Estimation for Neural Test Models with Missing Data

SHOJIMA Kojiro

October 2007

Department of Test Analysis and Evaluation, Research Division, The National Center for University Entrance Examinations

Estimation for neural test models with missing data

Kojiro Shojima

Abstract

We propose two estimation procedures for dealing with missing data in neural test theory (NTT). One is for use with the square of the Euclidian distance, and the other is for use with the maximum likelihood (ML) method. The second procedure should be useful when some data is missing at random (MAR) because it is based on the full-information maximum likelihood method. In addition, the weighted and unweighted observation ratio profiles of each item can be used to analyze the behavior of the item selection ratio through latent ranks. We show an example of analyzing an earth science test.

Key words: neural test theory, grade neural test model, maximum likelihood estimation, observation ratio profile, missing mechanism, missing at random.

欠測データがあるときのニューラルテストモデルの推定

荘島宏二郎

要約

本研究では，ニューラルテスト理論 (neural test theory, NTT) において，欠測処理のための 2 つの方法を論じた．1 つは，ユークリッド距離を用いたときの欠測処理方法であり，もう 1 つは，最尤推定法を用いたときの欠測処理方法である．最尤推定法のときの欠測処理方法は，完全情報最尤法を参考にしているため，欠測構造が MAR のときに有効であることが期待される．また，各潜在ランクにおける欠測率を見るのに，非重み付き観測率プロファイルと重み付き観測率プロファイルが有効である．最後に，地学テストの分析例を示した．

キーワード: ニューラルテスト理論，段階ニューラルテストモデル，最尤推定法，観測率プロファイル，欠測構造，ランダム欠測．

Department of Test Analysis and Evaluation, Research Division, The National Center for University Entrance Examinations

1 Introduction

Neural test theory (NTT; Shojima, 2007a, 2007b) is a statistical model, and it is useful for analyzing test data. The latent scale assumed in NTT is rank-ordered, and examinees are graded on the scale. The NTT model was first developed as a remodeled one-dimensional self-organizing map (SOM; Kohonen, 1995). That is, it was mathematical rather than statistical. Shojima (2007c) implemented the maximum likelihood method for estimating the latent rank and selecting the winner node. Shojima (2007d) later proposed a statistical testing method for examining the goodness-of-fit of NTT models to data by using the χ^2 statistic. Therefore, the statistical features of the NTT model have been discussed in terms of these methods.

Missing responses are frequently observed in test data. Although missing data are often regarded as false answers in achievement tests, it is very important to be able to destinguish missing data from false responses, for example, when the selected items differ by subgroup of examinees. In such a case, the correct answer ratios are unduly low if missing data are coded as 0s. In this study, we discuss methods for dealing with missing data in neural test models.

2 Method

In NTT, a treatment of missing data is relevant to latent rank estimation, winner node selection, and reference vector updating. Hereafter, the treatments for the dichotomous neural test (DNT; Shojima, 2007a) and graded neural test (GNT; Shojima, 2007b) models are described separately. The GNT model is an NTT model for ordered polytomous data, which reduces to the DNT model when all items are binary. Although it is difficult to give a simultaneous explanation for the two models, the missing data treatments for these models are basically identical.

2.1 Dichotomous neural test model

Assume that the sample size is N, the number of items is n, and the number of latent ranks is Q. The reference matrix is

$$\boldsymbol{V} = \{v_{qj}\} = \begin{bmatrix} v_{11} & \cdots & v_{1n} \\ \vdots & \ddots & \vdots \\ v_{Q1} & \cdots & v_{Qn} \end{bmatrix} \quad (Q \times n), \tag{1}$$

where the j-th column vector in $\boldsymbol{V}$, $\boldsymbol{v}_j$ $(j = 1, \cdots, n)$, is the item reference profile (IRP) of item j, and the q-th row vector in $\boldsymbol{V}$, $\boldsymbol{v}_q$ $(q = 1, \cdots, Q)$, is the rank reference vector (RRV) of latent rank R_q.

Let us assume that the response data of the examinees is $\boldsymbol{U} = \{u_{ij}\}$ $(N \times n)$, where u_{ij} is the response data of examinee i for item j, and it is a dichotomous variable which is coded 1 if the response is correct, and 0 otherwise. Assume further that $\boldsymbol{Z} = \{z_{ij}\}$ is the missing indicator matrix, where z_{ij} is coded 1 when the response of examinee i for item j is observed, and 0 when the response is missing.

The outline of the statistical learning procedure for the DNT model is as follows:

For (t=1; $t \leq T$; $t = t + 1$) (2)

— $\boldsymbol{U}^{(t)} \Leftarrow$Randomly sort the row vectors of $\boldsymbol{U}$. (3)

For (h=1; $h \leq N$; $h = h + 1$) (4)

— Obtain $\boldsymbol{z}_h^{(t)}$ from $\boldsymbol{u}_h^{(t)}$. (5)

— Select the winner for $\boldsymbol{u}_h^{(t)}$ by d. (6)

— Obtain $\boldsymbol{V}^{(t,h)}$ by updating $\boldsymbol{V}^{(t,h-1)}$. (7)

— $\boldsymbol{V}^{(t+1,0)} \Leftarrow \boldsymbol{V}^{(t,N)}$ (8)

Lines (30)-(7) make up a routine which is repeatedly applied when the value of h is from 1 to N, and the counter t is then incremented by one. This process continues until t approaches T.

In Line (6), there are two methods of selecting the winner node for the h-th row vector of $\boldsymbol{U}^{(t)}$, $\boldsymbol{u}_h^{(t)} = \{u_{hj}^{(t)}\}$ $(n \times 1)$: the square of the Euclidian distance method (ED^2; Shojima, 2007a) and the maximum likelihood method (ML; Shojima, 2007c). With missing data, these methods are reformulated as

$$R_w^{(ED^2)} : \ w = \arg\min_{q \in Q} \sum_{j=1}^{n} z_{hj}^{(t)} (u_{hj}^{(t)} - v_{qj}^{(t,h-1)})^2, \tag{9}$$

and

$$R_w^{(ML)} : \ w = \arg\max_{q \in Q} \sum_{j=1}^{n} z_{hj}^{(t)} \{u_{hj}^{(t)} \ln v_{qj}^{(t,h-1)} + (1 - u_{hj}^{(t)}) \ln(1 - v_{qj}^{(t,h-1)})\}, \tag{10}$$

where $z_{hj}^{(t)}$ is a dichotomous variable which is coded 1 when $u_{hj}^{(t)}$ is observed, and 0 when it is missing. Also, $\boldsymbol{v}_q^{(t,h)} = \{v_{qj}^{(t,h)}\}$ $(n \times 1)$ is the RRV of latent rank R_q in $\boldsymbol{V}^{(t,h)}$. These methods are useful for estimating the latent ranks of the examinees after obtaining the estimate of

the reference matrix $\hat{\boldsymbol{V}}$. Let r_i denote the latent rank of examinee i; it is estimated as

$$R_{r_i}^{(ED^2)}: \ r_i = \arg\min_{q \in Q} \sum_{j=1}^{n} z_{ij}(u_{ij} - \hat{v}_{qj})^2 \tag{11}$$

and

$$R_{r_i}^{(ML)}: \ r_i = \arg\max_{q \in Q} \sum_{j=1}^{n} z_{ij}\{u_{ij} \ln \hat{v}_{qj} + (1 - u_{ij}) \ln(1 - \hat{v}_{qj})\} \tag{12}$$

for the ED^2 and the ML methods, respectively. The rank membership profiles of the examinees and the rank membership distribution (Shojima, 2007c) can be calculated from the above equations.

Next, in (7), the method for updating the reference vectors become

$$\begin{aligned} &\text{For } (q{=}1;\ q \leq Q;\ q = q + 1) \\ &\text{— } \boldsymbol{v}_q^{(t,h)} = \boldsymbol{v}_q^{(t,h-1)} + h_{qw}(t)\{\boldsymbol{z}_h^{(t)} \odot (\boldsymbol{u}_h^{(t)} - \boldsymbol{v}_q^{(t,h-1)})\}, \end{aligned} \tag{13}$$

where $\odot$ is the Hadamard product. In addition, the factor h_{qw} in the above equation is given by

$$h_{qw}(t|\alpha_t, \sigma_t^2) = \alpha_t \exp\Big\{-\frac{(R_q - R_w)^2}{2\sigma_t^2}\Big\}, \tag{14}$$

where

$$\alpha_t = \frac{T - t + 1}{T}\alpha_1, \tag{15}$$

and

$$\sigma_t = \frac{(T - t)\sigma_1 + (t - 1)\sigma_0}{T - 1}. \tag{16}$$

Equation (13) means that the elements in the reference vector corresponding to the missing data are not updated. The updating method can be used with both ED^2 and ML selection methods.

2.2 Graded neural test model

The graded neural test (GNT; Shojima, 2007b) is a polytomous NTT model for analyzing polytomously ordered data. Let us assume that the sample size is N, the number of items is n, the number of latent ranks is Q, and that the number of categories of item j is C_j

$(0, 1, \cdots, C_j - 1)$. Let us also suppose that the examinees with higher abilities select the higher categories. Then, the reference matrix is

$$\boldsymbol{V} = \{v_{qjk}\} = \begin{bmatrix} v_{1,1,1} & \cdots & v_{1,1,C_1-1} & v_{1,2,1} & \cdots & v_{1,n,C_n-1} \\ v_{2,1,1} & \cdots & v_{2,1,C_1-1} & v_{2,2,1} & \cdots & v_{2,n,C_n-1} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ v_{Q,1,1} & \cdots & v_{Q,1,C_1-1} & v_{Q,2,1} & \cdots & v_{Q,n,C_n-1} \end{bmatrix} \quad \Big(Q \times \Big(\sum_{j=1}^{n} C_j - 1\Big)\Big), \quad (17)$$

where each column vector in $\boldsymbol{V}$, $\boldsymbol{v}_{jk} = \{v_{qjk}\}$ $(j = 1, \cdots, n; k = 1, \cdots, C_j - 1)$, is the boundary category reference profile (BCRP; Shojima, 2007b) of category k in item j, and v_{qjk} is the selection ratio of the examinees in latent rank R_q for item j's category k or higher. The expanded reference matrix is obtained from the reference matrix as

$$\boldsymbol{P} = \{p_{qjk}\} = \begin{bmatrix} p_{1,1,0} & \cdots & p_{1,1,C_1-1} & p_{1,2,0} & \cdots & p_{1,n,C_n-1} \\ p_{2,1,0} & \cdots & p_{2,1,C_1-1} & p_{2,2,0} & \cdots & p_{2,n,C_n-1} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ p_{Q,1,0} & \cdots & p_{Q,1,C_1-1} & p_{Q,2,0} & \cdots & p_{Q,n,C_n-1} \end{bmatrix} \quad \Big(Q \times \sum_{j=1}^{n} C_j\Big), \quad (18)$$

where each column vector of $\boldsymbol{P}$, $\boldsymbol{p}_{jk} = \{p_{qjk}\}$ $(j = 1, \cdots, n; k = 0, \cdots, C_j - 1)$, is the item category reference profile (ICRP; Shojima, 2007b), and each p_{qjk} is computed as

$$p_{qjk} = v_{qjk} - v_{qjk+1} \quad (k = 0, 1, \cdots, C_j - 1), \quad (19)$$

provided that

$$v_{qj0} = 1, \quad (20)$$

and

$$v_{qjC_j} = 0. \quad (21)$$

Next, let $\boldsymbol{X} = \{x_{ij} | x_{ij} \in \{0, 1, \cdots, C_j\}\}$ denote the response data of the examinees, where x_{ij} is the response of examinee i to item j. The variables $\boldsymbol{Z}$, $\boldsymbol{U}$ and $\boldsymbol{Y}$ required in the statistical learning process are generated from $\boldsymbol{X}$ as follows:

$$\boldsymbol{z}_i = \{z_{ij}\} \quad (n \times 1), \quad (22)$$

$$z_{ij} = \begin{cases} 1 & \text{if } x_{ij} \text{ is observed} \\ 0 & \text{otherwise} \end{cases}, \quad (23)$$

$$\boldsymbol{u}_i = \{u_{ijk}\} = [u_{i,1,1}\ u_{i,1,2} \cdots u_{i,1,C_1-1}\ u_{i,2,1} \cdots u_{i,n,C_n-1}]' \quad \Big\{\sum_{j=1}^{n}(C_j - 1) \times 1\Big\} \quad (24)$$

$$u_{ijk} = \begin{cases} 1 & \text{if } x_{ij} \geq k \\ 0 & \text{otherwise} \end{cases} \quad (k = 1, \cdots, C_j - 1), \quad (25)$$

$$\boldsymbol{y}_i = \{y_{ijk}\} = [y_{i,1,0}\ y_{i,1,1} \cdots y_{i,1,C_1-1}\ y_{i,2,0} \cdots y_{i,n,C_n-1}]' \quad \Big(\sum_{j=1}^{n} C_j \times 1\Big) \tag{26}$$

$$y_{ijk} = \begin{cases} 1 & \text{if } x_{ij} = k \\ 0 & \text{otherwise} \end{cases} \qquad (k = 0, \cdots, C_j - 1), \tag{27}$$

where z_{ij} is the missing indicator as in the previous section, u_{ijk} is a dichotomous variable which is coded 1 when the response of examinee i for item j is k or higher, and 0 otherwise, and y_{ijk} is also a dichotomous variable coded 1 if the response is k, and 0 otherwise. The outline of the statistical learning for the GNT model is as follows:

For (t=1; $t \leq T$; $t = t + 1$) (28)

— $\boldsymbol{X}^{(t)} \Leftarrow$Randomly sort the row vectors of $\boldsymbol{X}$. (29)

For (h=1; $h \leq N$; $h = h + 1$) (30)

— Obtain $\boldsymbol{z}_h^{(t)}$, $\boldsymbol{u}_h^{(t)}$, and $\boldsymbol{y}_h^{(t)}$ from $\boldsymbol{x}_h^{(t)}$. (31)

— Select the winner node. (32)

— Obtain $\boldsymbol{V}^{(t,h)}$ by updating $\boldsymbol{V}^{(t,h-1)}$. (33)

— Obtain $\boldsymbol{P}^{(t,h)}$ from $\boldsymbol{V}^{(t,h)}$. (34)

— $\boldsymbol{V}^{(t+1,0)} \Leftarrow \boldsymbol{V}^{(t,N)}$. (35)

— $\boldsymbol{P}^{(t+1,0)} \Leftarrow \boldsymbol{P}^{(t,N)}$. (36)

In (32), the winner node can be selected by using the ED^2 method or the ML method. That is,

$$R_w^{(ED^2)} : \ w = \arg\min_{q \in Q} \sum_{j=1}^{n} z_{hj}^{(t)} \frac{\sum_{k=1}^{C_j-1} (u_{hjk}^{(t)} - v_{qjk}^{(t,h-1)})^2}{C_j - 1}, \tag{37}$$

and

$$R_w^{(ML)} : \ w = \arg\max_{q \in Q} \sum_{j=1}^{n} z_{hj}^{(t)} \sum_{k=1}^{C_j-1} y_{hjk}^{(t)} \ln v_{qjk}^{(t,h-1)}. \tag{38}$$

These methods are also useful for estimating the latent ranks of the examinees after obtaining the estimate of the reference matrix $\hat{\boldsymbol{V}}$ and the expanded reference matrix $\hat{\boldsymbol{P}}$. Assume that the latent rank of examinee i is r_i; r_i is estimated as

$$R_{r_i}^{(ED^2)} : \ r_i = \arg\min_{q \in Q} \sum_{j=1}^{n} z_{ij} \frac{\sum_{k=1}^{C_j-1} (u_{ijk} - \hat{v}_{qjk})^2}{C_j - 1} \tag{39}$$

and

$$R_{r_i}^{(ML)}: \ r_i = \arg\max_{q \in Q} \sum_{j=1}^{n} z_{ij} \sum_{k=1}^{C_j - 1} y_{ijk} \ln \hat{v}_{qjk} \tag{40}$$

under the ED^2 and the ML methods, respectively.

Next, in (33), the rank reference vectors (RRVs), $\boldsymbol{v}$s (not $\boldsymbol{p}$s), should be updated to numerically approach the input data, $\boldsymbol{u}$s (not $\boldsymbol{y}$s), as the supervising signals of the RRVs. That is,

$$\begin{aligned} &\text{For } (q{=}1;\ q \leq Q;\ q = q+1) \\ &\text{— } \boldsymbol{v}_q^{(t,h)} = \boldsymbol{v}_q^{(t,h-1)} + h_{qw}(t)\{\boldsymbol{g}_h^{(t)} \odot (\boldsymbol{u}_h^{(t)} - \boldsymbol{v}_q^{(t,h-1)})\}, \end{aligned} \tag{41}$$

where the factor $h_{qw}(t)$ in the above equation is identical to that in (14), and

$$\boldsymbol{g}_h^{(t)} = [z_{h1}^{(t)} \mathbf{1}'_{C_1 - 1} \cdots z_{hn}^{(t)} \mathbf{1}'_{C_n - 1}]' \quad \Big\{ \sum_{j=1}^{n} (C_j - 1) \times 1 \Big\}. \tag{42}$$

2.3 Observation ratio profile

The observation ratio profile (ORP) is useful for examining the behavior of the selection-missing response ratio of each item. Two kinds of ORPs are possible: weighted and unweighted ORPs. That is,

$$\boldsymbol{z}_{Wj} = \{z_{Wqj}\} \quad (Q \times 1), \tag{43}$$

$$z_{Wqj} = \frac{\sum_{i=1}^{N} z_{ij} p_{iq}}{\sum_{i=1}^{N} p_{iq}}, \tag{44}$$

and

$$\boldsymbol{z}_{Uj} = \{z_{Uqj}\} \quad (Q \times 1), \tag{45}$$

$$z_{Uqj} = \frac{\sum_{i=1}^{N} z_{ij} f_{iq}}{\sum_{i=1}^{N} f_{iq}}, \tag{46}$$

where p_{iq} is the rank membership profile (RMP; Shojima, 2007c) of examinee i to rank R_q, and f_{iq} is a dichotomous variable that is coded 1 when examinee i belongs to rank R_q, and coded 0 otherwise.

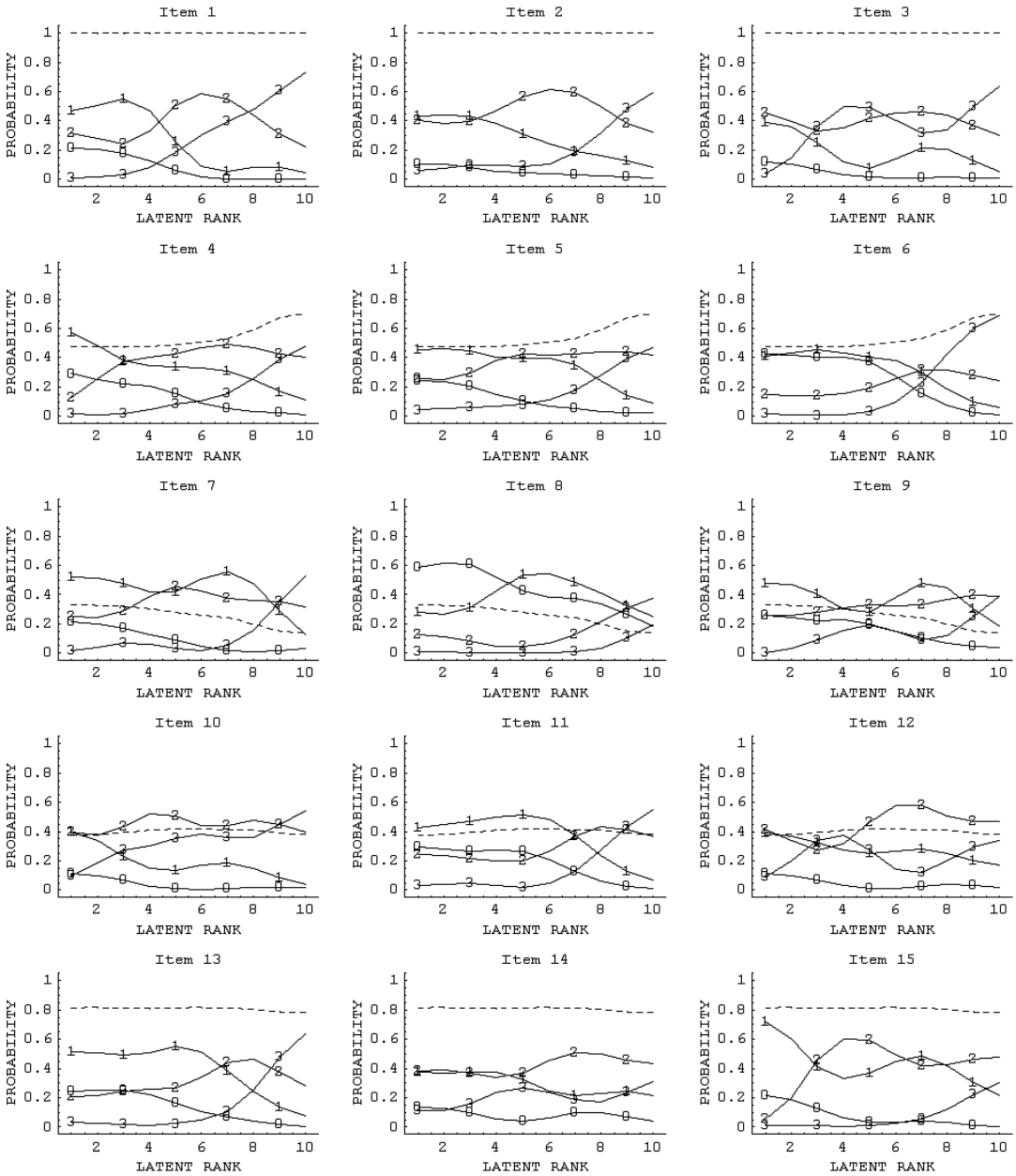

Figure 1: ICRPs and Weighted Observation Ratio Profiles

3 Analysis

An earth science test was analyzed using the proposed method. The sample size of the test was 3,810, the number of items was 15, and the number of categories of each item was $4(= 0, 1, 2, 3)$. Therefore, the test data was analyzed under the GNT model. Each examinee was required to answer Items 1-3, and selected two item sets from the four sets: Items 4-

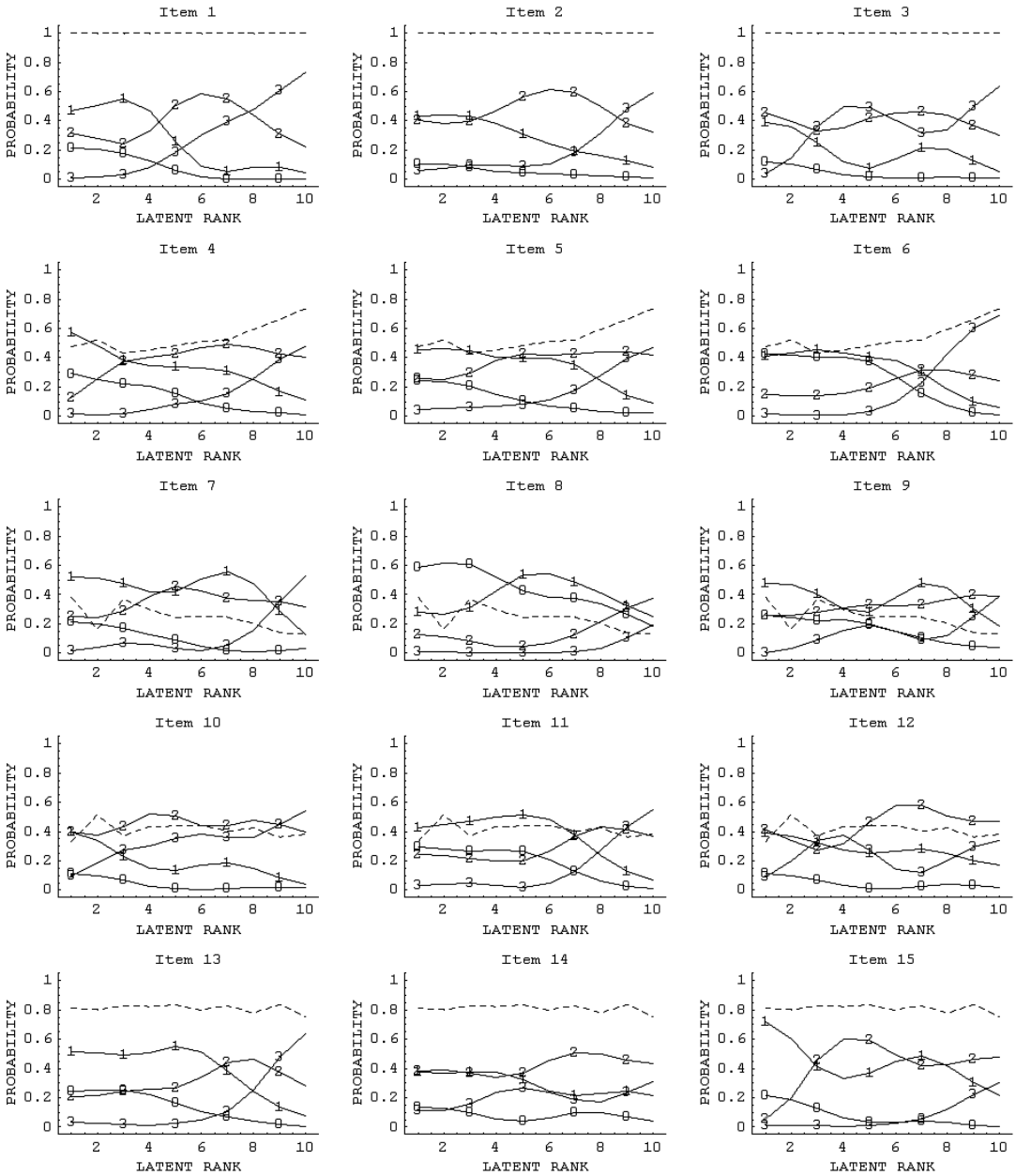

Figure 2: ICRPs and Unweighted Observation Ratio Profiles

6, 7-9, 10-12, and 13-15. Accordingly, the responses of nine items were observed for each examinee, and those for the other six items were missing.

In addition, the ML method was used for the latent rank estimation and the winner node selection. The parameters Q, T, α_1, σ_1, and σ_0 were set to 10, 500, 0.1, 10, and 1.0, respectively.

Figure 1 shows the item category reference profiles (ICRPs) of the 15 items, and the dashed line in each panel stands for the weighted observation ratio profile (WORP) of the item. Figure 2 also plots the ICRPs and the unweighted ORPs (UORPs) with solid and dashed lines, respectively.

The WORP and the UORP are useful for monitoring the transition of the response ratio for each item through latent ranks. In the test, the WORPs and the UORPs were obtained to be identical within each item set. In addition, the figures indicate that Items 4-6 were inclined to be selected by the examinees with higher abilities, whereas Items 7-9 were preferred by the examinees with lower abilities. Furthermore, the shapes of the WORPs were smoother than those of the UORPs. The WORPs can be said to show the features of the population, while the UORPs show those of the sample.

4 Discussion

This study described two methods of dealing with missing data in NTT: one for use with the square of the Euclidian distance and the other for use with the maximum likelihood (ML) method. In addition, it showed that the weighted and the unweighted observation ratio profiles are effective means for analysts and test administrators to examine the transition of the item response ratio through latent ranks.

In fact, the ML method described in this study is based on the concept of the full-information maximum liklihood method (FIML; e.g., Finkbeiner, 1979; Muthén, Kaplan & Hollis, 1987; Bock, Gibbons & Muraki, 1994; Muraki & Carlson, 1995; Arbuckle, 1996). Accordingly, it may have features derived from the FIML method. For example, it might be useful when the missing data mechanism is MAR (missing at random; Rubin, 1976; Little & Rubin, 1987). Further researches or simulation studies are required to clarify the characteristics of the proposed method.

In the analysis of test data, we should not confuse missing data with false answers if we are to accurately estimate statistical features such as the item difficulty. The response-missing ratio is also an important statistical feature of individual items because the ratio sometimes shows the preference of the examinees for selective items.

References

Arbuckle, J. L. (1996) Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques*. NJ: Lawrence Erlbaum Associates, Inc. (pp.243-277).

Bock, R. D., Gibbons, R., & Muraki, E. (1994) Full-information item factor analysis. *Applied Psychological Measurement*, **12**, 261-280.

Finkbeiner, C. (1979) Estimation for the multiple factor model when data are missing. *Psychometrika*, **44**, 409-420.

Kohonen, T. (1995) *Self-organizing maps*. Springer.

Little, R. J. A. & Rubin, D. B. (1987) *Statistical analysis with missing data*. John Wiley & Sons.

Lord, F. M. (1980) *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates.

Muraki, E., & Carlson, J. E. (1995) Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement*, **19**, 73-90.

Muthén, B., Kaplan, D. & Hollis, M. (1987) On structural equation modeling with data that are not missing completeley at random. *Psychometrika*, **52**, 431-462.

Rubin, D. B. (1976) Inference and missing data. *Biometrika*, **63**, 581-592.

Shojima, K. (2007a) Neural test theory. *DNC Research Note*, 07-02.

Shojima, K. (2007b) The graded neural test model: A neural test model for ordered polytomous data. *DNC Research Note*, 07-03.

Shojima, K. (2007c) Maximum likelihood estimation of latent rank under the neural test model. *DNC Research Note*, 07-04.

Shojima, K. (2007d) Chi-square goodness-of-fit test under the neural test model. *DNC Research Note*, 07-05.