# Latent Rank Theory:

## Estimation of Item Reference Profile by Marginal Maximum Likelihood Method with EM Algorithm

SHOJIMA Kojiro

October 2007

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

# Latent rank theory:
## Estimation of item reference profile by marginal maximum likelihood method with EM algorithm

Kojiro Shojima

### Abstract

A method of estimating the item reference profiles (IRPs) of the latent rank model by the marginal maximum likelihood method with the EM (expectation maximization) algorithm (MML-EM) is presented. The latent rank theory is a generic theory, which includes the neural test theory, in which the latent scale assumed in the model is not continuous but rank-ordered. Three methods for estimating IRPs are discussed: simple MML-EM, MML-EM with a monotonically increasing constraint, and MML-EM with moving-averaged rank membership profiles.

Key words: latent rank theory, neural test theory, item reference profile, marginal maximum likelihood method, EM algorithm.

# 潜在ランク理論:
## EM アルゴリズムを用いた周辺最尤法による項目参照プロファイルの推定

荘島宏二郎

### 要約

本研究では，潜在ランクモデルの項目参照プロファイルを EM アルゴリズムによる周辺最尤法を用いて推定する方法を論じた．潜在ランク理論は，ニューラルテスト理論を含む，潜在尺度が順序尺度が仮定されている理論の総称である．単純な MML-EM による推定，単純増加制約のもとでの MML-EM 推定，ランクメンバーシッププロファイルを移動平均化したときの MML-EM 推定という 3 つの方法を提案した．

キーワード: 潜在ランク理論，ニューラルテスト理論，項目参照プロファイル，周辺最尤法，EM アルゴリズム．

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 theory based on the mechanism of the self-organizing map (SOM; Kohonen, 1995). It is a test theory in which the assumed latent scale is rank-ordered. In NTT, each latent rank has its own rank reference vector (RRV), and the item reference profile (IRP) that expresses the transition of the correct answer rate for each item through latent ranks is estimated by RRVs repeatedly learning the state of the input data and numerically approaching it.

However, estimation by the SOM algorithm is only an option, and the IRPs can be obtained without it. In this study, a method of estimating IRPs by marginal maximum likelihood method with the EM (expectation maximization) algorithm (MML-EM) was investigated. The EM algorithm includes the process of updating the IRPs, so the algorithm can be said to be statistical learning. However, the model estimated by the method might not be a "neural" test model because nodes (latent ranks) do not mutually monitor the states of their neighboring nodes in the process under the simple application of the EM algorithm. This method has the potential to be developed into a slightly different theory from NTT. Therefore, it is called the latent rank theory (LRT) as a generic theory that includes NTT and the method investigated in this study. Accordingly, NTT is positioned as LRT with the SOM algorithm (LRT-SOM).

# 2 Method

Let us assume that the number of items is $n$ and that the number of latent ranks is $Q$. Then, the reference matrix is $\boldsymbol{V} = \{v_{qj}\}$ $(Q \times n)$, where the $q$-th row vector $\boldsymbol{v}_q$ $(q = 1, \cdots, Q)$ is the RRV of latent rank $R_q$, and the $j$-th column vector $\boldsymbol{v}_j$ $(j = 1, \cdots, n)$ is the IRP of item $j$. Let us also assume that the sample size is $N$ and that the response matrix of examinees is $\boldsymbol{U} = \{u_{ij}\}$ $(N \times n)$, where $u_{ij}$ is a dichotomous variable that is coded 1 if the response of examinee $i$ to item $j$ is correct and 0 otherwise. In addition, let $\boldsymbol{Z} = \{z_{ij}\}$ $(N \times n)$ be the missing indicator matrix (Shojima, 2007e), where $z_{ij}$ is also a dichotomous variable that is coded 1 when the response of examinee $i$ to item $j$ is observed and 0 when the response is missing. Furthermore, $\boldsymbol{F} = \{f_{iq}\}$ $(N \times Q)$ is the membership indicator matrix, where $f_{iq}$ is a dichotomous variable coded 1 if examinee $i$ is located in latent rank $R_q$ and 0 otherwise. Then, the probability that the response matrix $\boldsymbol{U}$ is observed under the assumption of local independence is

$$p(\boldsymbol{U}|\boldsymbol{F},\boldsymbol{V})=\prod_{i=1}^{N}\prod_{q=1}^{Q}\prod_{j=1}^{n}\{v_{qj}^{u_{ij}}(1-v_{qj})^{1-u_{ij}}\}^{z_{ij}\times f_{iq}}, \tag{1}$$

This equation is the likelihood of $\boldsymbol{U}$ with unknown parameters $\boldsymbol{V}$ and $\boldsymbol{F}$. Because the number of elements in $\boldsymbol{V}$ is invariant as the sample size increases, while that in $\boldsymbol{F}$ is under the influence of the sample size, $\boldsymbol{V}$ and $\boldsymbol{F}$ are called the structural and nuisance parameters, respectively.

In multivariate analysis, such as item response theory (Lord, 1980) and latent class analysis (Everitt & Hand, 1981; Titterington, Smith, & Makov, 1985), it is known that estimates of the structural parameters are biased when the nuisance parameters are simultaneously estimated in the process. Therefore, it is logical to avoid estimating the membership indicator $\boldsymbol{F}$ when the reference matrix $\boldsymbol{V}$ is estimated. In this case, the EM algorithm (Dempster, Laird, & Rubin, 1977) is useful because the nuisance parameters are integrated over the log-likelihood, so the expected log-likelihood is optimized.

By the Bayes theorem, the expected log-likelihood in which the nuisance parameters are integrated out is decomposed as follows:

$$\begin{aligned}\ln p(\boldsymbol{V}|\boldsymbol{U}) =& E_{\boldsymbol{F}|\boldsymbol{U},\boldsymbol{V}^{(t)}}[\ln p(\boldsymbol{V}|\boldsymbol{U},\boldsymbol{F})]\\ =& E_{\boldsymbol{F}|\boldsymbol{U},\boldsymbol{V}^{(t)}}[\ln p(\boldsymbol{U}|\boldsymbol{F},\boldsymbol{V})]+E_{\boldsymbol{F}|\boldsymbol{U},\boldsymbol{V}^{(t)}}[\ln p(\boldsymbol{V}|\boldsymbol{F})]-E_{\boldsymbol{F}|\boldsymbol{U},\boldsymbol{V}^{(t)}}[\ln p(\boldsymbol{U}|\boldsymbol{F})],\end{aligned} \tag{2}$$

where $p$ is used to denote both the probability and the probability density. Also, $\boldsymbol{V}^{(t)}$ is the estimate of $\boldsymbol{V}$ obtained in the $t$-th EM cycle. The first term in (2) is

$$E_{\boldsymbol{F}|\boldsymbol{U},\boldsymbol{V}^{(t)}}[\ln p(\boldsymbol{U}|\boldsymbol{F},\boldsymbol{V})]=\sum_{i=1}^{N}\sum_{q=1}^{Q}f_{iq}^{(t)}\ln p(\boldsymbol{u}_i|\boldsymbol{v}_q), \tag{3}$$

where $\boldsymbol{f}_i^{(t)}=\{f_{iq}^{(t)}\}$ $(Q\times 1)$ is the posterior distribution of $\boldsymbol{f}_i$ given $\boldsymbol{V}^{(t)}$ and $\boldsymbol{u}_i$ in the $t$-th EM cycle. That is,

$$f_{iq}^{(t)}=p(f_{iq}|\boldsymbol{u}_i,\boldsymbol{v}_q^{(t)})=\frac{p(\boldsymbol{u}_i|\boldsymbol{v}_q^{(t)})p(f_{iq}|\boldsymbol{\phi})}{\sum_{q'=1}^{Q}p(\boldsymbol{u}_i|\boldsymbol{v}_{q'}^{(t)})p(f_{iq'}|\boldsymbol{\phi})}, \tag{4}$$

where $p(f_{iq}|\boldsymbol{\phi})$ is the prior distribution of $f_{iq}$ with hyper parameter $\boldsymbol{\phi}$. This $\boldsymbol{f}_i^{(t)}$ is identical to rank membership profile (RMP; Shojima, 2007c). In addition, the second term in (2) can be regarded as the prior distribution of $\boldsymbol{V}$. It is given by

$$E_{\boldsymbol{F}|\boldsymbol{U},\boldsymbol{V}^{(t)}}[\ln p(\boldsymbol{V}|\boldsymbol{F})]=\ln p(\boldsymbol{V}). \tag{5}$$

Finally, the third term in (2) is a constant. Therefore, (2) is rewritten as

$$\begin{aligned}\ln p(\boldsymbol{V}|\boldsymbol{U}) =& \sum_{i=1}^{N}\sum_{q=1}^{Q} f_{iq}^{(t)} \ln p(\boldsymbol{u}_i|\boldsymbol{v}_q) + \ln p(\boldsymbol{V}) + \text{const.} \\ =& \sum_{i=1}^{N}\sum_{q=1}^{Q}\sum_{j=1}^{n} f_{iq}^{(t)} z_{ij}\{u_{ij}\ln v_{qj} + (1-u_{ij})\ln(1-v_{qj})\} + \ln p(\boldsymbol{V}) + \text{const.}\end{aligned} \tag{6}$$

In the above equation, each $v_{qj}$ can be optimized individually. The first derivative of (6) with respect to $v_{qj}$ becomes

$$\frac{\partial \ln p(\boldsymbol{V}|\boldsymbol{U})}{\partial v_{qj}} = \sum_{i=1}^{N} f_{iq}^{(t)} z_{ij}\Bigl(\frac{u_{ij}}{v_{qj}} - \frac{1-u_{ij}}{1-v_{qj}}\Bigr) + \frac{\partial \ln p(v_{qj})}{\partial v_{qj}} = 0. \tag{7}$$

Consequently, the maximum a posteriori (MAP) estimate of $v_{qj}$ in the $(t+1)$-st EM cycle is given by solving the above equation. When the prior distribution of $v_{qj}$ is a constant, the maximum likelihood estimate of $v_{qj}$ can be obtained explicitly as follows:

$$v_{qj}^{(t+1)} = \frac{\sum_{i=1}^{N} f_{iq}^{(t)} z_{ij} u_{ij}}{\sum_{i=1}^{N} f_{iq}^{(t)} z_{ij}}. \tag{8}$$

The cycle described above is repeated until a certain convergence criterion is satisfied. The $\chi^2$ statistic (Shojima, 2007d) or the value of the expected log-likelihood in (6) is useful for the criterion.

# 3 Analysis

## 3.1 Example 1

A world history test was analyzed and the results are reported in this section. The sample size was 2049 and the number of items was 36. All items were binary. This data is identical to that analyzed in Shojima (2007a, 2007c). The number of latent ranks was set to 10, and the initial values of the RRV of latent rank $R_q$ were set to $\boldsymbol{v}_q^{(1)} = q\boldsymbol{1}/(Q+1)$ $(q = 1, \cdots, Q)$. In addition, the stopping rule of the EM cycle was

$$|\ln p(\boldsymbol{V}^{(t+1)}|\boldsymbol{U}) - \ln p(\boldsymbol{V}^{(t)}|\boldsymbol{U})| < 0.0001 \times |\ln p(\boldsymbol{V}^{(t)}|\boldsymbol{U})| + 0.01. \tag{9}$$

The IRPs obtained for the 36 items are shown in Figure 1. The number of EM cycles required for convergence was 21, and the goodness-of-fit of the estimated model was $\chi^2_{(324)} = 277.78$ $(p = 0.970)$. The $\chi^2$ value was very satisfactory for this size of test data. However,

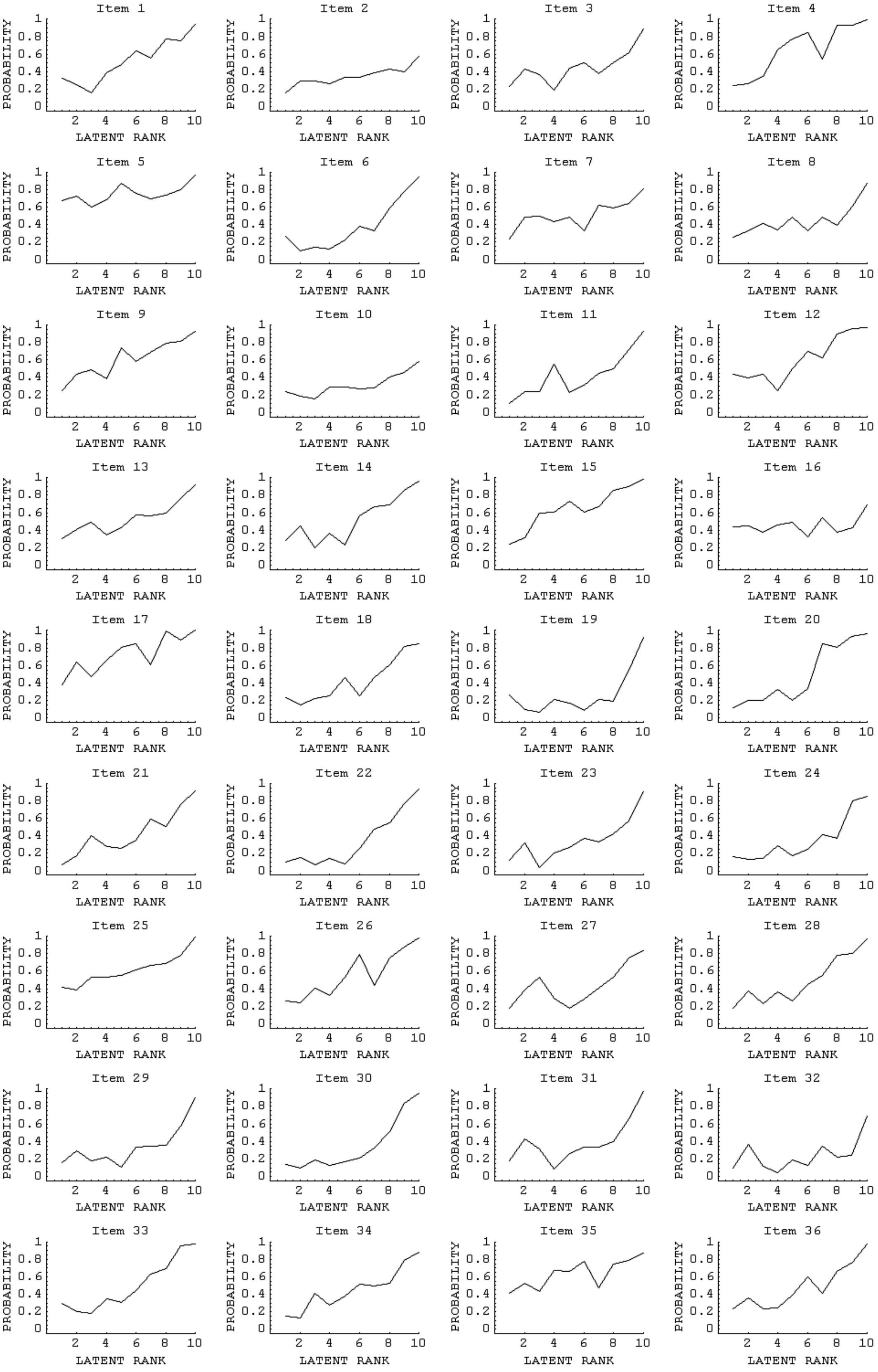
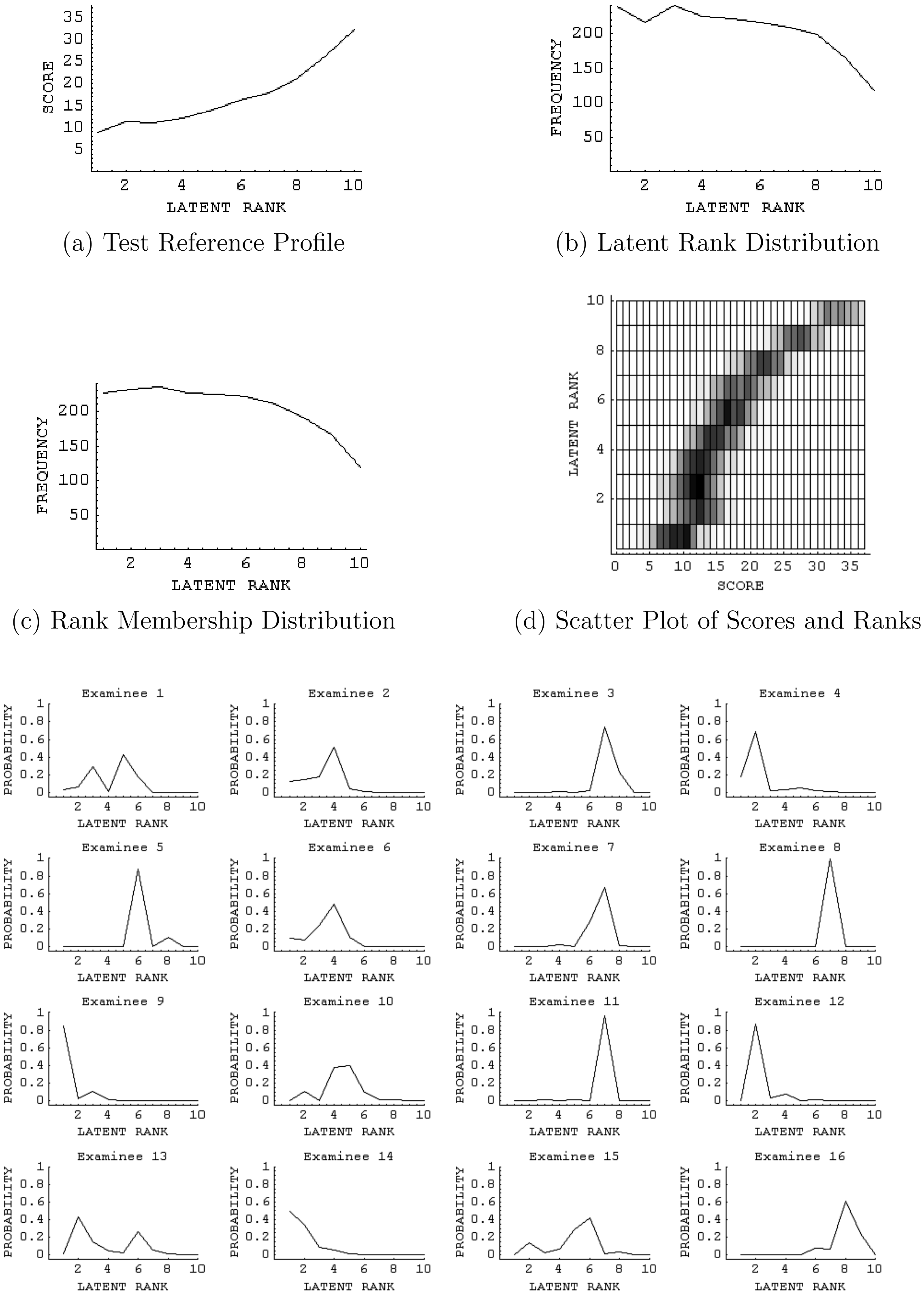


Figure 1: Item Reference Profiles



(a) Test Reference Profile

(b) Latent Rank Distribution

(c) Rank Membership Distribution

(d) Scatter Plot of Scores and Ranks

(e) Rank Membership Profiles of Examinees 1–16

Figure 2: TRP, LRD, RMD, Scatter Plot, and RMPs

this goodness of the model-fit indicates that the IRPs overfit the data. The IRPs of the model are merely describing the data without succeeding in extracting the structure from it. Each panel in Figure 1 is not smoother than that in Shojima (2007a, Figure 3; 2007c, Figure 1) because each node (latent rank) does not monitor the states of their neighboring nodes in the process of the EM algorithm.

In addition, Figure 2 shows the test reference profile (TRP; Shojima, 2007a), the latent rank distribution (LRD; Shojima, 2007a), the rank membership distribution (RMD; Shojima, 2007c), the scatter plot of the scores and the estimated latent ranks, and the rank membership profiles of examinees 1–16.[1] The TRP in Figure 2(a) is the sum of the 36 IRPs and it expresses the transition of the expected score through latent ranks. It is obvious from the figure that the TRP is not always obtained to be monotonically increasing. In NTT with the SOM algorithm, the TRP shape almost exclusively monotonically increases even if some IRPs do not because the statistical learning process in the SOM algorithm is known as a kind of nonlinear and nonparametric principal component analysis (Ritter, Martinetz, & Schulten, 1992; Kohonen, 1995; Mulier, & Cherkassky, 1995). This provides strong evidence that the latent scale obtained by NTT is rank-ordered. However, the TRP obtained from the IRPs estimated by the MML-EM is not always monotonically increasing, as described above, which must be self-contradictory for a model of the latent rank theory. Therefore, some modification of the estimation procedure is necessary.

The LRD and RMD in Figures 2(b) and 2(c) are very different to those in Shojima (2007c, Figures 2(b) and 2(c)). As for Figure 2(d), Spearman's rank correlation coefficient between the scores and the latent ranks was 0.889. In addition, the RMPs in 2(e) were not smoother than those in Shojima (2007c, Figure 3(a)), and not a few RMPs had bimodality in their shapes.

### 3.2 Example 2

The shape of the TRP must be monotonically increasing for a latent rank model. Therefore, a constraint was added in EM cycles to make the IRPs monotonically increasing. The IRPs under the monotonically increasing constraint (MIC) can be obtained by inserting the following step after each EM cycle.

$$\begin{aligned}&\text{For } (j{=}1;\ j \leq n;\ j = j+1) \\ &\text{— Sort}(\boldsymbol{v}_j^{(t)}).\end{aligned} \tag{10}$$

[1] The LRDs in Shojima (2007a, 2007b, 2007c) were plotted in left-right inversion by mistake. Accordingly, the dots in the scatter plots were also inversely plotted with respect to the vertical axes.

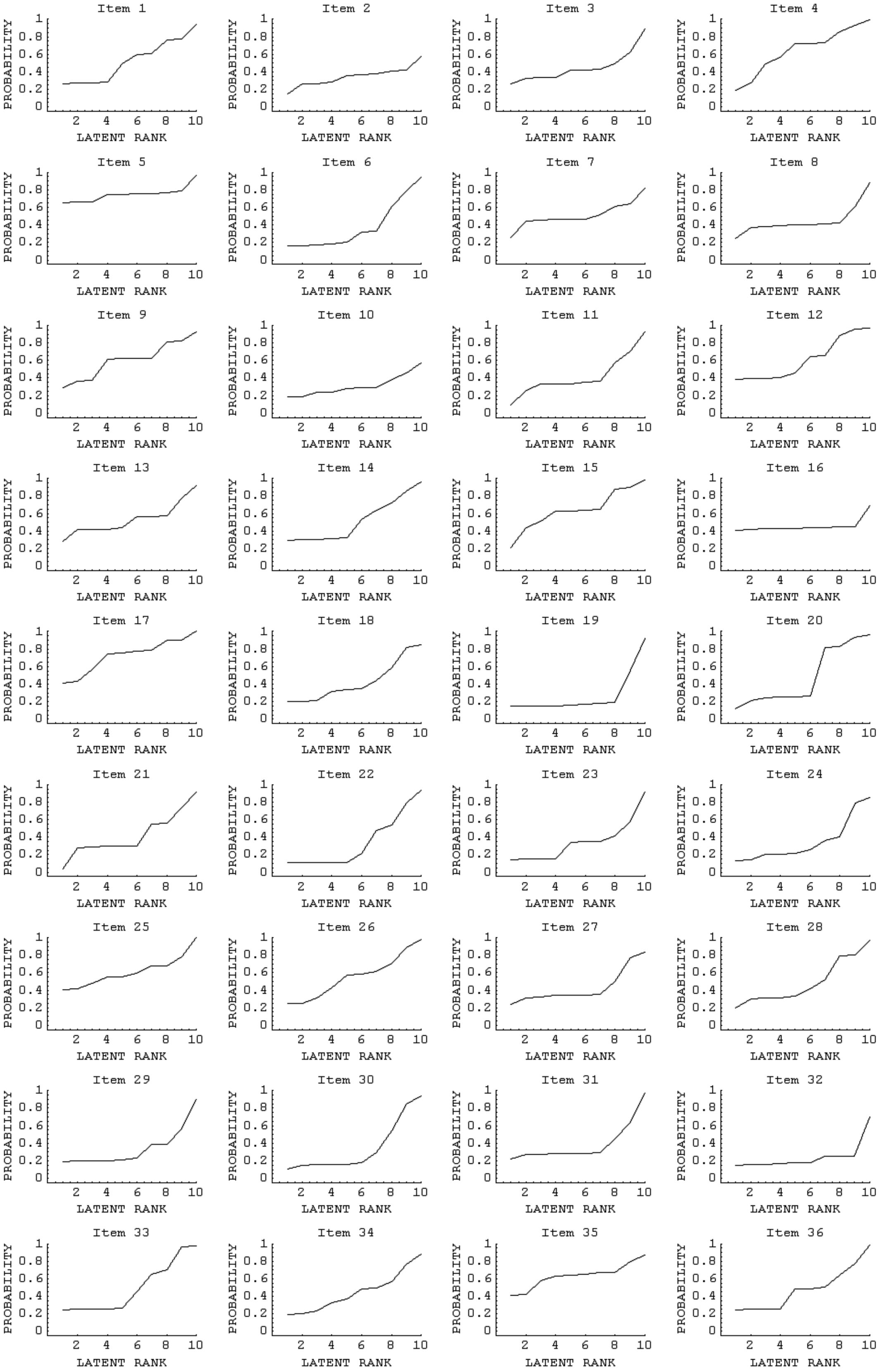
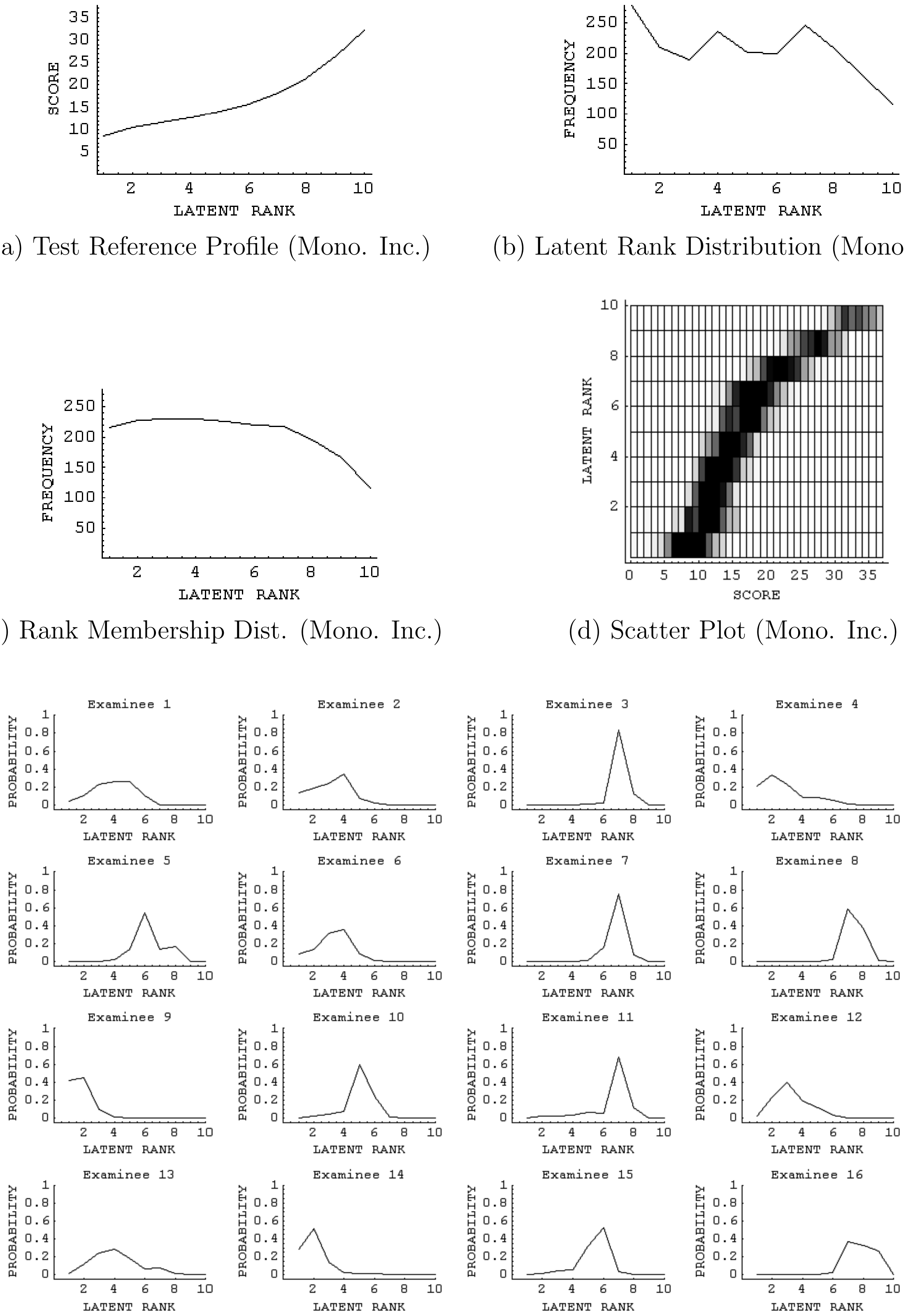


Figure 3: Monotonically Increasing Item Reference Profiles



(a) Test Reference Profile (Mono. Inc.)

(b) Latent Rank Distribution (Mono. Inc.)

(c) Rank Membership Dist. (Mono. Inc.)

(d) Scatter Plot (Mono. Inc.)

(e) Rank Membership Profiles of Examinees 1–16 (Mono. Inc.)

Figure 4: TRP, LRD, RMD, Scatter Plot, and RMPs with Monotonically Increasing IRPs

The other conditions were the same as in Example 1. The number of EM cycles until convergence was 21, and the goodness-of-fit was $\chi^2_{(324)} = 1007.92$ ($p < 0.000$). The model-fit became worse than that of Example 1 because the model artificiality was raised by inserting the MIC step. The IRPs of the 36 items under the MIC are shown in Figure 3. In addition, Figure 4 shows the TRP, LRD, RMD, the scatter plot of the scores and the latent ranks, and the RMPs of examinees 1–16.

Although the IRPs are still not smooth because they have merely been sorted, Figure 4(a) shows that the obtained TRP is monotonically increasing. The MIC is necessary for test administrators in practice because it is natural that the correct answer rates of items monotonically increase as the latent ranks (ability levels) of examinees become higher, although the phenomenon (data) is not always monotonically increasing.

The LRD and RMS in Figures 4(b) and 4(c) are still not very different from those in Shojima (2007c, Figures 2(b) and 2(c)). In addition, Spearman's rank correlation coefficient was 0.910 in the scatter plot of Figure 4(d). Furthermore, in Figure 4(e), there are fewer RMPs with bimodality than in Figure 2(e).

## 3.3 Example 3

Whether or not the real phenomenon (data) is smooth, it cannot be helped that the predictive performance for the future data of the model deteriorates when the model overfits the present data. Such a model is said to lack predictive validity. That is, the model should not overlearn the present data. In other words, it is necessary to weaken the sensitivity of the model in the statistical learning process. Here, information interchange among nodes is required to prevent the RRVs from locally overadapting to the input data.

As a method for making the model locally insensitive, it is effective to smooth $\boldsymbol{f}_i^{(t)}$ in (4), the posterior distribution of the RMP of examinee $i$ in the $t$-th EM cycle, because the IRPs are then smoothed by the effect of the smoothness of $\boldsymbol{f}_i^{(t)}$ by (8). Although there are many smoothing methods, the moving average method is used here. The method of smoothing $\boldsymbol{f}_i^{(t)}$ by the $(2m+1)$-term simple moving average (SMA$(2m+1)$) is given by

$$g_{iq}^{(t)*} = \frac{\sum_{q'=q-m}^{q+m} f_{iq'}^{(t)}}{2m+1} \quad (q = 1, \cdots, Q), \tag{11}$$

where

$$f_{il}^{(t)} = f_{i1}^{(t)} \quad (l \leq 0) \tag{12}$$

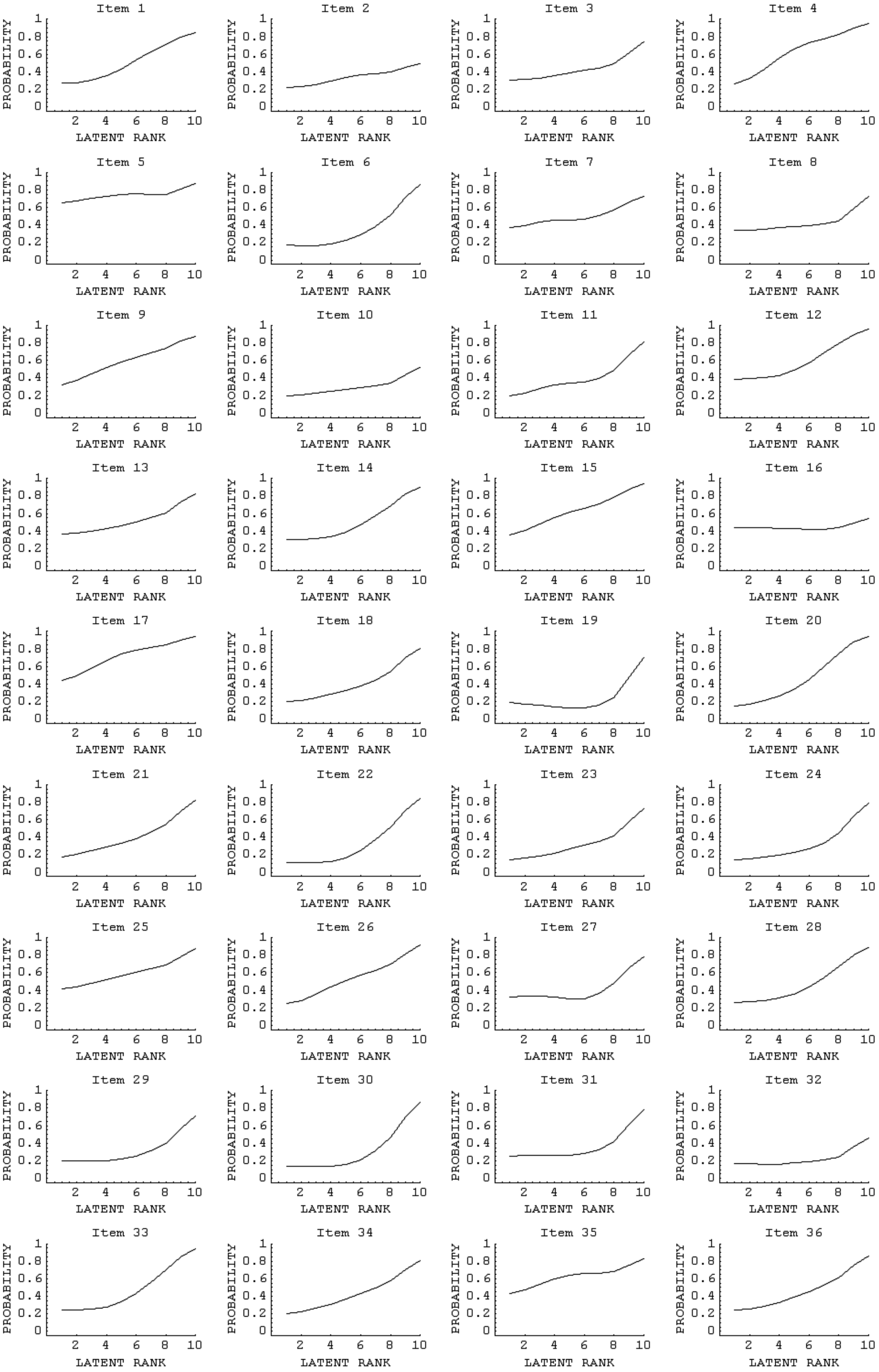
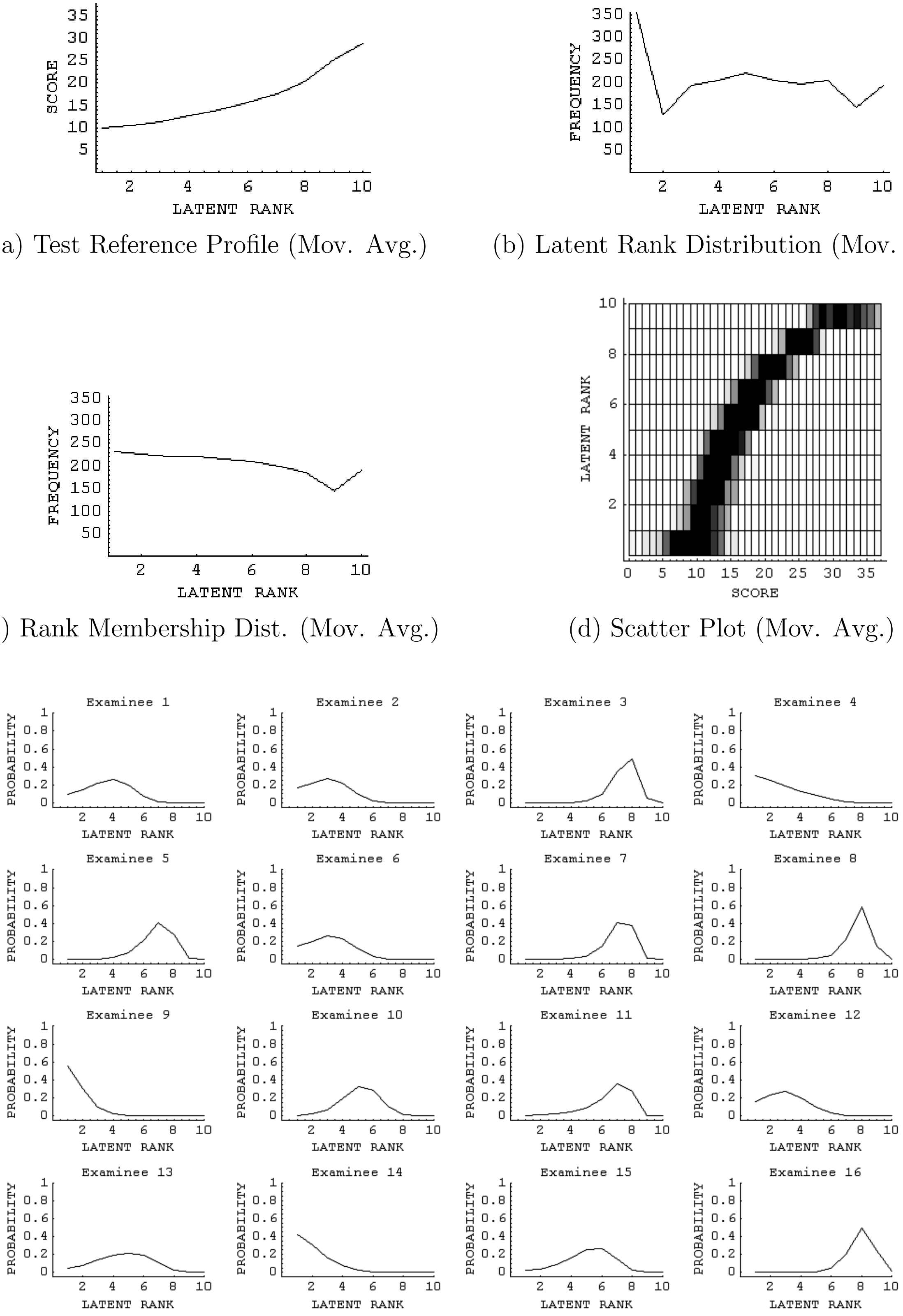


Figure 5: Item Reference Profiles with Moving-averaged RMPs



(a) Test Reference Profile (Mov. Avg.)

(b) Latent Rank Distribution (Mov. Avg.)

(c) Rank Membership Dist. (Mov. Avg.)

(d) Scatter Plot (Mov. Avg.)

(e) Rank Membership Profiles of Examinees 1-16 (Mov. Avg.)

Figure 6: TRP, LRD, RMD, Scatter Plot, and RMPs with Moving-averaged RMPs

and

$$f_{il'}^{(t)} = f_{iQ}^{(t)} \quad (l' \geq Q+1). \tag{13}$$

In addition, the standardization

$$g_{iq}^{(t)} = \frac{g_{iq}^{(t)*}}{\sum_{q=1}^{Q} g_{iq}^{(t)*}} \tag{14}$$

is required because $\sum_q g_{iq}^{(t)*}$ is not always one. Then, the IRPs in the $(t+1)$-st EM cycle are reweighted as

$$v_{qj}^{(t+1)} = \frac{\sum_{i=1}^{N} g_{iq}^{(t)} z_{ij} u_{ij}}{\sum_{i=1}^{N} g_{iq}^{(t)} z_{ij}}. \tag{15}$$

The data was analyzed by the MML-EM method with SMA(3). The other settings were identical to those in Examples 1 and 2. The number of EM cycles was 12, and the goodness-of-fit was $\chi^2_{(324)} = 880.02$ ($p < 0.000$). The model-fit was better that for the model estimated in Example 2. It is clear from Figure 5 that the IRPs with SMA(3) are much smoother than those in Figures 1 and 3, although their expressions seem to be slightly monotoic and poor compared with those obtained by the SOM algorithm in Shojima (2007c, Figure 1). In addition, the expected log-likelihood in each EM cycle monotonically increases in the conventional use of the EM algorithm, but that obtained by the MML-EM with SMA was not always monotonically increasing in the analysis. Although this feature can also be observed in the NTT model (the latent rank model with the SOM algorithm, LRT-SOM), the log-likelihood does not decrease very often.

The TRP, LRD, RMD, the scatter plot of the scores and the latent ranks, and the RMPs of examinees 1–16 are shown in Figure 6. Figure 6(a) shows that the TRP is monotonically increasing even if some IRPs are not. The LRD and RMD in Figures 6(b) and 6(c) are very similar to those in Shojima (2007c, Figures 2(b) and 2(c)). In addition, Spearman's rank correlation coefficient in Figures 6(d) is 0.919. Furthermore, the shapes of the RMPs in Figure 6(e) appear to be very similar to those in Shojima (2007c, Figure 3(a)).

# 4 Discussion

This study found that IRPs can be estimated without using the SOM algorithm. Specifically, the IRPs were estimated by the marginal maximum likelihood method with the EM algorithm (MML-EM). Therefore, some models in this study cannot be explained in the NTT framework. Here, such models are called latent rank theory (LRT) models. They include the NTT models and the models presented in this study. Consequently, the NTT models are positioned as LRT models estimated by the SOM algorithm (LRT-SOM models), and the models presented in this study are the LRT-EM models.

The LRT-EM model can easily be extended to a model for polytomous data. However, the LRT-EM is a little boring because it is a simple application of the conventional statistics. In addition, the TRP summed by the IRPs estimated by the simple LRT-EM does not always increase monotonically, as seen in Example 1. This fact is self-contradictory for the latent rank model. Therefore, it is necessary to impose the constraint of monotonic increase on IRPs or to smooth the rank membership profiles, as seen in Examples 2 and 3, respectively.

The goodness-of-fit of the simple LRT-EM model to the data was excellent in Example 1, but this was because the model overfit the data. The simple LRT-EM model merely described the data rather than extracted the structure from it. The LRT-EM model with the monotonically increasing constraint (LRT-EM with MIC) in Example 2 was not a natural way to express the structure underlying the data, but the MIC itself is sometimes necessary to administer tests. The IRPs of the LRT-EM with SMA models obtained in Example 3 were very smooth, but those of the LRT-SOM model were a little smoother. Further research is required to develop other smoothing methods.

Generally speaking, although the convergence speed of the EM algorithm is not very high, it is much higher than that of NTT or LRT-SOM. In addition, the IRPs estimated by LRT-EM are invariant, while those estimated by LRT-SOM are slightly different in each calculation. They are outstanding features of the LRT-EM models. In a broad sense, the MML-EM method can be interpreted in the context of the statistical learning method because it contains the process of updating the IRPs. However, the models in Examples 1 and 2 might not be called "neural" test models because the nodes that represent the latent ranks are merely linked and do not monitor the states of their neighboring nodes, while the model in Example 3 can be said to be neural because the elements of each RRV are affected by those of the neighboring nodes.

# References

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, **39**, 1-38.

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

Everitt, B. S., & Hand, D. J. (1981) *Finite mixture distributions.* Chapman and Hall.

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

Mulier, F. & Cherkassky, V. (1995) Self-organization as an iterative kernel smoothing process. *Neural Computation*, **7**, 1165-1177.

Ritter, H., Martinetz, T. & Schulten, K. (1992) *Neural computation and self-organizing maps: An introduction.* Addison-Wesley.

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.

Shojima, K. (2007e) Estimation for neural test models with missing data. *DNC Research Note*, 07-09.

Titterington, D. M., Smith, A. F. M., & Makov, U. E. (1985) *Statistical analysis of finite mixture distributions.* Wiley.