Neural Test Theory:

A Latent Rank Theory for Analyzing Test Data

SHOJIMA Kojiro

April 2008

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

Neural test theory:
A latent rank theory for analyzing test data

Kojiro Shojima

Abstract

Neural test theory (NTT) is a latent rank theory for analyzing test data using the mechanism of a self-organizing map. Although the conventional method for estimating item reference profiles is the least squares method (the minimum Euclidian distance method), we introduce a distance based on likelihood and apply the maximum likelihood (ML) method and the Bayesian method. This enables the NTT model to compute the rank membership profile, the posterior distribution of each examinee's latent rank, which is useful for reviewing the probabilities of examinees belonging to respective latent ranks. In addition, by introducing the ML method, the discrepancy between the NTT model and data can be evaluated, and the model fitness using several goodness-of-fit indices can be examined. The indices are also helpful for determining an adequate number of latent ranks.

Key words: neural test theory, latent rank theory, item reference profile, rank membership profile, goodness-of-fit indices, ordinal alignment condition, self-organizing map.

ニューラルテスト理論:
テストデータ分析のための潜在ランク理論

荘島宏二郎

要約

ニューラルテスト理論(NTT)は，自己組織化マップのメカニズムを利用したテストデータを分析するための潜在ランク理論である．本研究では，項目参照プロファイルを推定する上で尤度を用いて勝者ランクを推定する方法を提案する．これにより，NTTモデルは，潜在ランクの事後分布であるランク・メンバーシップ・プロファイル(RMP)を推定することができる．RMPは，各受験者がどの潜在ランクに所属するかを評価するのに便利である．また，最尤法を導入したことにより，NTTモデルの適合度指標を検討することができる．適合度指標は，潜在ランク数を決定する上でも役に立つ．

キーワード: ニューラルテスト理論, 潜在ランク理論, 項目参照プロファイル, ランク・メンバーシップ・プロファイル, 適合度指標, 順序配置条件, 自己組織化マップ.

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

1 Introduction

Neural test theory (NTT; Shojima, 2008) is a test theory using the mechanism of a self-organizing map (SOM; Kohonen, 1995). In the NTT model, the item reference profile (IRP) which represents statistical characteristics of each item is estimated. In addition, NTT assumes that the latent scale is ordinal, and each examinee is located on the latent rank scale. To begin, we describe the importance of the latent rank scale, that is, why a model with such a scale is needed.

The latent scale must be ordinal for at least two reasons, one passive and one positive. The passive reason is that tests do not have enough resolution to continuously evaluate human ability or attitude; the most that tests can do is classify examinees into several grades. In other words, the reliability of tests is too low to locate examinees on a continuous scale. Tests cannot distinguish two examinees whom have nearly equal abilities, unlike, for example, weighing machines that can detect the difference between two persons of almost the same weight; tests cannot correctly line up examinees in order of ability, while weighing machines can correctly array people in order of weight. Therefore, it is safer to prepare an ordinal scale and roughly array examinees on it rather than use a continuous scale, because there is a high risk of the continuous scale incorrectly sorting the examinees given the low reliability of the test.

The positive reason for an ordinal scalie being necessary is more pedagogical and sociological, while the passive reason is metrical and methodological. Negative consequences of using continuous scales in the field of education, for example, are apparent. In Japan, most tests administered in schools are scored on continuous scales. Such tests include almost every test students experience from admission examinations to daily short tests. Japanese school students are thus continuously evaluated every day. Students are strongly motivated to get the highest scores they can, which is a natural response but one that can be exploited by people acting on either good or bad intentions. For example, many *juku* (private tutoring school) teachers and even school teachers coach their students on “test techniques” to help students get their highest possible scores. Such techniques include skills such as how to find the correct answers without reading the lead sentences of questions. Many books about these techniques can be purchased in bookstores, and the extremely provocative book covers that are typical are intended to fuel the impatience to improve their test scores that students feel. must fan the feelings of impatience of the students. Such a situation can be found in many other countries, especially those in East and South East Asia. We should try to prevent students from swinging between extremes of gladness and sadness in response to fluctuations in unreliable continuous test score because human ability cannot be drastically changed overnight. Using an ordinal scale to evaluate students will encourage them to develop attitudes more conductive to gradual academic improvement over a longer period, a more realistic goal, because ordinal evaluation is more robust than continuous one.

Shojima (2007) discussed three contexts in which testing plays important roles in our society: the contexts of measurement, explanation, and existence. The passive and positive reasons given above for rank ordering corresponding to the first and the third contexts, respectively. The first context is the most standard role of testing. Testing socially undertakes to measure the ability or psychological composition of members of our society. The second context refers to accountability as to what each test measures and why measuring it is necessary. Generally speaking, preparation of a test is nearly equal to a declaration that the society administering the test consideres what the test measures to be an important ability or skill in the society. Finally, the third context stresses that the form in which a test exists must have salutary repercussions on the society. That is, administering the test itself must be a worthwhile practice. For example, a test is expected to maintain and improve the ability level of members of the society and provide opportunities for them to achieve self-discipline and self-realization. At present, we believe the use of continuous scales in testing has negative consequences, as explained above. The use of an ordinal scale to evaluate students and examinees is an effective way to make the practice of testing more beneficial to our society.

In Japan, reports by the Central Education Council in 1999 and the University Council in 2000 stated that the National Center Test for University Admissions (NCT) should be used as a qualifying test. The NCT is composed of 30 subjects from 10 subject areas. It is taken by about 500,000 high school graduates and graduands each year, and NCT scores are used for admission to all national and public universities, as well as some private universities. The gist of the reports by the councils is that use of the NCT scores should be restricted to judging whether the examinees have the minimum academic achievements to enter the universities and that admission decisions should refer to other information such as that from school records and interviews. The NTT is compatible with such a movement. The ordinal scale of the NTT grades examinees into about 10 ranks. Inevitably, more examinees than the number of enrollment spaces must be passed through a test administered using the NTT model. The surplus number should then be reduced by using school records and interviews. Therefore, NTT can encourage the use of information other than test scores by relatively reducing the excessively authoritative status of testing in our society.

The subtitle of this study, the latent rank theory (LRT; Shojima, 2007b), is a generic theory including NTT, and the latent rank model or LRT model is a collective term indicating statistical models whose latent scale is ordinal. Generally, the latent variable model (LVM; e.g., Loehlin, 2003) stands for statistical models dealing with variables which are not directly observable. Factor analysis (FA; e.g., Harman, 1976; Gorsuch, 1983), structural equation modeling (SEM; e.g., Jöreskog & Sörbom, 1993), and item response theory (IRT; e.g., Lord, 1980) are representative models of the LVM when the latent variables are continuous. In addition, latent profile analysis and latent class analysis (e.g., Titterington, Smith, & Makov, 1985; McLachlan, & Peel, 2000) are LVMs in which the latent variables are nominal-categorical. According to this framework, the LRT model or the latent rank model (LRM) is a subcategory of the LVM when the latent variable is ordinal.

Statistical models in which latent variables are ordinal are especially necessary in the fields of psychology, pedagogy, sociology, and behavioral sciences. For example, psychological questionnaires are the tools often used to measure characteristics which cannot be directly observed, such as depression, anxiety, sense of inferiority, and so on. Like ability tests, these questionnaires are not reliable enough for a continuous evaluation, in this case of the psychological status of subjects. Thus, it is better to assume the latent variables are ordinal when analyzing psychological questionnaire and social survey data.

The purpose of this study is to add new features and improve the NTT model provided by Shojima (2008). He applied the least-squares method in the winner rank selection and the latent rank estimation in the statistical learning process. This was the result of directly introducing the SOM mechanism when the NTT was developed. We propose the application of new methods based on stochastic theory — the maximum likelihood method and the Bayesian method — for selecting the winner rank and estimating the latent rank. Accordingly, we can compute the model-fitness and evaluate the posterior probabilities that each examinee belongs to respective latent ranks.

2 Method

2.1 Reference Matrix

Figure 1 shows the latent rank scale of the NTT. The white belt represents the scale, and black circles stand for the individual latent ranks ($R_1, \cdots, R_Q$), where Q is the number of latent ranks. In the figure, the number of latent ranks is 10. The matrix composed of gray hexagonal cells is the reference matrix $\boldsymbol{V}$ with size ($n \times Q$), where n is the number of items which is usually larger than Q. In addition, the element in the j-th row and q-th column

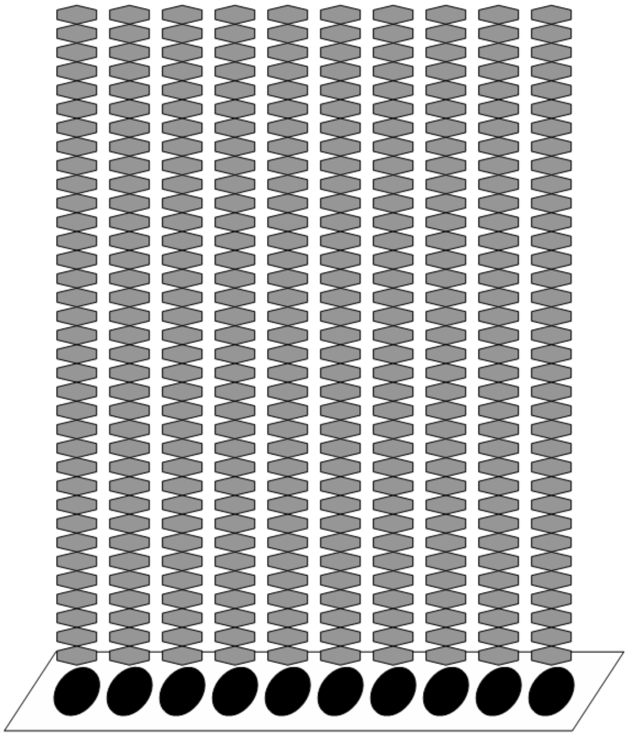


Figure 1: Latent Rank Scale

in the reference matrix, v_{jq}, is the rank reference element (RRE) representing the correct answer ratio of the examinees in latent rank R_q to item j. That is,

$$p(U_j = 1|R_q) = v_{jq}. \tag{1}$$

In addition, the j-th row vector in $\boldsymbol{V}$, $\boldsymbol{v}_j$ $(Q \times 1)$, is the item reference profile (IRP; Shojima, 2008), and the q-th column vector in $\boldsymbol{V}$, $\boldsymbol{v}_q$ $(n \times 1)$, is the rank reference vector (RRV).

2.2 Estimation of Item Reference Profile

Let us suppose that the sample size is N and the response matrix of the examinees is $\boldsymbol{U} = \{u_{ij}\}$ $(N \times n)$, where u_{ij} is a dichotomous variable coded 1 if the response of examinee i to item j is correct, and otherwise coded 0. Let us also assume that $\boldsymbol{Z} = \{z_{ij}\}$ $(N \times n)$ is the missing indicator matrix, where z_{ij} is also a dichotomous variable coded 1 when u_{ij} is observed and 0 when it is missing. The missing indicators are necessary when the responded-to items vary from examinee to examinee. Note that missingness (absence) is different from non-response which is generally treated as an incorrect answer in achievement tests. When an examinee does not respond to a submitted item, the response is not missing data but non-response. Therefore, u and z of the examinee are coded 0 (incorrect) and 1 (observed), respectively. It is important to regard non-responses as the incorrect answers. Otherwise, the item difficulty would be underestimated.

2.2.1 Statistical Learning Process

The NTT uses the SOM mechanism when estimating the reference matrix $\boldsymbol{V}$, and the matrix or the IRPs are obtained through the statistical learning process as follows:

```
For ($t$=1; $t \leq T$; $t = t + 1$)                                        (2)
— $\boldsymbol{U}^{(t)}$ ⇐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 rank for $\boldsymbol{u}_h$.                        (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)
```

Line (2) indicates that the procedure causes Lines (3)–(8) to be executed repeatedly T times. Line (4) causes Lines (5)–(7) to run until counter h reaches N. Line (3) is to cancel the input order effect on the statistical learning, and $\boldsymbol{U}^{(t)}$ is the input data of the t-th period. In addition, $\boldsymbol{u}_h^{(t)}$ is the h-th row vector of $\boldsymbol{U}^{(t)}$, and $\boldsymbol{z}_h^{(t)}$ is the missing indicator vector corresponding to $\boldsymbol{u}_h^{(t)}$. Furthermore, $\boldsymbol{V}^{(t,h)}$ is the reference matrix after learning the input data $\boldsymbol{u}_h^{(t)}$, where $V^{(1,0)}$ is the initial value of the reference matrix, and the recommended value is

$$\boldsymbol{V}^{(1,0)} = \left\{ v_{jq}^{(1,0)} \middle| v_{jq}^{(1,0)} = \frac{q}{Q+1} \right\} \ (n \times Q). \tag{9}$$

2.2.2 Winner Rank Selection

Line (6) is the process to determine the winner rank for the input data $\boldsymbol{u}_h^{(t)}$. The winner rank $R_w^{(t,h)}$ is the latent rank at which RRV is closest to the input data $\boldsymbol{u}_h^{(t)}$, and it is determined by

$$R_w^{(t,h)} : \ w = \arg\min_{q \in Q} d(t, h, q), \tag{10}$$

where d is a certain measure representing the distance between the RRV of latent rank R_q and the input data $\boldsymbol{u}_h^{(t)}$. Shojima (2008) proposed a method for determining the winner rank from the square of the Euclidean distance (the least-squares (LS) method). That is,

$$d_{LS}(t,h,q) = ||\boldsymbol{z}_h^{(t)} \odot (\boldsymbol{u}_h^{(t)} - \boldsymbol{v}_q^{(t,h-1)})||^2 = \sum_{j=1}^{n} z_{hj}^{(t)}(u_{hj}^{(t)} - v_{jq}^{(t,h-1)})^2. \tag{11}$$

The winner rank selected by the above equation is the closest latent rank to the input data in terms of (the square of) the Euclidean distance.

In this study, a method of determining the winner rank by a stochastic distance, likelihood, is proposed. Under the assumption of local independence (Hambleton & Swaminathan, 1985) which requires that the probabilities of correctly answering any of two items are independent provided that latent rank R_q is given, the probability of observing $\boldsymbol{u}_h^{(t)}$ given latent rank R_q becomes

$$p(\boldsymbol{u}_h^{(t)}|R_q) = \prod_{j=1}^{n} \Big\{ p(U_j = 1|R_q)^{u_{hj}^{(t)}} p(U_j = 0|R_q)^{1-u_{hj}^{(t)}} \Big\}^{z_{hj}^{(t)}}. \tag{12}$$

Then, the closest latent rank in terms of the stochastic distance is the maximum likelihood (ML) estimate of the winner rank, and the RRV of the latent rank minimizes

$$d_{ML}(t,h,q) = -\sum_{j=1}^{n} z_{hj}^{(t)} \big\{ u_{hj}^{(t)} \ln v_{jq} + (1 - u_{hj}^{(t)}) \ln(1 - v_{jq}) \big\}. \tag{13}$$

In addition, using the prior probability π_q that the winner rank is latent rank R_q, the winner rank can also be determined as the latent rank for which RRV minimizes the following distance:

$$d_{MAP}(t,h,q) = -\ln \pi_q - \sum_{j=1}^{n} z_{hj}^{(t)} \big\{ u_{hj}^{(t)} \ln v_{jq} + (1 - u_{hj}^{(t)}) \ln(1 - v_{jq}) \big\}. \tag{14}$$

The closest rank in terms of the above distance is the maximum a posteriori (MAP) estimate of the winner rank.

2.2.3 Updating the Reference Matrix

Line (8) is to update the reference matrix. Its specification is

$$\boldsymbol{V}^{(t,h)} = \boldsymbol{V}^{(t,h-1)} + (\boldsymbol{1}_n \boldsymbol{h}^{(t)\prime}) \odot (\boldsymbol{z}_h^{(t)} \boldsymbol{1}_Q') \odot (\boldsymbol{u}_h^{(t)} \boldsymbol{1}_Q' - \boldsymbol{V}^{(t,h-1)}), \tag{15}$$

where

$$\boldsymbol{h}^{(t)} = \Big\{ h_{qw}^{(t)} \Big| h_{qw}^{(t)} = \frac{\alpha_t Q}{N} \exp\Big(-\frac{(q-w)^2}{2Q^2\sigma_t^2} \Big) \Big\} \quad (Q \times 1), \tag{16}$$

and

$$\alpha_t = \frac{(T-t)\alpha_1 + (t-1)\alpha_T}{T-1} \tag{17}$$

$$\sigma_t = \frac{(T-t)\sigma_1 + (t-1)\sigma_T}{T-1}. \tag{18}$$

The factor $h_{qw}^{(t)}$ is called "tension" and controls the learning size or the amount of change of the RRV of latent rank R_q. The tension becomes greater as the latent rank locates geographically closer to the winner. Therefore, the more closely the rank is located adjacent to the winner, the more the RRV of the rank numerically approximates the input data. The constant α regulates the size of the tension, and it linearly decreases from the initial value α_1 to the final value α_T as t increases. In addition, σ specifies the region size that the learning propagates, and it also reduces from σ_1 to σ_T as t reaches T.

While Equation (15) shows the method of updating the reference matrix, individual RRVs are updated as follows:

$$\boldsymbol{v}_q^{(t,h)} = \boldsymbol{v}_q^{(t,h-1)} + h_{qw}^{(t)} \boldsymbol{z}_h^{(t)} \odot (\boldsymbol{u}_h^{(t)} - \boldsymbol{v}_q^{(t,h-1)}) \quad (q = 1, \cdots, Q). \tag{19}$$

Clearly, the rank reference elements (RREs) corresponding to the missing data remain invariant.

The finally obtained $\boldsymbol{V}^{(T,0)}$ is the estimate of the reference matrix, $\hat{\boldsymbol{V}}$, and the j-th row vector in the estimated matrix, $\hat{\boldsymbol{v}}_j$ $(Q \times 1)$, is the item reference profile (IRP) of item j.

2.2.4 Test Reference Profile

The test reference profile (TRP) is the weighted sum of the IRPs and is given by

$$\boldsymbol{t} = \{t_q | t_q = E[T|R_q]\} \quad (Q \times 1), \tag{20}$$

$$E[T|R_q] = \sum_{j=1}^{n} w_j E[U_j|R_q] = \sum_{j=1}^{n} w_j p(U_j = 1|R_q) = \sum_{j=1}^{n} w_j v_{jq}, \tag{21}$$

where w_j is the weight for item j. The q-th element of the TRP is the expected score of examinees who belong to latent rank R_q. The TRP becomes the profile for the expected number-right score through the latent rank scale when every weight is 1.0.

2.3 Latent Rank Estimation

In test theories, describing the characteristics of each examinee's measured object is as important as clarifying the statistical characteristics of each item. The method for estimating the latent rank is the same as that for selecting the winner rank. Applying the LS method proposed by Shojima (2008), we obtain the latent rank estimate of examinee i, R_{r_i} by

$$R_{r_i}^{(LS)}: \; r_i = \arg\min_{q \in Q} ||\boldsymbol{z}_i \odot (\boldsymbol{u}_i - \hat{\boldsymbol{v}}_q)||^2 = \arg\min_{q \in Q} \sum_{j=1}^{n} z_{ij}(u_{ij} - \hat{v}_{jq})^2. \tag{22}$$

In this study, we apply the ML method and the Bayesian method to estimate the latent ranks of examinees. By referring to Equations (13) and (14), the ML and Bayesian methods can be obtained as follows:

$$\begin{aligned} R_{r_i}^{(ML)}: \; r_i &= \arg\min_{q \in Q}\{-\ln p(\boldsymbol{u}_i|R_q)\} \\ &= \arg\max_{q \in Q} \sum_{j=1}^{n} z_{ij}\big\{u_{ij}\ln \hat{v}_{jq} + (1-u_{ij})\ln(1-\hat{v}_{jq})\big\}, \end{aligned} \tag{23}$$

and

$$\begin{aligned} R_{r_i}^{(MAP)}: \; r_i &= \arg\min_{q \in Q}\{-\ln \pi_q - \ln p(\boldsymbol{u}_i|R_q)\} \\ &= \arg\max_{q \in Q}\Big[\ln \pi_q + \sum_{j=1}^{n} z_{ij}\big\{u_{ij}\ln \hat{v}_{jq} + (1-u_{ij})\ln(1-\hat{v}_{jq})\big\}\Big]. \end{aligned} \tag{24}$$

2.3.1 Rank Membership Profile

The rank membership profile (RMP) is useful for evaluating the possibilities of each examinee belonging to respective ranks. The RMP of examinee i is obtained by

$$\boldsymbol{p}_i = \{p_{iq}|p_{iq} = p(R_q|\boldsymbol{u}_i)\} \; (Q \times 1), \tag{25}$$

$$p(R_q|\boldsymbol{u}_i) = \frac{p(\boldsymbol{u}_i|R_q)\pi_q}{\sum_{q'=1}^{Q} p(\boldsymbol{u}_i|R_{q'})\pi_{q'}}. \tag{26}$$

The q-th element of the RMP is the posterior probability of examinee i's belonging to latent rank R_q given the examinee's response vector $\boldsymbol{u}_i$. Accordingly, the RMP is the posterior distribution of the latent rank.

2.3.2 Latent Rank Distribution and Rank Membership Distribution

The latent rank distribution (LRD; Shojima, 2008) is the frequency distribution of the latent rank estimates of the examinees. Let us assume that f_{iq} is a dichotomous variable coded 1 if the latent rank estimate of examinee i is R_q, and otherwise is 0. The LRD then becomes

$$\boldsymbol{f} = \left\{ f_q \middle| f_q = \sum_{i=1}^{N} f_{iq} \right\} \quad (Q \times 1). \tag{27}$$

In addition, the rank membership distribution (RMD) is the simple sum of the RMPs of the examinees, and is given by

$$\boldsymbol{f}^* = \left\{ f_q^* \middle| f_q^* = \sum_{i=1}^{N} p_{iq} \right\} \quad (Q \times 1). \tag{28}$$

The LRD expresses the characteristics of the frequency distribution of the latent ranks for the sample, while the RMD expresses that for the population.

2.3.3 Observation Ratio Profile

The observation ratio profile (ORP) is useful for reviewing the response ratios through the latent rank scale. The ORPs are effective for clarifying that some items are selected by higher latent rankers and some items by lower rankers when a test system allows examinees to select and answer several items. Such tests are not frequently used, but neither are they rare. For example, Mathematic II & B of the NCT usually has six testlets. All examinees are required to do the first and second testlets, but they can select any two testlets from testlets 3–6.

The unweighted and weighted ORPs are given by

$$\boldsymbol{z}_j = \left\{ z_{jq} \middle| z_{jq} = \frac{\sum_{i=1}^{N} z_{ij} f_{iq}}{f_q} \right\} \quad (Q \times 1), \tag{29}$$

and

$$\boldsymbol{z}_j^* = \left\{ z_{jq}^* \middle| z_{jq}^* = \frac{\sum_{i=1}^{N} z_{ij} p_{iq}}{f_q^*} \right\} \quad (Q \times 1). \tag{30}$$

The unweighted ORP is computed based on the latent rank estimates, while the weighted one is computed based on the RMP. That is, the unweighted ORP represents the profile of the sample, while the weighted one stands for that of the population.

2.4 Goodness of Fit

It is necessary to examine how well the present model fits the data. Since the ML method has been introduced in the previous sections, the condition to statistically evaluate the NTT model is ready. In addition, the model fitness for the item response theory (IRT) is discretely computed by partitioning its latent continuous scale into several quadrature points (Mislevy & Bock, 1990). Therefore, the method under the IRT is directly applicable to the NTT model because the NTT scale is discrete from the beginning. The model fitness can be evaluated item by item.

To begin, the log-likelihood necessary for calculating the fitness of item j is

$$LL_{Mj} = \sum_{q=1}^{Q} \{g_{jq} \ln v_{jq} + (f_{jq} - g_{jq}) \ln(1 - v_{jq})\}, \tag{31}$$

where f_{jq} is the frequency of examinees who have responded to item j out of the examinees who belong to latent rank R_q, and g_{jq} is the frequency of examinees who have correctly answered the item out of the f_{jq} examinees. When using the latent rank estimates, the frequencies f_{jq} and g_{jq} are computed by

$$f_{jq} = \sum_{i=1}^{N} z_{ij} f_{iq}, \tag{32}$$

$$g_{jq} = \sum_{i=1}^{N} z_{ij} u_{ij} f_{iq}. \tag{33}$$

In addition, the frequencies can be obtained by using the RMPs as follows:

$$f^*_{jq} = \sum_{i=1}^{N} z_{ij} p_{iq}, \tag{34}$$

$$g^*_{jq} = \sum_{i=1}^{N} z_{ij} u_{ij} p_{iq} \tag{35}$$

The log-likelihood of Equation (31) computed using the RMP-based frequencies becomes the expected log-likelihood.

Next, it is necessary to define a saturated model and null model in the NTT for the sake of model comparison. The saturated model is the model that best fits the data, and the null model is the one which extremely misfits the data. The relative fitness of the present model or the tested model is determined by comparison to that of the saturated model or the null model. Although many statistical models define their saturated model and null model, the definitions vary according to how the states "best fit to the data" and "extremely misfit to the data" are regarded. For example, most software for the structural equation model (SEM) considers the saturated model as the model in which every element in the variance-covariance matrix is estimated, and the null model as the model in which the covariance elements in the matrix are fixed at zero, although the variance elements are estimated.

Considering the saturated model in the NTT, the number of latent ranks of the model must equal the number of response patterns. Although such a model could almost completely explain all the data, the log-likelihood of the model would then become close to zero. Such a model is too close to ideal and it is unrealistic for use as the saturated model. Therefore, the saturated model in the NTT should be more restricted so that it does not completely fit the data.

Here, a good choice is to set the number of latent ranks for the saturated model as equal to the number of items n. The number of items is usually much smaller than both the sample size N and the number of response patterns, and n is generally larger than the number of latent ranks of the tested model Q. In general, a model with a larger number of latent ranks better fits the data. Therefore, it is valid to set the number of latent ranks of the saturated model to n because the saturated model must fit the data better than the tested model does. We refer to this model as the benchmark model because the term "saturated" is inappropriate for a model with n ranks. In addition, the batch-type NTT model proposed by Shojima (2007b) which is estimated by an EM algorithm (Dempster, Laird, & Rubin, 1977) is used as the benchmark model. The model estimated by the EM algorithm is said to be the NTT model with the mechanism of generative topographic mapping (GTM; Bishop, Svensen, & Williams, 1998), and the model estimated by the GTM mechanism generally better fits the data than the model by the SOM mechanism. Therefore, the model with n ranks and the GTM mechanism is valid as the benchmark model instead of the saturated model.

The reference matrix of the benchmark model, $\boldsymbol{B} = \{b_{jq}\}$ $(n \times n)$, can be obtained through the EM algorithm. First, the RMP of examinee i at the t-th period is computed by

$$\pi_{iq}^{(t)} = \frac{p(\boldsymbol{u}_i|\boldsymbol{b}_q^{(t)})\pi_q}{\sum_{q'=1}^{Q} p(\boldsymbol{u}_i|\boldsymbol{b}_{q'}^{(t)})\pi_{q'}} \quad (i = 1, \cdots, N;\ q = 1, \cdots, n), \tag{36}$$

where

$$p(\boldsymbol{u}_i|\boldsymbol{b}_q^{(t)}) = \prod_{j=1}^{n} \{b_{jq}^{(t)}\}^{u_{ij}} \{1 - b_{jq}^{(t)}\}^{1-u_{ij}}. \tag{37}$$

The reference matrix estimate of the t-th period is then obtained by

$$b_{jq}^{(t+1)} = \frac{\sum_{i=1}^{N} \pi_{iq}^{(t)} z_{ij} u_{ij}}{\sum_{i=1}^{N} \pi_{iq}^{(t)} z_{ij}}. \tag{38}$$

The reference matrix of the benchmark model is obtained after repeatedly running the cycle composed of the above two equations until a certain convergence criterion is satisfied. The log-likelihood of the benchmark model is then given by

$$LL_{Bj} = \sum_{q=1}^{n} \{\gamma_{jq} \ln b_{jq} + (\phi_{jq} - \gamma_{jq}) \ln(1 - b_{jq})\}, \tag{39}$$

where ϕ_{jq} is the frequency of examinees who have responded to item j out of the examinees belonging to latent rank R_q, and γ_{jq} is the frequency of examinees who have correctly answered the item out of the ϕ_{jq} examinees under the benchmark model. There are two ways to compute these, based on the latent rank estimates or based on the RMPs, and these frequencies can be obtained by

$$\phi_{jq} = \sum_{i=1}^{N} z_{ij}\phi_{iq}, \tag{40}$$

$$\gamma_{jq} = \sum_{i=1}^{N} z_{ij}u_{ij}\phi_{iq}, \tag{41}$$

and

$$\phi^*_{jq} = \sum_{i=1}^{N} z_{ij}\pi_{iq}, \tag{42}$$

$$\gamma^*_{jq} = \sum_{i=1}^{N} z_{ij}u_{ij}\pi_{iq}, \tag{43}$$

respectively.

Next, the null model in the NTT can be easily determined, and it should be the model with one rank. The reference matrix of the null model then reduces to a vector with size n in which the j-th element is $v_{j1} = \sum_{i=1}^{N} z_{ij}u_{ij} / \sum_{i=1}^{N} z_{ij}$. That is, the log-likelihood of the null model becomes

$$LL_{Nj} = \sum_{i=1}^{N} z_{ij}\{u_{ij} \ln v_{j1} + (1 - u_{ij}) \ln(1 - v_{j1})\}. \tag{44}$$

Finally, the log-likelihood ratio chi-square for evaluating the model fitness of item j is obtained by

$$G_{Mj} = 2(LL_{Bj} - LL_{Mj}). \tag{45}$$

The statistic G_{Mj} is also called deviance and follows a χ^2 distribution with $df_j = n - Q$ degrees of freedom. Although the statistic must be positive, Equation (45) is rarely negative in the cases where the tested model better fits the data than the benchmark model does.

In cases where it is negative, the negative value should be replaced by zero. The statistic represents the extent of the deviance or the discrepancy by evaluating how badly the tested model for item j fits the data in comparison to the state where the benchmark model for item j better fits the data. Therefore, the smaller the statistic becomes, the better the model fitness will be. In addition, the model fitness of the whole test is obtained from the simple sum of the item model fitness. That is,

$$G_M = \sum_{j=1}^{n} G_{Mj} \tag{46}$$

This statistic also follows a χ^2 distribution with $df = \sum_{j=1}^{n} df_j$ degrees of freedom.

In addition, other indices are available for evaluating the model fitness. For example, the normed fit index (NFI; Bentler & Bonett, 1980), the relative fit index (RFI; Bollen, 1986), the incremental fit index (IFI; Bollen, 1989), the Tucker-Lewis index (TLI; Bollen, 1989), the comparative fit index (CFI; Bentler, 1990), the root mean square error of approximation (RMSEA; Browne & Cudeck, 1993), the Akaike information criterion (AIC; Akaike, 1987), the consistent AIC (CAIC; Bozdogan, 1987), and the Bayes information criterion (BIC; Schwarz, 1978) are computed by

$$NFI_j = 1 - \frac{G_{Mj}}{G_{Nj}}, \tag{47}$$

$$RFI_j = 1 - \frac{G_{Mj}/df_j}{G_{Nj}/df_{Nj}}, \tag{48}$$

$$IFI_j = 1 - \max\left\{0, \frac{\max(0, G_{Mj} - df_j)}{G_{Nj} - df_j}\right\}, \tag{49}$$

$$TLI_j = 1 - \max\left\{0, \frac{\max(0, G_{Mj}/df_j - 1)}{G_{Nj}/df_{Nj} - 1}\right\}, \tag{50}$$

$$CFI_j = 1 - \max\left\{0, \frac{\max(0, G_{Mj} - df_j)}{G_{Nj} - df_{Nj}}\right\}, \tag{51}$$

$$RMSEA_j = \sqrt{\frac{\max(0, G_{Mj} - df_j)}{df_j(N-1)}}, \tag{52}$$

$$AIC_j = G_{Mj} - 2df_j, \tag{53}$$

$$CAIC_j = G_{Mj} - df_j(\ln N + 1), \tag{54}$$

$$BIC_j = G_{Mj} - df_j(\ln N), \tag{55}$$

where

$$G_{Nj} = LL_{Bj} - LL_{Nj}, \tag{56}$$

and

$$df_{Nj} = n - 1. \tag{57}$$

These indices are popular in the SEM field. Using G_M, G_N $(= \sum_j G_{Nj})$, df, and df_N $(= \sum_j df_{Nj})$ in Equations (47)–(55), the fit indices for the whole test can be calculated.

3 Analysis

In this section, we look at four examples of NTT analysis applied to data from a geography test. The sample size was 5000, the number of test items was 35, and all the items were multiple-choice single-answer questions. The test was an achievement test, so missing data can be regarded as non-response in all cases. Accordingly, all missing data was treated as incorrect responses.

Figure 2 shows the frequency distribution for the number-right scores of the 5000 examinees, and Table 1 shows the descriptive statistics for the number-right scores. The figure along with the skewness and kurtosis from the table indicate that the number-right scores were close to normally distributed. In addition, Cronbach's (1951) alpha coefficient was about 0.7, and reliability cannot be said to be low for a test in which every item was binary.

3.1 Example 1: Result with $Q = 10$ by the ML method

In this subsection, we consider results for the ML method when the number of latent ranks was 10. The parameters required for the statistical learning were set to $(T,\ \alpha_1,\ \alpha_T,\ \sigma_1,\ \sigma_T) = (100,\ 1.0,\ 0.1,\ 1,\ 0.12)$. Table 2 shows the reference matrix estimate. The j-th row vector of the matrix is the item reference profile (IRP) of item j. Figure 3 shows the IRP of the respective items. The IRP of each item indicates the characteristics of the correct answer ratio of the item, and the ratio generally becomes larger with higher latent ranks, although we found that the IRP did not always monotonically increase, as shown by the IRPs of items 7 and 10. In addition, items 2, 5, 20, 32, and 33 were found to be very difficult, while items 3, 6, and 30 were very easy. In addition, the IRP of item 15 has a plateau around the mid-ranks.

In applications of the NTT as a test theory, the process to obtain estimates of the IRPs is essential because this process directly represents the test scaling. The scaling is an important part of the test standardization as well as the test equating. Each IRP expresses an individual characteristic of the item, and the IRP or the row vector of the reference matrix estimate

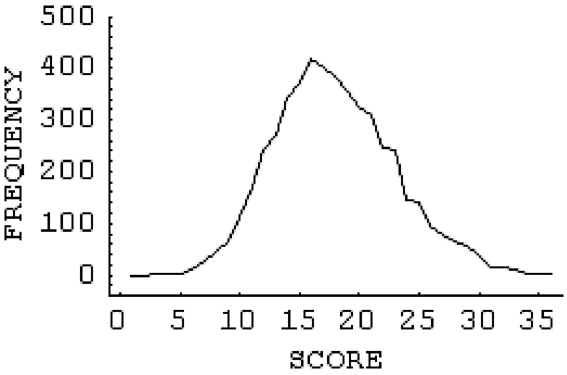


Figure 2: Number-Right Score Distribution

Table 1: Marginal Statistics of Number-Right Scores

Statistic	Value
N	5000
n	35
Median	17
Max	35
Min	2
Range	33
Mean	16.911
SD	4.976
Skew.	0.313
Kurt.	−0.074
Alpha	0.704

can be regarded as item parameters in the IRT. That is, each item is independent under the local independence assumption. Therefore, each item can be treated individually in the test editing. For example, if we select very difficult items, such as items 2, 5, 20, 32, and 33, from an item bank administered under the NTT, we can create a very difficult test.

Table 3 shows the model-fit indices based on the expected log-likelihood (or the RMP-based log-likelihood). Some items fit the data well, but others did not. The indices for each item are information that must also be maintained in the item bank along with the item's IRP, and items with very unsatisfactory fitness should be excluded from the candidates in the test editing. The NFI, RFI, IFI, TLI, and CFI become larger as the χ^2 statistic becomes smaller, and the maximum value of each index is 1.0. In addition, the IFI and CFI values tend to be larger than the NFI, RFI, and TLI values. Furthermore, the RMSEA

Table 2: Reference Matrix Estimate (ML, $Q = 10$)

Item	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	R_9	R_{10}
1	0.262	0.257	0.255	0.269	0.303	0.344	0.380	0.416	0.460	0.497
2	0.271	0.255	0.240	0.242	0.261	0.286	0.308	0.319	0.320	0.317
3	0.597	0.624	0.669	0.725	0.782	0.826	0.847	0.856	0.867	0.880
4	0.210	0.204	0.202	0.212	0.245	0.302	0.377	0.460	0.539	0.592
5	0.227	0.219	0.214	0.218	0.226	0.238	0.264	0.319	0.390	0.445
6	0.747	0.784	0.836	0.881	0.905	0.912	0.912	0.914	0.921	0.928
7	0.352	0.326	0.296	0.290	0.320	0.371	0.417	0.439	0.440	0.436
8	0.229	0.234	0.238	0.246	0.269	0.317	0.391	0.490	0.593	0.667
9	0.444	0.491	0.562	0.627	0.671	0.707	0.744	0.778	0.802	0.816
10	0.287	0.254	0.210	0.193	0.235	0.328	0.440	0.548	0.648	0.719
11	0.541	0.579	0.622	0.637	0.620	0.601	0.606	0.645	0.701	0.744
12	0.403	0.420	0.437	0.449	0.474	0.534	0.626	0.723	0.803	0.850
13	0.240	0.263	0.276	0.266	0.257	0.280	0.346	0.442	0.541	0.609
14	0.169	0.200	0.240	0.273	0.302	0.333	0.372	0.424	0.483	0.523
15	0.385	0.446	0.517	0.556	0.556	0.544	0.556	0.607	0.677	0.729
16	0.248	0.263	0.281	0.294	0.304	0.318	0.352	0.418	0.495	0.549
17	0.167	0.210	0.277	0.344	0.388	0.402	0.401	0.411	0.437	0.462
18	0.554	0.583	0.626	0.677	0.736	0.797	0.851	0.895	0.931	0.953
19	0.331	0.355	0.379	0.402	0.439	0.500	0.577	0.665	0.752	0.814
20	0.195	0.212	0.226	0.228	0.229	0.241	0.271	0.319	0.371	0.406
21	0.300	0.346	0.410	0.468	0.503	0.519	0.533	0.561	0.599	0.627
22	0.349	0.369	0.402	0.451	0.511	0.569	0.613	0.645	0.674	0.699
23	0.313	0.330	0.364	0.408	0.460	0.517	0.571	0.616	0.647	0.663
24	0.198	0.269	0.375	0.485	0.574	0.633	0.660	0.670	0.684	0.704
25	0.360	0.414	0.509	0.615	0.700	0.754	0.779	0.786	0.789	0.794
26	0.173	0.183	0.205	0.250	0.323	0.413	0.493	0.554	0.604	0.644
27	0.398	0.444	0.534	0.648	0.743	0.793	0.802	0.800	0.808	0.820
28	0.210	0.262	0.337	0.407	0.456	0.489	0.526	0.592	0.682	0.753
29	0.389	0.400	0.410	0.422	0.453	0.513	0.591	0.672	0.749	0.804
30	0.561	0.610	0.691	0.773	0.832	0.864	0.881	0.896	0.912	0.924
31	0.315	0.349	0.395	0.429	0.441	0.447	0.468	0.520	0.595	0.657
32	0.189	0.170	0.157	0.172	0.212	0.254	0.281	0.302	0.332	0.360
33	0.168	0.188	0.221	0.254	0.279	0.294	0.306	0.333	0.376	0.414
34	0.407	0.413	0.424	0.443	0.468	0.499	0.535	0.566	0.585	0.593
35	0.481	0.522	0.569	0.595	0.606	0.628	0.669	0.719	0.765	0.794

represents the non-centrality per degree of freedom and indicates a better model-fit as its value moves closer to 0. The index tends to be smaller as the χ^2 statistic becomes closer to 0. The information criteria — the AIC, CAIC, and BIC — are used for comparing the

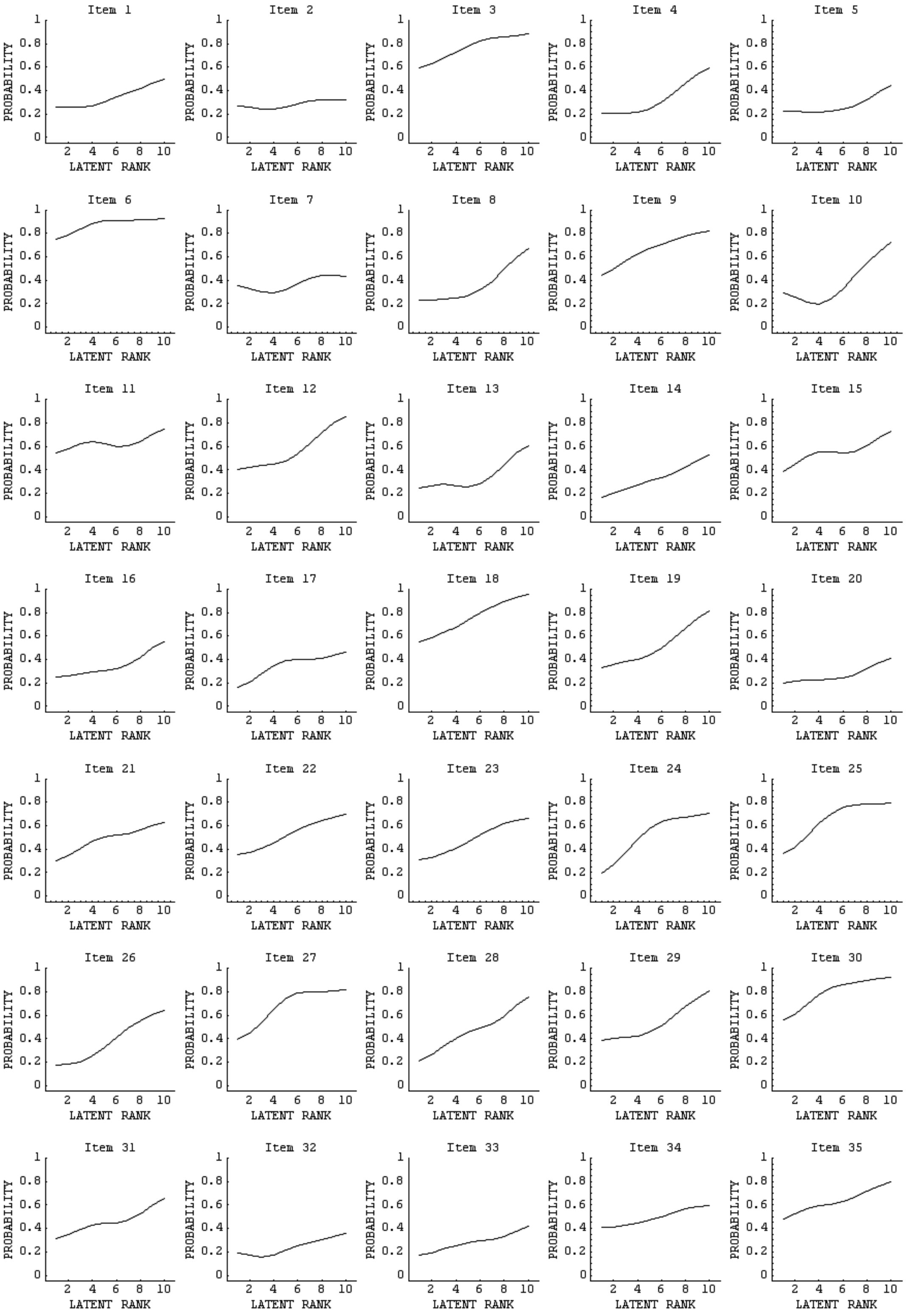


Figure 3: Item Reference Profile (ML, $Q = 10$)

Table 3: Item Fit Indices (ML, $Q = 10$)

Item	χ^2_{25}	NFI	RFI	IFI	TLI	CFI	RMSEA	AIC	CAIC	BIC
1	79.24	0.696	0.587	0.770	0.675	0.761	0.021	29.24	−158.69	−133.69
2	1.55	0.925	0.897	1.000	1.000	1.000	0.000	−48.45	−236.38	−211.38
3	45.63	0.871	0.824	0.937	0.912	0.935	0.013	−4.37	−192.30	−167.30
4	32.72	0.935	0.912	0.984	0.978	0.984	0.008	−17.28	−205.21	−180.21
5	81.49	0.694	0.584	0.766	0.669	0.757	0.021	31.49	−156.44	−131.44
6	31.02	0.854	0.801	0.968	0.954	0.966	0.007	−18.98	−206.91	−181.91
7	0.53	0.991	0.988	1.000	1.000	1.000	0.000	−49.47	−237.40	−212.40
8	46.49	0.926	0.899	0.964	0.951	0.964	0.013	−3.51	−191.44	−166.44
9	39.90	0.904	0.870	0.962	0.947	0.961	0.011	−10.10	−198.03	−173.03
10	13.94	0.981	0.975	1.000	1.000	1.000	0.000	−36.06	−223.99	−198.99
11	39.40	0.729	0.631	0.880	0.824	0.871	0.011	−10.60	−198.53	−173.53
12	67.67	0.898	0.862	0.933	0.908	0.932	0.018	17.67	−170.26	−145.26
13	26.09	0.940	0.919	0.997	0.996	0.997	0.003	−23.91	−211.84	−186.84
14	44.28	0.883	0.841	0.945	0.924	0.944	0.012	−5.72	−193.65	−168.65
15	61.18	0.808	0.739	0.877	0.827	0.873	0.017	11.18	−176.75	−151.75
16	50.77	0.832	0.771	0.907	0.869	0.904	0.014	0.77	−187.16	−162.16
17	22.24	0.916	0.886	1.000	1.000	1.000	0.000	−27.76	−215.69	−190.69
18	40.33	0.938	0.916	0.975	0.966	0.975	0.011	−9.67	−197.60	−172.60
19	83.80	0.885	0.844	0.917	0.885	0.915	0.022	33.80	−154.13	−129.13
20	39.16	0.782	0.704	0.909	0.868	0.903	0.011	−10.84	−198.77	−173.77
21	38.83	0.865	0.816	0.947	0.926	0.945	0.011	−11.17	−199.10	−174.10
22	28.84	0.919	0.890	0.988	0.984	0.988	0.006	−21.16	−209.09	−184.09
23	19.59	0.944	0.924	1.000	1.000	1.000	0.000	−30.41	−218.34	−193.34
24	61.93	0.918	0.889	0.950	0.931	0.949	0.017	11.93	−176.00	−151.00
25	10.59	0.981	0.975	1.000	1.000	1.000	0.000	−39.41	−227.34	−202.34
26	53.76	0.926	0.899	0.959	0.944	0.959	0.015	3.76	−184.17	−159.17
27	0.00	1.000	1.000	1.000	1.000	1.000	0.000	−50.00	−237.93	−212.93
28	119.89	0.853	0.801	0.880	0.835	0.879	0.028	69.89	−118.04	−93.04
29	84.86	0.856	0.804	0.894	0.854	0.892	0.022	34.86	−153.07	−128.07
30	67.08	0.883	0.841	0.923	0.894	0.922	0.018	17.08	−170.85	−145.85
31	102.37	0.727	0.629	0.779	0.692	0.773	0.025	52.37	−135.56	−110.56
32	24.86	0.848	0.794	1.000	1.000	1.000	0.000	−25.14	−213.07	−188.07
33	38.04	0.820	0.755	0.930	0.900	0.926	0.010	−11.96	−199.89	−174.89
34	34.75	0.742	0.649	0.911	0.868	0.903	0.009	−15.25	−203.18	−178.18
35	63.18	0.800	0.727	0.868	0.815	0.864	0.017	13.18	−174.75	−149.75

nested models, as explained in Example 2.

Next, Figure 4 shows the rank membership profiles (RMPs; Equation 25) of examinees 1–15 out of the 5000 examinees. The RMP is not possible without use of the ML method

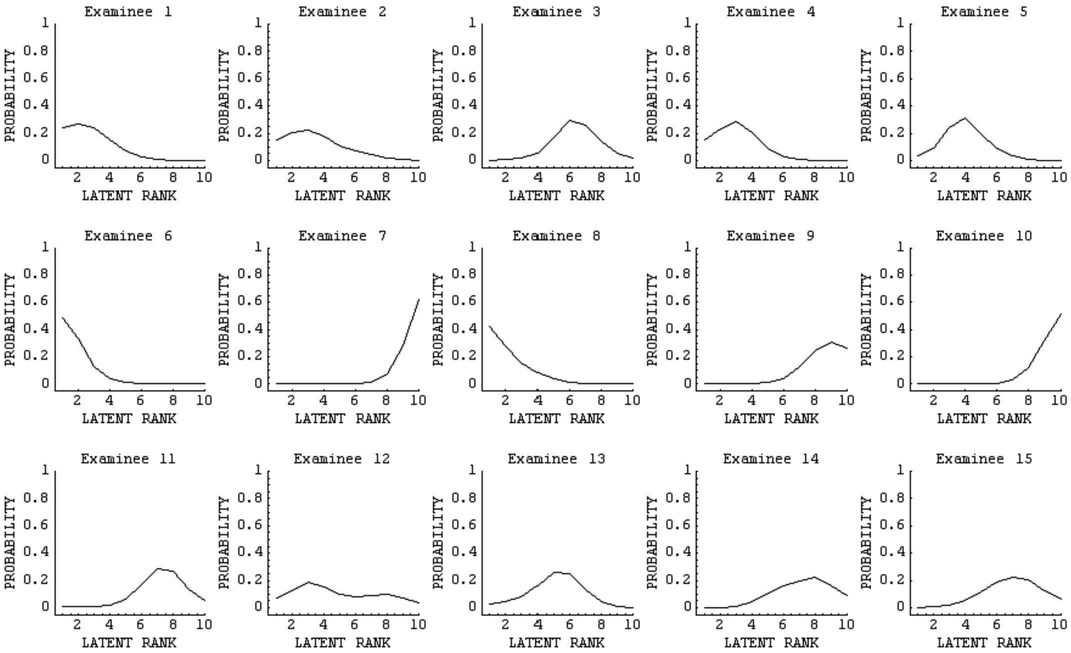
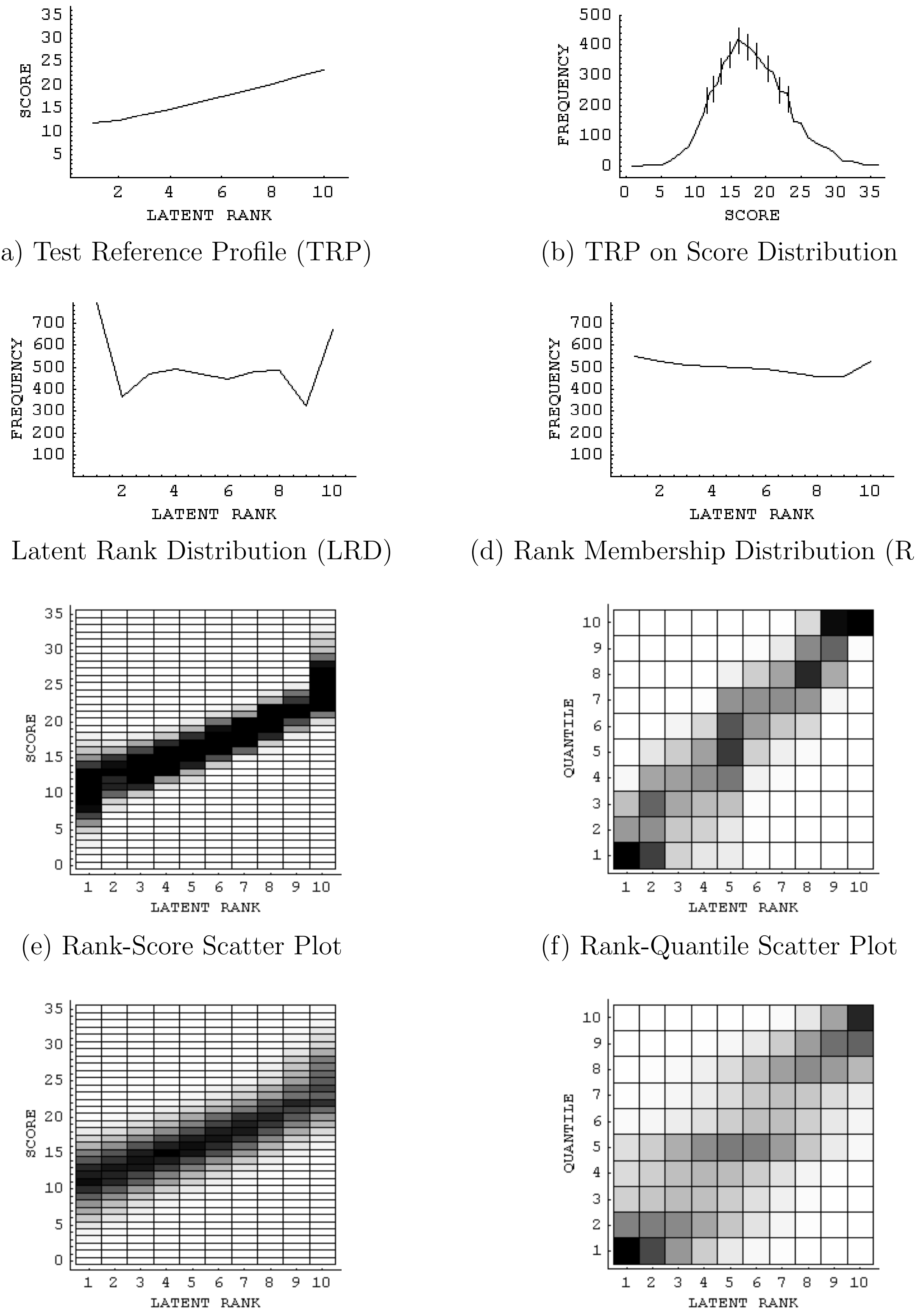


Figure 4: Rank Membership Profiles of Examinees 1–15 (ML, $Q = 10$)

as introduced in this study. The profile is useful for minutely examining the ability level of each examinee and provides information about the possibility of the examinee's latent rank estimate. That is, the RMP implies a certain reliability of the NTT scale. As the figure shows, the memberships of the same latent rank estimates were not always identical. For example, although the latent rank estimates of examinees 7 and 10 were the same, R_{10}, the probability that examinee 7 belongs to the latent rank is greater than that of examinee 10. In fact, the number-right score of examinee 7 is 28, while that of examinee 10 is 24. That is, the ability level of examinee 7 is also slightly higher than that of examinee 10 on the number-right score scale.

The results shown so far have concerned information about the respective items; Figure 5 shows analysis results regarding the whole test. Figure 5(a) shows the test reference profile (TRP; Equation 20). The obtained TRP was to be monotonically increasing, although some IRPs did not monotonically increase as shown in Figure 3. The monotonicity of the TRP is absolutely necessary to ensure the NTT scale ranks are ordered because without the monotonicity the NTT scale has no evidence showing that the scale is ordinal. In other words, the NTT scale can be ordinal if and only if the monotonicity requirement is satisfied. The state of the TRP being monotonic means that the "weakly ordinal alignment condition" is satisfied.



(a) Test Reference Profile (TRP)

(b) TRP on Score Distribution

(c) Latent Rank Distribution (LRD)

(d) Rank Membership Distribution (RMD)

(e) Rank-Score Scatter Plot

(f) Rank-Quantile Scatter Plot

(g) Membership-Score Scatter Plot

(h) Membership-Quantile Scatter Plot

Figure 5: TRP, LRD, RMD, and Scatter Plots (ML, $Q = 10$)

The condition is generally realized in most cases because the SOM mechanism used in the NTT is known to be similar to nonparametric and nonlinear principal component analysis (Ritter, Martinetz, & Schulten, 1992; Kohonen, 1995; Mulier & Cherkassky, 1995). However, the condition cannot be attained in rare cases. In such cases, the analysis is no longer successful because the obtained scale is not ordinal. Therefore, the analysis must be re-executed with different parameters. A directly effective way to satisfy the weak condition is to increase the values of parameters σ_1 and σ_T, which regulate the region of the learning propagation, because this change prevents the model from locally learning the data. Another effective way is to reduce the number of iterations, T, and lower the factors of the tension, α_1 and α_T, to realize the monotonic TRP.

Figure 5(b) shows the TRP ticked on the number-right score distribution. The ticks are clearly concentrated in the high density zone of the distribution. This zone also implies the target ability of the test. Every test has its own target ability to minutely measure because a test cannot cover the entire ability range. Although the target ability of the test is never clarified in detail unless each item content is carefully investigated, Figure 5(b) provides a key to revealing it.

Figure 5(c) shows the latent rank distribution (LRD; Equation 27). As stated by Shojima (2008), the frequencies of the ranks at both ends of the scale (R_1 and R_{10} in the analysis) become larger than those of the intermediate ranks. This property can be observed in the SOM field (e.g., Amari, 1980; Ritter & Schulten, 1986; Kohonen, 1995), so the NTT derives the property from the SOM. Figure 5(d) shows the rank membership distribution (RMD; Equation 28) which represents the distribution of the population, while the LRD shows that of the sample. Although the frequencies of the ranks at both ends in the RMD were also larger than those of the intermediate ranks, the tendency is suppressed in comparison to the LRD.

Figure 5(e) shows a scatter plot of the latent rank estimates and the number-right scores. Darker areas indicate that a larger frequency was observed in the respective areas. The higher latent rankers were generally higher number-right scorers, but the latent rank estimates of examinees with the same score were not always the same, and Spearman's rank correlation coefficient between the ranks and the scores was 0.929. Figure 5(f) shows a scatter plot of the latent rank estimates and the decile scores (the 10 percentile scores). It is obvious from the figure that the ranks of examinees with larger decile scores were estimated to be higher, although the ranks of the examinees with the same decile score were not always identical, as seen in Figure 5(e), and Spearman's correlation was 0.925. The largeness of the two correlation coefficients suggests a certain type of validity for the NTT scale. That is, the NTT scale is not very different from the number-right score and the quantile score scales. However, the direction of the NTT scale is not totally identical to those of the number-right score and the quantile score scales. This is because the weights of individual items in the NTT differ as expressed in the variety of the IRPs, while the weight of every item in the number-right score scale is one. Furthermore, Figures 5(g) and 5(h) show the RMDs stratified by, respectively, the number-right scores and the decile scores. These figures express the states for the population, while Figures 5(e) and 5(f) indicate those of the sample.

Table 4 shows the model-fit indices for the whole test. In general, these indices indicate that the fitness of the model to the data was satisfactory. The information criteria are discussed in Example 2 below.

Table 4: Test Fit Indices (ML, $Q = 10$)

Index	Value
χ^2_{875}	1596.01
NFI	0.889
RFI	0.848
IFI	0.946
TLI	0.925
CFI	0.945
RMSEA	0.013
AIC	−153.99
CAIC	−6731.53
BIC	−5856.53

3.2 Example 2: Result with $Q = 5$ by the ML method

The number of latent ranks is up to the analyst or the administrator of the test. For example, a school teacher will empirically know how many groups, based on learning progress, there are in a class. In addition, $Q = 3$ is desirable when a placement test is used to roughly grade the enrollees for the purpose of classifying them. The latent rank scale can be regarded as a continuous scale when the number of latent ranks becomes larger. However, the maximum number of latent ranks cannot be more than around 20 because a test is not reliable enough to minutely measure human abilities. Therefore, the realistic range of Q is $3 \leq Q \leq 20$. A junior high school teacher told the authors that $Q = 5$ is practical for explaining and objectifying the ranks. That is, the ranks of Excellent, Very Good, Good, Below Average, and Needs Improvements are generally used. Therefore, it is worthwhile to show the results with five ranks.

The data was analyzed using parameters identical to those in Example 1, except for the number of latent ranks. Figure 6 shows the IRPs of items 1–35. The IRPs with $Q = 5$ were more monotonic than those with $Q = 10$ (Figure 3) because the IRPs become simpler as Q decreases.

Tables 5 and 6 show the model-fit indices for the items and the test, respectively. The indices from the χ^2 statistic to the RMSEA where examined in the same way as for Example 1. The information criteria — the AIC, CAIC, and BIC — were exclusively used to relatively compare models, and are useful in the NTT to determine an adequate number of latent ranks. The models in Examples 1 and 2 are those with $Q = 10$ and 5, respectively. In general, a model with larger Q can more flexibly fit the data, but is more likely to overfit the data. In contrast, a model with very small Q does not fit the data very well. The information criteria are applied to judge which of the models with $Q = 10$ and 5 better fits the data from the viewpoint of efficiency because these criteria penalize both overfitness and underfitness to the data. The model with lower information criteria values will have a more appropriate number of latent ranks.

From Tables 4 and 6, we see that the AIC supported the model with $Q = 10$ (Example 1), and the CAIC and BIC judged the model with $Q = 5$ (Example 2) as more efficient. Therefore, the model with $Q = 5$ would be better in terms of the information criteria. In fact, the other model-fit indices can also be used for model comparison, and all the indices from the NFI to the RMSEA favor the model with $Q = 10$. It is difficult to choose only one model from the two, so the model-fit indices of the models with $Q = 6$, 7, 8, and 9 should also be examined. However, we recommended against determining the number of latent ranks based exclusively on the statistical indices. If an analyst judges that the examinees should be classified into five ranks, the data must be analyzed by the model with $Q = 5$. The statistical indices should be restricted to be used as a reference for model comparison.

Figure 7 shows the RMPs of examinees 1–15. In addition, Figure 8 shows the TRP, LRD, RMD and four scatter plots obtained from the analysis. These figures can be understood in in the same way as for the corresponding figures in Example 1.

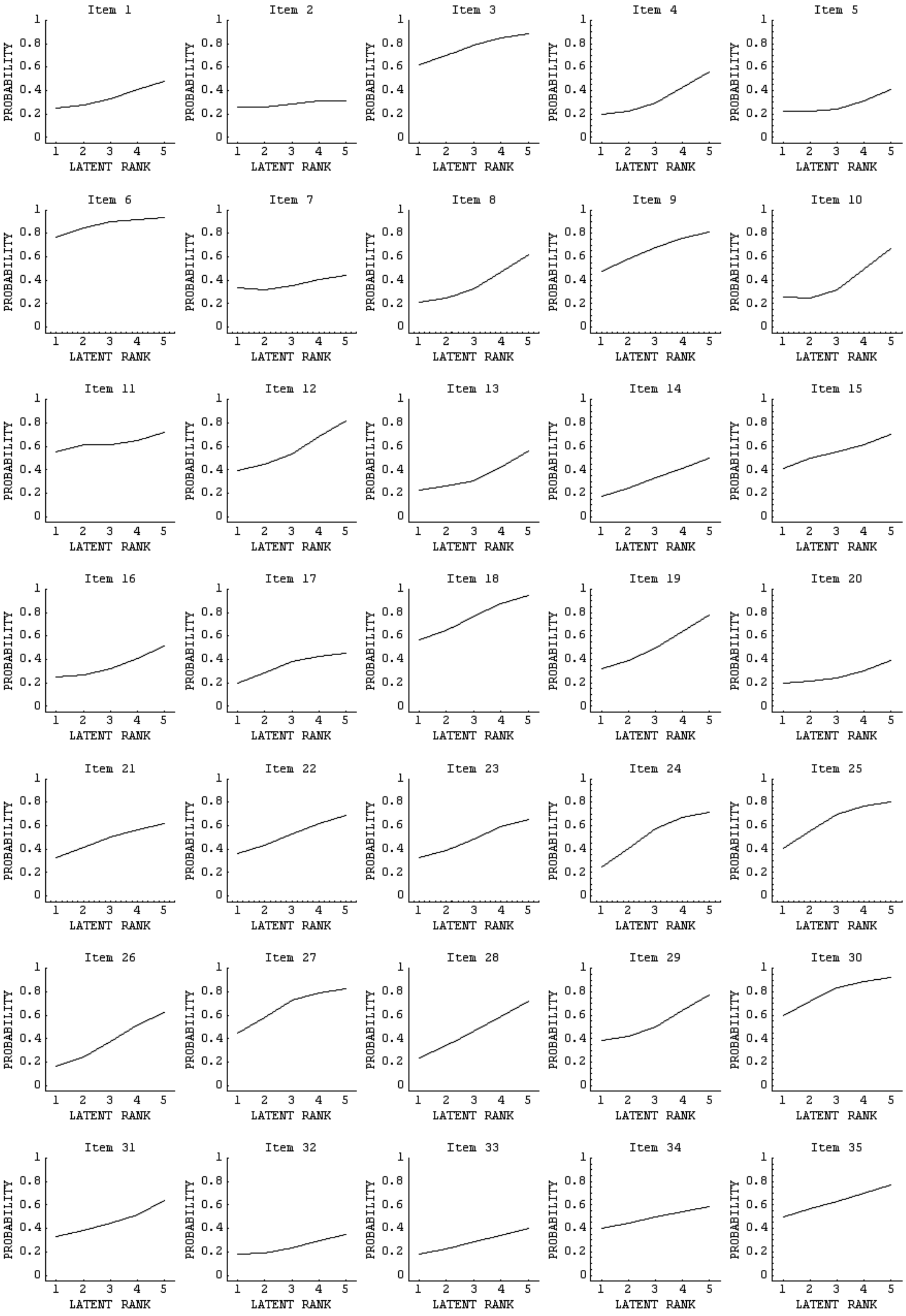


Figure 6: Item Reference Profile (ML, $Q = 5$)

Table 5: Item Fit Indices (ML, $Q = 5$)

Item	χ^2_{30}	NFI	RFI	IFI	TLI	CFI	RMSEA	AIC	CAIC	BIC
1	83.42	0.680	0.638	0.769	0.733	0.765	0.019	23.42	−202.09	−172.09
2	7.88	0.617	0.566	1.000	1.000	1.000	0.000	−52.12	−277.63	−247.63
3	60.68	0.828	0.805	0.905	0.891	0.904	0.014	0.68	−224.83	−194.83
4	56.01	0.889	0.874	0.945	0.937	0.945	0.013	−3.99	−229.50	−199.50
5	102.17	0.616	0.565	0.695	0.648	0.689	0.022	42.17	−183.35	−153.35
6	38.97	0.816	0.791	0.951	0.943	0.950	0.008	−21.03	−246.55	−216.55
7	13.81	0.773	0.743	1.000	1.000	1.000	0.000	−46.19	−271.71	−241.71
8	85.80	0.863	0.845	0.906	0.893	0.906	0.019	25.80	−199.71	−169.71
9	55.24	0.867	0.850	0.935	0.925	0.934	0.013	−4.76	−230.28	−200.28
10	112.41	0.850	0.830	0.886	0.870	0.885	0.023	52.41	−173.10	−143.10
11	54.80	0.623	0.573	0.785	0.747	0.777	0.013	−5.20	−230.71	−200.71
12	95.56	0.856	0.837	0.897	0.882	0.896	0.021	35.56	−189.96	−159.96
13	58.16	0.867	0.849	0.931	0.921	0.930	0.014	−1.84	−227.36	−197.36
14	39.38	0.896	0.882	0.973	0.969	0.973	0.008	−20.62	−246.13	−216.13
15	76.78	0.759	0.727	0.838	0.814	0.836	0.018	16.78	−208.74	−178.74
16	59.74	0.802	0.776	0.891	0.874	0.889	0.014	−0.26	−225.78	−195.78
17	40.39	0.847	0.827	0.956	0.949	0.955	0.008	−19.61	−245.12	−215.12
18	41.46	0.936	0.928	0.982	0.979	0.981	0.009	−18.54	−244.05	−214.05
19	89.09	0.878	0.862	0.916	0.904	0.915	0.020	29.09	−196.43	−166.43
20	40.15	0.777	0.747	0.932	0.921	0.930	0.008	−19.85	−245.36	−215.36
21	41.26	0.856	0.837	0.956	0.950	0.955	0.009	−18.74	−244.25	−214.25
22	41.02	0.885	0.870	0.966	0.961	0.966	0.009	−18.98	−244.49	−214.49
23	24.13	0.932	0.922	1.000	1.000	1.000	0.000	−35.87	−261.39	−231.39
24	90.86	0.880	0.864	0.916	0.905	0.916	0.020	30.86	−194.66	−164.66
25	57.23	0.899	0.886	0.949	0.942	0.949	0.013	−2.77	−228.29	−198.29
26	71.62	0.902	0.888	0.940	0.932	0.940	0.017	11.62	−213.90	−183.90
27	46.69	0.915	0.903	0.968	0.963	0.967	0.011	−13.31	−238.83	−208.83
28	120.98	0.852	0.832	0.884	0.868	0.884	0.025	60.98	−164.54	−134.54
29	100.97	0.829	0.806	0.873	0.855	0.872	0.022	40.97	−184.54	−154.54
30	91.20	0.841	0.820	0.888	0.872	0.887	0.020	31.20	−194.31	−164.31
31	102.33	0.727	0.691	0.791	0.760	0.788	0.022	42.33	−183.18	−153.18
32	41.42	0.747	0.714	0.915	0.900	0.912	0.009	−18.58	−244.09	−214.09
33	41.21	0.805	0.779	0.938	0.928	0.937	0.009	−18.79	−244.31	−214.31
34	33.90	0.748	0.715	0.963	0.956	0.961	0.005	−26.10	−251.61	−221.61
35	67.88	0.785	0.756	0.867	0.847	0.865	0.016	7.88	−217.64	−187.64

Table 6: Test Fit Indices (ML, $Q = 5$)

Index	Value
χ^2_{1050}	2184.61
NFI	0.847
RFI	0.827
IFI	0.914
TLI	0.902
CFI	0.914
RMSEA	0.015
AIC	84.61
CAIC	-7808.44
BIC	-6758.44

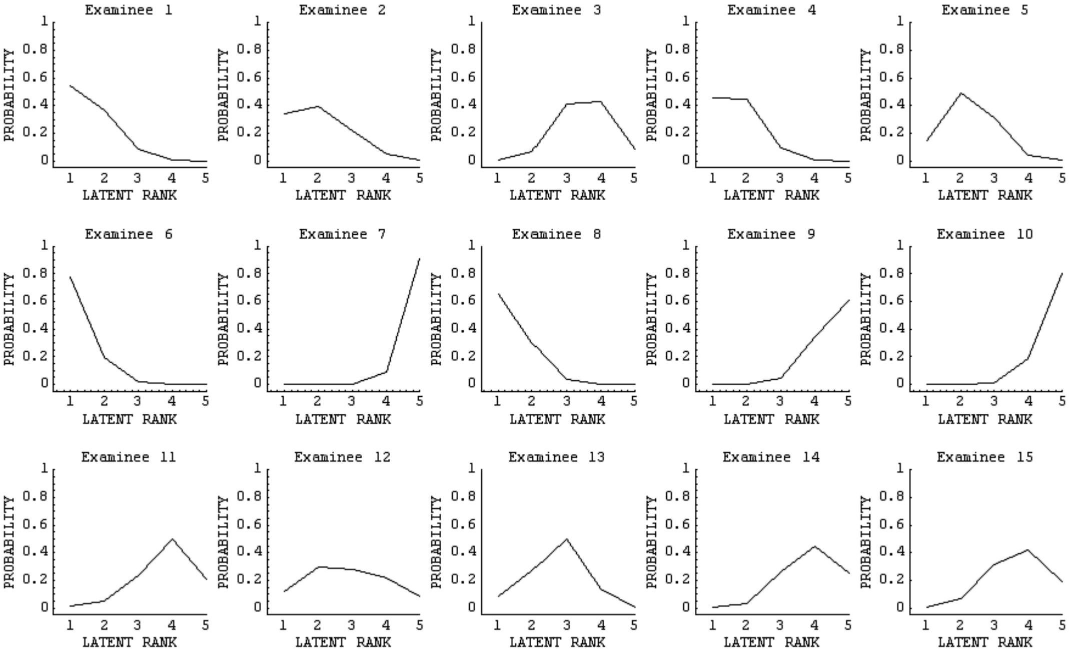
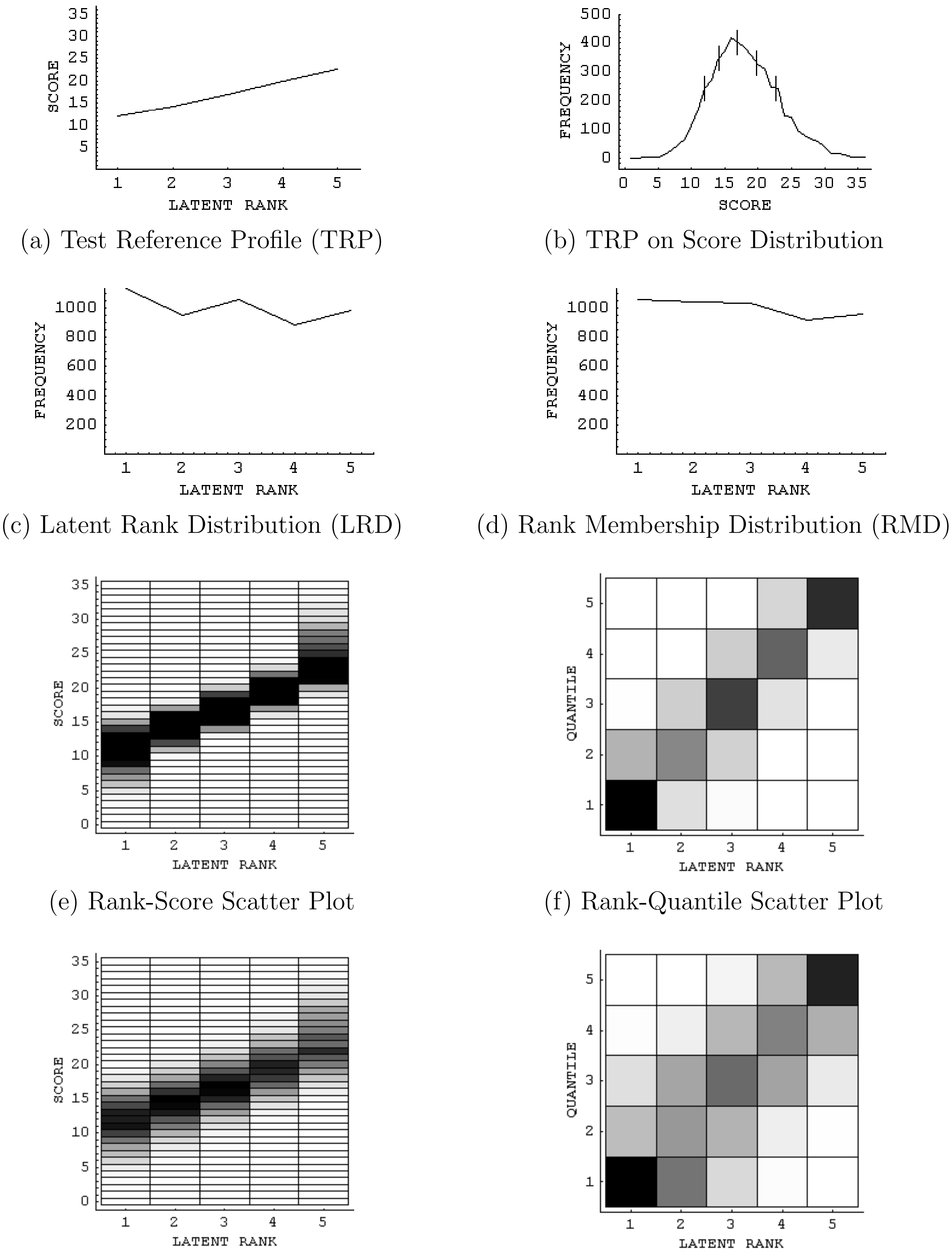


Figure 7: Rank Membership Profiles of Examinees 1–15 (ML, $Q = 5$)



(a) Test Reference Profile (TRP)

(b) TRP on Score Distribution

(c) Latent Rank Distribution (LRD)

(d) Rank Membership Distribution (RMD)

(e) Rank-Score Scatter Plot

(f) Rank-Quantile Scatter Plot

(g) Membership-Score Scatter Plot

(h) Membership-Quantile Scatter Plot

Figure 8: TRP, LRD, RMD, and Scatter Plots (ML, $Q = 5$)

3.3 Example 3: Result with $Q = 10$ and the MIC by the ML method

As seen in Figure 3, the IRP does not always monotonically increase. Although Figure 6 shows that the IRP tended to be monotonic as the number of latent ranks was small, the IRP cannot always be monotonic even in such a case. This is because the simplicity of the IRP shape does not always lead to monotonicity, and whether an obtained IRP is monotonic also depends on the behavior and nature of the data itself. However, some analysts or test administrators might feel a strong need to avoid the use of such non-monotonic items in practice. Accordingly, it will be useful to have a statistical learning procedure to obtain monotonic IRPs by imposing the monotonic increasing constraint (MIC). The simplest method is to sort the rank reference elements (RREs) or the IRP elements of each item according to their size. That is, adding the statements

$$\text{For } (j{=}1;\ j \leq n;\ j = j+1) \tag{58}$$

$$\text{— Sort } \boldsymbol{v}_j^{(t+1,0)}. \tag{59}$$

after Line (8) easily ensures the IRPs are monotonic. The test reference profile (TRP) always satisfies the monotonicity requirement when all the IRPs monotonically increase because the TRP is the weighted sum of the IRPs. The state in which both the TRP and all the IRPs are monotonic means that the “strongly ordinal alignment condition” is satisfied.

We did an analysis with the MIC is imposed on the IRPs and all other settings is identical to those in Example 1. Figure 9 shows the IRPs with the MIC and confirmed that all the IRPs monotonically increased. In addition, the MIC can be imposed on some selected items by slightly changing Line (59). The model-fit indices, RMP, LRD, and RMD were not significantly changed from those in Example 1, and further discussion of them is omitted.

3.4 Example 4: Result with $Q = 10$ by the Bayesian method

Examples 1–3 were results obtained using the ML method. As shown in Figure 5(c), the frequencies of the lowest and highest ranks tended to be larger than those of the intermediate ranks, although this tendency was weakened as the number of latent ranks became smaller as shown in Figures 8(c). In other words, the NTT scale obtained through the ML method was not an “equiprobability scale”. This phenomenon could have been due to clustering at both ends of the latent rank scale of examinees with ability levels significantly above or below the target ability. As mentioned, a test generally has a specific target ability, and

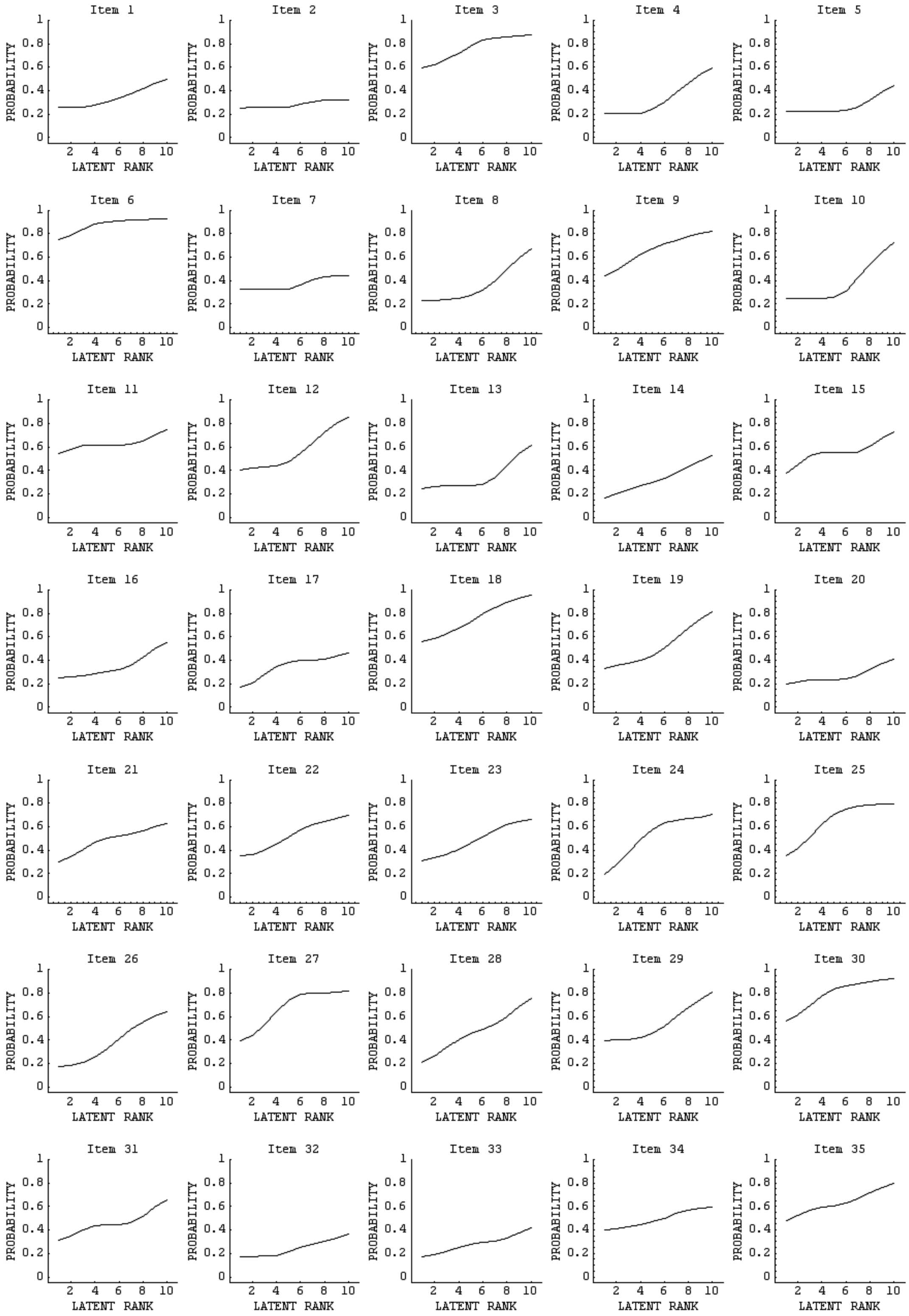


Figure 9: Item Reference Profile (ML, $Q = 10$, MIC)

it is impossible for a test to cover the entire ability range. However, some analysts or test administrators would like to classify examinees into nearly equal frequencies.

In fact, the non-equiprobability characteristic is derived from the SOM mechanism, and several researchers who have noticed the problem have sought to obtain an equiprobability map (e.g., Kohonen, 1995; Van Hulle, 2000). These efforts have mainly focused on not producing dead units in the map. Similar approaches do not appear promising under the NTT, though, because the dead units or dead ranks are rarely produced in the NTT scale. The appearance possibility of dead ranks containing no examinee increases as the number of latent ranks Q becomes larger, but Q is usually much smaller than the sample size. Consequently, another approach, different from theose proposed regarding the SOM, is needed to solve the non-equiprobability problem.

The Bayesian method is useful to obtain the equiprobability rank scale by assuming the prior distribution of each examinee's RMP to control the shape of the LRD and RMD. Assuming a trapezoidal prior distribution in which the prior probabilities for the lowest and highest latent ranks are smaller than those of the intermediate ranks is an effective way to lower the frequencies at both ends of the scale. That is, the prior distribution of each examinee's latent rank is given by

$$\boldsymbol{\pi} = \{\pi_q\} \ (Q \times 1) \tag{60}$$

$$\pi_q = \begin{cases} \pi & \text{if } q = 1, Q \\ (1-2\pi)/(Q-2) & \text{otherwise} \end{cases} \quad (\forall i \in N) \tag{61}$$

provided that $\pi \leq 1/Q$.

The data was analyzed under the same settings as in Example 1, except that we used the Bayesian method with $\pi = 0.085$. That is, the prior probabilities of latent rank R_1 and R_{10} were 0.0850, and those for the latent ranks from R_2 to R_9 were 0.1035. The IRPs and the model-fit indices were not drastically changed from those in Example 1, and they are not discussed here.

Figure 10 shows the RMPs of examinees 1–15 obtained by the Bayesian method. Although the differences between the RMPs by the ML and the Bayesian methods are very small, the posterior probabilities of the latent ranks at both ends of the scale in Figure 10 are certainly smaller than those in Figure 10. This is the effect by the trapezoidal prior distribution.

Fig. 11

Figure 11 shows the marginal distributions when the latent ranks are estimated by the Bayesian method. In contrast to the LRD and RMD obtained by the ML method (Figures 5(c) and 5(d)), the frequencies of the latent ranks at both ends of the scale in the LRD

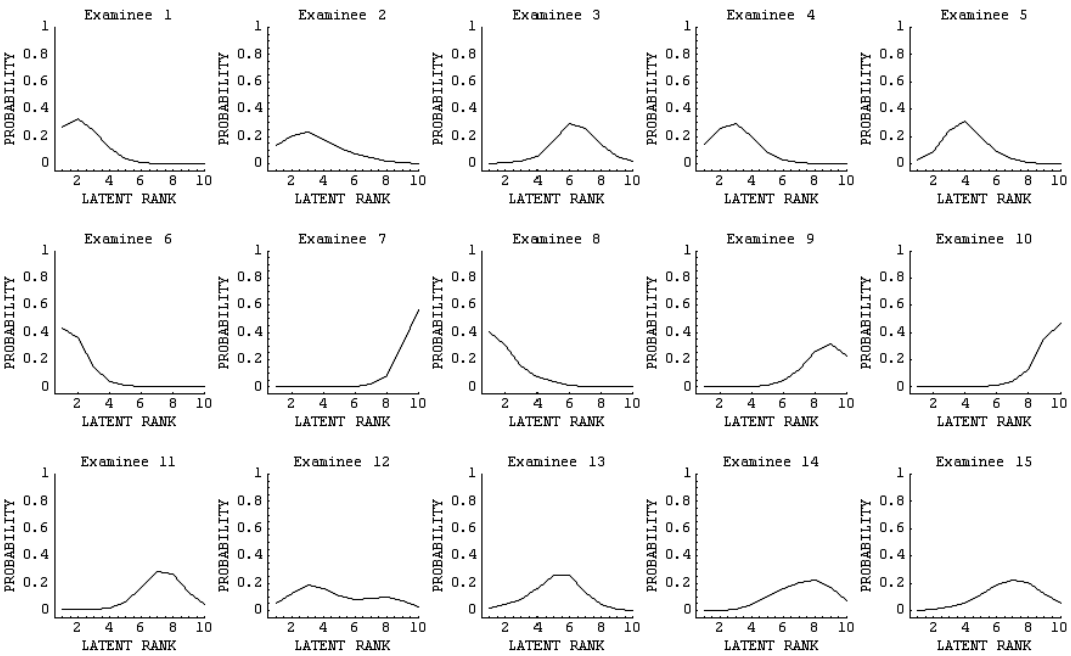


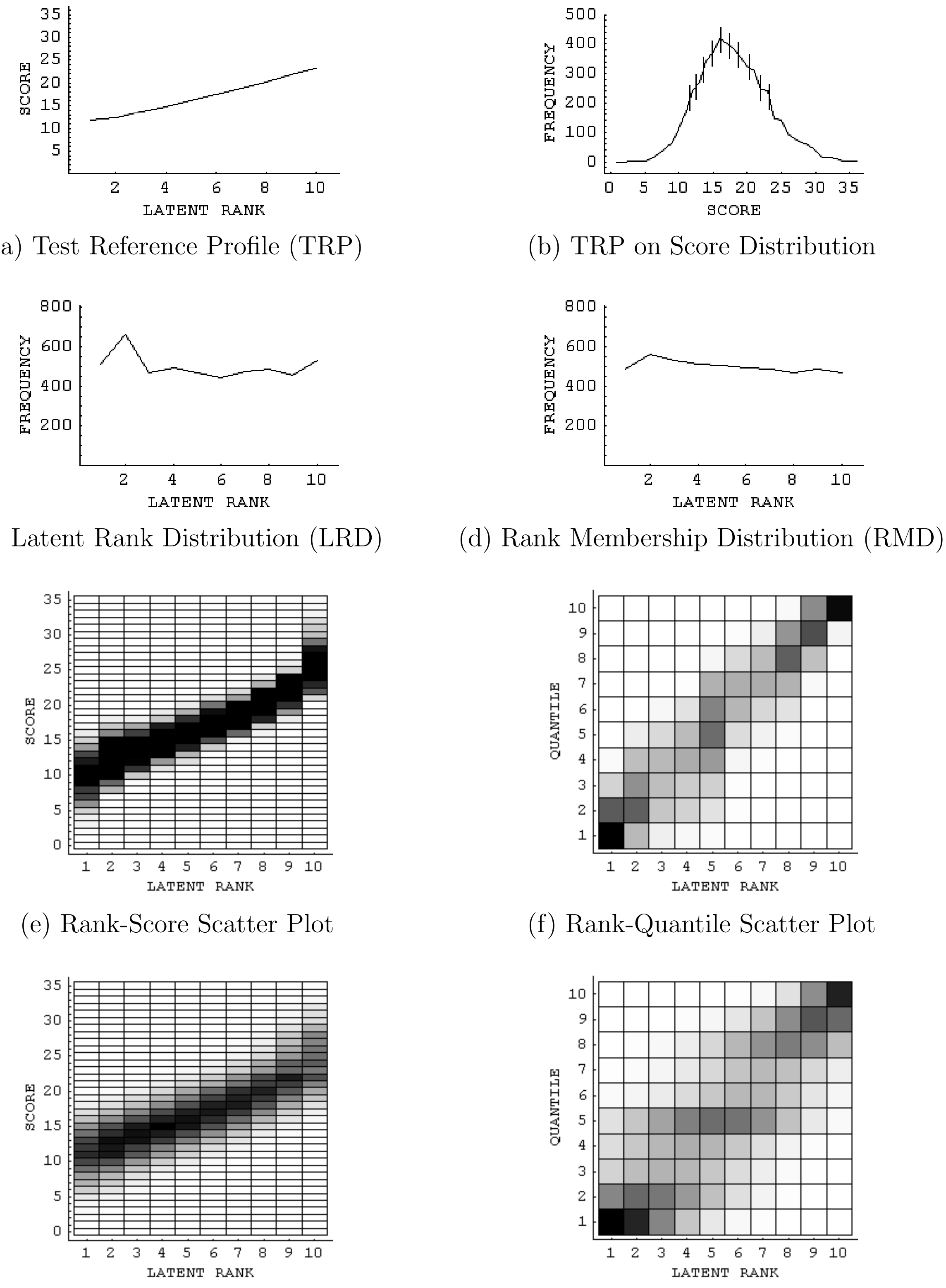
Figure 10: Rank Membership Profiles of Examinees 1–15 (Bayes, $Q = 10$)

and RMD obtained by the Bayesian method (Figures 11(c) and 11(d)) became smaller. The effect of the prior distribution became relatively larger as the number of items decreased.

4 Discussion

In this study, we introduced use of the ML and Bayesian methods in the statistical learning process, although the conventional SOM mechanism and the NTT estimation procedure by Shojima (2008) used the least-squares method. The ML and Bayesian methods are probabilistic measures between the model and the data. Accordingly, the rank membership profile (RMP), rank membership distribution (RMD), observation ratio profile (OR), and the model-fit indices could be defined. The model-fit indices are also useful for determining the number of latent ranks. Instead of introducing the statistical measure, the assumption of local independence was required to construct the likelihood, and research to examine the effect of this assumption is needed. A method for examining the standard error of the IRP is also required.

In addition, the NTT model is a statistical learning model, so the reference matrix estimate is different in each calculation. A batch-type model such as the NTT model estimated by the EM algorithm (Shojima, 2007b) must be a candidate for solving the problem. Fur-



(a) Test Reference Profile (TRP)

(b) TRP on Score Distribution

(c) Latent Rank Distribution (LRD)

(d) Rank Membership Distribution (RMD)

(e) Rank-Score Scatter Plot

(f) Rank-Quantile Scatter Plot

(g) Membership-Score Scatter Plot

(h) Membership-Quantile Scatter Plot

Figure 11: TRP, LRD, RMD, and Scatter Plots (Bayes, $Q = 10$)

thermore, the equiprobability scale could be obtained by the Bayesian method. Further researches is needed to investigate the appropriate strength and shape of the prior distribution in accordance with the number of items, the number of latent ranks, and the sample size.

The NTT is a test theory in which the latent scale is ordinal, and it can be relevant to the ordinal latent class model (Croon, 2002).[1] Research is also needed to investigate the theoretical relationship between the models.

The NTT scale is not a continuous scale divided into several regions; it is developed as ordinal from the beginning. Although the psychological variables might be continuously distributed, the tools for measureing them are not reliable enough for measurement ona continuous scale. Therefore, a latent rank model whose scale is ordinal, like that of the NTT model, will be valuable for such measurement.

References

Akaike, H. (1987) Factor analysis and AIC. *Psychometrika*, **52**, 317-332.

Amari, S. (1980) Topographic organization of nerve fields. *Bulletin of Mathematical Biology*, **42**, 339-364.

Bentler, P. M. (1990) Comparative fit indexes in structural models. *Psychological Bulletin*, **107**, 238-246.

Bentler, P. M. & Bonett, D. G. (1980) Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, **88**, 588-606.

Bishop, C. M., Svensen, M., & Williams, C. K. I. (1998) GTM: The generative topographic mapping. *Neural Computation*, **10**, 215-234.

Bollen, K. A. (1986) Sample size and Bentler and Bonnet's nonnormed fit index. *Psychometrika*, **51**, 375-377.

Bollen, K. A. (1989) A new incremental fit index for general structural equation models. *Sociological Methods & Research*, **17**, 303-316.

Bozdogan, H. (1987) Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, **52**, 345-370.

Browne, M. W. & Cudeck, R. (1993) Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.) *Testing structural equation models*. Sage. (pp. 136-162).

[1] The authors are grateful to Dr. Junker (Carnegie Mellon University) and Dr. Shigemasu (The University of Tokyo) for suggesting this during the meeting of Current Issues on Testing held at the National Center for University Entrance Examinations (Tokyo, Japan) in March 2008.

Cronbach, L. J. (1951) Coefficient alpha and the internal structure of tests. *Psychometrika*, **16**, 297-334.

Croon, M. (2002) Ordering the classes. In J. A. Hagenaars & A. L. McCutcheon (Eds.) *Applied latent class analysis*. Cambridge University Press. (pp. 137-162).

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.

Gorsuch, R. L. (1983) *Factor analysis*. Lawrence Erlbaum.

Hambleton, R. K. & Swaminathan, H. (1985) *Item response theory*. Kluwer-Nijhoff.

Harman, H. H. (1976) *Modern factor analysis*. University of Chicago Press.

Hastie, T., Tibshirani, R. & Friedman, J. (2001) *The elements of statistical learning:* . Springer-Verlag.

Jöreskog, K. G. & Sörbom, D. (1993) *LISREL 8: Structural Equation Modeling with the SIMPLIS Command Language*. Lawrence Erlbaum Associates.

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

Loehlin, J. C. (2003) *Latent variable models: An introduction to factor, path, and structural equation analysis*. Lawrence Erlbaum.

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

McLachlan, G. J. & Peel, D. (2000) *Finite mixture models*. Wiley.

Mislevy, R. J. & Bock, R. D. (1990) *Bilog 3: Item analysis and test scoring with binary logistic models*. Scientific Software International, Inc.

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.

Ritter, H. & Schulten, K. (1986) On the stationary state of Kohonen's self-organizing sensory mapping. *Biological Cybernetics*, **54**, 99-106.

Schwarz, G. (1978) Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461-464.

Shojima, K. (2007a) Artimage test theory. T. Otsu (Ed.) *Theory and practice for data analysis*. (pp. 57-85). (in Japanese)

Shojima, K. (2007b) Latent rank theory: Estimation of item reference profile by marginal maximum likelihood method with EM algorithm. *DNC Research Note*, 07-12.

Shojima, K. (2008) Neural test theory. In K. Shigemasu, A. Okada, T. Imaizumi, & T. Hoshino (Eds.) *New trends in psychometrics*. Universal Academy Press, Inc. (in press).

Titterington, D., Smith, A., & Makov, U. (1985) *Statistical analysis of finate mixture distributions*. John Wiley & Sons.

Van Hulle, M. M. (2000) *Faithful representations and topographic maps*. John Wiley & Sons, Inc.