

相関構造を仮定したテストレットモデルのベイズ推論
—国立大学法人等の教育研究評価データへの適用—

A Bayesian Inference Assuming a Correlation Structure for Testlet Models :
Application for Evaluation Data of Education and
Research of Japanese National Universities

森 一将, 大森 拓哉, 繁桝 算男

MORI Kazumasa, OHMORI Takuya, SHIGEMASU Kazuo

1. はじめに.....	3
2. テストレットに対する IRT モデルの提案.....	4
3. シミュレーション研究.....	6
4. 事例分析.....	8
4.1 北アイルランドの標準テストデータ.....	8
4.2 日本における国立大学法人等の教育研究評価データ.....	10
4.3 事例分析の考察.....	13
5. まとめと今後の展開.....	13
補足：マルコフ連鎖モンテカルロ法による母数推定について.....	14
参考文献.....	14
ABSTRACT.....	16

相関構造を仮定したテストレットモデルのベイズ推論

—国立大学法人等の教育研究評価データへの適用—

森 一将*, 大森 拓哉**, 繁樹 算男***

要 旨

大問形式のテストに代表されるテストレットに対して新しいIRT（項目反応理論）モデルを提案した。提案モデルはテストレット群内の関係性をより詳細に反映させるように能力母数の分散行列に相関構造を導入したものである。提案モデルはシミュレーション研究により推定精度が評価され既存モデルよりも全体として適切な推定を行うことが検証された。

事例分析として、モデルは学力テストデータと国立大学法人等の教育研究評価データに適用され、既存モデルと比較した母数推定の特徴が明らかにされた。同時に国立大学法人等の教育研究評価データがテストレットとみなせることを明らかにし、より合理的・統合的な評価手法としてIRTの適用可能性を示した。提案モデルは学力テストの結果に限らず、授業評価や人事評価など様々なデータへの適用が期待される。

キーワード

テストレット, 局所独立性, 自己回帰共分散構造, 教育評価データ, 研究評価データ, ベイズ推論

1. はじめに

特定の知識領域についての1つの文脈に関連した項目群のことをテストレットという(Wainer and Kiely, 1987)。テストレットは、大問形式からなるテストがその典型例であり、近年 Computer Based Test における適応型テストの発達と共に日本においても注目されるようになってきた。

このテストレットにおける大きな問題の1つが局所独立性の仮定の問題である。局所独立性は、Rosenbaum (1988) に基づくと以下のような仮定として定義される。あるテストのある項目への反応について、誤答を0, 正答を1とし、このような2種類の値を取る確率変数を $X = \{X_1, \dots, X_M\}$ とする。この確率変数 X はその背後に存在する潜在変数 θ と共に構造化されるとする。このとき、局所独立性は以下のように定義される。

$$X_1 \perp X_2 \perp \dots \perp X_M \mid \theta \quad (1.1)$$

つまり、潜在変数 θ を所与とした場合における項目反応の発生確率は独立であるという仮定である。これは、ある被験者が同質の問題を複数回答したとき、それぞれの正答・誤答の反応は、被験者の能力や問題の内容に依存せず、ランダムに発生するという状況に対応する。テストレットからなる項目反応データに項目反応理論 (Item Response Theory; IRT) モデルを適用すると、推定が過剰に行われることは様々な研究で指摘されている。例えば、Thissen et al. (1989) においては読解力テストの事例を用い、(1.1) の局所独立性に関する評価を行っている。この研究では読解力テストをいくつかのテストレットから構成されるものとみなし、局所独立性を仮定するモデルと仮定しないモデルによる情報関数曲線の比較を行った。同時にこれらのモデルより算出された母数の推定値と、より妥当性の高い言語テストの相関を計算した。その結果、テストレットにおいては局所独立性を仮定しないモデルは言語テストとの相関が比

* 東京大学大学院 総合文化研究科

** 多摩大学 経営情報学部

*** 帝京大学 文学部 心理学科

較的低い一方で、情報関数曲線の変化がより大きく推定されてしまうという過剰推定の状態にあることを指摘した。より以前の研究における同様の指摘については、Wainer et al. (2007) に詳しく見られる。

この局所独立の問題に対して、Rosenbaum (1988) は以下のようなテストレットに適した独立性に関する仮定を提案した。項目反応 $X_1, \dots, X_M$ に対して互いに要素が重複しない $L \leq M$ 個の部分集合 $V_1, \dots, V_L$ を考え、次のような独立性を定義する。

$$V_1 \perp V_2 \perp \dots \perp V_L \mid \Theta \quad (1.2)$$

部分集合 $V_1, \dots, V_L$ をテストレットとして扱うと、この仮定は潜在変数 Θ 所与の下において、テストレット群間の独立のみを仮定し、群内における項目については (1.1) のような局所独立の仮定を行わない制約条件が緩和されたモデルであることがわかる。これは、ある被験者が同質の問題を複数回答したとき、問題が同じテストレットに属さない場合には局所独立の場合と同様に正答・誤答の反応がランダムに発生する一方で、同じテストレットに属する場合にはこれらの反応に、相関関係が発生してしまうという状況に対応する。この仮定はテストレットにおける状況をより適切にあらわしており、その後の多くの IRT モデルに用いられている。

一方で、従来の IRT テストレットモデルにおいてはテストレット群内の項目間の関係については比較的注意が払われてこられず、多くの研究において群内における相関が一定であるとみなされてきた。ところが、この設定は大問形式テストなど、ある分野の学力テストにおいては実際的なものとはいえない。加えて、学力テスト以外の評価データにおいてもこの設定が実際的でない場合がある。例えば、大学の教育研究活動評価においては、各評価項目の背後に特定の知識領域に関する文脈（例えば「教育」、「研究」）を仮定したテストレットとみなすことが妥当な場合がある。加えて、実際的には、このテストレット群内においても「教育実施体制」などの基礎的な評価項目と「学業の成果」などの応用的な評価項目に分類できることがある。この場合において、基礎的な評価項目間の相関と、基礎的な評価項目と応用的な評価項目

の間の相関は必ずしも等しくないことが容易に考えられる。このように、一般的な状況においてはテストレット群内の相関は一定でないと考えるのが妥当である。

これらの問題に対して、本研究ではテストレット群内における能力母数に関して潜在構造を設定し、より適切な母数推定を行えるモデルの提案を行う。加えて事例分析を行い、大学評価に代表されるより広範な範囲の実例への適用について検討を行う。

本論は4節から成る。2節において従来モデルの検討と新しいIRTの提案を行う。提案されたモデルは3節においてシミュレーション研究によりその推定精度が評価・考察される。4節では事例分析が示される。対象は、従来のIRTの主な適用範囲である学力テストデータと、拡張的事例としての大学評価データの2つである。全体的なまとめと今後の課題は5節で示される。

2. テストレットに対するIRTモデルの提案

まず、IRTモデルに対しての定式化を行う。 i 番目の被験者の j 番目のテストレットの k 番目の項目に対する反応を $x_{ijk} = \{0, 1\}$ ($i = 1, \dots, I, j = 1, \dots, J, k = 1, \dots, n_j$) とおき、以下のようなモデルを設定する。

$$P(x_{ijk} = 1 \mid \Theta) = \Phi(u_{ijk}) \quad (2.1)$$

ここで、 P は確率、 Φ は正規累積関数、 u_{ijk} は被験者 i の項目 (j, k) に対する潜在得点、 Θ は推定の対象となる母数全体の集合を示す。このモデルをIRTにおける正規累積モデル (Lord and Novick, 1968) という。(2.1)において、項目反応 x_{ijk} が発生する確率は、他の項目への反応や他の要因は所与とされておらず、 Θ のみが所与とされている。つまり、母数集合 Θ が所与の下で、項目反応 x_{ijk} が独立と仮定されていることがわかる。したがって、このモデルは、(1.1) で定義された局所独立の仮定を満たすモデルであることがわかる。このモデルは、(2.1) の潜在得点 u_{ijk} と被験者 i の能力を示す潜在変数 θ_i の間に線形関係を設定し利用される。例えば一般的によく利用されるものの1つである2母数モデルでは u_{ijk} に以下のような仮定をおく。

$$u_{ijk} = \alpha_{jk}(\theta_i - \beta_{jk}), \theta_i \sim N(0, 1) \quad (2.2)$$

ここで, N は正規分布を示し, θ_i は被験者 i の真の能力を示す母数を示す。また, α_{jk} , β_{jk} は項目 jk の反応傾向を制御する母数でそれぞれ識別力母数, 困難度母数という。また, この 2 母数正規累積モデルの応用として, 正答反応がなされる確率 (2.1) を以下のロジスティック関数に置き換えたモデルも提案されている (Birnbaum, 1968)。

$$P(x_{ijk}=1 | \Theta) = \frac{1}{1 + \exp\{-D\alpha_{jk}(\theta_i - \beta_{jk})\}} \quad (2.3)$$

ここで, $D = 1.7$ は正規近似のための定数である。一方, 2 母数正規累積モデルの拡張として, 次のような誤差を仮定した変量効果モデルも提案されている (Albert (1990), Baker (1998) など)。

$$u_{ijk} = \alpha_{jk}(\theta_i - \beta_{jk}) + \varepsilon_{ijk}, \theta_i \sim N(0, 1), \\ \varepsilon_{ijk} \sim N(0, \sigma_\varepsilon^2) \quad (2.4)$$

ここで, ε_{ijk} は誤差項を示す。

テストレットにおいては, 同一テストレット群内に属する項目の反応傾向に関して相関関係が観察される場合が多い。したがって, この様な場合, 潜在的に仮定する潜在得点 u_{ijk} についても (1.1)

(2.1) のような局所独立の仮定が成立しないと考えられ, 従来から様々なモデルが提案されてきた。代表的なテストレット IRT モデルに, Bladlow et al. (1999) が (2.4) の変量効果モデルを元にして提案した次のようなモデルがある。

$$u_{ijk} = \alpha_{jk}(\theta_i - \beta_{jk} - \gamma_{ik}) + \varepsilon_{ijk}, \theta \sim N(0, 1), \\ \varepsilon_{ijk} \sim N(0, \sigma_\varepsilon^2) \quad (2.5)$$

このモデルでは, 2 母数モデル (2.5) に対して新たに γ_{ik} という被験者 i , テストレット群 j に依存する母数が追加される。これは, 特定のテストレットが被験者に及ぼす効果を困難度に反映したものとも考えられ, 特定被験者に対するテストレット効果と呼ばれる。Bladlow et al. (1999) は, (2.5) の設定の下で, 被験者 i に対する 2 つの異なる項目の潜在得点 u_{ijk} の相関を解析的に計算し, 同一テストレット群 j に属する場合には高く, 属さない場合には低くなることを示した。この場合にはテストレット群内における各項目の相関はすべて一定の値をとる。ところが, 大問形式のテストや

大学教育における専門科目のように, 一般的な状況においては, 同一カテゴリ (テストレット) 内においても相関の高低に差が生じてしまうことがしばしば発生する。これは同一カテゴリにおいても, 近接する問題や科目については類似する特有知識や能力を利用することにより相関が高くなり, 近接しないものに対してはあまり類似する知識や能力を利用しないため相関が低くなるというごく妥当な状況により生じていると考えられる。この場合 (2.5) のような同一テストレット内の相関を一定にするモデル設定は必ずしも適切であるとは言えず, あらたなモデルの枠組みが必要となってくる。

そこで, 本研究では (2.1) の正規累積モデルの潜在得点 u_{ijk} に対し, 2 段階モデルを提案する。

$$u_{ij}^* \sim N(\theta_i 1_{n_j}, \Sigma_j), \theta_i \sim N(0, 1) \quad (2.6)$$

$$u_{ijk} = \alpha_{jk}(u_{ij}^* - \beta_{jk}) + \varepsilon_{ijk}, \varepsilon_{ijk} \sim N(0, \sigma_\varepsilon^2) \quad (2.7)$$

ここで, 1_{n_j} は全要素が 1 である $n_j \times 1$ の列ベクトル, u_{ijk} は被験者の真の能力を補正する母数である。これは, 先ほどの例における特有知識や技能を考慮した, テストレットで測定する「部分的な能力」に相当すると考える。また, (2.7) において, 任意の 2 つの潜在得点 u_{ijk}^* と $u_{i'j'k}^*$ は $i \neq i'$ か $j \neq j'$ であるときに独立であることが明らかであるので, (1.2) のテストレットにおける独立性の仮定を満たしていることがわかる。加えて, (2.6) の分散共分散行列に以下のような制約を入れる。

$$\Sigma_j = \begin{bmatrix} \sigma^2 & \sigma_{12} & \cdots & \sigma_{1n_j} \\ \sigma_{21} & \cdot & \cdot & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \sigma_{n_j 1} & \cdots & \cdots & \sigma^2 \end{bmatrix} = \sigma^2$$

$$\begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{n_j-1} & \rho_{n_j} \\ \rho_1 & 1 & \rho_1 & \cdots & \cdot & \rho_{n_j-1} \\ \rho_2 & \rho_1 & \vdots & \vdots & \cdot & \rho_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho_{n_j-1} & \cdot & \vdots & \cdot & \cdot & \rho_1 \\ \rho_{n_j} & \rho_{n_j-1} & \cdots & \rho_2 & \rho_1 & 1 \end{bmatrix} = \sigma^2 R_j \quad (2.8)$$

$$\rho_1 \geq \cdots \geq \rho_{n_j}$$

ただし, R_j は $n_j \times n_j$ の行列であり, 以下のように定義される相関行列である。

$$\begin{cases} r_{kk} = 1 & k' = k \\ r_{kk} = \rho_{|k' - k|} & k' \neq k \end{cases} \quad (2.9)$$

(2.8), (2.9) で示されたような特殊な分散共分散行列の構造を自己回帰共分散構造 (Autoregressive Covariance Structure) という。自己回帰共分散構造は成長曲線分析を中心とした時系列解析モデルでしばしば利用される仮定であるが (例えば, Lee (1988), Fujikoshi et al. (1990), Lee and Chang (2000) など), 本研究ではこの構造をテストレットに適用し, 同一群内における部分的能力 u_{ijk}^* の相関関係のモデル化を行う。

本モデルに類似するものとして多次元 IRT (Multidimensional IRT; MIRT), がある。しかし MIRT は能力母数 θ_i を多次元化してしまう設定をとったため, (例えば, Beguin and Glas (2001) や Zheng (2000)), 推定された能力母数の評価が比較的難しくなるという側面を持っている。また, このような MIRT のテストレットへの適用においても多次元化された θ_i の分散共分散行列を対角行列として扱い, 必ずしも本研究で指摘されているテストレット群内における反応傾向の相関関係を適切にモデル化していないと思われるものも存在する (Wang and Wilson (2005))。本研究においては, 自己回帰共分散構造の導入により, 近接する項目への反応の類似性の高さを適切にモデル化し, かつこのような類似性の影響を真の能力に反映させない, より柔軟な分析を可能にした。

また, 本研究ではベイズアンプローチを導入する。ベイズアンプローチでは, まず推定対象の母数に対して観測データ (テストデータ) を得る前の知識を事前分布によって表現する。この事前分布と母数の尤度を表すモデル関数を用いて, 観測データ所与の下での母数の確率分布 (事後分布) を求め, これを基にして母数の推定を行う (繁桝, 1985)。ベイズアンプローチは, テストデータを得る前に分析者が持つ知識を事前分布という形で取り入れることができるため, 柔軟にモデル構築・推定が行えるという利点を持つ。ベイズアンプローチの導入のために, (2.6), (2.8), (2.9) をモデル関数とみなし, 推定対象の母数 α_{jk} , β_{jk} , ρ_k に対して以下の事前分布を設定する。

$$\begin{aligned} \alpha_{jk} &\sim N(\mu_\alpha, \sigma_\alpha^2), \beta_{jk} \sim N(\mu_\beta, \sigma_\beta^2), \\ \rho_k &\sim \text{TN}_{a, b}(\mu_\rho, a, b, \sigma_\rho^2). \end{aligned} \quad (2.10)$$

ここで, μ_α , σ_α^2 , μ_β , σ_β^2 , μ_ρ , a , b , σ_ρ^2 は事前分布それぞれの母数, $\text{TN}_{a, b}$ は区間 $[a, b]$ における切断正規分布を示す。

また, 一般的に (2.8) のような複雑な構造を持つ母数の推定は計算が複雑になるため困難である場合が多い。本研究では, このような困難性を軽減するために, ベイズアンプローチにおける数値計算アルゴリズムの 1 種である Gibbs Sampler (Geman and Geman, 1984) をはじめとしたマルコフ連鎖モンテカルロ法を導入する。マルコフ連鎖モンテカルロ法を用いた場合における推定方法については本論末の補足に記す。

3. シミュレーション研究

前節で提案されたモデルに対してシミュレーション研究によりその母数推定精度を評価した。まずは, シミュレーション研究による評価のためにデータセットを用意した。このデータセットは $\sum_{j=1}^J n_j$ 個の項目より構成されている。そのうち, 50% のデータを (2.1) (2.2) により, 残りの 50% のデータを (2.1) (2.6) (2.7) (2.8) より発生させた。このとき, 後者の 50% のデータには局所独立は仮定できず, 1 種のテストレットとみなすことが出来るため, 通常の単問テストとテストレットの混合という一般的な状況を仮定したものといえる。また, データにおいては, 単問テストとテストレットが等比率で含まれているため, モデルの構造による有利・不利が極力存在しないような設定となっている。このようなデータセットを 100 セット生成した。なお, 各データセットについて $I=100$, $J=8$, $n_j=8$ (すべてのテストレットで一定) とした。また, テストレットにおける各項目の相関は以下のように設定した。

$$R_j = \begin{bmatrix} 1 & & & & & & & \\ 0.6 & 1 & & & & & & \\ 0.5 & 0.6 & 1 & & & & & \\ 0.4 & 0.5 & 0.6 & 1 & & & & \\ 0.3 & 0.4 & 0.5 & 0.6 & 1 & & & \\ 0.2 & 0.3 & 0.4 & 0.5 & 0.6 & 1 & & \\ 0.15 & 0.2 & 0.3 & 0.4 & 0.5 & 0.6 & 1 & \\ 0.1 & 0.15 & 0.2 & 0.3 & 0.4 & 0.5 & 0.6 & 1 \end{bmatrix}$$

これらのデータセットに対して，先行研究（Bladlow et al., 1999）に基づき 4 つのモデルを適用し，能力母数 θ ，識別力母数 α ，困難度母数 β に対する推定精度の評価を行った。適用したモデルは以下の通り。

モデル 1 (2.1) (2.7) (2.8) を用いた本研究で提案された新しいテストレットモデル。

モデル 2 (2.1) (2.6) を用いた Bladlow et al. (1999) におけるテストレットモデル。

モデル 3 (2.1) (2.5) を用いた 2 母数正規累積 IRT における変量効果モデル。

モデル 4 (2.1) (2.3) を用いた 2 母数ロジスティックモデル。

なお，モデル 1，2，3 についてはソフトウェア Open BUGS (Spiegelhalter et al., 2007) を用いて Gibbs Sampler による推定を行い，モデル 4 については統計ソフトウェア R のパッケージ ltm (Rizopoulos, 2006) を用いて最尤法による推定を行った。

推定精度評価は平均絶対誤差 (Mean Absolute Error; MAE¹) と相関係数により行った。推定精度の評価結果を表 1 に示す。IRT における主な目的は被験者の能力母数 θ を正しく測定することであるが，提案モデルはテストレットが混在する状

況下において他のモデルよりも正しく θ を推定していることがわかる。また同様に困難度母数 β についても推定精度も他のモデルに比べて高い。識別力母数 α については，Bladlow et al. (1999) によるテストレットモデル (モデル 2)，及び 2 母数正規累積 IRT における変量効果モデル (モデル 3) の 2 つのベイズ推定よりは精度が低くなっているが，よく用いられる最尤法による 2 母数ロジスティックモデル (モデル 4) よりは精度が高くなった。

また，各モデルの推定値と真値の相関行列を表 2～表 4 に示す。表 2 より，能力母数 θ に関しては，各モデルとも真値—推定値の相関が非常に高いことが分かる。その一方で，識別力母数 α はどの推定値も真値との相関は高いとはいえない (表 3)。その中でも提案モデル (モデル 1) は比較となる他のモデルに比べて，比較的高い相関を示している。このことより，提案モデルは識別力母数 α の推定においても限定的ではあるがある程度の推定力を持っているとも考えられる。困難度母数 β についてはテストレットを前提としたモデル (モデル 1，2) は 0.6 以上と高い相関を示しているが，一般的なモデル (モデル 3，4) は相関が 0.5 前後と比較的低くなっている (表 4)。

以上より，提案したモデルは群内に相関関係を想定したテストレットが存在するような通常のテ

表 1 シミュレーションによる推定精度の評価

推定値対象の母数	推定値の MAE			
	モデル 1	モデル 2	モデル 3	モデル 4
能力母数 θ	109.48	124.80	110.03	115.66
識別力母数 α	15.99	7.34	13.71	20.72
困難度母数 β	16.97	19.38	17.04	29.68

表 2 シミュレーションによる能力母数 θ に関する真値—推定値の相関行列

	真値	モデル 1	モデル 2	モデル 3	モデル 4
真値	1.00	0.97	0.97	0.97	0.96
モデル 1	0.97	1.00	0.99	0.99	0.98
モデル 2	0.97	0.99	1.00	0.99	0.98
モデル 3	0.97	0.99	0.99	1.00	0.98
モデル 4	0.96	0.98	0.98	0.98	1.00

¹ MAE は $|(\text{母数の推定値}) - (\text{母数の真値})|$ の平均として示され，その値が小さい方がより良い推定を行っているといみなされる指標である。

表3 シミュレーションによる識別力母数 α に関する真値—推定値の相関行列

	真値	モデル 1	モデル 2	モデル 3	モデル 4
真値	1.00	0.29	0.24	0.23	0.13
モデル 1	0.29	1.00	0.44	0.46	0.22
モデル 2	0.24	0.44	1.00	0.29	0.16
モデル 3	0.23	0.46	0.29	1.00	0.15
モデル 4	0.13	0.22	0.16	0.15	1.00

表4 シミュレーションによる困難度母数 β に関する真値—推定値の相関行列

	真値	モデル 1	モデル 2	モデル 3	モデル 4
真値	1.00	0.72	0.60	0.50	0.49
モデル 1	0.72	1.00	0.52	0.44	0.38
モデル 2	0.60	0.52	1.00	0.31	0.28
モデル 3	0.50	0.44	0.31	1.00	0.20
モデル 4	0.49	0.38	0.28	0.20	1.00

ストにおいて全体的により適切な推定を行うということがいえる。

4. 事例分析

提案モデル及び既存のモデルをテストレット型の事例データに適用し、その母数推定の特長について考察する。適用される事例は、北アイルランドの標準テストデータと日本の国立大学法人等の教育研究評価データである。

4.1 北アイルランドの標準テストデータ

第1の事例分析は、北アイルランドの小学校における標準テスト導入時の得点データ（標準テストデータ）を用いる（Airasian et al., 1981）。このデータは、1973—76年に渡り、北アイルランドにおいて標準テストを導入した際の教師、生徒、父兄などへの影響を評価するために取得され、175の小学校、のべ 40,000人以上の学生が参加した。本節ではそのうち、調査4年目（1976年）における4年生の評価テストデータを用いる。評価テストは、算数、アイルランド語、英語の3科目であり、それぞれは以下のような複数の単元項目から構成される。

- 算数：計算，概念テスト，総合問題。
- アイルランド語：読解（語彙），読解（理解力），用法，スペリング。
- 英語：読解（単語分析），読解（理解力），文

法（大文字小文字の区別，句読点の用法），文法（用法，会話文の選択），スペリング

この複数の単元項目は算数，アイルランド語，英語という各科目それぞれごとについて関連性を持っている。したがって，各科目を単元項目から構成される1種のテストレットとみなすことが可能である。

データセットでは，上記のそれぞれの単元項目について各受験者（学生）の得点データが与えられている。分析ではこれを上位20%以内の得点（1）とそれ以下の得点（0）と変換し，単元項目ごとの2値データを作成した。対象となったデータは3,850（人）である。

変換後の2値データに対して，単元項目ごとの相関係数（ ϕ 係数）を計算した結果を表5に示す。この表に示された相関係数について代表値を計算すると，テストレット群内の相関係数の中央値は0.53である一方，テストレット群内の相関係数は0.41と差が生じていることがわかる。項目反応の相関係数が高いということは，(2.1) (2.6) (2.7)で導かれるように，その背後に存在する潜在得点 u_{ijk} の相関が高いことを意味する。したがって，群内の相関係数が高いことはテストレットにおいて背後に存在する潜在得点 u_{ijk} に相関関係が存在することを示唆している。加えて，テストレット群内の各項目についても， ϕ 係数が一定ではなく，適切な並べ替えを行うことで(2.8)のような構造

に近くなることが分かる。したがって、前述の議論と同様に、テストレット群内において事例データ (2.8) の構造を仮定することは一定の妥当性を持つ。

この2値データに3節で適用した4つのモデルを適用した。識別力母数 α と困難度母数 β の推定結果を表6，表7に示す。困難度母数 β の推定値が最尤法による2母数ロジスティックモデル（モデル4）で他よりも低く推定されている。これはテストレット構造を仮定しないモデル構造である

こと，比較的外れ値の影響を受けやすい固定効果モデルであること，推定に最尤法を用いている事などの理由から推定が不安定になったものと考え

る。また，能力母数 θ ，識別力母数 α ，困難度母数 β に対する各モデルの推定値の相関を表8～表10に示す。能力母数 θ ，困難度母数 β についてはシミュレーションと同様に各モデルとも非常に強い相関を示している。識別力母数 α については，提案モデルを含むテストレットを前提としたモデ

表5 標準テストデータにおける単元項目間の相関

評価項目		算数			アイルランド語				英語				
		1	2	3	1	2	3	4	1	2	3	4	5
算数	1. 総合問題	1.00											
	2. 計算	0.57	1.00										
	3. 概念テスト	0.55	0.45	1.00									
アイルランド語	1. 読解（語彙）	0.43	0.41	0.41	1.00								
	2. 読解（理解力）	0.44	0.42	0.41	0.65	1.00							
	3. スペリング	0.37	0.40	0.36	0.61	0.53	1.00						
	4. 用法	0.35	0.35	0.34	0.58	0.54	0.52	1.00					
英語	1. 読解（理解力）	0.49	0.40	0.43	0.47	0.49	0.42	0.41	1.00				
	2. 読解（単語分析）	0.49	0.38	0.42	0.48	0.47	0.42	0.42	0.63	1.00			
	3. スペリング	0.41	0.38	0.36	0.50	0.48	0.56	0.43	0.53	0.53	1.00		
	4. 文法（用法・会話）	0.37	0.33	0.36	0.41	0.40	0.38	0.38	0.49	0.52	0.49	1.00	
	5. 文法（大文字小文字・句読点）	0.48	0.42	0.44	0.45	0.46	0.41	0.38	0.49	0.49	0.46	0.43	1.00

網掛け部分はテストレット群内をしめし，それ以外の部分はテストレット群間を示す。

また，テストレット群内の相関構造を分かりやすく見せるために一部項目の並べ替えを行った。並べ替えは，次のような手順によって行った。

1. テストレット群内の各単元項目ごとの ϕ 係数の合計値を計算する。
2. ϕ 係数の合計値が一番大きい各単元項目を一番右に並べ替える。
3. その他の項目をこの項目との ϕ 係数が降順になるように並べ替える。

表6 標準テストデータにおける識別力母数 α に関する各モデルによる推定値

		モデル1	モデル2	モデル3	モデル4
算数	計算	2.28	1.37	1.57	1.84
	概念テスト	2.16	1.31	1.55	1.87
	総合問題	3.78	2.74	1.88	2.26
アイルランド語	読解（語彙）	3.85	3.50	2.72	3.18
	読解（理解力）	2.23	1.43	1.80	2.10
	用法	3.33	1.98	2.54	2.94
	スペリング	2.74	1.55	2.04	2.42
英語	読解（単語分析）	3.18	2.44	2.21	2.80
	読解（理解力）	3.25	1.97	2.26	2.83
	文法（用法・会話）	1.91	1.20	1.60	1.94
	文法（大小文字・句読点）	1.85	1.20	1.88	2.26
	スペリング	2.52	1.42	2.16	2.59

表7 標準テストデータにおける困難度母数 β に関する各モデルによる推定値

		モデル 1	モデル 2	モデル 3	モデル 4
算数	計算	-0.09	-0.11	-0.07	-0.16
	概念テスト	-0.27	-0.36	-0.26	-0.50
	総合問題	-0.03	-0.03	-0.03	-0.08
アイルランド語	読解 (語彙)	-0.16	-0.18	-0.12	-0.41
	用法	-0.05	-0.05	-0.04	-0.09
	読解 (理解力)	-0.12	-0.14	-0.10	-0.30
	スペリング	-0.16	-0.19	-0.13	-0.32
英語	読解 (単語分析)	-0.19	-0.24	-0.17	-0.52
	読解 (理解力)	-0.14	-0.17	-0.13	-0.39
	文法 (用法・会話)	-0.33	-0.44	-0.30	-0.62
	文法 (大小文字・句読点)	-0.19	-0.25	-0.17	-0.39
	スペリング	-0.09	-0.11	-0.09	-0.23

表8 標準テストデータにおける能力母数 θ に関する各モデル推定値の相関行列

	モデル 1	モデル 2	モデル 3	モデル 4
モデル 1	1.00	0.97	0.98	0.98
モデル 2	0.97	1.00	0.97	0.98
モデル 3	0.98	0.97	1.00	0.99
モデル 4	0.98	0.98	0.99	1.00

表9 標準テストデータにおける識別力母数 α に関する各モデル推定値の相関行列

	モデル 1	モデル 2	モデル 3	モデル 4
モデル 1	1.00	0.92	0.75	0.75
モデル 2	0.92	1.00	0.70	0.69
モデル 3	0.75	0.70	1.00	0.99
モデル 4	0.75	0.69	0.99	1.00

表10 標準テストデータにおける困難度母数 β に関する各モデル推定値の相関行列

	モデル 1	モデル 2	モデル 3	モデル 4
モデル 1	1.00	1.00	0.99	0.94
モデル 2	1.00	1.00	1.00	0.93
モデル 3	0.99	1.00	1.00	0.93
モデル 4	0.94	0.93	0.93	1.00

ル (モデル 1, 2) が非常に強い相関を示していることが見て取れる。このことから提案モデルは既存のテストレットモデルと整合的な推定を行っていることが標準テストデータの分析からも示された。

4.2 日本における国立大学法人等の教育研究評価データ

第2の事例分析は独立行政法人大学評価・学位

授与機構が2008年に発表した「国立大学法人等の中期目標期間に係る教育研究の状況 (平成16年度～平成19年度)」に関するデータ (国立大学法人等の教育研究評価データ) である (大学評価・学位授与機構, 2008)。この評価データは, 国立大学法人等の学部・研究科を対象にして, 教育水準, 研究水準の状況についての評価を第3者機関である大学評価・学位授与機構が行ったものであり, 「中期目標の達成状況」と「学部・研究科等の現況分

析」の2つから構成される。事例分析では、学部・研究科等の現況分析に含まれる「教育の分析項目ごとの水準」と「研究の分析項目ごとの水準」を抜き出し、データ化したものを用いる。

学部・研究科等の現況分析において利用した評価項目は以下の7項目である。

- 教育水準：教育実施体制，教育内容，教育方法，学業の成果，進路・就職の状況。
- 研究水準：研究活動の状況，研究成果の状況。

この複数の評価項目は，教育水準，研究水準というより総合的な評価項目ごとに関連性を持っており，テストレットの1種とみなすことが可能である。

データセットにおいて，各評価項目は，「期待される水準を大きく上回る」，「期待される水準を上回る」，「期待される水準にある」，「期待される水準を下回る」の4段階評価を行っている。本研究における分析ではこの評価水準を以下のように加工し2値データを作成した²。

- 「期待される水準を大きく上回る」，「期待される水準を上回る」⇒期待される水準より優れている (1)
- 「期待される水準にある」，「期待される水準を下回る」⇒期待される水準以下である (0)

教育水準の評価結果（5項目）及び研究水準の評価結果（2項目）のすべての項目に欠損値がない組織のみをその分析対象とした。対象となったデータは225（学科・研究科）である。

標準テストの事例と同様に変換後の2値データに対して，評価項目ごとの相関係数（ ϕ 係数）を計算した結果を表11に示す³。テストレット群間の相関に一部高い値が見られるが，この表に示された相関係数について代表値を計算すると，テストレット群内の相関係数の中央値は0.34である一方，テストレット群内の相関係数は0.25と差が生じていることがわかる。したがって，前述の議論と同様にテストレット群内において事例データ（2.8）の構造を仮定することは一定の妥当性を持つ。

この2値データに3節で適用した4つのモデルを適用した。識別力母数 α と困難度母数 β の推定結果を表12，表13に示す。困難度母数 β の推定値が最尤法による2母数ロジスティックモデル（モデル4）で他よりも高く推定されている。これは標準テストデータの分析結果と同様に推定の不安定さから来るものと考えられる。加えて， β の推定において教育水準に関するテストレットの方が研究水準に関するテストレットよりも値が大きくなっている。これは，国立大学法人の真の能力（教育・研究などを行うための潜在的能力）が第

表11 国立大学法人等の教育研究評価データにおける項目間の相関

評価項目		教育水準					研究水準	
		1	2	3	4	5	1	2
教育水準	1. 教育実施体制	1.00						
	2. 教育内容	0.58	1.00					
	3. 教育方法	0.37	0.39	1.00				
	4. 学業の成果	0.36	0.34	0.25	1.00			
	5. 進路・就職の状況	0.23	0.22	0.21	0.28	1.00		
研究水準	1. 研究活動の状況	0.31	0.44	0.21	0.18	0.23	1.00	
	2. 研究成果の状況	0.28	0.37	0.18	0.20	0.29	0.66	1.00

² 実際のデータは，大学評価・学位授与機構のホームページ（http://www.niad.ac.jp/n_hyouka/kokuritsu/hyoukakekka/index.html）より，各法人の評価データを入手し，集計・加工を行った。また，2値データの分割方法については，元のデータが4段階であること，変換後のデータによる比率に著しい偏りが生じないことを考慮し，本文の通りに決定した。

³ 網掛け部分はテストレット群内をしめし，それ以外の部分はテストレット群間を示す。また，テストレット群内において，標準テストデータの分析事例と同様の手順により項目の並べ替えを行った。

表12 国立大学法人等の教育研究評価データにおける識別力母数 α に関する各モデルによる推定値

		モデル 1	モデル 2	モデル 3	モデル 4
教育水準	教育実施体制	1.86	1.77	2.28	1.95
	教育内容	2.71	1.84	3.13	2.71
	教育方法	1.10	1.04	1.09	1.16
	学業の成果	1.72	1.03	1.16	1.26
	進路・就職の状況	1.26	0.74	1.10	1.27
研究水準	研究活動の状況	4.00	2.33	1.99	3.70
	研究成果の状況	4.25	2.84	1.66	3.01

表13 国立大学法人等の教育研究評価データにおける困難度母数 β に関する各モデルによる推定値

		モデル 1	モデル 2	モデル 3	モデル 4
教育水準	教育実施体制	1.26	1.33	1.03	2.32
	教育内容	1.00	1.11	0.80	2.45
	教育方法	1.43	1.53	1.35	1.71
	学業の成果	1.57	2.11	1.74	2.39
	進路・就職の状況	1.85	2.55	1.84	2.50
研究水準	研究活動の状況	0.54	0.54	0.39	1.42
	研究成果の状況	0.49	0.49	0.42	1.17

表14 国立大学法人等の教育研究評価データにおける能力母数 θ に関する各モデル推定値の相関行列

	モデル 1	モデル 2	モデル 3	モデル 4
モデル 1	1.00	0.94	0.94	0.96
モデル 2	0.94	1.00	0.95	0.97
モデル 3	0.94	0.95	1.00	0.97
モデル 4	0.96	0.97	0.97	1.00

表15 国立大学法人等の教育研究評価データにおける識別力母数 α に関する各モデル推定値の相関行列

	モデル 1	モデル 2	モデル 3	モデル 4
モデル 1	1.00	0.94	0.40	0.94
モデル 2	0.94	1.00	0.50	0.89
モデル 3	0.40	0.50	1.00	0.59
モデル 4	0.94	0.89	0.59	1.00

表16 国立大学法人等の教育研究評価データにおける困難度母数 β に関する各モデル推定値の相関行列

	モデル 1	モデル 2	モデル 3	モデル 4
モデル 1	1.00	0.98	0.98	0.77
モデル 2	0.98	1.00	0.99	0.76
モデル 3	0.98	0.99	1.00	0.70
モデル 4	0.77	0.76	0.70	1.00

3 者から評価されやすい項目は教育分野よりも研究分野であると解釈される。ただし、この傾向が国立大学法人以外のすべての大学組織に共通する一般的傾向であるかどうかは、今後より広範な対

象による検討を行う必要がある。

また、能力母数 θ 、識別力母数 α 、困難度母数 β に対する各モデルの推定値の相関を表14～表16に示す。能力母数 θ についてはシミュレーショ

ンや標準テストデータの事例と同様に各モデルとも非常に強い相関を示している(表14)。ところが、識別力母数 α においては、2母数正規累積IRTにおける変量効果モデル(モデル3)が他の3つのモデルに比べて低い相関を示している(表15)。同様に、困難度母数 β においては2母数ロジスティックモデル(モデル4)が他の3つに比べてやや低い相関を示している(表16)。以上より、3つの母数 θ , α , β すべてにおいて整合的な推定を行っているモデルは、提案モデル(モデル1)と既存のテストレットモデル(モデル2)であることがわかる。

4.3 事例分析の考察

本節においては、テストレット型の事例データとして北アイルランドの標準テストデータと日本の国立大学法人等の教育研究評価データを適用・分析した。2つの事例における推定では、能力母数 θ の推定値が共通して高い相関を持つ結果を示した。これは前節で示されたシミュレーション研究の結果と整合的である。

識別力母数 α , 困難度母数 β については、テストレットを仮定しないモデル(2母数正規累積IRTにおける変量効果モデル, 2母数ロジスティックモデル)についてはその相関の度合いにばらつきが生じた。これについては、対象データがテストレットの形式を持ち、局所独立性の仮定が成立しないため、推定値が安定していないことにより生じているとも考えられる。その一方でテストレットを仮定して構築されたモデル(提案モデル, 既存のテストレットモデル)は一貫して相関の高い推定を行っている。このことより、提案モデルは実際のテストレットデータにおいて、先行研究(Bradlow et al., 1999)におけるモデルが類似した推定を行っていることがわかる。

また、本節においては、IRTモデルの適応範囲の拡張として国立大学法人等の教育研究評価データを分析した。結果は前述の通り、シミュレーション研究や学力テストの分析結果と整合的であり、テストレットデータとしての扱いが可能であることが明らかになった。IRTに関する代表的なテキストである豊田(2002a, b)で示されているように従来のIRTの適用は人間(個人)を主な対象としており、組織や組織が提供するサービスに

対する適用はまだ少ない。特に、大学評価は近年多面化かつ詳細化してきており、多くの項目に基づく複雑性の高い評価が行なわれており、項目を合理的に統合するための分析手法の要請も高い。これに対し、国立大学法人等の教育研究評価データをテストレットとみなし、IRTを用いて分析することにより、大学自体の総合的な評価を能力母数 θ を用いて推定する方法は単純で分かりやすく、学内で行われる学力テストの分析結果などの既存のほかのIRTによる分析結果との統合も行いやすいという利点を持つ。以上より、事例分析で適用可能性が明らかになったテストレットIRTの国立大学法人等の教育研究評価データへの適用は評価の統合的理解に対しより合理的かつ有効な手段であると考えられる。

5. まとめと今後の展開

本論文ではテストレット群内の各項目間の背後に一定でない相関が存在するという大問形式のテストにおけるデータ構造を想定し、これに対する母数推定を適切に行うモデルを提案した。このモデルにはベイジアンアプローチを導入し、複雑な潜在構造をより柔軟に分析することを可能とした。またこのモデルは、シミュレーション研究により既存モデルと比較してより全体としてより適切な推定を行うことが検証された。

事例分析では北アイルランドにおける標準テストのデータ及び日本の国立大学法人の評価データを分析した。結果としてシミュレーション研究と同様の結果が出たことに加え、提案モデルと先行研究(Bradlow et al., 1999)におけるモデルが類似した推定を行っていることが示唆された。加えて、従来個人の学力テストをその対象としていたIRTの適用範囲を大学評価という組織に対する評価データにまで拡張することについて検討を行い、提案モデルに代表されるテストレットIRTが有効な分析方法の1つになる可能性を示した。

今後の展開としては、まず提案モデルの事例適用が挙げられる。テストレットは実際上の適用が多い一方で、本論文で検討されているような群内の相関関係まで考慮に入れられたIRTモデルはまだ考えられていない。このような状況は、事例で挙げられた学力テストや大学評価のほか、大学カリキュラムにおける単位評価、企業における人材

育成の評価測定など応用の分野は非常に多いと考える。今後、事例適用を進めていくことで、モデルの有効性の検証や応用モデルの構築を行うことが必要になってくると考えられる。

また、本研究で提案されたモデルは (2.10) で示されたように相関係数 ρ の事前分布に区間を指定した切断正規分布を用いることにより、より数値計算時の負荷を減らしている。その一方、切断正規分布の区間を指定することにより、相関係数の推定範囲が限定されてしまい、テストレット内の項目数 n_j が大きくなると柔軟性に欠ける推定を行う可能性がある。この問題への対処に関して、Lee and Chang (2000) では、成長曲線モデルを対象として (2.9) で設定されたような順序性の制約のみを用いた母数推定を行っている。今後はこの研究や関連の研究を参考にしながら、テストレット IRT においてもより制約の少ないモデルの構築法について検討を行っていくことが必要である。

補足：マルコフ連鎖モンテカルロ法による母数推定について

マルコフ連鎖モンテカルロ法では、事前分布とモデル関数によって求められた事後分布について、完全条件付事後分布、そして提案分布を定義する。完全条件付事後分布は推定対象となる母数の1つのみに注目し、それ以外の母数と観測データを所与にした場合の条件付確率である。また、提案分布は完全条件付事後分布が既知の確率分布に従わないときに用いられる代替的な役割を果たす分布である。マルコフ連鎖モンテカルロ法についてのより詳細な理論的説明については伊庭ほか (2005) の第I部及び第III部を参照のこと。本論文で提案されたモデルにおける推定手順の概略は以下の通りである。

1. 能力母数 θ , 識別力母数 α , 困難度母数 β , 相関係数 ρ の初期値を設定し、それぞれ $\theta^{(0)}$, $\alpha^{(0)}$, $\beta^{(0)}$, $\rho^{(0)}$ とする。
2. 以下の計算を適切な回数 (I 回) 繰り返す ($i = 1, \dots, I$)。
 - (a) 完全条件付事後分布 $\pi(\theta | X, \alpha^{(i-1)}, \beta^{(i-1)}, \rho^{(i-1)})$ により i 番目のサンプル $\theta^{(i)}$ を発生させる。

(b) 完全条件付事後分布 $\pi(\alpha | X, \theta^{(i-1)}, \beta^{(i-1)}, \rho^{(i-1)})$ により i 番目のサンプル $\alpha^{(i)}$ を発生させる。

(c) 完全条件付事後分布 $\pi(\beta | X, \theta^{(i-1)}, \alpha^{(i-1)}, \rho^{(i-1)})$ により i 番目のサンプル $\beta^{(i)}$ を発生させる。

(d) i 番目のサンプル $\rho^{(i)}$ を発生させる。発生は単変数 ρ_{jk} ごとに以下の手順を繰り返す。

i. i 番目のサンプル $\rho^{(i)}$ の候補 ρ' を提案分布 $q(\rho' | X, \theta^{(i-1)}, \alpha^{(i-1)}, \beta^{(i-1)})$ より発生させる。

ii. 採択率

$$a = \frac{\pi(\rho' | X, \theta^{(i-1)}, \alpha^{(i-1)}, \beta^{(i-1)}) q(\rho' | X, \theta^{(i-1)}, \alpha^{(i-1)}, \beta^{(i-1)})}{\pi(\rho^{(i-1)} | X, \theta^{(i-1)}, \alpha^{(i-1)}, \beta^{(i-1)}) q(\rho^{(i-1)} | X, \theta^{(i-1)}, \alpha^{(i-1)}, \beta^{(i-1)})}$$

を計算する。

iii. 計算された値を基に、確率 a で $\rho^{(i)} = \rho'$ とし、確率 $1 - a$ で $\rho^{(i)} = \rho^{(i-1)}$ とする。

3. 適切な回数の繰り返し後、それぞれの母数に対して得られたサンプルに対して初期値に依存する最初の部分を取り除き、残りの部分の平均をとる。これを推定値とする。

ここで手順2の (a)-(c) までは Gibbs Sampler といい、(d) を Metropolis-Hastings アルゴリズムという。特に (d) では相関係数 ρ に関する推定を行っているが、(2.10) で示されているとおり、 ρ に関する事前分布は切断正規分布であり、値域が制限されているため、Metropolis-Hastings アルゴリズムにより発生した値やこれを基に計算される推定値もその取り得る値が制限を受けることになる。

参考文献

- [1] Airasian, P. W., Madaus, G. F., and Kellaghan, T (1981), Consequences of Introducing Educational Testing in Northern Ireland 1973-1977, Inter-university Consortium for Political and Social Research.
- [2] Albert, J. H. (1992), Bayesian estimation of

- normal ogive response curves using Gibbs sampling, *Journal of Educational Statistics*, **17**, 251-269
- [3] Albert, J. H. and Chib, S. (1993), Bayesian analysis of binary and polychotomous response data, *Journal of American Statistical Association*, **88**, 669-679.
- [4] Beguin A. A. and Glas A. A. W., (2001), MCMC estimation and some model-fit analysis of multidimensional IRT models, *Psychometrika*, **66**, 541-562.
- [5] Birnbaum, A. (1968), Some latent trait models and their use in inferring an examinee's ability, In *Statistical Theory of Mental Test Scores*, eds. Lord, F. M. and Novick, M. R., Reading, MA: Addison-Wesley.
- [6] Bladlow, E., Wainer, H. and Wang. X. (1999), A Bayesian random effects model for testlets, *Psychometrika*, **64**, 153-168.
- [7] 大学評価・学位授与機構 (2008), 「国立大学法人等の中期目標期間に係る教育研究の状況 (平成16年度～平成19年度)」, http://www.niad.ac.jp/n_hyouka/kokuritsu/hyoukakekka/index.html.
- [8] Fujikoshi, Y., Kanda, T. and Tanimura, N. (1990), The growth curve model with an autoregressive covariance structure, *Annals of the Institute of Statistical Mathematics*, **42**, 533-542.
- [9] Geman, A. and Geman, D. (1984), Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, *IEEE transactions on Pattern Analysis and Machine Intelligence*, **6**, 721-741.
- [10] 伊庭幸人, 種村正美, 大森裕浩, 和合肇, 佐藤整尚, 高橋明彦, 「計算統計Ⅱ—マルコフ連鎖モンテカルロ法とその周辺—」, 統計科学のフロンティア12, 岩波書店.
- [11] Lee, J. C. (1988), Prediction and estimation of growth curves with special covariance structures, *Journal of the American Statistical Association*, **83**, 432-440.
- [12] Lee, J. C. and Chang C. H. (2000), Bayesian analysis of a growth curve model with a general autoregressive covariance structure, *Scandinavian Journal of Statistics*, **27**, 703-713.
- [13] Lord, F.M. and Novick, M.R. (1968), *Statistical theories of mental test scores*. MA: Addison-Wesley.
- [14] Rizopoulos, D. (2006), ltm: An R package for latent variable modeling and item response theory analyses, *Journal of Statistical Software*, **17**, issue 5, 1-25.
- [15] Rosenbaum, P. R. (1988), Item bundles, *Psychometrika*, **53**, 349-359.
- [16] 繁樹算男 (1985), 「ベイズ統計学入門」, 東京大学出版会.
- [17] Spiegelhalter, D., Thomas, A., Best, N. and Lunn, D. (2007), OpenBUGS software, <http://mathstat.helsinki.fi/openbugs/>.
- [18] Thissen, D., Steinberg, L. and Mooney, J. A. (1989), Trace lines for testlets -A use of multiple-categorical-response models, *Journal of educational measurement*, **26**, 247-260
- [19] 豊田秀樹 (2002a), 「項目反応理論 [入門編] —テストと測定の科学—」, 朝倉書店.
- [20] 豊田秀樹 (編著) (2002b), 「項目反応理論 [事例編] —新しい心理テストの構成法—」, 朝倉書店.
- [21] Wainer, H., Kiely, G. L. (1987), Item clusters and computerized adaptive testing: A case for testlets, *Journal of Educational Measurement*, **24**, 185-201.
- [22] Wainer, H., Bladlow, E., and Wang. X. (2007), *Testlet response theory and its applications*, Cambridge University Press.
- [23] Wang, W.C. and Wilson, M. (2005), The Rasch testlet model, *Applied Psychological Measurement*, **29**, 126-149.
- [24] Zheng, B. (2000), Bayesian estimation of multidimensional bibitem response theory modeling using Gibbs sampling, *Communications in Statistics-Theory and Methods*, **29**, 1405-1417.

(受稿日 平成22年1月4日)

(受理日 平成22年6月24日)

[ABSTRACT]

A Bayesian Inference Assuming a Correlation Structure for Testlet Models :
Application for Evaluation Data of Education and Research of Japanese National Universities

MORI Kazumasa*, OHMORI Takuya**, SHIGEMASU Kazuo***

This paper proposes a new IRT model for a testlet is proposed. In the proposed model, correlation structure is assumed to a covariance matrix of ability parameters. Some performances of the proposed model and conventional models are compared using simulation studies. It was found that proposed model estimates parameters more properly than conventional models in the case of testlet data. In case studies, two cases are applied: Irish achievement test data (achievement test data) and evaluation data of education and research of Japanese national universities (evaluation data of national universities). We found some characteristics of the proposed IRT model compared with conventional models. We also found estimation results of the evaluation data of national universities resembled that of achievement test data. We therefore concluded that the evaluation data of national universities can be regarded as a kind of testlet and can be analyzed using testlet IRT models.

* Graduate School of Arts and Sciences, The University of Tokyo

** School of management and information sciences, Tama University

*** Department of psychology, Teikyo University