotes the difficulty parameter when the answer to the parent item was correct.

These parameters were used to generate 10,000 examinees' responses randomly. In this way, 1000 sets of data were generated for each case.

In all cases, X_1 and X_2 were locally independent. Local independence between X_1 and X_2 was tested by Yen's Q_3 [9] and Chen and Thissen's G^2 [10].

4.2 Results

Frequency distributions of Q_3 in cases (a), (b), and (c) are shown in Fig. 8. The solid line is the distribution when items other than the target pairs were locally independent, which is a necessary assumption for Q_3. Compared with the distribution in case (a), values of Q_3 in cases (b) and (c) were larger.

When the number of items was 20 (Fig. 9), the distributions in cases (e) and (f) did not fit the distribution in case (d).

Frequency distributions of G^2 in cases (a), (b), and (c) are shown in Fig. 10. In cases (b) and (c), excessively large values were obtained. Distributions in cases (b) and (c) did not fit the distribution in case (a).

However, when the number of items was 20 (Fig. 11), the distribution in case (e) fitted the distribution in case (d). On the other hand, the distribution in case (f) did not fit the distribution in case (d).

These results show that, when the number of locally dependent items other than the targets increases, the statistics of traditional CI tests are seriously biased. Therefore, this bias might cause bias of CI detection.

5. Proposed Method

The traditional CI tests implicitly assume that all items except the two target items are conditionally independent given a latent variable. However, our simulation experiment revealed that these CI tests are biased when this assumption does not hold. Here, we propose a new method to solve this problem. Our method uses the Bayesian network IRT model [14], which alleviates the local independence assumption of traditional IRT models.

5.1 Bayesian Network IRT

Bayesian network IRT [14] is an IRT model that relaxes the CI assumption given a latent variable. This model introduces different item parameters for responses to other items. An item whose response changes the item parameter value

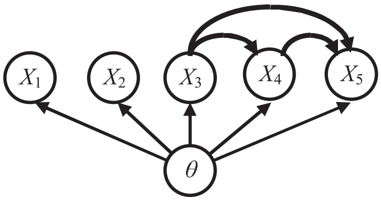

Fig. 12 Example of the structure of items and the latent variable.

of the i-th item is designated the parent item of the i-th item. Consequently, the Bayesian network IRT model is regarded as an IRT model containing parent items. This model is described in detail below.

Let a certain examinee's response pattern to m items be

$$\mathbf{x} = (x_1, x_2, \cdots, x_i, \cdots, x_m)^t,$$

where

$$x_i = \begin{cases} 0 & \text{for an incorrect response} \\ 1 & \text{for a correct response} \end{cases}. \tag{6}$$

When B_s encodes the CI assertions in a model, the joint probability of scores is given by

$$\begin{aligned} &P(X_1 = x_1, \cdots, X_m = x_m|\theta, \xi, B_s) \\ &= \prod_{i=1}^{m} \prod_{j=0}^{2^{p_i}-1} \Big\{ P(X_i = 1|\theta, \xi_i, \tilde{\mathbf{X}}_{ij})^{x_i u_{ij}} \cdot \\ &\qquad \left[1 - P(X_i = 1|\theta, \xi_i, \tilde{\mathbf{X}}_{ij})\right]^{(1-x_i)u_{ij}} \Big\}, \end{aligned} \tag{7}$$

where

$$u_{ij} = \begin{cases} 1 & \text{for the } j\text{–th response pattern to parent} \\ & \text{items of the } i\text{–th item} \\ 0 & \text{for the other patterns} \end{cases}$$

p_i : number of parent items of the i–th item,

$\tilde{\mathbf{X}}_{ij}$: j–th response pattern to parent items of the i–th item,

ξ_i : parameter vector for $\tilde{\mathbf{X}}_{ij}$

$\xi = \left(\xi_1^t, \xi_2^t, \cdots, \xi_i^t, \cdots, \xi_m^t\right)$

B_s : conditional dependence structure among items

The dependence structure among items in the Bayesian network IRT model can be expressed as a directed graph. In the graph, two conditionally dependent items are linked by a directed arc, whereas conditionally independent items are not linked. The direction of the arc indicates the parent and child: the arc's tail is the parent item and its head is the child item. For example, in the structure shown in Fig. 12, dependencies exist between X_3-X_4, X_3-X_5, and X_4-X_5. X_3 is the parent of X_4 and X_5, and X_4 is the parent of X_5.

Bayesian network IRT can express a conditional item characteristic curve given the responses to other items. For example, item 3 in Fig. 12 is the parent item of item 4, and the item characteristic curve of item 4 $P(X_4 = 1|\theta, \xi_4, \tilde{\mathbf{X}}_{4j})$ can be written as

$$\begin{aligned} &P(X_4 = 1|\theta, \xi_4, \tilde{\mathbf{X}}_{4j}) \\ &= \prod_{j=0}^{1} P(X_4 = 1|\theta, \xi_4, \tilde{\mathbf{X}}_{4j})^{u_{4j}} \\ &= P(X_4 = 1|\theta, a_4, b_{40}, b_{41}, X_3 = 0)^{u_{40}} \cdot \\ &\quad P(X_4 = 1|\theta, a_4, b_{40}, b_{41}, X_3 = 1)^{u_{41}}, \end{aligned} \tag{8}$$

where a_4 is the discrimination parameter of item 4 and b_{4x_3} is the difficulty parameter of item 4 when the response to item 3 is x_3 ($x_3 = 0, 1$).

In this paper, we use the concept of this model to propose a new CI test given a latent variable.

Furthermore, it should be noted that in Bayesian network IRT, the order of test items should be fixed because the parameter estimates are affected by the order. However, our purpose in this paper is to detect the local dependency of two items, so the detection results are not affected by varying the order.

5.2 Conditional Mutual Information Measure

The conditional mutual information measure is a measure of dependence between two random variables. It is used for learning the Bayesian network skeletons, for example, in the PC algorithm [15] and MMPC algorithm [16]. When X, Y, and $\mathbf{Z}$ are random variables, the conditional mutual information between X and Y given $\mathbf{Z}$, which is written as $I(X; Y|\mathbf{Z})$, is calculated as follows.

$$\begin{aligned} &I(X; Y|\mathbf{Z}) \\ &= \sum_{\mathbf{z}} P(\mathbf{Z} = \mathbf{z}) \cdot \\ &\quad \sum_{x} \sum_{y} P(X = x, Y = y|\mathbf{Z} = \mathbf{z}) \cdot \\ &\quad \log_2 \frac{P(X = x, Y = y|\mathbf{Z} = \mathbf{z})}{P(X = x|\mathbf{Z} = \mathbf{z}) P(Y = y|\mathbf{Z} = \mathbf{z})} \end{aligned} \tag{9}$$

Moreover, in the Bayesian network IRT model [14], conditional dependence between two items given a latent variable is measured using the following conditional mutual information measure.

$$\begin{aligned} &I(X_i; X_{i'}|\theta, \xi, B_s) \\ &= \int p(\theta) \cdot \\ &\quad \sum_{x_i=0}^{1} \sum_{x_{i'}=0}^{1} P(X_i = x_i, X_{i'} = x_{i'}|\theta, \xi, B_s) \cdot \\ &\quad \log_2 \frac{P(X_i = x_i, X_{i'} = x_{i'}|\theta, \xi, B_s)}{P(X_i = x_i|\theta, \xi, B_s) P(X_{i'} = x_{i'}|\theta, \xi, B_s)} d\theta. \end{aligned} \tag{10}$$

To calculate Eq. (10), we need to know the structure B_s and item parameters ξ. In this study, we assumed a structure B_s in which all items are mutually dependent (Fig. 13).

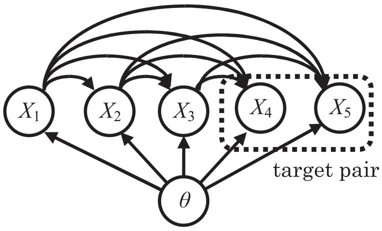

Fig. 13 Structure of B_s^c (completely dependent structure given a latent variable).

Such a structure is designated as a completely dependent structure (complete graph) and denoted B_s^c. The direction of the arc from a previously shown item to a later shown item is determined by the test item order.

Let the target items be X_i and $X_{i'}$. Assuming B_s^c means that all items except the targets, which are denoted $\mathbf{X}^{\neg ii'}$, are regarded as parents of the targets. In this case, if the topology order of items is given, the conditional probabilities given $\mathbf{X}^{\neg ii'}$ correspond to the conditional probabilities given the estimated latent ability variable $\hat{\theta}$ because, according to Neyman factorization theorem [17], an examinee's response pattern is sufficient for estimation of the latent variable θ when the number of items is sufficiently large. Consequently, in this study we define the following $I(X_i; X_{i'}|\mathbf{X}^{\neg ii'}, \xi, B_s^c)$, in which an examinee's response pattern $\mathbf{X}^{\neg ii'}$ substitutes for his/her latent ability variable θ in Eq. (10).

Definition 1: $I(X_i; X_{i'}|\mathbf{X}^{\neg ii'}, \xi, B_s^c)$

$$= \sum_{j=0}^{2^{m-2}-1} P(\mathbf{X}^{\neg ii'} = \mathbf{x}_j^{\neg ii'}|\xi, B_s^c) \cdot \sum_{x_i=0}^{1}\sum_{x_{i'}=0}^{1} P(X_i = x_i, X_{i'} = x_{i'}|\mathbf{X}^{\neg ii'} = \mathbf{x}_j^{\neg ii'}, \xi, B_s^c) \cdot \log_2 \frac{P(X_i = x_i, X_{i'} = x_{i'}|\mathbf{X}^{\neg ii'} = \mathbf{x}_j^{\neg ii'}, \xi, B_s^c)}{P(X_i = x_i|\mathbf{X}^{\neg ii'} = \mathbf{x}_j^{\neg ii'}, \xi, B_s^c) P(X_{i'} = x_{i'}|\mathbf{X}^{\neg ii'} = \mathbf{x}_j^{\neg ii'}, \xi, B_s^c)}, \tag{11}$$

where

m : number of items
X_i : $i-$th item
$X_{i'}$: $i'-$th item
$\mathbf{X}^{\neg ii'}$: all items except X_i and $X_{i'}$
$\mathbf{x}_j^{\neg ii'}$: $j-$th response pattern to $\mathbf{X}^{\neg ii'}$
$(j = 0, \cdots, (2^{m-2} - 1))$
ξ : item parameters of the model
B_s^c : parent variable set of the $i-$th and $i'-$th items with a completely dependent structure

■

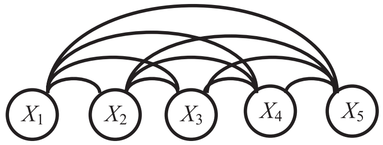

Fig. 14 Graph showing when θ is integrated out from the graph of Fig. 12.

Namely, $I(X_i; X_{i'}|\mathbf{X}^{\neg ii'}, \xi, B_s^c)$ is the conditional mutual information measure between X_i and $X_{i'}$ given all items except X_i and $X_{i'}$.

Using Eq. (11), we can derive the following theorem.

Theorem 1: When $I(X_i; X_{i'}|\mathbf{X}^{\neg ii'}, \xi, B_s^c) = 0$ for $\forall i, i'$, the i-th and i'-th items are conditionally independent given a latent variable.

When $I(X_i; X_{i'}|\mathbf{X}^{\neg ii'}, \xi, B_s^c) > 0$ for $\forall i, i'$, the i-th and i'-th items are conditionally dependent given a latent variable.

■

Proof 1: For $\forall i, i', j$,

$$\sum_{x_i=0}^{1}\sum_{x_{i'}=0}^{1} P(X_i = x_i, X_{i'} = x_{i'}|\mathbf{X}^{\neg ii'} = \mathbf{x}_j^{\neg ii'}, \xi, B_s^c) \log_2 \frac{P(X_i = x_i, X_{i'} = x_{i'}|\mathbf{X}^{\neg ii'} = \mathbf{x}_j^{\neg ii'}, \xi, B_s^c)}{P(X_i = x_i|\mathbf{X}^{\neg ii'} = \mathbf{x}_j^{\neg ii'}, \xi, B_s^c) P(X_{i'} = x_{i'}|\mathbf{X}^{\neg ii'} = \mathbf{x}_j^{\neg ii'}, \xi, B_s^c)} \geq 0. \tag{12}$$

Then, $I(X_i; X_{i'}|\mathbf{X}^{\neg ii'}, \xi, B_s^c) = 0$ when and only when

$$\begin{aligned} &P(X_i = x_i, X_{i'} = x_{i'}|\mathbf{X}^{\neg ii'} = \mathbf{x}_j^{\neg ii'}, \xi, B_s^c) \\ =\ &P(X_i = x_i|\mathbf{X}^{\neg ii'} = \mathbf{x}_j^{\neg ii'}, \xi, B_s^c) \cdot \\ &P(X_{i'} = x_{i'}|\mathbf{X}^{\neg ii'} = \mathbf{x}_j^{\neg ii'}, \xi, B_s^c). \end{aligned} \tag{13}$$

When all items are assumed to depend on the latent variable θ, as shown in Fig. 12, for $\forall i, i'$, both the i-th and i'-th items are made marginally dependent by integrating out θ: the dependence structure of the items is a probability network model of a completely dependent structure, as shown in Fig. 14. An examinee's response pattern is sufficient for estimation of the ability variable θ. Neyman factorization theorem [17] states that, when $\mathbf{x} = (x_1, \ldots, x_n)$ is a random variable with probability density function $f(\mathbf{x}; \theta)$, a necessary and sufficient condition for a statistic $t(\mathbf{x})$ to be sufficient for θ is that

$$f(\mathbf{x}; \theta) = q(t(\mathbf{x}); \theta) \cdot r(\mathbf{x}), \tag{14}$$

where $q(t(\mathbf{x}); \theta)$ is the probability density function of $t(\mathbf{x})$ and $r(\mathbf{x})$ is a function of $\mathbf{x}$ that does not depend on θ. In Eq (7), the values of the variables $\tilde{\mathbf{X}}_{ij}$, x_i, and u_{ij} are fixed when an examinee's response pattern $\mathbf{x}$ is given. Consequently, Eq (7) can be regarded as the probability function of the response pattern $\mathbf{x}$ given parameters θ, ξ, and B_s. Therefore, the response pattern is sufficient for estimating θ when ξ and B_s are given, so θ in (10) can be replaced by the corresponding response pattern.

For sufficiently large m,

$$p(\theta|\mathbf{X}, \xi, B_s^c) \approx p(\theta|\mathbf{X}^{\neg ii'}, \xi, B_s^c). \tag{15}$$

These are the main ideas of this paper.

Accordingly, if

$$\begin{aligned} & P(X_i = x_i, X_{i'} = x_{i'}|\mathbf{X}^{\neg ii'} = \mathbf{x}_j^{\neg ii'}, \xi, B_s^c) \\ &= P(X_i = x_i|\mathbf{X}^{\neg ii'} = \mathbf{x}_j^{\neg ii'}, \xi, B_s^c) \cdot \\ &\quad P(X_{i'} = x_{i'}|\mathbf{X}^{\neg ii'} = \mathbf{x}_j^{\neg ii'}, \xi, B_s^c) \end{aligned} \tag{16}$$

for $\forall i, i', j$, then

$$\begin{aligned} & P(X_i = x_i, X_{i'} = x_{i'}|\theta, \xi, B_s^c) \\ &= P(X_i = x_i|\theta, \xi, B_s^c) \cdot P(X_{i'} = x_{i'}|\theta, \xi, B_s^c). \end{aligned} \tag{17}$$

Consequently, when $I(X_i; X_{i'}|\mathbf{X}^{\neg ii'}, \xi, B_s^c) = 0$ for $\forall i, i'$, the i-th and i'-th items are conditionally independent. In the same manner, when $I(X_i; X_{i'}|\mathbf{X}^{\neg ii'}, \xi, B_s^c) > 0$ for $\forall i, i'$, the i-th and i'-th items are conditionally dependent.

■

Equation (11) includes four conditional probabilities, which are parameters of Bernoulli distributions. The maximum likelihood estimates of those parameters are obtainable as

$$\hat{P}(\mathbf{X}^{\neg ii'} = \mathbf{x}_j^{\neg ii'}|\xi, B_s^c) = \frac{N_j}{N} \tag{18}$$

$$\hat{P}(X_i = x_i, X_{i'} = x_{i'}|\mathbf{X}^{\neg ii'} = \mathbf{x}_j^{\neg ii'}, \xi, B_s^c) = \frac{N_{x_i x_{i'} j}}{N_j}$$

$$\hat{P}(X_i = x_i|\mathbf{X}^{\neg ii'} = \mathbf{x}_j^{\neg ii'}, \xi, B_s^c) = \frac{N_{x_i j}}{N_j} \tag{19}$$

$$\hat{P}(X_{i'} = x_{i'}|\mathbf{X}^{\neg ii'} = \mathbf{x}_j^{\neg ii'}, \xi, B_s^c) = \frac{N_{x_{i'} j}}{N_j}, \tag{20}$$

where the number of examinees whose response pattern is $\mathbf{x}_j^{\neg ii'}$ is defined as $N_{x_i x_{i'} j}$, and

$$\begin{aligned} N_{x_i j} &= N_{x_i 0 j} + N_{x_i 1 j} \\ N_{x_{i'} j} &= N_{0 x_{i'} j} + N_{1 x_{i'} j} \\ N_j &= N_{00j} + N_{01j} + N_{10j} + N_{11j} \\ N &= \sum_{j=0}^{2^{m-2}-1} N_j. \end{aligned}$$

Substituting them into Eq. (11), we get the proposed conditional mutual information measure between the i-th and i'-th items as follows.

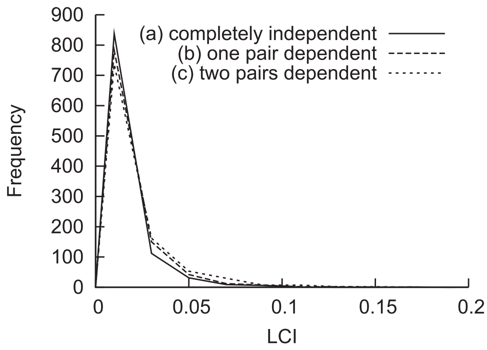

Fig. 15 Frequency distributions of LCI test statistic of locally independent pairs (number of items: 7).

$$\begin{aligned} & I(X_i; X_{i'}|\mathbf{X}^{\neg ii'}, \xi, B_s^c) \\ &= \sum_{j=0}^{2^{m-2}-1} \frac{N_j}{N} \sum_{x_i=0}^{1} \sum_{x_{i'}=0}^{1} \frac{N_{x_i x_{i'} j}}{N_j} \log_2 \frac{\dfrac{N_{x_i x_{i'} j}}{N_j}}{\dfrac{N_{x_i j}}{N_j} \dfrac{N_{x_{i'} j}}{N_j}} \\ &= \frac{1}{N} \sum_{j=0}^{2^{m-2}-1} \sum_{x_i=0}^{1} \sum_{x_{i'}=0}^{1} N_{x_i x_{i'} j} \log_2 \frac{N_{x_i x_{i'} j} N_j}{N_{x_i j} N_{x_{i'} j}}. \end{aligned} \tag{21}$$

The response pattern $\mathbf{X}^{\neg ii'}$ contains a lot of missing data. If we ignore the missing data, then Eq (21) can be transformed into

$$\begin{aligned} & I(X_i; X_{i'}|\mathbf{X}^{\neg ii'}, \xi, B_s^c) \\ &= \frac{1}{N} \sum_{j'=0}^{J'-1} \sum_{x_i=0}^{1} \sum_{x_{i'}=0}^{1} N_{x_i x_{i'} j'} \log_2 \frac{N_{x_i x_{i'} j'} N_{j'}}{N_{x_i j} N_{x_{i'} j'}}, \end{aligned} \tag{22}$$

where J' is the number of observed patterns. Therefore, even though the number of items m is large, the actual amount of computation is $O(N)$. According to Eqs. (3) and (4), the amount of computation for traditional CI tests is also $O(N)$.

As mentioned in Proof 1, the main idea of this paper is to obtain the conditional probability given a latent variable, with replacing θ by $\mathbf{X}^{\neg ii'}$ using Neyman's theorem. This reduces the computational costs to $O(N)$ from $O(2^m)$. That is, the amount of computation for our method is the same as that for traditional methods.

Using $I(X_i; X_{i'}|\mathbf{X}^{\neg ii'}, \xi, B_s^c)$ in Eq. (22), we can define the following latent conditional independence (LCI) test given a latent variable.

Definition 2: (Latent conditional independence (LCI) test)
If $I(X_i; X_{i'}|\mathbf{X}^{\neg ii'}, \xi, B_s^c) \geq \varepsilon$
→ the i-th and i'-th items are conditionally dependent when a latent variable is given
else
→ the i-th and i'-th items are conditionally independent when a latent variable is given,
where ε is a certain threshold.

■

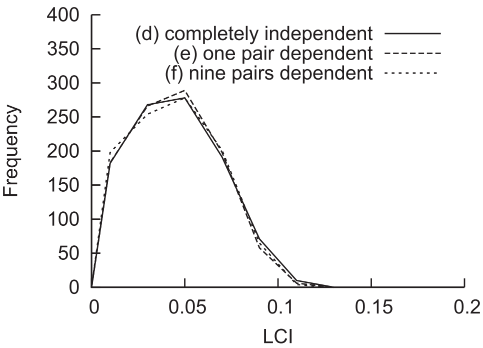

Fig. 16 Frequency distributions of LCI test statistic of locally independent pairs (number of items: 20).

The LCI test can correctly detect CI given a latent variable even when items other than the target pairs are mutually dependent. When the LCI test was applied to the same simulation data as used in Sect. 4, differences between the distributions were reduced (Figs. 15 and 16). It means that the LCI test was not affected by other item dependencies.

6. Evaluation of LCI Test

This section evaluates how correctly the LCI test can detect CI between two items given a latent variable.

6.1 Method

When a data set contains locally dependent items and a CI test is applied to such a data set, item pairs are classified into one of four categories:

- dependent pairs incorrectly judged as independent (a)
- dependent pairs correctly judged as dependent (b)
- independent pairs correctly judged as independent (c)
- independent pairs incorrectly judged as dependent (d)

To evaluate the performances of detecting local dependencies, we obtained ratio $a/(a+b)$. Moreover, we also obtained ratio $c/(c+d)$ to evaluate performances of detecting local independencies.

In Sect. 4, we generated data that contained locally dependent pairs of items. We applied the proposed LCI test and two traditional CI tests to these data and calculated the two abovementioned ratios.

Furthermore, we generated data using the structure (Fig. 17) estimated from an actual test [14], which is called “case (g)”. The estimated item parameters are given in Appendix A. For details of item contents, see [14]. The number of examinees was 10,000, and 10 sets of data were generated.

For all the experiments, 0.01, 0.05, and 0.10 were used as the LCI test thresholds ε, and performances were compared.

6.2 Results

The results for cases (b) and (c) are given in Table 1. All CI tests detected more than 90 percent of the local dependencies, and G^2 detected the greatest number of local dependencies. However, traditional CI tests often failed to detect local independencies, whereas the LCI test detected local independencies with high accuracy. This means that the LCI test could avoid overfitting problems which the traditional CI tests suffered from in these cases.

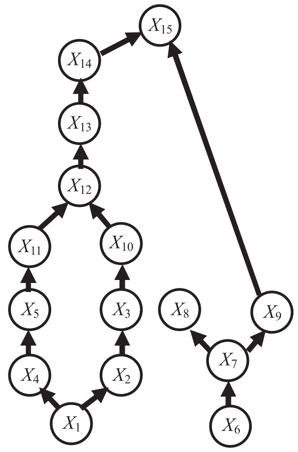

Fig. 17 Structure from actual test (Note: although the latent variable θ is implicit in this graph, all items depend on θ.).

Table 1 Average ratio of correctly detected dependencies and independencies (number of items: 7).

Case (b) (one pair dependent case)		
CI test	dependency	independency
LCI (ε = 0.01)	0.977	0.737
LCI (ε = 0.05)	0.959	0.976
LCI (ε = 0.10)	0.902	0.996
G^2	0.992	0.176
Q_3	0.922	0.214
Case (c) (two pairs dependent case)		
CI test	dependency	independency
LCI (ε = 0.01)	0.998	0.853
LCI (ε = 0.05)	0.963	0.988
LCI (ε = 0.10)	0.912	0.998
G^2	0.999	0.139
Q_3	0.936	0.246

The results for cases (e) and (f) are given in Tables 2. Although the overfitting problem of G^2 was mitigated in case (e), G^2 still suffered from this problem in case (f). When ε was 0.01 or 0.05, the LCI test also failed to detect local independencies. However, when ε was 0.10, the LCI test kept high accuracy for detecting both dependencies and independencies in cases (e) and (f).

The results for case (g) is given in Table 3. Whereas G^2 and Q_3 often failed to detect local independencies, the LCI test detected both dependencies and independencies accurately when ε was 0.01. However, when ε was 0.10, the LCI test failed to detect local dependencies. Therefore, for the LCI test, a method of investigating an appropriate ε must be explored.

Summarizing the results, we can say that traditional CI tests suffer from the overfitting. On the other hand, when

Table 2 Average ratio of correctly detected dependencies and independencies (number of items: 20).

Case (e) (one pair dependent case)		
CI test	dependency	independency
LCI (ε = 0.01)	0.995	0.164
LCI (ε = 0.05)	0.953	0.659
LCI (ε = 0.10)	0.845	0.995
G^2	1.000	0.779
Q_3	0.977	0.392
Case (f) (nine pairs dependent case)		
CI test	dependency	independency
LCI (ε = 0.01)	0.997	0.641
LCI (ε = 0.05)	0.966	0.981
LCI (ε = 0.10)	0.889	1.000
G^2	1.000	0.452
Q_3	0.981	0.298

Table 3 Average ratio of correctly detected dependencies and independencies (case (g)).

CI test	dependency	independency
LCI (ε = 0.01)	0.860	1.000
LCI (ε = 0.05)	0.530	1.000
LCI (ε = 0.10)	0.270	1.000
G^2	1.000	0.095
Q_3	0.947	0.344

an appropriate threshold is determined, the LCI test can correctly detect both local independence and local dependence. However, since the performance of the LCI test is highly sensitive to the choice of threshold ε, a suitable method of determining the threshold, for example, by bootstrapping or cross validation, must be used.

7. Application to Real Data

In this section, our method is applied to real data.

7.1 Method

A mathematics test answered by 367 freshmen at five national universities in Tokyo was analyzed. The test contained seven items. Items Q1, Q2, and Q3 were questions about inequalities of the second degree, and Q3 required the correct response to Q2. Items Q4, Q5, Q6, and Q7 were questions about logical expressions. See Appendix B for details.

7.2 Results

The values of the LCI test statistics are given in Table 4. When the threshold ε was 0.10, only Q2 and Q3 were locally dependent. In contrast, when ε was 0.01, 17 out of 21 pairs were locally dependent. In this section, we interpret the result when ε was 0.05. Pairs of items that were judged to be locally dependent are shown linked in Fig. 18.

Table 4 LCI test statistics for seven items of a real test.

items	Q1	Q2	Q3	Q4	Q5	Q6	Q7
Q1		0.012	0.012	0.005	0.017	0.000	0.019
Q2			0.159	0.009	0.022	0.004	0.031
Q3				0.029	0.032	0.018	0.048
Q4					0.066	0.020	0.053
Q5						0.010	0.052
Q6							0.056
Q7							

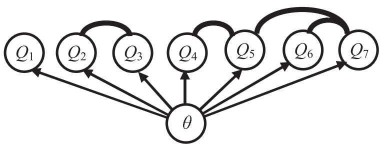

Fig. 18 Pairs whose LCI statistics were greater than 0.05.

Between the second-degree-inequality items (Q1, Q2, and Q3) and the logical-expression items (Q4, Q5, Q6, and Q7), almost all values of LCI statistics were smaller than ε. In contrast, within the logical-expression items, most of the values were larger than ε. These results indicate that items belonging to different areas were locally independent given a latent variable.

The largest LCI was between Q2 and Q3. Since Q3 explicitly required the correct response to Q2, this local dependence might reflect the item makers' intentions.

Within the logical-expression items (Q4, Q5, Q6, and Q7), four pairs were judged to be locally dependent. These four items share the same alternatives. Such sharing of alternatives might cause local dependence.

8. Conclusion

In this study, we investigated the latent conditional independence (LCI) test given a latent variable to detect conditionally independent items. The performances were compared with those of traditional conditional independence (CI) tests such as Q_3 and G^2. There were two main findings.

First, when some items that are not targets are conditionally dependent given a latent variable, traditional CI test statistics are seriously biased. On the other hand, the LCI test statistic is robust irrespective of other items. Secondly, when an appropriate threshold ε is chosen, the LCI test can detect both local independencies and local dependencies, whereas traditional CI tests often fail to detect local independencies. The application of the LCI test to actual data suggests that the sharing of alternatives might cause conditional dependence.

However, some problems remain unsolved. As described in this paper, we knew which item was the parent because items in the test data were sequentially arrayed. For cases in which directions are unknown, a method of determining the parent item is necessary. In addition, the performance of the LCI test is highly sensitive to the choice of ε. Therefore, methods of determining an appropriate value of ε, for example, by bootstrapping or, cross validation, should be used. When these problems have been solved, the LCI test should be more useful.

References

[1] F.M. Lord and M.R. Novick, Statistical theories of mental test scores, Addison-Wesley, MA, 1968.

[2] G. Rasch, "An item analysis which takes individual differences into account," British Journal of Mathematical and Statistical Psychology, vol.19, pp.49–57, 1966.

[3] A. Birnbaum, "Efficient design and use of tests of a mental ability for various decision-making problems," (Series Report 58-16, no.7755-23). USAF School of Aviation Medicine, Randolph Air Force Base, Texas, 1957.

[4] A. Birnbaum, "Some latent trait models," in Statistical Theories of Mental Test Scores, ed. F.M. Load and M.R. Novick, pp.397–424, Reading, Addison-Wesley, MA, 1968.

[5] F. Samejima, "Estimation of latent ability using a response pattern of graded scores," Psychometrika Monograph, no.17, 1969.

[6] F. Samejima, "A general model for free-response data," Psychometrika Monograph, no.18, 1972.

[7] G.N. Masters, "A Rasch model for partial credit scoring," Psychometrika, vol.35, pp.43–50, 1982.

[8] R.D. Bock, "Estimating item parameters and latent ability when responses are scored in two or more nominal categories," Psychometrika, vol.37, pp.29–51, 1972.

[9] W.M. Yen, "Effects of local item dependence on the fit and equating performance of the three-parameter logistic model," Applied Psychological Measurement, vol.8, pp.125–145, 1984.

[10] W.H. Chen and D. Thissen, "Local dependence indexes for item pairs using item response theory," Journal of Educational and Behavioral Statistics, vol.22, pp.265–289, 1997.

[11] L.M. Reese, "The impact of local dependencies on some LSAT outcomes," Law School Admission Council Statistical Report, vol.95-02, 1995.

[12] M. Sano, "Detecting overestimation of discrimination parameter applying mutual information," Japanese Journal for Research on Testing, vol.5, pp.3–21, 2009.

[13] S.G. Sireci, D. Thissen, and H. Wainer, "On the reliability of testlet-based tests," Journal of Educational Measurement, vol.28, pp.237–247, 1991.

[14] M. Ueno, "An extension of the IRT to a network model," Behaviormetrika, vol.29, pp.59–79, 2002.

[15] P. Spirtes, C. Glymour, and R. Scheines, Causation, prediction, and search, Springer Verlag, 1993.

[16] I. Tsamardinos, L. Brown, and C. Aliferis, "The max-min hill-climbing Bayesian network structure learning algorithm," Mach. Learn., vol.65, pp.31–78, 2006.

[17] S.S. Wilks, Mathematical Statistics, 2nd. ed., pp.355–356, Wiley, 1962.

Appendix A: Item Parameters of Case (g) in Sect. 6.

item	a	b
X_1	0.511	−3.243
$X_2 \vert X_1 = 1$	0.847	−2.154
$X_2 \vert X_1 = 0$		3.463
$X_3 \vert X_2 = 1$	0.905	−1.139
$X_3 \vert X_2 = 0$		3.124
$X_4 \vert X_1 = 1$	0.537	−1.322
$X_4 \vert X_1 = 0$		3.463
$X_5 \vert X_4 = 1$	0.826	−0.277
$X_5 \vert X_4 = 0$		2.679
X_6	1.281	−1.326
$X_7 \vert X_6 = 1$	1.176	−1.022
$X_7 \vert X_6 = 0$		3.269
$X_8 \vert X_7 = 1$	1.127	−1.010
$X_8 \vert X_7 = 0$		3.246
$X_9 \vert X_7 = 1$	0.981	0.697
$X_9 \vert X_7 = 0$		3.969
$X_{10} \vert X_3 = 1$	1.129	−0.440
$X_{10} \vert X_3 = 0$		2.217
$X_{11} \vert X_5 = 1$	1.313	0.212
$X_{11} \vert X_5 = 0$		2.647
$X_{12} \vert X_{10} = 1, X_{11} = 1$	1.534	0.330
$X_{12} \vert X_{10} = 0, X_{11} = 1$		3.366
$X_{12} \vert X_{10} = 1, X_{11} = 0$		2.216
$X_{12} \vert X_{10} = 0, X_{11} = 0$		3.366
$X_{13} \vert X_{12} = 1$	1.226	−0.130
$X_{13} \vert X_{12} = 0$		2.896
$X_{14} \vert X_{13} = 1$	1.372	−0.153
$X_{14} \vert X_{13} = 0$		3.004
$X_{15} \vert X_9 = 1, X_{14} = 1$	1.227	2.649
$X_{15} \vert X_9 = 0, X_{14} = 1$		3.370
$X_{15} \vert X_9 = 1, X_{14} = 0$		3.384
$X_{15} \vert X_9 = 0, X_{14} = 0$		3.844

Appendix B: Test Used in Section 7

Please answer Q1 through Q7.

[1] In a rectangle ABCD, AB = CD = 8 and BC = DA = 12. For a point P on side AB, a point Q on side BC, and a point R on side CD, the following relation holds.

AP = BQ = CR

Let AP = x ($0 < x < 8$).

Q1. The area of the trapezoid PBCR is $\boxed{\;?\;}$.

Q2. The area of △PQR is

$$S = x^2 - \boxed{\;?\;}x + \boxed{\;?\;}.$$

Q3. If $S < 24$ holds, then x must be in the range of

$$\boxed{\;?\;} < x < \boxed{\;?\;}.$$

[2] Fill in boxes $\boxed{\text{A}}$ through $\boxed{\text{D}}$ selecting for each box one option from ⓪ through ③ below. You may select the same options as many times as you wish.

m and *n* are natural numbers. There are three conditions: *p*, *q*, and *r*.

p: $m + n$ is divisible by 2.
q: *n* is divisible by 4.
r: *m* is divisible by 2, and *n* is divisible by 4.

Let the negation of condition *p* be $\bar{p}$ and let the negation of condition *r* be $\bar{r}$. Then,

Q4. *p* is $\boxed{\text{A}}$ for *r*.

Q5. $\bar{p}$ is $\boxed{\text{B}}$ for $\bar{r}$.

Q6. "*p* and *q*" is $\boxed{\text{C}}$ for *r*.

Q7. "*p* or *q*" is $\boxed{\text{D}}$ for *r*.

⓪ a necessary and sufficient condition
① a necessary condition but not a sufficient condition
② a sufficient condition but not a necessary condition
③ neither a necessary condition nor a sufficient condition

Takamitsu Hashimoto received an M.A. degree from the University of Tokyo in 2002. He was a research associate at the National Center for University Entrance Examinations in 2004 and has been an assistant professor since 2007. Since 2009, he has been a student at the University of Electro-Communications. His research interests include Bayesian statistics and test theory.

Maomi Ueno received an M.Ed. degree from Kobe University in 1992 and a Ph.D. degree from Tokyo Institute of Technology in 1994. He was a research associate at Tokyo Institute of Technology in 1994 and an associate professor at Nagaoka University of Technology in 2000. Since 2006, he has been an associate professor at the University of Electro-Communications. His research interests include e-Learning, Bayesian statistics, and data mining. He received a prize from the Behaviormetric Society of Japan in 2004, an outstanding paper award from e-Learn 2004, a prize from the Japanese Society for Information and Systems in Education in 2005, and an outstanding paper award from e-Learn 2007. He is an executive board member of the Japanese Society for Educational Technology, an executive of the Behaviormetric Society of Japan, and a member of the council of Japanese Society for Information and Systems in Education. He was an IEEE Computer society ICALT2007 Chair.