Neural Test Theory

SHOJIMA Kojiro

September 2007

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

Neural test theory

Kojiro Shojima

Abstract

A new test theory, called the neural test theory (NTT), is proposed. Its model is based on the mechanism used in the self-organizing map (SOM), which is one of the neural network models. In the NTT model, latent space is expressed as a stream of connected nodes like a chain. Therefore, this latent scale is not a continuous one but more like a rank-ordered one. The number of ranks is the same as the number of nodes, and each examinee is ranked on the latent rank scale. The item reference profile (IRP) obtained by NTT for each item is very effective for assessing the probability of a correct answer, which is the counterpart of the item characteristic curve in item response theory. In addition, the monotonically increasing IRP can be obtained by imposing adequate constraints on the statistical learning process. An analysis example that illustrates some usages and features of NTT is shown.

Key words: test theory, neural network model, self-organizing map, nonparametric item response theory.

ニューラルテスト理論

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) という新しいテスト理論を提案した．それは，自己組織化マップのメカニズムを応用している．NTT モデルでは，潜在空間は，ノードが一本の連結した鎖のように表現されているため，連続尺度というよりも順序尺度である．順位数は，ノードの数であり，各サンプルは，潜在順序尺度のいずれかに布置される．NTT によって得られる項目参照プロフィール (item reference profile, IRP) は，ノンパラメトリック項目反応理論のような柔軟性をもち，各順位での当該項目の正答率を表現しており，項目反応理論における項目反応関数に相当する．また，統計的学習過程に適切な制約を課すことで，単純増加の IRP を得ることもできる．さらに，実データ解析を行い，NTT のいくつかの使い方と特性を示した．

キーワード: テスト理論，ニューラルネットワークモデル，自己組織化マップ，ノンパラメトリック項目反応理論．

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

1 Introduction

The latent scale assumed in item response theory (IRT; e.g., Lord, 1980; Hambleton & Swaminathan, 1985) is continuous. The continuous scale used in almost all IRT applications is a 1-dimensional space on which the transition of the correct answer rate for each item is expressed, and the latent trait or ability level of each examinee is located.

However in Japan, educational evaluation on a continuous scale is driving students to become adept at getting the highest possible score. For example, such skills include techniques for finding the correct answer without reading the lead sentences. In fact, some school teachers advise such approaches in their lessons. A test is a public tool (Shojima, 2007a), so its existence is expected to have salutary repercussions on society. For example, tests should motivate members of society in terms of self-discipline and self-realization. Shojima (2007b) called the "context of existence" the third role that a test must play.

In addition, as many test practitioners already know, a test does not have high enough sensitivity to discriminate the difference between two persons who have nearly equal abilities, whereas a weighing machine can distinguish the slight difference between two persons who are almost the same weight. In other words, the resolution of a test is lower than that of a weighing machine. Generally speaking, the reliability of a test cannot be very high. The most that a test can do is to rank examinees into several groups. Therefore, we need a test theory that can locate examinees not on a continuous scale but on a rank-ordered one.

In this study, we propose a test theory in which the assumed latent scale is rank-ordered. This theory, called the neural test theory (NTT), uses the mechanism of the self-organizing map (SOM; Kohonen, 1995) which is a kind of neural network model. Furthermore, SOM is a statistical learning theory that is frequently used in market research, and the conventional one is a clustering method where similar samples are located adjacent to each other on a 2-dimensional lattice.

2 Method

First of all, let us assume a latent rank scale, where the number of ranks is Q. Each latent rank is represented by Node R_q $(q = 1, \cdots, Q)$, where the ability of the examinee located at Node R_{q+1} is higher than that of the examinee at R_q $(q = 1, \cdots, Q-1)$. Let us also assume that the number of items is n and that Node R_q has a n-dimensional vector $\boldsymbol{v}_q$, called the reference vector. The latent rank scale for $(Q, n) = (7, 12)$ is shown in Figure 1. Here, big black circles are the nodes that represent latent ranks, and small gray circles

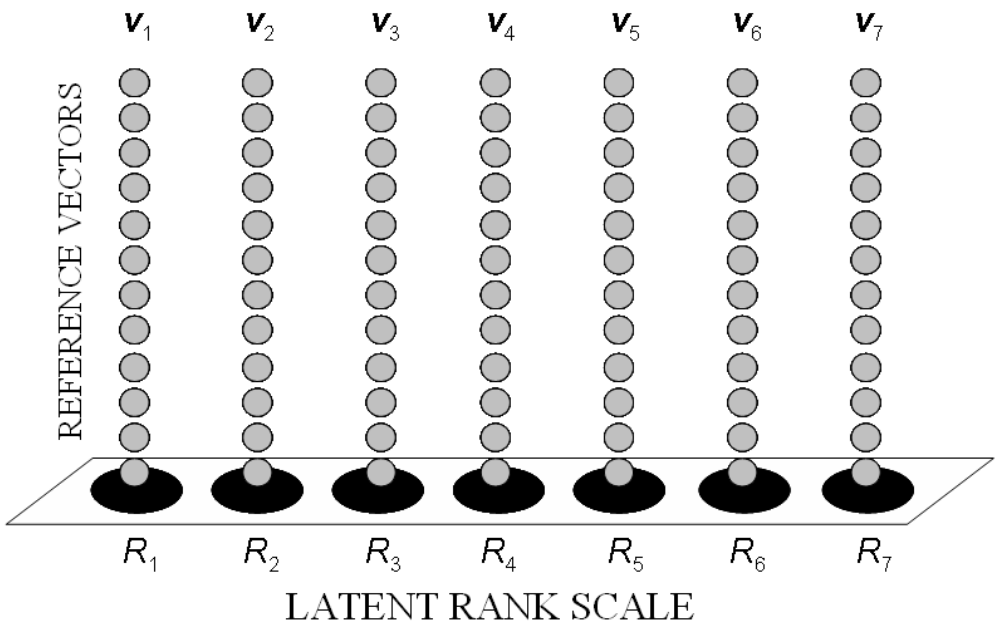


Figure 1: Latent Rank Scale and Reference Vectors

above each node is the reference vector.

Let us further suppose that the sample size is N and that the response data of examinees is $\boldsymbol{U} = \{\boldsymbol{u}_i\}$ $(i = 1, \cdots, N)$. In addition, assume that $\boldsymbol{v}_q^{(t)}$ is the reference vector of R_q at the t-th period, and the recommended initial value is

$$\boldsymbol{v}_q^{(1)} = q\mathbf{1}/Q. \tag{1}$$

The basic idea of the computational procedure is identical to that of SOM. The procedure for NTT is outlined below.

For (t=1; $t \leq T$; $t = t + 1$) (L1)

— Obtain $\boldsymbol{U}^{(t)}$ by randomly sorting the row vectors of $\boldsymbol{U}$. (L2)

For (h=1; $h \leq N$; $h = h + 1$) (L3)

— Input $\boldsymbol{u}_h^{(t)}$, the h-th row vector of $\boldsymbol{U}^{(t)}$, and select the winner with the closest reference vector in terms of the discrepancy function d. (L4)

— Obtain $\boldsymbol{V}_h^{(t)}$ after updating the reference vectors of the winner and neighboring nodes. (L5)

— $\boldsymbol{V}^{(t+1)} \Leftarrow \boldsymbol{V}_N^{(t)}$ (L6)

(L1) requires a series of Lines (L2)–(L6) to be repeatedly executed until t approaches

T. Similarly, (L3) indicates that Lines (L4) and (L5) should be repeatedly computed while $h \leq N$. (L3) is nested in (L1), and once Lines (L4) and (L5) have been executed N times (h counts from 1 to N), the counter t is then incremented by one. This process continues until $t = T$.

The essence of the statistical learning in NTT is reflected in the process of updating the reference vectors. First, to randomly sort the row vectors of $\boldsymbol{U}$ in Line (L2) is to cancel the sequential effect of the input data. Second, the winner node determined in Line (L4) depends on what discrepancy function d is selected. We recommend the square of the Euclidian distance as the discrepancy function because it is frequently used as d in many SOM applications. When $\boldsymbol{u}_h^{(t)}$ is input, the winner node R_w by the square distance is determined as follows:

$$R_w : \ w = \arg\min_{q \in Q} ||\boldsymbol{v}_q^{(t)} - \boldsymbol{u}_h^{(t)}||^2. \tag{L4'}$$

Next, the idea of updating the reference vectors is basically the same as the process used in the conventional SOM. That is, the reference vectors of the nodes that are geographically closer to the winner should be designed to become numerically closer to the input data. For updating $\boldsymbol{v}_{qh}^{(t)}$, the reference vector of Node R_q when $\boldsymbol{u}_h^{(t)}$ is input at the t-th period, one of the valid candidates is as follows:

$$\begin{aligned} &\text{For } (q{=}1;\ q \leq Q;\ q = q+1) \qquad\qquad (\text{L5a}) \\ &\text{— } \boldsymbol{v}_{qh}^{(t)} = \boldsymbol{v}_{qh-1}^{(t)} + h_{qw}(t)(\boldsymbol{u}_h^{(t)} - \boldsymbol{v}_{qh-1}^{(t)}) \end{aligned}$$

where

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

$$\alpha_t = \frac{T-t+1}{T}\alpha_1, \tag{3}$$

and

$$\sigma_t = \frac{(T-t)\sigma_1 + (t-1)\sigma_0}{T-1}. \tag{4}$$

The second term of (L5a) is the size of the reference vector's approach to the input data. In addition, h_{qw} in (2) is called the "tension"; it regulates the geographically closer nodes to the winner to have a greater size of updating the values of the reference vectors. The factor α_t in (3) is the parameter regulating the impact of the tension at the t-th period: the larger the value of α_t, the more sensitive the reference vector updates. As t increases, α_t, which is α at the t-th period, decreases linearly from α_1 (> 0), the initial value of α, to $\alpha_T = \alpha_1/T$.

In addition, σ_t in (4) is the function determining the neighborhood area around the winner, and the larger the value of σ_t, the further from the winner the statistical learning propagates. As t increases, the function σ_t decreases linearly from its initial value σ_1 (> 0), and approaches σ_0 at the T-th period ($\sigma_1 > \sigma_0 > 0$).

The parameters $(T, \alpha_1, \sigma_1, \sigma_0)$ must be set before the computation begins, and the calculation is ended when t approaches T. Alternatively, we can use a certain stopping rule to exit the statistical learning process. An efficient candidate is as follows:

$$C^{(t)} = \sum_{h=1}^{N} ||\boldsymbol{u}_h^{(t)} - \boldsymbol{v}_{wh}^{(t)}||^2, \tag{5}$$

where $\boldsymbol{v}_{wh}^{(t)}$ is the reference vector of the winner for $\boldsymbol{u}_h^{(t)}$. For example, we can stop computing when a certain criterion is satisfied, such as $C^{(t)} < crit.$ or $|C^{(t+1)} - C^{(t)}| < crit.$

3 Analysis

3.1 Example 1

The result of a world history test analyzed by NTT are shown in this section. The number of items was 36 and the sample size was 2049. All items were multiple-choice single-answer questions. The score distribution of the test is shown in Figure 2. The parameters were set to $(T, \alpha_1, \sigma_1, \sigma_0) = (100, 0.1, Q, 1.0)$ and the number of latent ranks was 10.

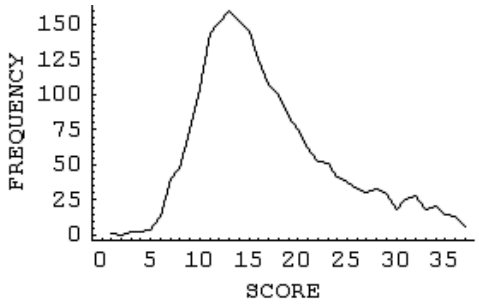


Figure 2: Score Distribution

The results are shown in Figure 3. These plots, called item referenece profiles (IRPs), are of the itemwise reference vectors of the finally obtained $\boldsymbol{V}^{(T)}$, $\boldsymbol{v}_j^{(T)}$ $(j = 1, \cdots, n)$. Each IRP is useful to interpret the behavior of the correct answer rate at each latent rank. The IRPs are not always monotonically increasing, but in general, the higher the latent rank, the higher the correct answer rate. Also, Item 2 was found to be difficult because the correct

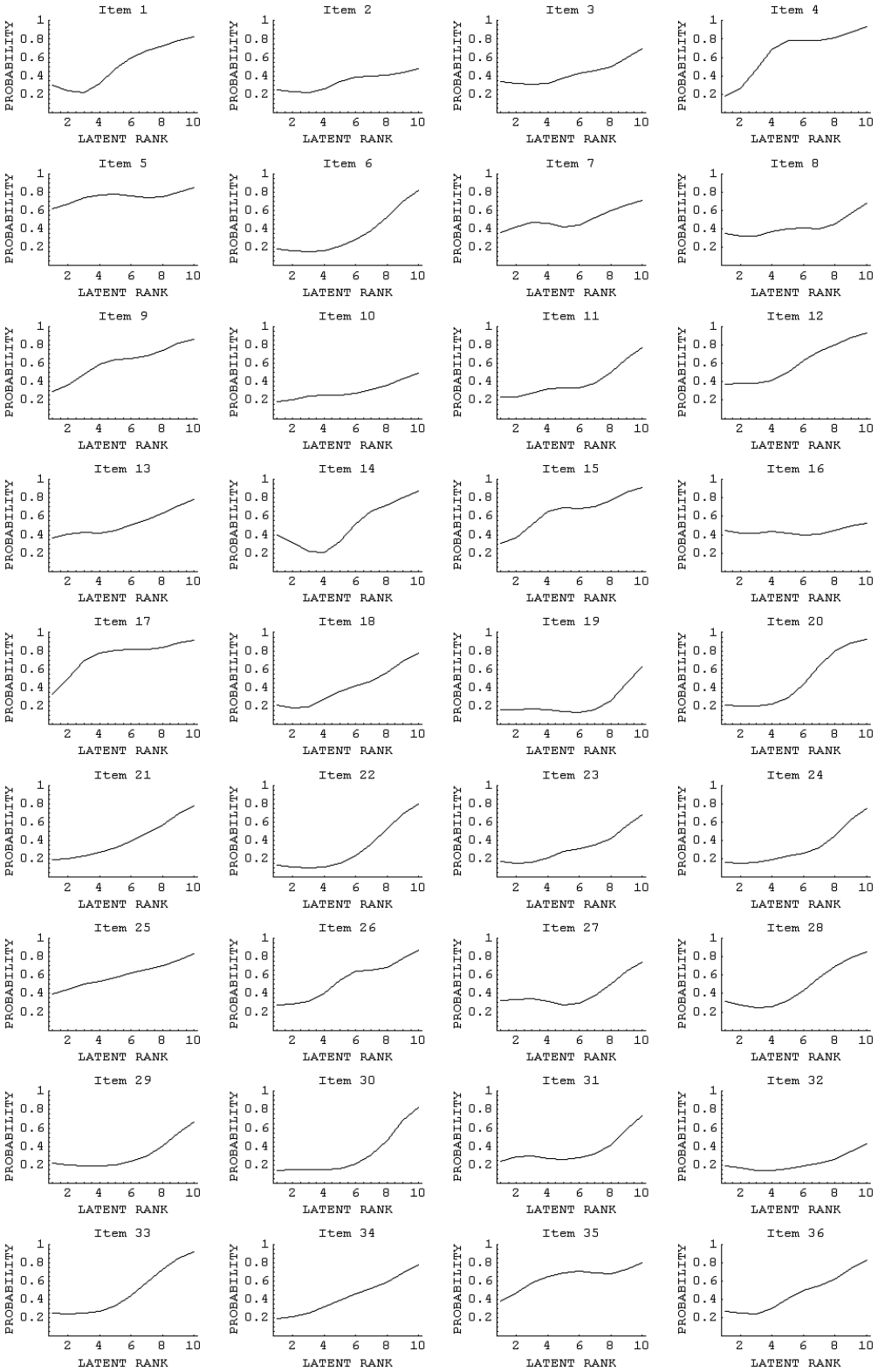
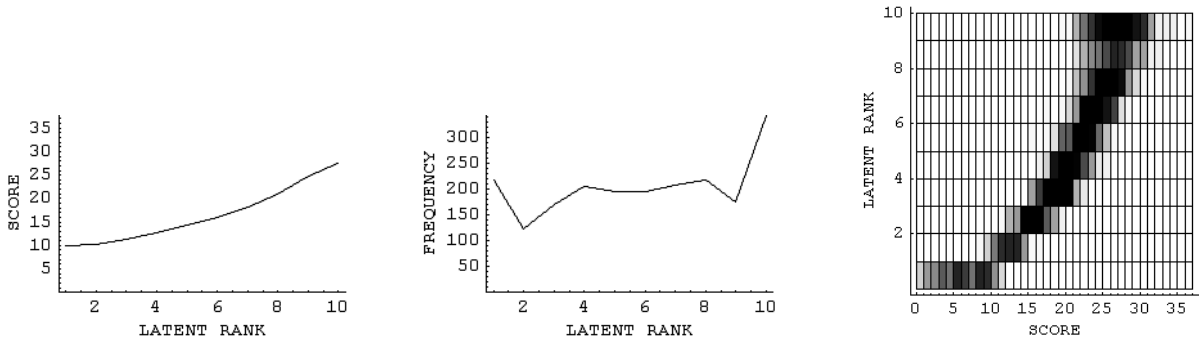


Figure 3: Item Reference Profiles



(a) Test Reference Profile (b) Latent Rank Distribution (c) Scatter Plot

Figure 4: TRP, LRD, and Scatter Plot of Scores and Ranks

answer rate of the examinees, even those located in the highest rank (Rank 10), did not exceed 0.5. On the other hand, Item 5 was very easy because the IRP was higher than 0.6 throughout the latent ranks. In addition, the slopes of Items 4 and 20 were sharp. Such items can be said to have high discriminancy.

Figure 4(a) is the simple sum of the 36 IRPs, called the test reference profile (TRP); it is useful for understanding the expected score at each latent rank. The TRP monotonically increases as the latent rank becomes higher. The TRP obtained by the NTT is computed to be almost exclusively monotonically increasing because the algorithm of the NTT is based on that of the SOM, and the principal dimension is extracted by the SOM algorithm (Ritter, Martinetz, & Schulten, 1992; Kohonen, 1995; Mulier, & Cherkassky, 1995). That is, SOM is a kind of nonlinear principal components analysis. However, the distances between nodes are not necessarily found to be equal. Therefore, the latent scale of the NTT is posteriori formed to be rank-ordered.

Figure 4(b) is the latent rank distribution. It is natural for latent ranks not to be distributed normally. The latent rank of each examinee is identical to the winner node computed by (L4'). Every test has its own target ability, and they cannot measure precisely outside the target ability. The test analyzed here was not very difficult for the examinees, so, the latent ranks of examinees whose scores were higher than around 26 were equally estimated to be 10. That is, the NTT minutely ranked the examinees whose abilities were within the target ability of the test. On the other hand, it put the examinees whose abilities were outside the target ability of the test into nodes at the two ends. This tendency was also observed in the SOM, as reported by Amari (1980), Ritter & Schulten (1986), and Kohonen (1995).

In addition, Figure 4(c) is the scatter plot of the scores and latent ranks. The darker

the area, the higher the density. Clearly, the estimated ranks of examinees whose scores are equal are not always the same.

3.2 Example 2

This section shows results when the number of ranks was five. The number of latent ranks is up to the analyst or test administrator. The IRPs were obtained as shown in Figure 5 provided that the other parameters were set to be equal to those in Section 3.1. Due to space limitation, IRPs of only Items 1–12 are shown in the figure.

Although the basic shapes are similar, it is clear that IRPs for $Q = 5$ are smoother than those for $Q = 10$ in Figure 3. The obtained IRPs tended to be monotonically increasing when the number of latent ranks was smaller. In addition, the TRP, the latent rank distribution, the scatter plot of the scores and the latent ranks are shown in Figure 6.

3.3 Example 3

It is clear from Figure 3 that not all the IRPs are monotonically increasing. Such items might be difficult for test administrators to use in practice even if they express the real state of the data. Therefore, it is useful to impose the constraint of monotonic increase in the process of the statistical learning. For example, we can simply add the following steps after Line (L6) to make BCRPs monotonic. That is,

$$\begin{aligned} &\text{For } (j{=}1;\ j \leq n;\ j = j+1) \qquad \text{(L5b)}\\ &\quad \text{For } (q{=}1;\ q \leq Q-1;\ q = q+1)\\ &\quad \text{— If } v_{q+1,j}^{(t+1)} \leq v_{qj}^{(t+1)}, \text{ then } v_{q+1,j}^{(t+1)} = v_{qj}^{(t+1)}. \end{aligned}$$

or

$$\begin{aligned} &\text{For } (j{=}1;\ j \leq n;\ j = j+1) \qquad \text{(L7)}\\ &\text{— Sort}(\boldsymbol{v}_j^{(t+1)}). \end{aligned}$$

With the same parameter setting as used in Sections 3.1 and 3.2, the world history test data were analyzed for $Q = 10$ and (L7). The IRPs obtained for Items 1–12 are shown in Figure 7. All IRPs were computed to be monotonically increasing, and the higher the latent rank was, the higher the correct answer rate became. As a practical matter, the IRPs under the monotonic increase constraint are preferable for test practitioners. The NTT can satisfy such a practical demand. In addition, Figure 8 shows the TRP, the latent rank distribution, the scatter plot of the scores and the latent ranks under the monotonic increase constraint.

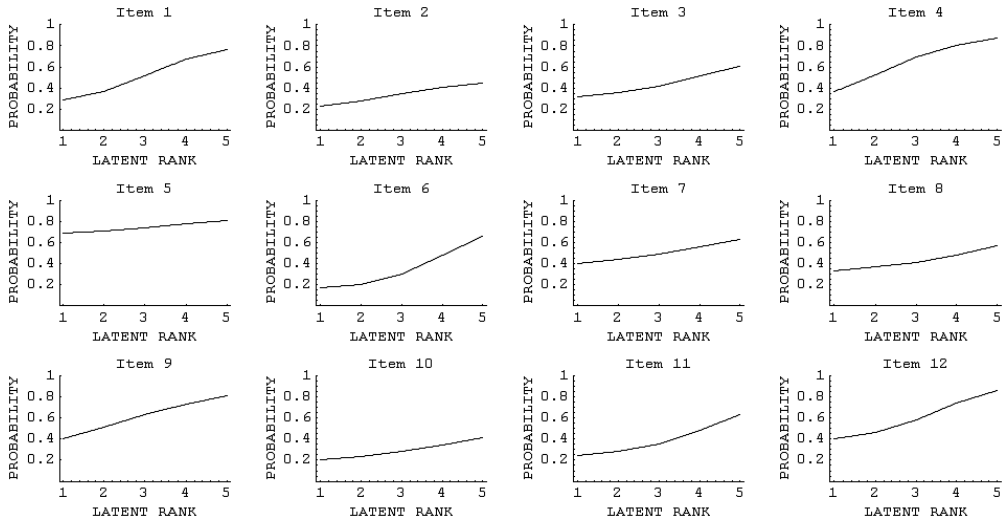
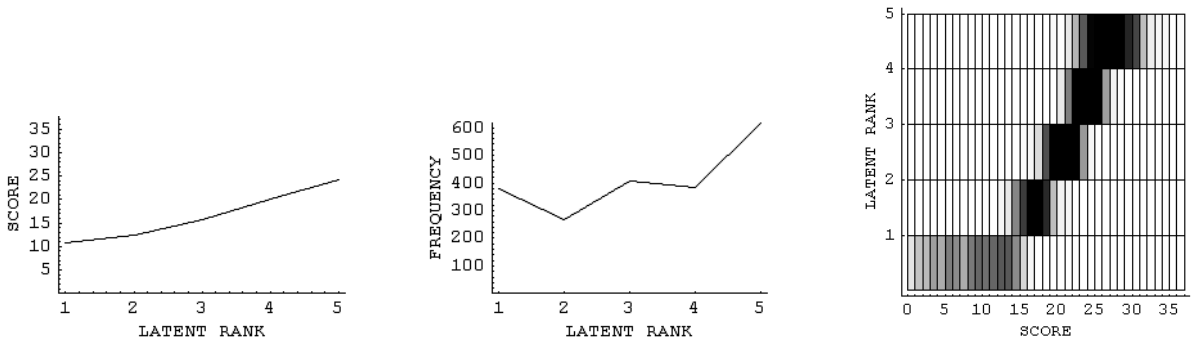


Figure 5: IRPs with $Q = 5$ of Items 1–12



(a) TRP ($Q = 5$) (b) LRD ($Q = 5$) (c) Scatter Plot ($Q = 5$)

Figure 6: TRP, LRD, and Scatter Plot of Scores and Ranks ($Q = 5$)

4 Discussion

The neural test theory (NTT) is a kind of nonparametric test theory because no mathematical formulation are imposed to shape item reference profiles. It is also a neural network model because nodes lined up like a chain are artificial neurons that represent the latent ranks. NTT uses the computation algorithm of the self-organizing map (SOM). The SOM usually prepares a 2-dimensional lattice when mapping samples, but NTT plots examinees on a 1-dimensional lattice (chain). That is, NTT can be said to be a reformed 1-dimensional SOM. The 1-dimensional SOM was theoretically studied by Ritter & Schulten (1986) and Kohonen (1995), but it has rarely been investigated since 2000.

The most significant difference from the existing test theories, such as item response

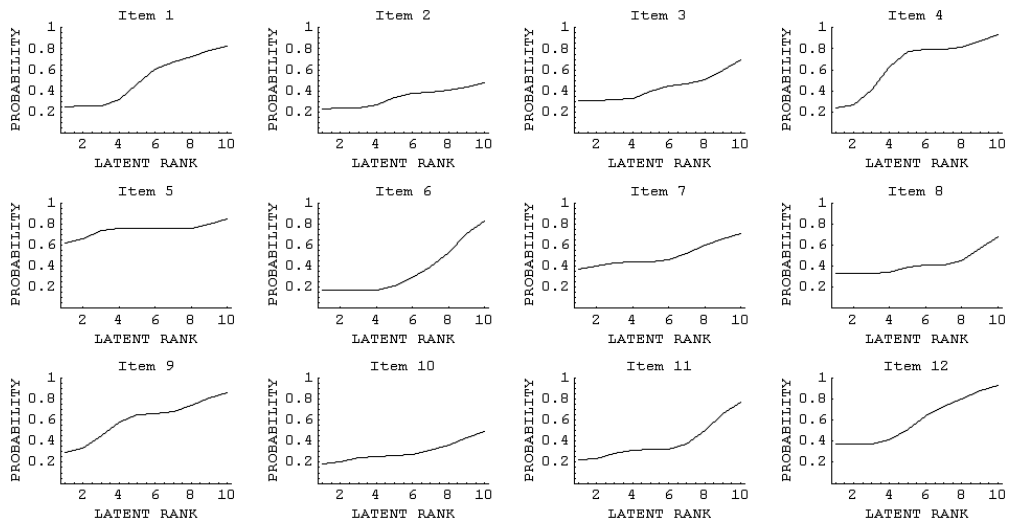
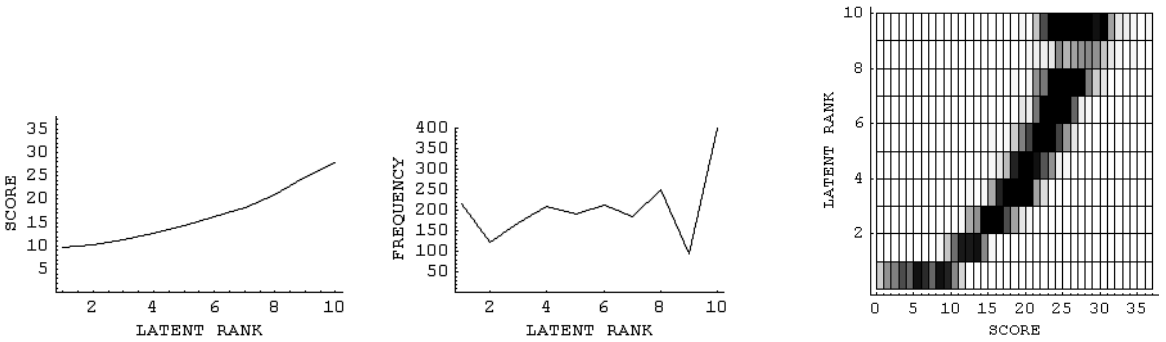


Figure 7: Monotonically Increasing IRPs of Items 1–12



(a) TRP (Mono. Inc.) (b) LRD (Mono. Inc.) (c) Scatter Plot (Mono. Inc.)

Figure 8: TRP, LRD, and Scatter Plot of Scores and Ranks (Mono. Inc.)

theory (IRT) or classical test theory (CTT), is that the latent scale assumed in NTT is rank-ordered. The continuous scale assumed in IRT or CTT can be used like a rank-ordered scale when it is partitioned into several zones. Such a usage of the latent scale is not always irrational; it may sometimes be useful to feed the test results back to examinees. However, the continuous scale itself is not necessary in the first place because tests cannot discriminate the slight difference between examinees who have nearly equal abilities. The most that a test can do is to rank examinees into several grades.

NTT can be easily extended to be multidimensional and to deal with missing data and polytomous response data. In addition, NTT is useful in test editing, equating, and computerized adaptive testing. However, further studies to gather experience and knowledge about NTT are required to make the theory useful for test practitioners.

References

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

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

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

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

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.

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

Shojima, K. (2007a) Scholastic achievement structure of National Center Test 2006 by self-organizing map. *Japanese Journal for Research on Testing*, **3**, 161-178. (in Japanese with English abstract)

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