一般研究

ニューラルテスト理論の応用可能性
－方法論的課題の考察と小論文データへの多値型モデルの適用例－

Applicability of neural test theory
–Discussing methodological problems and applying a polytomous model to essay test–

宇佐美　慧
（東京大学大学院教育学研究科・日本学術振興会）

Satoshi Usami
Graduate School of Education, University of Tokyo
Japan Society for the Promotion of Science

ニューラルテスト理論の応用可能性
－方法論的課題の考察と小論文データへの多値型モデルの適用例－

宇佐美　慧
（東京大学大学院教育学研究科・日本学術振興会）

近年，教育測定・心理計量の領域で話題になっている統計手法の一つにニューラルテスト理論（Neural Test Theory：NTT）がある．NTT では，受験者を離散的な潜在ランクに割り当てる点が一つの大きな特徴である．NTT は現段階では，潜在ランクおよび項目参照プロファイルの推定精度，項目サンプリングを超えた潜在ランクの推定値の一貫性，異なる最適化基準に基づく推定値の比較，分析モデルの構築・改良などの方法論的課題を中心に検討すべき課題があった．本論文ではまず，項目サンプリングを超えた潜在ランクの推定値の一貫性について，順序データ用の多値型 NTT を用いた方法と，項目和得点を基準に潜在ランクの度数分布が一様になるように分割した方法との比較をシミュレーションにて検討した．そして，実際の小論文データへの多値型 NTT の適用例を示し，また項目和得点や項目反応理論に基づく分析結果との簡単な比較検討を行った．
キーワード　：　ニューラルテスト理論　順序データ　小論文試験　教育測定

Applicability of neural test theory
–Discussing methodological problems and applying a polytomous model to essay test–

Satoshi Usami
Graduate School of Education, University of Tokyo/ Japan Society for the Promotion of Science

Neural test theory (NTT) is a data analysis method that has gradually become popular in educational measurement and psychometrics. In NTT, examinees are clustered to a discrete latent rank. The author noted several technical issues for future researches of NTT, such as accuracy of estimates for latent ranks and ICRP, consistency of estimates for latent ranks over item sampling, comparison among several optimizing criteria, and construction and improvement of algorithm for NTT models. In the present research, the author performed a simulation study by using polytomous NTT for ordered data, to compare consistency of estimates for latent ranks over between NTT and another method using total test score. Finally, a real data example for essay test data was shown by using polytomous NTT, and the author compared these results with methods based on item response theory and total test score.
Keywords：neural test theory, ordered data, essay test, educational measurement

1 問題と目的

1.1 はじめに

ニューラルテスト理論（Neural Test Theory：NTT）は，自己組織化マップ（Self-organizing Map：SOM, Kohonen, 1995）のメカニズムを利用した，テストデータを分析する為の潜在ランク理論であり，荘島(2007)によって提案された．NTT モデルにおいては，各項目の各カテゴリに対する選択確率を表す項目参照プロファイル（Item Category Reference Profiles：ICRP）が潜在ランクごとに推定され，同時に受験者は離散的な潜在ランク上に配置される(e.g., Shojima, 2008)．

この手法を用いる根拠として，荘島(2007)や Shojima(2008)では，受験者を連続尺度上に正確に配置できるほど一般にテストの信頼性は高くないという方法論的な観点からの理由を挙げている．また，Shojima(2008)では心理学や教育学，社会学や行動科学においては，学力や性格特性，態度などの潜在特性を測定する場合が多く，それらの信頼性を考慮すると，潜在変数に対して潜在ランクを念頭に置いた分析手法が望ましいと述べている．

NTT は当初，ICRP や潜在ランクの推定のための最適化基準として，最小二乗基準を用いた方法が提案されていたが（荘島, 2007），Shojima(2007b, 2007c, 2008)では最尤基準とベイズ基準を用いた推定方法が示されている．最尤基準の利用により，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)を受験者ごとに推定することができ，受験者がどの潜在ランクに所属するかを確率的に評価できるのが利点である．また，最尤基準を導入したことにより NTT モデルの適合度を多角的な指標を用いて検討することができ，これは潜在ランク数を決定する上でも有用である．このように，NTT は推定法の議論を中心に徐々に理論的な発展を遂げ，実用的にも注目されつつある手法と言えよう．

1.2 NTT の方法論的課題

NTT は既に一部で実用化がなされているが，NTT に関する研究自体の歴史はまだ極めて浅い．これまでの議論も推定法に関する議論が主体であり，その理論や実践において生じうる様々な課題については，今後多くの研究者や実務家を交えたより広汎な議論が待たれるところである．NTT において今後検討すべき課題としては，例えば潜在ランクおよび ICRP の推定精度の検証，項目サンプリングを超えた潜在ランクの推定値の一貫性の検証，異なる最適化基準から得られる推定値の比較，分析モデルの構築・改良などに関する方法論的課題がまず挙げられる．

NTT のアルゴリズムは自己組織化マップの手法を援用しており，ICRP の要素について適当な初期値を設定した上で，ある最適化基準のもとに反復的に要素の更新を行い，任意の反復回数に達した場合か一定の収束基準に達したと判断された場合に計算を終了する．このような反復的なアルゴリズムを用いた中で，推定される潜在ランクや ICRP が，データの性質に応じてどの程度の推定精度を有しているかという問題がある．この点については，NTT モデルから発生されたシミュレーションにより検討することが可能ではあるが，潜在ランクの数や度数分布，反応カテゴリ数や ICRP の要素をいかに設定するかという問題があり，シミュレーションの実行そのものの難しさがある．

また，推定精度の問題だけでなく，潜在ランクの推定値が項目サンプリングを超えてどの程度一貫しているのか，という問題もある．すなわち，同じ項目領域からサンプリングされたデータから推定される潜在ランクが，データの性質に応じてどの程度一貫しているかという問題である．この点の検証についても，上と同様の問題が生じうるが，例えば項目反応データを用いたシミュレーションによる間接的な検討を行うことは比較的容易である．

他にも，上記２つの問題において，最適化基準の違いと推定精度や一貫性の関連も重要な問題である．さらに，そもそも NTT を用いずに，項目和得点や項目反応理論を用いて得られた θ を任意のパーセンタイルを基準に分割してランクを形成する簡便法では対処できないかという疑問も生じうるが，この点も応用上検討すべき重要な課題であると言えよう．

分析モデルについても，これまで二値型データ・多値データを扱う為のモデルが検討されており（荘島, 2007；Shojima, 2007a），プログラムも一部公開されているが，利用が手軽で，データ数や潜在ランク数・カテゴリ数・最適化基準・初期値などの設定も柔軟に対応できるプログラムが必要であり，既存の分析アルゴリズムについても若干の改良の余地がある．

他にも，NTT においては学習させるデータの順番がランダムに設定されていることから，同じデータを用いたときの潜在ランクの推定値の一貫性の問題や（橋

本・荘島, 2008)，また等化法に関する議論（荒井・橋本・荘島, 2008）も，上述の議論と平行して検討すべき方法論的課題と言えるだろう．

1.3 本論文の目的

前小節を踏まえ，本論文では，項目サンプリングを超えた潜在ランクの推定値の一貫性の検証を，多値型の NTT モデルを用いて行うことを一つ目の目的とする．具体的には，まず既存の順序データ用の多値型 NTT モデルを，より汎用的に利用できるように若干の改良を加える．そして，項目反応モデルから発生させた多値型データをもとに，データに影響を与えるノイズの大きさや項目数・被験者数・母集団の分散・項目識別力などを変化させた上で，潜在ランクの推定値の一致率の比較を行う．比較は，NTT モデルにおける最小二乗基準・最尤基準・ベイズ基準の三つの最適化基準と，項目和得点の度数分布が一様になるように均等に割り当てた均等分割基準で行う．また，項目反応理論を用いて得られた θ に対して均等分割する基準との比較も興味深いが，今回のシミュレーションでは項目反応モデルから発生させたデータを用いる為に θ に対する均等分割に対して有利なシミュレーションになる恐れがあること，また項目反応理論を用いた推定には項目反応モデルや推定法の設定によっていくらかの任意性が残ることから，この方法に対しては別途適切なシミュレーションのもとで行うのが妥当であると判断し，今回のシミュレーションでは直接比較を行わないことにした．

次に，実際の小論文データに対して多値型 NTT の適用例を示すことを二つ目の目的とする．具体的にはまだ応用事例の少ない多値型 NTT において，分析結果の解釈の具体例を示し，また NTT と項目和得点や θ の均等分割基準に基づく分析結果の比較検討を行い，NTT を用いた推定値の特徴を明らかにしたい．

2 推定法

本節では，順序データ用の NTT の推定アルゴリズムについて述べる．Shojima(2007a)では，順序データ用のアルゴリズムが既に構築されているが，本論文においてはそのより汎用的な利用を目的として，若干の改良を行っている．本節ではアルゴリズムを先に紹介し，具体的な改良点については後で説明することとする．また，二値型のアルゴリズムや，最尤基準を用いた場合の適合度指標に関する議論は Shojima(2008)を参照のこと．

まず以下に，後の議論に登場する各記号の意味をここで前もって記しておく．

Q ：潜在ランク数　　N ：受験者数
n ：項目数　　K ：カテゴリ数
$\boldsymbol{D}$ ：サイズ $N \times n$ の順序データ行列
$\boldsymbol{V}$ ：サイズ $(K \times n) \times Q$ の参照行列
$\boldsymbol{U}$ ：サイズ $N \times (K \times n)$ の，反応したカテゴリを示すダミー変数からなる行列
$\boldsymbol{Z}$ ：サイズ $N \times (K \times n)$ の，$\boldsymbol{U}$ に対応する要素の欠損の有無を示す行列
T ：学習回数

本論文では，議論を単純化する為，カテゴリ数は全ての項目において等しく K とする．まず最初に，その要素に 1 から K の値が含まれている多値の順序データ行列 $\boldsymbol{D}$ をもとに，列の要素の数を K 倍したサイズ $N \times (K \times n)$ の行列 $\boldsymbol{U}$ を，全ての $i(1 \cdots i \cdots N)$，$j(1 \cdots j \cdots n)$，$k(1 \cdots k \cdots K)$ において

$$U_{ijk} = \begin{cases} 1 & (D_{ij} = k) \\ 0 & (D_{ij} \neq k) \end{cases} \qquad (1)$$

の規則に基づいて生成する．この操作により $\boldsymbol{U}$ の形は具体的には以下のようになる．

$$\begin{pmatrix} D_{11} & \cdots & D_{1j} & \cdots & D_{1n} \\ \vdots & \ddots & \vdots & & \vdots \\ D_{i1} & \cdots & D_{ij} & \cdots & D_{in} \\ \vdots & & \vdots & \ddots & \vdots \\ D_{N1} & \cdots & D_{Nj} & \cdots & D_{Nn} \end{pmatrix} \rightarrow$$

$$\begin{pmatrix} U_{111} \cdots U_{11K} & \cdots & U_{1j1} \cdots U_{1jK} & \cdots & U_{1n1} \cdots U_{1nK} \\ \vdots \quad \vdots & & \vdots \quad \vdots & & \vdots \quad \vdots \\ U_{i11} \cdots U_{i1K} & \cdots & U_{ij1} \cdots U_{ijK} & \cdots & U_{in1} \cdots U_{inK} \\ \vdots \quad \vdots & & \vdots \quad \vdots & & \vdots \quad \vdots \\ U_{N11} \cdots U_{N1K} & \cdots & U_{Nj1} \cdots U_{NjK} & \cdots & U_{Nn1} \cdots U_{Nnk} \end{pmatrix} \qquad (2)$$

この操作によって，列数がK倍になった行列$\boldsymbol{U}$に対して，二値データの場合のアルゴリズムを応用することができる．以下，$\boldsymbol{U}$を用いた推定アルゴリズムについて説明する．

STEP 1.1 $\boldsymbol{V}$の初期値を定める．

全てのk,qに対して，$\boldsymbol{V}$のq番目の潜在ランクにおけるカテゴリkに対応するベクトル
$\boldsymbol{v}_{kq}\{=v_{\{n(k\text{-}1)+1\}q}, v_{\{n(k\text{-}1)+2\}q}, \cdots v_{\{n(k\text{-}1)+j\}q}, \cdots v_{\{nk\}q}\}$
の要素を，項目jに依存しない形で以下のように定める．

$$v_{\{n(k\text{-}1)+j\}q} = (1-\frac{2q-1}{2Q})(\frac{K-k}{K-1}) + \frac{(2q-1)}{2Q}(1-\frac{K-k}{K-1}) \quad (3)$$

ここで$0 < v_{\{n(k\text{-}1)+j\}q} < 1$である．(3)式では全ての項目に対して，番号の小さい潜在ランク(e.g., $q=1$)ほど，番号の小さいカテゴリ(e.g., $k=1$)に高い確率で，また番号の大きいカテゴリ(e.g., $k=K$)には低い確率で反応し，そして潜在ランクの値が大きいほど(e.g., $q=Q$)その逆になるように設定されている．より具体的には，隣接する潜在ランクの要素の差が$1/2Q$となり，さらに$k=1$かつ$q=1$，および$k=K$かつ$q=Q$のときに要素が$(2Q-1)/2Q$に，また$k=1$かつ$q=Q$，および$k=K$かつ$q=1$のときに要素が$1/2Q$となるように設定したときに，$1-(2q-1)/2Q$と$(2q-1)/2Q$はそれぞれ$k=1$と$k=K$における一般項となる．そして残りのカテゴリに対する要素については，kの値を基準に，先ほどの一般項にそれぞれ$(K-k)/(K-1)$と$1-(K-k)/(K-1)$を重みづけしている．

STEP 1.2 潜在ランクの事前分布の設定

STEP3以降でベイズ基準を利用する場合に必要な設定であり，具体的には事前確率$\pi = (\pi_1 \cdots \pi_q \cdots \pi_Q)$を定める．Shojima(2008)でも指摘されているように，SOMの性質上，最小二乗基準や最尤基準を用いた場合に，両端の潜在ランク($q=1,Q$)の度数に偏りが生じることから，事前確率の設定はその影響を緩和する効果がある．本論文では以下のように，事前確率を設定する．

$$\pi_q = \begin{cases} \dfrac{1}{LQ} & q = 1, Q \\ \dfrac{LQ-2}{LQ(Q-2)} & q \neq 1, Q \end{cases} \quad (4)$$

Lが大きいほど，両端の潜在ランクの事前確率が小さくなり，$L=1$の場合に最尤基準と同値になる．本論文では$L=1.5$とする．これは，離散一様分布で求められる確率に比べ$(1/L=)2/3$倍の大きさを設定することを意味する．

STEP 2 $\boldsymbol{U}$の行要素をランダムにソートする．

これは，STEP3以降の学習において各行（受験者）で行われる学習の順番を公平にする為である．

STEP 3.1 ある$i\,(1\cdots i\cdots N)$に対して，最適化基準である距離関数dが最も小さくなるランク$w(1\cdots w\cdots Q)$を選ぶ．

dについては，最小二乗基準(LS)・最尤基準(ML)・ベイズ基準(MAP)があり，それぞれ以下で示される最適化基準を最小にするwが選ばれる．

$$d_{LS} = \left\| \boldsymbol{Z}_i \bullet (\boldsymbol{u}_i - \boldsymbol{v}_q) \right\|^2 = \sum_{j=1}^{n} z_{ij}(u_{ij} - v_{jq})^2 \quad (5)$$

$$d_{ML} = -\sum_{j=1}^{n} z_{ij}(u_{ij} \log v_{jq}) \quad (6)$$

$$d_{MAP} = -\log \pi_q - \sum_{j=1}^{n} z_{ij}(u_{ij} \log v_{jq}) \quad (7)$$

$$w = \operatorname{argmin} \; d \quad (8)$$

$\bullet$はアダマール積を意味する．$\boldsymbol{Z}_i, \boldsymbol{u}_i$はそれぞれ$\boldsymbol{Z}$と$\boldsymbol{U}$の中で受験者$i$のデータに対応するサイズ$(K \times n) \times 1$のベクトルであり，$\boldsymbol{v}_q$は$q$番目の潜在ランクに対応する要素を含むサイズ$(K \times n) \times 1$のベクトルである．(6)と(7)式に関しては，$u_{ij}$の中に被験者$i$の選択したデータが含まれているので，$(1-u_{ij})\log(1-v_{jq})$の項を設ける必要がないという

点には注意が必要である.

STEP 3.2 $\boldsymbol{V}$ の要素を更新する.

更新の仕方は必ずしも一義的でないが，本論文では以下のようにする.

$$\boldsymbol{v}_q = \boldsymbol{v}_q + h_{qw}\boldsymbol{Z}_i \bullet (\boldsymbol{u}_i - \boldsymbol{v}_q) \tag{9}$$

$$h_{qw} = \frac{\alpha_t Q}{N}\exp(-\frac{(q-w)^2}{2\sigma_t^2}) \tag{10}$$

$$\alpha_t = \frac{(T-t)\alpha_1 + (t-1)\alpha_T}{T-1} \tag{11}$$

$$\sigma_t = \frac{(T-t)\sigma_1 + (t-1)\sigma_T}{T-t} \tag{12}$$

$\alpha_1, \alpha_T, \sigma_1, \sigma_T$ は超母数であり，本論文では Shojima(2008)に倣って，$(\alpha_1, \alpha_T, \sigma_1, \sigma_T) = (1.00, 0.10, 1.00, 0.12)$とそれぞれ設定した.

STEP 3.3 $\boldsymbol{V}$ の要素の調整（1）

$\boldsymbol{V}$ の要素について，すべてのk,q に対して

$$\boldsymbol{v}^*_{kq} = (\sum_{k=1}^{K}\widetilde{\boldsymbol{V}}_{kq})^{-1}\boldsymbol{v}_{kq} \tag{13}$$

と変換することで，全ての項目における各カテゴリへの反応確率の和が１になるように調整する．ただし，$\widetilde{\boldsymbol{V}}_{kq}$ は j 番目の対角要素に$\boldsymbol{v}_{kq}$の j 番目の要素が入ったサイズ$n \times n$ の対角行列である.

STEP 3.4 $\boldsymbol{V}$ の要素の調整（2）

両端のカテゴリ（$k = 1, K$)に対応する$\boldsymbol{V}^*$ の要素 $(\boldsymbol{v}^*_{1q}, \boldsymbol{v}^*_{Kq})$ を更新し，１番目のカテゴリに対して単調減少制約を，K 番目のカテゴリに対しては単調増加制約を施す．具体的にはある q $(2 \cdots q \cdots Q)$ に対して，その要素が以下のように定義されるベクトル
$\boldsymbol{c}_{Kq} = (c_{K1q}, \cdots c_{Kjq}, \cdots c_{Knq})$ と
$\boldsymbol{c}_{1q} = (c_{11q}, \cdots c_{1jq}, \cdots c_{1nq})$ を計算し，

$$c_{Kjq} = sgn[v^*_{\{n(K-1)+j\}q} - v^*_{\{n(K-1)+j\}\{q-1\}}] \tag{14}$$

$$c_{1jq} = sgn[v^*_{j\{q-1\}} - v^*_{jq}] \tag{15}$$

そしてこの$\boldsymbol{c}_{Kq}$ ならびに$\boldsymbol{c}_{1q}$ を用いて，$\boldsymbol{v}^*_{Kq}, \boldsymbol{v}^*_{1q}$ をそれぞれ以下のように更新する.

$$\boldsymbol{v}^{**}_{Kq} = \frac{(\boldsymbol{c}_{Kq}+1)}{2}\bullet\boldsymbol{v}^*_{Kq} - \frac{(\boldsymbol{c}_{Kq}-1)}{2}\bullet\boldsymbol{v}^*_{K\{q-1\}} \tag{16}$$

$$\boldsymbol{v}^{**}_{1q} = \frac{(\boldsymbol{c}_{1q}+1)}{2}\bullet\boldsymbol{v}^*_{1q} - \frac{(\boldsymbol{c}_{1q}-1)}{2}\bullet\boldsymbol{v}^*_{1\{q-1\}} \tag{17}$$

STEP 3.5 $\boldsymbol{V}$ の要素の調整（3）

$\boldsymbol{v}_{1q} = \boldsymbol{v}^{**}_{1q}, \boldsymbol{v}_{kq} = \boldsymbol{v}^*_{kq}$ $(k = 2,3,\cdots K-1), \boldsymbol{v}_{Kq} = \boldsymbol{v}^{**}_{Kq}$

と更新する.

STEP 4 STEP 3 を全てのi に対して繰り返す.

STEP 5 STEP 2,3,4 をT 回繰り返す.

アルゴリズムの改良点について

順序データの為の NTT は Shojima(2007a)ですでにアルゴリズムの検討がなされていたが，より汎用的に使用する為に今回若干の改良を行った．改良を行った点は主に三点である．一点目は(3)式で表された，$\boldsymbol{V}$ の初期値の設定である．適切な初期値を与えない場合に，不適解や$\boldsymbol{V}$ の非順序性及びその解釈についての問題を引き起こす可能性があるが，(3)式は潜在ランクに応じた初期値をカテゴリと潜在ランクの関数で表しており，一般的な表現を試みている．二点目は(4)式による潜在ランクの事前分布の設定である．Shojima(2008)では事前確率は任意の定数で定められており，潜在ランク数の関数で表現されていなかった．その点，今回のアルゴリズムでは潜在ランク数に応じた事前分布が設定され，後のシミュレーションも実行しやすいというメリットがある．三点目は，(14)～(17)式に表されている，要素の単調増加及び単調減少の順序制約である．これは，$\boldsymbol{V}$ の要素および潜在ランクの推定値の解釈を容易にし，また ICRP の推定値を安定させる上でも有用である.

他にも， Shojima(2007a)ではカテゴリ K に対応す

る要素が$\boldsymbol{U}$に含まれていないが，今回のアルゴリズムには(2)式にあるように含まれている．その結果，(13)式のような調整を行わなくてはならない反面，これにより各項目に対応する$\boldsymbol{V}$の要素の和について直接制約をかけることができるので，不適解の発生を抑えることが期待される．

3 シミュレーション

3.1 方法

本シミュレーションでは，項目反応モデルに基づいて発生させたデータをもとに，項目サンプリングを超えた潜在ランクの推定値の一貫性を，NTT（最適化基準は最小二乗基準と最尤基準とベイズ基準の三種類）と潜在ランクの度数分布が離散一様分布になるように分割した均等分割基準の間で比較検討する．シミュレーションは，被験者数($N = 100,400,1600$)，項目数($n = 10,20,40$)，被験者パラメタの母集団分散（$\sigma^2 = 1,2$)，潜在ランク数（$Q = 5,10,15$)，データに与えるノイズの大きさ（$\sigma_e{}^2 = 0,1,2$)，識別力パラメタの高さ（高・低・混合条件)，カテゴリ数（$K = 2,5$）をそれぞれ変化させた上で検討した．なお，学習回数TについてはT=150 と固定した．シミュレーションは以下の手順によって行った．

STEP1 被験者パラメタ・項目パラメタを以下の分布からそれぞれ抽出する．

- N 個の被験者パラメタ……$\theta_i \sim N(0, \sigma^2)$
- n 個の識別力パラメタ
  高条件…$\log\alpha_j \sim N(0.20, 0.25)$
  低条件…$\log\alpha_j \sim N(-0.70, 0.25)$
  混合条件… $n/2$ 個が$\log\alpha_j \sim N(0.20, 0.25)$で，
  残りの$n/2$ 個は$\log\alpha_j \sim N(-0.70, 0.25)$
- n 個の困難度パラメタ…… $\delta_j \sim N(0, 1)$
- $(K$-1)$\times n$ 個の閾値パラメタ……
  $\tau_{j1} \sim N(-1.20, 0.25)$, $\tau_{j2} \sim N(-0.50, 0.25)$
  $\tau_{j3} \sim N(0.50, 0.25)$, $\tau_{j4} \sim N(1.20, 0.25)$

STEP2 STEP1 で発生させた被験者パラメタと項目パラメタを用いて，各被験者iが項目jにおいてカテゴリkを選択する確率$P_{jk}(\theta_i)$を，以下の式で示される，Muraki(1992)のGPCM を用いて計算する．

$$P_{jk}(\theta_i) = \frac{\exp[\alpha_j(k(\theta_i - \delta_j) + \sum_{m=0}^{k}\tau_{jm}]}{\sum_{l=0}^{K-1}\exp[\alpha_j(l(\theta_i - \delta_j) + \sum_{m=0}^{l}\tau_{jm}]} \quad (19)$$

STEP3 サイズ$N \times n$の標準一様乱数行列$\boldsymbol{R}$を発生させ，以下の規則に基づいて生データ行列$\boldsymbol{D}$を生成し，前節の方法により各被験者の潜在ランクを推定する．

$$D_{ij} = \begin{cases} 5 & (R_{ij} > P_{j1}(\theta_i) + P_{j2}(\theta_i) + P_{j3}(\theta_i) + P_{j4}(\theta_i)) \\ 4 & (P_{j1}(\theta_i) + P_{j2}(\theta_i) + P_{j3}(\theta_i) + P_{j4}(\theta_i) \geq R_{ij} \\ & > P_{j1}(\theta_i) + P_{j2}(\theta_i) + P_{j3}(\theta_i)) \\ 3 & (P_{j1}(\theta_i) + P_{j2}(\theta_i) + P_{j3}(\theta_i) \geq R_{ij} > P_{j1}(\theta_i) + P_{j2}(\theta_i)) \\ 2 & (P_{j1}(\theta_i) + P_{j2}(\theta_i) \geq R_{ij} > P_{j1}(\theta_i)) \\ 1 & (P_{j1}(\theta_i) \geq R_{ij}) \end{cases} \quad (20)$$

STEP4 STEP1～3 をS回繰り返す．ただし各回でθはSTEP 1 で得た値に対して，$\boldsymbol{N}(0, \sigma_e^2)$から独立に抽出した$\boldsymbol{N}$個の乱数を足した値を新たな$\theta$として用いる．すなわち，$\sigma_e{}^2$=0 の条件では各回で固定された$\theta$を用いる．

均等分割基準の場合は$\boldsymbol{D}$を利用して各々の被験者の項目和得点を求め，各潜在ランクの度数が等しくなるように潜在ランクを決定する．各々の$s(\leq S)$回目の結果を用いて，潜在ランクの完全一致率，±1 のズレを許容した一致率の平均値と標準偏差をそれぞれ求める．具体的には，${}_S C_2 = S(S-1)/2$回分の組み合わせがあるので，それらの値を利用して一致率の平均値と標準偏差を計算することになる．今回は$S = 10$と設定した．

3.2 結果と考察

紙面の都合により，$K=5, \sigma^2=1, \sigma_e{}^2=0$, 識別力高条件における，$\boldsymbol{N}=(100, 400, 1600)$, $\boldsymbol{n}=(10, 20, 40)$，$Q=(5,10,15)$それぞれの場合での，NTT（最小二乗基準と最尤基準とベイズ基準の三種類）および項目和得点に基づく均等分割基準による潜在ランクの推定値の完全一致率の平均値と±1も含めた一致率の平均値と標準偏差をTable 1に報告する.

①最適化基準の違いによる影響

完全一致率を基準としてNTTの各最適化基準を比較すると，全体として最小二乗基準に基づいた推定に比べ，最尤基準，ベイズ基準による推定の方がやや優れている傾向が見られた．これらは最小二乗法基準による推定が項目特性値の違いを考慮しない為と考えられる．また，最尤基準とベイズ基準の比較では，完全一致率の場合では最尤基準の方が同程度若しくは優れている傾向が見られるが，±１も含めた一致率による比較では，ベイズ基準の方が総じて優れている結果が見られた．これは最尤基準（および最小二乗基準）では，SOMの性質から両端の潜在ランクに度数が偏る傾向があるため，両端の潜在ランクの影響で完全一致率が上昇する一方，中間の潜在ランクへの度数が少なく見積もられ，この部分の一致率の推定が不安定になった結果，±１も含めた一致率が低くなった可能性が考えられる．ベイズ基準の場合は，最尤基準に比べて相対的に両端の潜在ランクへの偏りの影響が少なく，その結果，完全一致率はランクの偏りの影響がある最尤基準に比べ劣るものの，中間の潜在ランクにおける±１を含めた一致率の推定が安定していたことが予想される．これらの傾向は他の条件においても一貫して見られた傾向であった．このことはNTTの最適化基準の選択において，SOMの性質である両端の潜在ランクへの過度の偏りを抑える為に適切な事前分布を設定することができれば，一貫性の高さと分析の柔軟性の観点からは，ベイズ基準が最も望ましい最適化基準となる可能性を示唆している.

また, NTTの各最適化基準と均等分割基準の結果を比べると，均等分割基準において完全一致率は最尤基準やベイズ基準と同等程度であるが，±１を含む完全一致率では最も高い値が観察された. これについては，ベイズ基準においても今回は事前分布の設定が完全に適切ではなかった為という可能性と，事前分布の設定の妥当性に関係なく均等分割基準の方が安定的であるためという可能性が考えられる. また，$\boldsymbol{N}=100$でQ＝15のときはいずれもNTTにおける完全一致率の平均値は均等分割基準に比べ高かったが，これは受験者数$\boldsymbol{N}$の大きさに比べ事前分布の設定が極端であり，結果的にNTTのいずれの最適化基準においても潜在ランクの偏りが改善されなかったことに起因すると考えられる.

これらの結果をまとめると，各最適化基準では一致率という観点からそれらの優劣に大きな差があったとは言えないが，均等分割基準が全体的に最も安定的な推定値を示していたと言えるだろう．しかしこのことは常に均等分割基準の選択が最も優れていることを示しているわけではない. NTTは項目の識別力や困難度の違いから生じる，項目得点の度数分布の形状の歪みを考慮した潜在ランクの推定を行い，似た回答パタンをする受験者を，その度数の偏りに関係なくランク付けできることにその特性があることを考えると，均等分割基準とは根本的に目的が異なる手法であることは注意すべきであろう.

②変化させたパラメタによる影響

受験者数Nに関しては, その値が大きくなるほど一致率の推定値の標準誤差も小さくなるが，いずれの推定法の場合でも一致率の高さに大きな影響は与えていないこと，及びN=100の結果の一部では潜在ランクの度数分布の偏りの影響が強く影響したために一致率が高くなっていることが確認される．また，項目数$\boldsymbol{n}$についてはその値が大きくなるにつれて一致率が高まることが，いずれの条件においても見られた．これについては，被験者の潜在ランクを推定する為の情報が増えたことによる影響と考えられる.

また，潜在ランク数Qの影響については当然，その大きさが増えるにつれ正確な判別が難しくなる為にその一致率が低下していた．n＝40の場合であっても，Q＝15以上となると四種類全ての最適化基準において完全一致率は5割以下となった．また，Q＝5の場合でもnが20程度では，いずれの最適化基準においても完全一致率が7割程度であったという結果は，項目を超えたサンプリングとしての一貫性を維持することの難しさを認識することができる.

Table 1.1 $K=5, \sigma^2=1, \sigma_e{}^2=0$, 識別力高条件における，最小二乗基準および最尤基準による一致率の平均値と標準偏差

			最小二乗基準				最尤基準			
			完全	*SD*	±1	*SD*	完全	*SD*	±1	*SD*
$n=10$	$N=100$	$Q=5$	0.5471	(0.0388)	0.8689	(0.0296)	0.6057	(0.0364)	0.9164	(0.0233)
		$Q=10$	0.3821	(0.0342)	0.6268	(0.0465)	0.4436	(0.0337)	0.6975	(0.0458)
		$Q=15$	0.2750	(0.0350)	0.4404	(0.0347)	0.3218	(0.0389)	0.4854	(0.0516)
	$N=400$	$Q=5$	0.5218	(0.0219)	0.8859	(0.0189)	0.5797	(0.0191)	0.9131	(0.0127)
		$Q=10$	0.3063	(0.0139)	0.6039	(0.0217)	0.3976	(0.0217)	0.7265	(0.0196)
		$Q=15$	0.2264	(0.0199)	0.4424	(0.0242)	0.2796	(0.0142)	0.4977	(0.0212)
	$N=1600$	$Q=5$	0.5774	(0.0089)	0.9300	(0.0062)	0.5964	(0.0108)	0.9388	(0.0062)
		$Q=10$	0.3196	(0.0105)	0.6480	(0.0095)	0.3423	(0.0108)	0.6972	(0.0109)
		$Q=15$	0.2265	(0.0104)	0.4719	(0.0117)	0.2719	(0.0104)	0.5493	(0.0129)
$n=20$	$N=100$	$Q=5$	0.6732	(0.0384)	0.9618	(0.0216)	0.6964	(0.0330)	0.9754	(0.0174)
		$Q=10$	0.4739	(0.0363)	0.7814	(0.0370)	0.4957	(0.0379)	0.7793	(0.0377)
		$Q=15$	0.3682	(0.0364)	0.6107	(0.0456)	0.4004	(0.0365)	0.6496	(0.0276)
	$N=400$	$Q=5$	0.6174	(0.0212)	0.9572	(0.0120)	0.6676	(0.0164)	0.9746	(0.0089)
		$Q=10$	0.4002	(0.0207)	0.7572	(0.0164)	0.4708	(0.0183)	0.8379	(0.0175)
		$Q=15$	0.3015	(0.0211)	0.5834	(0.0198)	0.3557	(0.0166)	0.6574	(0.0212)
	$N=1600$	$Q=5$	0.6278	(0.0125)	0.9656	(0.0047)	0.6688	(0.0099)	0.9786	(0.0030)
		$Q=10$	0.4128	(0.0103)	0.8105	(0.0076)	0.4550	(0.0095)	0.8581	(0.0083)
		$Q=15$	0.2973	(0.0073)	0.6147	(0.0126)	0.3437	(0.0112)	0.6988	(0.0115)
$n=40$	$N=100$	$Q=5$	0.7114	(0.0447)	0.9836	(0.0121)	0.7332	(0.0365)	0.9846	(0.0128)
		$Q=10$	0.5114	(0.0547)	0.8586	(0.0258)	0.5586	(0.0427)	0.8721	(0.0339)
		$Q=15$	0.4393	(0.0411)	0.6968	(0.0424)	0.4904	(0.0465)	0.7736	(0.0404)
	$N=400$	$Q=5$	0.7167	(0.0161)	0.9925	(0.0434)	0.7396	(0.0175)	0.9946	(0.0029)
		$Q=10$	0.4943	(0.0188)	0.8760	(0.0213)	0.5313	(0.0185)	0.9180	(0.0142)
		$Q=15$	0.3801	(0.0235)	0.7313	(0.0229)	0.4270	(0.0214)	0.7842	(0.0219)
	$N=1600$	$Q=5$	0.7456	(0.0089)	0.9959	(0.0016)	0.7500	(0.0078)	0.9965	(0.0012)
		$Q=10$	0.5112	(0.0155)	0.9173	(0.0068)	0.5567	(0.0105)	0.9474	(0.0043)
		$Q=15$	0.3914	(0.0073)	0.7937	(0.0102)	0.4280	(0.0114)	0.8353	(0.0089)

*「完全」は完全一致率の，「±１」は±１の違いも含めた一致率の平均を意味する．

Table 1.2 $K=5, \sigma^2=1, {\sigma_e}^2=0$, 識別力高条件における，ベイズ基準および均等分割基準による一致率の平均値と標準偏差

			ベイズ基準				均等分割基準			
			完全	*SD*	±1	*SD*	完全	*SD*	±1	*SD*
$n=10$	N=100	Q=5	0.5986	(0.0344)	0.9593	(0.0232)	0.5814	(0.0334)	0.9604	(0.0149)
		Q=10	0.3704	(0.0423)	0.7475	(0.0473)	0.3536	(0.0329)	0.7829	(0.0403)
		Q=15	0.2904	(0.0467)	0.5332	(0.0499)	0.2411	(0.0421)	0.5518	(0.0468)
	N=400	Q=5	0.5635	(0.0270)	0.9551	(0.0098)	0.5735	(0.0233)	0.9563	(0.0082)
		Q=10	0.3732	(0.0210)	0.7680	(0.0232)	0.3518	(0.0252)	0.7386	(0.0194)
		Q=15	0.2400	(0.0175)	0.5469	(0.0235)	0.2528	(0.0226)	0.5813	(0.0231)
	N=1600	Q=5	0.5977	(0.0092)	0.9670	(0.0054)	0.6117	(0.0108)	0.9710	(0.0054)
		Q=10	0.3338	(0.0110)	0.7405	(0.0081)	0.3637	(0.0097)	0.7598	(0.0089)
		Q=15	0.2494	(0.0109)	0.5854	(0.0136)	0.2700	(0.0101)	0.6144	(0.0118)
$n=20$	N=100	Q=5	0.7007	(0.0450)	0.9900	(0.0111)	0.7139	(0.0378)	0.9989	(0.0031)
		Q=10	0.4525	(0.0408)	0.8311	(0.0323)	0.4707	(0.0407)	0.9125	(0.0251)
		Q=15	0.3725	(0.0324)	0.6843	(0.0303)	0.3579	(0.0328)	0.7604	(0.0080)
	N=400	Q=5	0.6562	(0.0205)	0.9840	(0.0060)	0.6721	(0.0203)	0.9895	(0.0041)
		Q=10	0.4469	(0.0213)	0.8653	(0.0177)	0.4444	(0.0225)	0.8682	(0.0134)
		Q=15	0.3369	(0.0197)	0.6842	(0.0183)	0.3266	(0.0227)	0.7235	(0.0205)
	N=1600	Q=5	0.6766	(0.0081)	0.9878	(0.0023)	0.6617	(0.0090)	0.9879	(0.0020)
		Q=10	0.4536	(0.0129)	0.8771	(0.0072)	0.4654	(0.0096)	0.8873	(0.0062)
		Q=15	0.3292	(0.0104)	0.7227	(0.0124)	0.3373	(0.0113)	0.7235	(0.0108)
$n=40$	N=100	Q=5	0.7400	(0.0434)	0.9904	(0.0106)	0.7729	(0.0492)	1.0000	(0.0000)
		Q=10	0.5354	(0.0495)	0.8914	(0.0293)	0.5675	(0.0490)	0.9596	(0.0200)
		Q=15	0.4746	(0.0563)	0.8057	(0.0433)	0.4350	(0.0478)	0.8540	(0.0273)
	N=400	Q=5	0.7369	(0.0186)	0.9972	(0.0024)	0.7602	(0.0146)	0.9988	(0.0016)
		Q=10	0.5305	(0.0182)	0.9275	(0.0120)	0.5545	(0.0246)	0.9472	(0.0109)
		Q=15	0.4178	(0.0180)	0.8021	(0.0189)	0.4104	(0.0197)	0.8279	(0.0184)
	N=1600	Q=5	0.7527	(0.0102)	0.9979	(0.0009)	0.7808	(0.0089)	0.9992	(0.0005)
		Q=10	0.5605	(0.0114)	0.9529	(0.0046)	0.5540	(0.0130)	0.9535	(0.0055)
		Q=15	0.4231	(0.0100)	0.8468	(0.0091)	0.4324	(0.0107)	0.8592	(0.0106)

*「完全」は完全一致率の，「±１」は±１の違いも含めた一致率の平均を意味する．

被験者パラメタの母集団分散（σ^2)については，いずれの最適化基準においても$\sigma^2=2$の方が$\sigma^2=1$に比べ全体的に1～2割ほど完全一致率が上昇していた．これは，σ^2が大きいと被験者の能力値のばらつきが大きくなり，粗い精度しか有していないテストでも被験者の能力の識別が十分可能であるためだと考えられる．

データに与えるノイズσ_e^2については，その値が大きくなるにつれ完全一致率が減少していることが観察された．実際$\sigma_e^2=2$といったかなり大きなノイズが与えられている条件では，$\sigma_e^2=0$の条件に比べて平均3割程度の完全一致率の減少が観察された．またこのノイズに対する抵抗性については，他のパラメタの場合と同様,最適化基準による違いは見られなかった．この結果については，今回は項目の困難度や識別性などの項目特性に関係なく全ての項目反応に対して同様のノイズを与えたことによる影響もあるかもしれない．すなわち，ノイズを与える項目を差別化することで，異なる結果が得られる可能性もある．

4 分析例

4.1 小論文試験について

本節では高校生の書いた小論文データを用いた分析例を示す．小論文試験の評価は，採点者の評価の信頼性の問題，構成概念妥当性の問題，系列効果や文字の美醜などのバイアスの問題など，多くの測定論的課題を孕んでいることはよく指摘されているところである（e.g., Brown, Glaswell, & Harland, 2004；Chase, 1986；平・江上, 1992；宇佐美, 2008）．

小論文試験の測定論的課題の中でも特に採点者内・採点者間の評価の信頼性の問題は重要である(e.g., 渡部他,1988）．このように評価の信頼性の点で危惧される小論文試験について，離散化した潜在ランクによる評価法が，その信頼性の維持という意味で一つの対処法となる．また，特に小論文が分析的に評価された場合，各分析的評価項目の困難度や識別力などの項目特性が一様でない為に単純な項目和得点をとることが危惧される一方で, NTT はそれらの違いを反映する可能性が高く，NTT を利用する意義があると思われる．

4.2 データ

受験者

秋田県の県立高校の二年生 148 名 4 クラス．平均年齢 16.3 歳．男子 70 名，女子 78 名．

課題

本論文では「小学校の授業における英語の早期教育は必要であるか否かについて，あなたの意見とその根拠が明確になるように，800 字以内で論述しなさい．」という小論文のテーマのみを与えられるテーマ型の小論文を実施した．

採点者

日常的に小論文の作成や評価を行っている専門家二名と現役の国語教師二名．

評価基準の設定

Remondino(1959)や渡部他(1988)を参考にし，小論文試験の作成や評価の専門家と協議をしながら，分析的評価の為の評価基準を作成した．一般に比べ評価基準を多目に作成してあるが，これは他の研究の目的で評価構造に関する因子分析を行う為である．Table 2 に示されているように，文章の誤字・脱字や語彙力といった言語能力を測る分析的評価項目と，表現力・構成力・説得力など文章の内容的な質を測る分析的評価項目を設定してある．採点は，全ての分析的評価項目において 5 点満点で，総合評価については 10 点満点で行った．

4.3 分析結果

4.3.1 記述統計

採点者四名分のデータを平均し，各分析的評価観点の基本統計量（平均・標準偏差）と，総合評価点との相関係数の値をまとめて Table 3 に示す．

各分析的評価観点の平均値は比較的高く，特に「語句」・「表現」・「課題」・「一貫」・「形式」では特に高い．またこれらの観点では度数分布が左に裾を引いており，歪度が小さかった．また,「説得」の観点は平均値が最も低く，そして総合評価点との相関も最も高い．

4.3.2 NTT による分析

総合評価点の採点者間相関の平均が 0.512 と比較的高かったことなど，採点者の違いによる評価傾向の違

Table 2. 分析的評価観点の名称とその定義

1, 語句：	誤字・脱字はないか. 送り仮名は正しく書かれているか.
2, 表現の正確さ：	言葉の表現が正確であり, 読み手に言葉の意味が適切に通じるか.
3, 語彙力:	年齢相応の語彙力があり, 表現が稚拙でないか.
4, 課題内容の解釈：	設問の意図を正しく理解できており, 小論文がそれに正しく答える内容となっているか.
5, 簡潔性：	文章は冗長でなく, 簡潔であるか.
6, 主張の明確性：	自己の主張が, 文章内に明確に盛り込まれているか.
7, 構成：	小論文の内容的構成が適切であり, 自然な順序になっているか.
8, 一貫性：	展開されている主張の論旨が一貫しており, 矛盾がないか.
9, 説得力：	展開されている主張が説得的であり, 納得できるか.
10, 独創性：	文章には書き手自身の独自な視点・発想が盛り込まれていたか.
11, 形式：	原稿用紙の正しい使い方, 及び段落の設定・回答字数について問題はなかったか.

Table 3. 各観点の基本統計量と総合評価との相関係数

	総合	語句	表現	語彙	課題	簡潔	明確	構成	一貫	説得	独創	形式
平均	5.98	4.37	4.23	3.96	4.48	3.74	3.93	3.72	4.29	3.12	3.67	4.33
標準偏差	1.16	0.70	0.55	0.51	0.62	0.63	0.60	0.63	0.65	0.62	0.52	0.96
相関		0.423	0.545	0.718	0.531	0.556	0.638	0.760	0.694	0.805	0.622	0.702

いが比較的小さいことが事前の分析により確認されている為, 以下では各分析的評価観点における採点者四名の平均値の小数点第一桁を四捨五入した値を順序データとした, NTT による分析を行う. 総合評価点については後述のように, NTT で求めた潜在ランクや IRT によって求めた θ などとの相関分析のために用いる. 今回はデータ数が少なく, また最小二乗基準や最尤基準では潜在ランクの偏りが出ることが懸念された為, 潜在ランク数については $Q=5, 10$ のもとでベイズ基準による推定を行った.

潜在ランクの度数分布

潜在ランクの度数分布を以下の Figure 1 に示す. $Q=5$, $Q=10$ いずれの場合においても, 両端の潜在ランクへの極端な偏りが抑えられていることがわかる. 度数分布の偏りの大きい分析的評価観点が多かった為か, いずれにおいても番号の大きな潜在ランクに度数がやや偏っている傾向が見られる.

項目参照プロファイル（ICRP）の解釈

紙面の都合により, $Q=5$ の場合のみにおける, 参照行列 $\boldsymbol{V}$ の推定値を利用して作成した ICRP を折れ線で示した図を Figure 2 に示す. 「説得」の評価を除いて, いずれの分析的評価観点においても 1 点の評価の度数がほとんどないため, $k=1$ においては要素がいずれの潜在ランクにおいても小さな値に推定されている. また, 両端のカテゴリには順序制約をかけているため, $k=1$ と $k=5$ において, 図のように潜在ランク数 Q の値に対応してそれぞれ単調減少或いは単調増加の折れ線が描かれている. 困難度が最も高く, また総合評価点との相関の高かった「説得」の観点はいずれのランクにおいても $k=5$

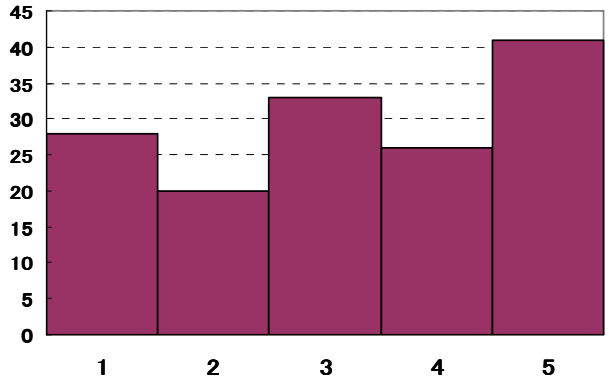

Figure 1.1　潜在ランクの度数分布(Q=5)

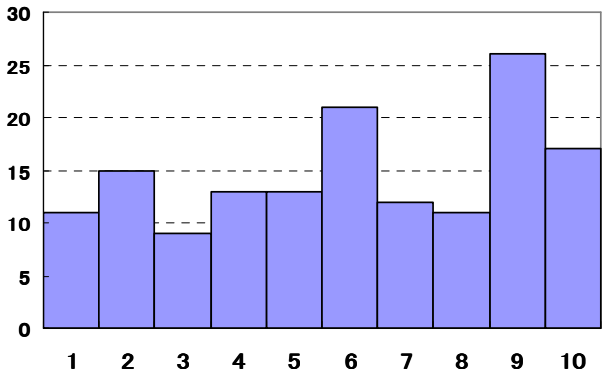

Figure 1.2　潜在ランクの度数分布(Q=10)

の折れ線は低調であり，k＝4においてもqが３以上になってようやく折れ線が上昇する様子が観察される．また，特に隣接するカテゴリの折れ線が交差する点に注目すると，q＝2はq＝1に比べ「簡潔」・「構成」の観点で，q＝3はq＝2に比べ「一貫」の観点で，q＝4はq＝3に比べ「表現」で，そしてq＝5はq＝4に比べ「説得」の評価点が優れていることがわかる．逆に言えば，他の分析的評価観点では潜在ランクの違いによる折れ線の変化が小さいことを示している．このように, NTT では被験者の所属する潜在ランクを推定しながら，同時に潜在ランクの違いに応じた各項目における各カテゴリの選択確率の推移を検討することができる．

θ および分析的評価和得点・総合評価点との相関

Q＝5とQ＝10の場合の各被験者の潜在ランクの推定値と, 分析的評価得点の項目和得点, 総合評価点, および分析的評価データを IRT（GPCM）を用いて推定したθとの相関係数を Table 4 に示す．(NTT の推定値との相関はいずれも Spearman のρを用いている)

NTT とθは分析的評価データを直接用いて推定している為，分析的評価点との相関は極めて高く，また NTT とθとの相関はかなり高い．NTT は，θと極めて類似した関係を持ちながら，一定の最適化基準のもとに潜在ランクの推定を行っていることが示唆される．

4.3.3 異なる手法間の採点者間相関係数の推定値

４名の採点者のローデータを用いて，異なる手法間の採点者間相関係数を比較する．分析的評価点から NTT によりQ＝5, 10 で推定した潜在ランク(NTT5, NTT10）・IRT（GPCM）により推定したθ(IRT)・θをQ＝5, 10 で均等分割して形成した潜在ランク(IRT5, IRT10)・分析的評価点の和得点（和得点）・分析的評価点の和得点をQ＝5, 10 で均等分割した形成した潜在ランク(和得点５，和得点 10)のそれぞれにおける, 4 名の採点者間相関の平均を示した表が Table 5 である（総合評価は前述の通り 0.512 であった）．

Table 4. 各手法の推定値間の相関係数

	NTT(Q=5)	NTT(Q=10)	総合評価点	分析和得点	θ
NTT(Q=5)	1	0.973	0.765	0.920	0.962
NTT(Q=10)	0.973	1	0.760	0.927	0.982
総合評価点	0.765	0.760	1	0.892	0.813
分析和得点	0.920	0.927	0.892	1	0.934
θ	0.962	0.982	0.813	0.934	1

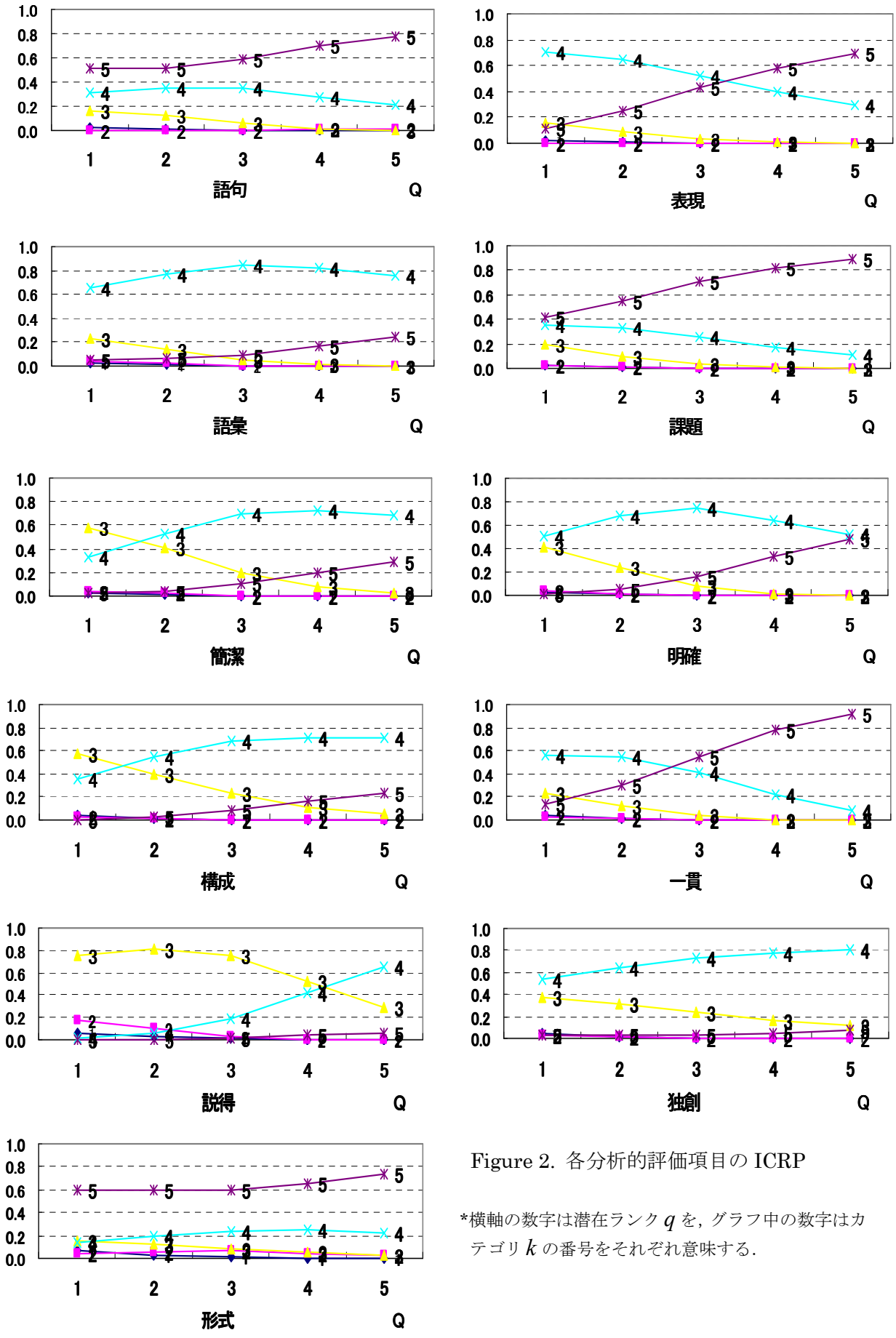

Figure 2. 各分析的評価項目の ICRP

*横軸の数字は潜在ランク q を，グラフ中の数字はカテゴリ k の番号をそれぞれ意味する．

Table 5. 各手法における，採点者間相関係数の平均値

NTT5	NTT10	IRT	IRT5	IRT10	和得点	和得点 5	和得点 10
0.340	0.346	0.458	0.358	0.374	0.557	0.402	0.423

この結果から，NTT と IRT に基づく方法のいずれにおいても，項目和得点を均等分割する方法に比べてその相関係数の平均が低いこと，および NTT と IRT では若干 IRT を用いた方法の方が高いことがわかる．特に前者の点については，NTT と IRT いずれの方法の場合においても推定誤差の影響を幾らか受けている可能性が高い．採点者間相関という観点からは，NTT と IRT を用いた方法の高さが似通っていることは興味深い．

5 総合考察

本論文では荘島(2007)の提案した NTT において，潜在ランクおよび ICRP の推定精度，項目サンプリングを超えた潜在ランクの推定値の一貫性，異なる最適化基準からの推定値の比較，分析モデルの構築・改良などの方法論的課題を指摘した．そして，項目サンプリングを超えた潜在ランクの推定値の一貫性について，NTT における三つの最適化基準と，項目和得点を利用した均等分割基準での違いを，若干の改良を加えた多値型 NTT モデルを用いたシミュレーションにて検証した．その結果，一貫性の高さと柔軟性の観点からは，NTT の手法の中ではベイズ基準が最も優れていること，またその一貫性は項目和得点を用いた均等分割基準と同等程度であることが示された．また，小論文データを用いた分析例を提示し，異なる分析手法に基づく分析結果についても比較検討を行った．

今後の検討課題としてはまず，潜在ランク数 Q の事前分布の設定が挙げられる．今回のシミュレーションでは受験者数 N が小さい場合に潜在ランクの度数の偏りが改善されなかったことから，事前分布を N の関数で表現する方法や，他にも事前分布を三角分布の形で表現するなどの工夫の余地があるだろう．また，今回は潜在ランクの推定精度に関する問題は直接扱うことができなかったため，今後適切な方法に基づいたシミュレーションにより検討していく必要があるだろう．さらに、今回のシミュレーションで検討した一貫性の問題における，IRT を用いた θ の均等分割基準による一致率の比較についても今後の検討課題と言える．

本論文では NTT の推定アルゴリズムや推定値の一貫性といった方法論的な内容に力点をおいてきたが，他にも検証すべき課題がある．それには例えば，潜在ランクに基づく評価が学習者に与える影響に関する，教育心理学・教育社会学的な課題が挙げられる．より具体的には，例えば荘島(2007)や Shojima(2008)では，NTT などによる離散的な潜在ランクによる評価法によって，受験者が過度な競争心に陥ることなく学習を進めることができると主張しているが，これは評価法自体に直結する問題であろうか．また，離散的な潜在ランクの利用により，学習者が一つ上の潜在ランクにステップアップする困難度が平均的に高まることによって，達成感を持続的に得ることができず，その結果動機づけが低下してしまう学習者がでてくる可能性もあるが，このような点に関する考察も必要である．

さらに，これは NTT 独自の問題ではないが，潜在ランク自体は，たとえば水泳の場合のように，「息継ぎができるレベル」，「10km の遠泳ができるレベル」といった質的で明確な内容を必ずしも意味しないため，潜在ランク自体にどのような意味を付与し，それを受験者にいかにフィードバックするかはテストのアカウンタビリティにも関わる問題である．このように，NTT の実際の利用に対しては受験者に与えうる心理的影響や社会的影響についても考慮する必要がある．

いずれにしても，現場で実用化されつつある NTT について，今回検討した方法論的課題や，他にも教育心理学・教育社会学的課題を含めた課題について，専門家と実務家を交えた議論が必要である．

謝辞

本論文を作成するにあたり様々な点でご助言くださった大学入試センターの荘島宏二郎先生に心から御礼申し上げます．

6 引用文献

荒井清佳・橋本貴充・荘島宏二郎 (2008). ニューラルテストモデルにおけるテスト等化の精度について　日本テスト学会第 6 回大会発表論文抄録集.

Brown, G.T., Glasswell, K. & Harland, D. (2004). Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system. *Assessing Writing*, **9**(2), 105-121.

Chase, C. I. (1986). Essay test scoring : interaction of relevant variables. *Journal of Educational Measurement,* **23**(1), 33-41.

橋本貴充・荘島宏二郎 (2008). 事前分布を伴うニューラルテスト理論を用いた選抜　日本テスト学会第 6 回大会発表論文抄録集.

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

Muraki, E. (1992). A generalized partial credit model : Application of an EM algorithm. *Applied Psychological Measurement*, **17**, 351-363.

Remondino, C. (1959). A factorial analysis of the evaluation of scholastic compositions in the mother tongue. *British Journal of Educational Psychology*, **30** 242-251.

荘島宏二郎 (2007) ニューラルテスト理論　日本テスト学会第 5 回発表論文集, 174-177.

Shojima, K. (2007a). The graded neural test model: A neural test model for ordered polytomous data. *DNC Research Note*, RN07-03.

Shojima, K. (2007b). Latent rank theory: Estimation of item reference profile by marginal maximum likelihood method with EM algorithm. *DNC Research Note*, RN07-12.

Shojima, K. (2007c). Bayesian estimation of latent rank in neural test theory. *DNC Research Note*, RN07-15.

Shojima, K. (2008). Neural test theory: A latent rank theory for analyzing test data. *DNC Research Note*, RN08-01.

平直樹・江上由実子 (1992). ESSAY TEST の方法論的諸問題に関する研究の動向について *教育心理学研究*, **40**(1), 108-117.

宇佐美慧 (2008). 小論文試験の採点における文字の美醜効果の規定因―メタ分析及び実験による検討―　*日本テスト学会誌,* **4**(1), 73-83.

渡部洋・平由実子・井上俊哉 (1988). 小論文評価データの解析　東京大学教育学部紀要, **28**, 143-164.