論　文

パフォーマンス評価における多次元項目反応モデル

八木　崇大[†a)]　　宇都　雅輝[†b)]

Multidimensional Item Response Theory Model for Performance Assessment

Shudai YAGI[†a)] and Masaki UTO[†b)]

あらまし　近年，受験者の実践的かつ高次な能力を測定する手法の一つとしてパフォーマンス評価が注目されている．しかし，パフォーマンス評価では，得られる能力測定値が評価者の特性に依存する問題が指摘されてきた．この問題を解決する手法の一つとして，評価者特性を考慮して受験者の能力を推定できる項目反応モデルが多数提案され，その有効性が示されている．他方で，これらのモデルは測定対象の能力に一次元性を仮定する．しかし，高次な能力の測定を目指すパフォーマンス評価では，複数の能力尺度で構成されるルーブリックを用いて採点を行うことが一般的であり，この場合には能力の一次元性は必ずしも満たされない．そこで，本論文では，評価者特性を考慮した多次元性項目反応モデルを提案する．提案モデルは，データから推定した最適な次元数の能力尺度上で，評価者特性を考慮した高精度な能力測定を実現できる．本論文では，提案モデルのパラメータ推定手法としてマルコフ連鎖モンテカルロ法に基づく手法を提案し，シミュレーション実験と実データ適用を通して提案モデルの有効性を示す．

キーワード　パフォーマンス評価，多次元項目反応理論，信頼性，評価者特性，マルコフ連鎖モンテカルロ

1. まえがき

近年，大学入試や資格試験，教育評価などの様々な評価場面において，受験者の実践的かつ高次な能力の測定を目指すパフォーマンス評価が注目されている [1]～[6]．パフォーマンス評価は，現実的な課題に対する受験者のパフォーマンスを評価者が直接採点する評価法であり [7]，論述式試験や面接試験，実技試験などの様々な形式で活用されてきた．

パフォーマンス評価の問題として，受験者の能力測定精度が評価者の特性に依存する点が指摘されてきた [5], [6], [8]～[15]．この問題を解決する手法の一つとして，評価者の特性を表すパラメータを付与した項目反応モデルが近年多数提案されている [6], [8], [10], [12], [13], [16]～[19]．これらのモデルでは評価者の甘さや厳しさなどの特性を考慮して受験者の能力を推定できるため，素点平均などの単純な得点化手法と比べて，高精度な能力測定を実現できることが報告されている [6], [8], [12], [14], [20]．

これらの項目反応モデルは測定対象の能力に一次元性を仮定している．しかし，高次な能力の測定を目指すパフォーマンス評価では，複数の能力尺度で構成されるルーブリックを用いて採点を行うことが一般的であり [21], [22]，この場合には能力の一次元性は必ずしも満たされないと考えられる．一次元性が満たされない場合に一次元性を仮定したモデルを適用すると，データに対するモデル適合が低下し，能力推定値にバイアスが生じることが知られている [23], [24]．

一方，能力の多次元性を仮定した項目反応理論として，多次元項目反応モデルが知られている [25], [26]．多次元項目反応モデルでは，テスト全体が複数の能力尺度を測定すると仮定し，多次元の尺度で受験者の能力を推定できる．しかし，既存の多次元項目反応モデルでは評価者の特性を考慮できないため，パフォーマンス評価に適用した場合には能力測定精度が評価者特性に依存する問題が残る．

以上の問題を解決するために，本研究では，評価者特性パラメータを付与した多次元項目反応モデルを提案する．具体的には，補償型多次元段階反応モデル [25], [26] に評価者の特性パラメータを付与したモデルとして定式化する．また，提案モデルのパラメータ

† 電気通信大学，調布市
The University of Electro-Communications, 1–5–1 Chofugaoka, Chofu-shi, 182–8585 Japan
a) E-mail: yagi@ai.lab.uec.ac.jp
b) E-mail: uto@ai.lab.uec.ac.jp
DOI:10.14923/transinfj.2019JDP7018

推定法として，メトロポリスヘイスティングスとギブスサンプリングを用いたマルコフ連鎖モンテカルロ法を提案する．提案モデルの特徴は以下のとおりである．

（1） 情報量規準を用いたモデル選択を適用することで，能力尺度の最適な次元数をデータから推定できる．

（2） モデルパラメータを解釈することで，得られた能力尺度の意味を分析できる．

（3） 評価者特性を考慮した多次元尺度での能力測定を行うことで，従来の多次元項目反応モデルより高精度な能力推定が可能となる．

本研究では，シミュレーション実験及び実データ実験により提案モデルの有効性を示す．

2. 評点データ

本研究では，パフォーマンス評価データ $\mathbf{U}$ として，受験者のパフォーマンスを評価者がルーブリックを用いて複数の評価項目で採点した「受験者」×「評価項目」×「評価者」の 3 相データを仮定する．ここで，受験者の集合を $\mathcal{I}=\{1,\cdots,I\}$，評価者の集合を $\mathcal{R}=\{1,\cdots,R\}$，ルーブリックの評価項目の集合を $\mathcal{J}=\{1,\cdots,J\}$，評価カテゴリーの集合を $\mathcal{K}=\{0,\cdots,K-1\}$ とおき，受験者 $i\in\mathcal{I}$ のパフォーマンスに対し，評価者 $r\in\mathcal{R}$ が評価項目 $j\in\mathcal{J}$ に基づいて与える評点を x_{ijr} とする．このとき，データ $\mathbf{U}$ は次のように定義できる．

$$\mathbf{U}=\{x_{ijr}|x_{ijr}\in\{-1\}\cup\mathcal{K},i\in\mathcal{I},j\in\mathcal{J},r\in\mathcal{R}\} \tag{1}$$

ここで，$x_{ijr}=-1$ は欠測データを表す．

本研究ではこの評価データ $\mathbf{U}$ に対して項目反応理論を適用する．

3. 項目反応理論

項目反応理論 [27] は，コンピュータ・テスティングの普及とともに，近年様々な分野で実用化が進められている数理モデルを用いたテスト理論の一つである．項目反応モデルの利点として，以下のような点が挙げられる．1）推定精度の低い異質項目の影響を小さくして能力推定を行うことができる．2）異なる項目への受験者の反応を同一尺度上で評価できる．3）欠測データから容易にパラメータを推定できる．

項目反応理論はこれまで，正誤判定問題や多肢選択式問題など，正誤が一意に判定できるような客観式テストへの利用が一般的であった．一方で，近年では，論述式試験などのパフォーマンス評価に多値型項目反応モデルを適用する研究も進められている [14]．

本研究で扱うようなリッカート型データに適用できる多値型項目反応モデルとして，段階反応モデル（GRM: Graded Response Model）[28] や一般化部分採点モデル（GPCM: Generalized Partial Credit Model）[29] が広く利用されてきた．次節では，本研究で基礎モデルとして利用する GRM について述べる．

3.1 段階反応モデル（GRM）

GRM では，受験者 i が項目 j にカテゴリー $k\in\mathcal{K}$ と反応する確率 P_{ijk} を次式で与える．

$$P_{ijk}=P^*_{ijk}-P^*_{ijk+1} \tag{2}$$
$$\begin{cases} P^*_{ijk}=\frac{1}{1+\exp[-\alpha_j(\theta_i-\beta_{jk})]} & k=1,\cdots,K-1\\ P^*_{ij0}=1\\ P^*_{ijK}=0 \end{cases}$$

ここで，θ_i は受験者 i の能力，α_j は項目 j の識別力，β_{jk} は項目 j において評価カテゴリー k と反応する困難度を表す．ただし，$\beta_{j1}<\beta_{j2}<\cdots<\beta_{jK-1}$ とする．このモデルでは，能力が低いほど低いカテゴリーへの反応確率が高くなり，能力が高いほど高いカテゴリーへの反応確率が高くなる．

一般にこのような既存の項目反応モデルで扱うデータは受験者のテスト項目への回答であり，「受験者」×「テスト項目」の 2 相データとなる．しかし，パフォーマンス評価で扱うデータは「受験者」×「評価項目」×「評価者」の 3 相データであり，通常の項目反応モデルを直接には適用できない．この問題を解決するために，評価者特性パラメータを加えた項目反応モデルが近年多数提案されている [6], [8], [10], [12], [13], [16]～[19]．次節では，Uto and Ueno [8] のモデルを紹介する．

3.2 評価者特性パラメータを付与した項目反応モデル

Uto and Ueno [8] は，評価者の厳しさと一貫性の特性を表すパラメータを付与した GRM を提案している．このモデルでは，受験者 i のパフォーマンスに対し，評価者 r が評価項目 j に基づいて評点 $k\in\mathcal{K}$ を与える確率 P_{ijrk} を次式で定義する．

$$P_{ijrk}=P^*_{ijrk}-P^*_{ijrk+1} \tag{3}$$
$$\begin{cases} P^*_{ijrk}=\frac{1}{1+\exp[-\alpha_j\alpha_r(\theta_i-\beta_{jk}-\epsilon_r)]} & k=1,\cdots,K-1\\ P^*_{ijr0}=1\\ P^*_{ijrK}=0 \end{cases}$$

ここで，α_j は評価項目 j の識別力，β_{jk} は評価項目 j において評点 k を得るための困難度を表す．ただし，$\beta_{j1} < \beta_{j2} < \cdots < \beta_{jK-1}$ とする．また，α_r は評価者 r の評価の一貫性，ϵ_r は評価者 r の評価の厳しさを表す．また，パラメータの識別性のために $\alpha_{r=1} = 1$，$\epsilon_1 = 0$ を仮定している．

このような評価者特性を考慮した項目反応モデルは，素点平均などの単純な得点化手法と比べて，高精度な能力測定を実現できることが報告されている [8], [20]．特に Uto and Ueno [8] のモデルでは，評価者数が多い場合にも高精度な能力推定が可能であり，大規模テストのように評価者数が多くなる場合に高い有効性が期待できる [6], [8], [16]．しかし，**1.** で述べたように，評価者特性を考慮した既存のモデルは測定対象の能力に一次元性を仮定しており，多次元尺度を想定した能力測定を行うことはできない．

3.3 多次元項目反応モデル

能力の多次元性を仮定した項目反応理論として多次元項目反応モデルが知られている [25], [26]．多次元項目反応理論は，補償型と非補償型のモデルに大別することができる [26]．補償型の多次元項目反応モデルは，いずれかの次元の能力が高ければ高い評点が得られると仮定したモデルであり，非補償型多次元項目反応モデルは，全ての次元の能力が高くなければ高い評点を得ることが難しいと仮定したモデルである．非補償型モデルは補償型モデルに比べてモデルパラメータ数が多いため，高精度なパラメータ推定に必要なデータが増加し，パラメータの解釈も困難になる [26]．そこで，本研究では，補償型の多次元項目反応モデルに着目する．

多値データに対する補償型多次元項目反応モデルとしては，補償型多次元段階反応モデル [26] が知られている．このモデルでは，受験者 i が項目 j において評点 k を得る確率 P_{ijk} を次式で定義する．

$$P_{ijk} = P^*_{ijk} - P^*_{ijk+1} \tag{4}$$

$$\begin{cases} P^*_{ijk} = \dfrac{1}{1+\exp\left[-(\sum_{l=1}^{L} \alpha_{jl}\theta_{il} - \beta_{jk})\right]} & k = 1, \cdots, K-1 \\ P^*_{ij0} = 1 \\ P^*_{ijK} = 0 \end{cases}$$

ここで，L は能力の次元数，θ_{il} は受験者 i の $l \in \{1, \cdots, L\}$ 次元目の能力，α_{jl} は項目 j の l 次元目の能力に対する識別力を表す．また，β_{jk} は項目 j において評点 k を得るための困難度を表す．ただし，$\beta_{j1} < \beta_{j2} < \cdots < \beta_{jK-1}$ とする．

このようなモデルを用いることで，多次元の能力尺度を仮定した能力測定が可能となる．更に，テスト項目の識別力を項目の内容と合わせて分析することで，個々の次元がどのような能力を測定しているかを解釈することができる．例えば，次元 l の識別力を項目間で比較したとき項目 j と項目 j' の値が突出して高かった場合，次元 l はテスト項目 j と j' に共通する尺度を測定していると解釈できる．したがって，これらの項目の内容的な共通性を分析することで，次元 l の意味を解釈できる．反対に，項目ごとにどの次元の識別力が高いかを分析することで，各項目がどのような尺度を測定しているかを分析することも可能である．

多次元段階反応モデルではこのような多次元尺度での能力測定が可能であるが，**3.1** で導入した段階反応モデルと同様に，「受験者」×「テスト項目」の 2 相データへの適用を仮定しており，本研究で扱う 3 相データに直接には適用できない．そこで本研究では，多次元項目反応モデルに評価者特性パラメータを付与した新たなモデルを提案する．

4. 提案モデル

提案モデルでは，受験者 i のパフォーマンスに対し，評価者 r が評価項目 j に基づいて評点 k を与える確率 P_{ijrk} を次式で定義する．

$$P_{ijrk} = P^*_{ijrk} - P^*_{ijrk+1} \tag{5}$$

$$\begin{cases} P^*_{ijrk} = \dfrac{1}{1+\exp\left[-\alpha_r(\sum_{l=1}^{L} \alpha_{jl}\theta_{il} - \beta_{jk} - \epsilon_r)\right]} & k = 1, \cdots, K-1 \\ P^*_{ijr0} = 1 \\ P^*_{ijrK} = 0 \end{cases}$$

ここで，α_{jl} は評価項目 j の l 次元目の能力に対する識別力を表し，β_{jk} は評価項目 j において評点 k を得るための困難度を表す．ただし，$\beta_{j1} < \beta_{j2} < \cdots < \beta_{jK-1}$ とする．また，モデルの識別性のために，$\alpha_{r=1} = 1$，$\epsilon_1 = 0$ を仮定する．

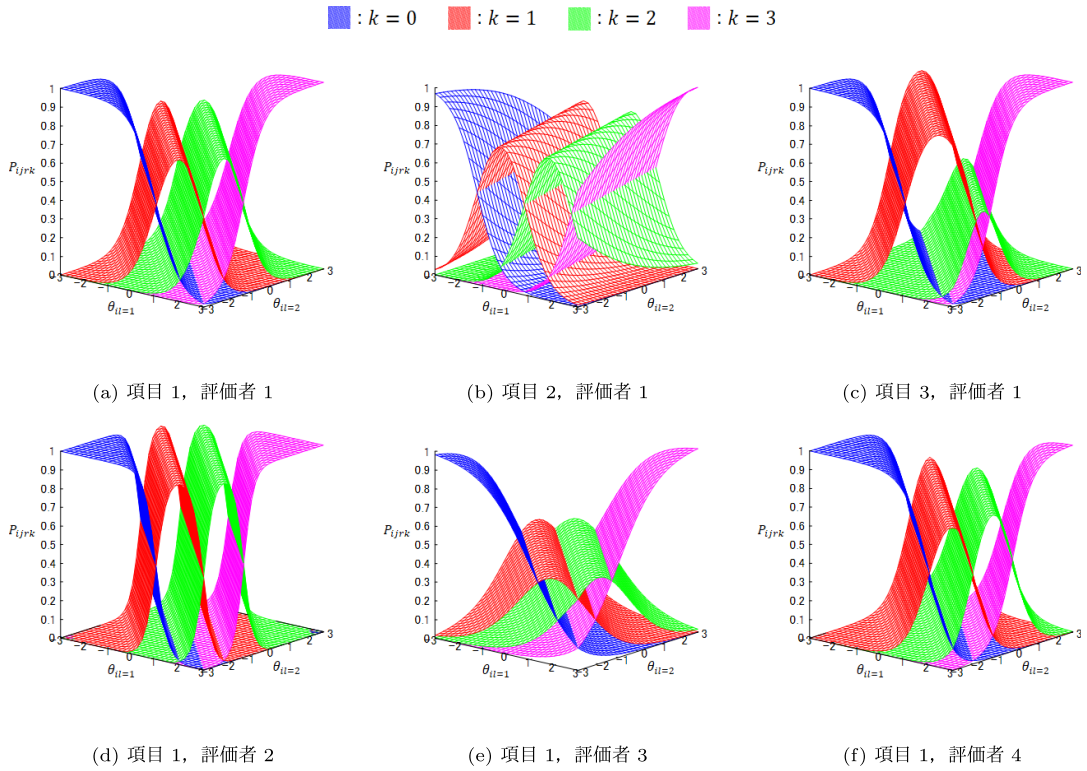
4.1 パラメータの解釈

ここで，提案モデルのパラメータの解釈を説明するために，次元数 $L = 2$，評価カテゴリー数 $K = 4$ において，表 1 のパラメータを所与としたときの，式

表 1　図 1 で使用するパラメータ
Table 1　Parameters used in Fig. 1.

	α_{j1}	α_{j2}	β_{j1}	β_{j2}	β_{j3}
項目 1	2.0	2.0	−4.0	0.0	4.0
項目 2	2.0	0.5	−4.0	0.0	4.0
項目 3	2.0	2.0	−4.0	2.0	4.0

	α_r	ϵ_r
評価者 1	1.0	0.0
評価者 2	2.0	0.0
評価者 3	0.5	0.0
評価者 4	1.0	2.0



(a) 項目 1，評価者 1　(b) 項目 2，評価者 1　(c) 項目 3，評価者 1

(d) 項目 1，評価者 2　(e) 項目 1，評価者 3　(f) 項目 1，評価者 4

図 1　表 1 のパラメータを適用した場合の項目反応曲面
Fig. 1　Item response surfaces given parameters in Table 1.

(5) で表される項目反応曲面（IRS: Item Response Surface）を図 1 に示す．例えば，図 1 (a) は，表 1 における項目 1 と評価者 1 のパラメータを適用したときの IRS を表す．各図は，横軸に受験者の能力 θ_{il} を次元ごとに示し，縦軸が各評点への反応確率 P_{ijrk} を表す．図 1 から，どの項目においても，各次元の能力が低いほど低い評点を得る確率が高くなり，各次元の能力が高いほど高い評点を得る確率が高くなっていることがわかる．以降ではパラメータごとの解釈を示すために，図 1 (a) を基準に各パラメータを個別に変化させた図 1 (b)～(f) について説明する．

図 1 (b) は，図 1 (a) から識別力 α_{j2} を小さくした場合の IRS である．図 1 (a) と比較すると，曲面の勾配の向きが変化しており，2 次元目の能力の変動に対する各評点への反応確率の変化が緩やかになっていることがわかる．これは，識別力 α_{jl} が小さい項目は，l 次元目の能力を精度よく測定できないことを表現している．

図 1 (c) は，困難度 β_{j2} が高い場合の IRS である．困難度パラメータ β_{jk} は，値が大きくなるほど評点 k 以上を取ることが難しくなる．項目 3 では，$k=1$ と $k=2$ の IRS の境界位置における能力値が項目 1 と比べて高くなっていることがわかる．これは，項目 3 で評点 $k=2$ を得るには，項目 1 で同じ評点を得るより高い能力が必要であることを意味する．また，困難度パラメータは，隣接する値 $\beta_{jk+1}-\beta_{jk}$ の差が大きくなるほど，評点 k への反応確率が高くなる．項目 3 は $\beta_{j2}-\beta_{j1}$ が大きいため，図 1 (a) と比べて評点 $k=1$ を得る確率が全体的に高くなっている．反対に，$\beta_{j3}-\beta_{j2}$ については相対的に差異が小さくなっているため，評点 $k=2$ を得る確率は図 1 (a) と比べて全体的に低く表現されている．

図 1 (d) は，評価者の一貫性 α_r が高い場合の IRS である．図 1 (a) と比べると，全てのカテゴリーに対して IRS の勾配が大きくなっており，θ_{il} の変動に対して反応確率が敏感に変動するようになっていることがわかる．これは，一貫性の高い評価者は，受験者の能力が高いほど高い得点を，能力が低いほど低い得点

を一貫して与えるとともに，同等の能力の受験者に対しては安定して同一の評点を与える傾向が強いことを表現している．逆に，図 1 (e) のように，α_r が低い場合には，能力の変化に伴う反応確率の変動が小さく，カテゴリー間での反応確率の差異が全体として小さくなっている．これは，一貫性の低い評価者は，評価のランダムネスが大きく，受験者の能力と必ずしも相関した評価を行わないことを表現している．したがって，一貫性が高い評価者ほど，受験者の能力を同一の基準のもとで安定して評価できる望ましい評価者と一般に判断できる．

図 1 (f) は，評価者の厳しさパラメータ ϵ_r が大きい場合の IRS である．図 1 (a) と比べて，全体として IRS のピークが θ_{il} の値が大きくなる方に移動していることがわかる．これは，厳しい評価者から高い評点を得るにはより高い能力が必要であることを表現している．

提案モデルでは，これらの評価項目と評価者の特性を考慮して多次元尺度での能力測定を行うことができるため，従来の多次元項目反応モデルと比べて高精度な能力測定が期待できる．

4.2 次元数の推定と次元の解釈

提案モデルを利用するためには，最適な次元数 L を決定する必要がある．一般に項目反応理論における次元数の分析は，因子分析に基づくスクリープロットを用いて行うことが多い [30]．しかし，因子分析は本研究で扱うような 3 相データには適用できない．他方で，次元数の選択はモデル選択として解釈することができる．一般に，モデル選択は BIC（Bayesian Information Criterion）[31] や AIC（Akaike Information Criterion）[32] などの情報量規準に基づいて行うことが多い．提案モデルでもこれらの情報量規準を用いることで，最適な次元数 L をデータから決定できる．

また，得られた多次元尺度において，個々の次元がどのような能力を測定しているかは，通常の多次元項目反応モデルと同様に，各評価項目の識別力と内容を分析することで解釈できる．具体的には，次元 l の識別力の値を項目間で比較したとき，評価項目 j と評価項目 j' の値が突出して高ければ，次元 l は評価項目 j と j' に共通する能力を測定していると解釈できる．

4.3 MCMC によるパラメータ推定手法

項目反応理論におけるパラメータ推定手法としては，EM アルゴリズムを用いた周辺最ゆう推定法やニュートンラフソン法による事後確率最大化推定法が広く用いられてきた [33]．一方で，本研究で扱うような複雑な項目反応モデルの場合には，マルコフ連鎖モンテカルロ（MCMC: Markov Chain Monte-Carlo）アルゴリズムを用いた期待事後確率 (EAP: Expected A Posteriori) 推定法が高精度であることが示されている [8], [34]．項目反応理論における MCMC アルゴリズムとしては，メトロポリスヘイスティングスとギブスサンプリングを組み合わせたアルゴリズム [8], [14], [35] が一般的である．そこで，本研究でも，提案モデルのパラメータ推定手法として，メトロポリスヘイスティングスとギブスサンプリングを組み合わせた MCMC 法を提案する．

ここで，各パラメータの集合を $\boldsymbol{\theta} = \{\theta_{11}, \cdots, \theta_{IL}\}$，$\boldsymbol{\alpha_j} = \{\alpha_{11}, \cdots, \alpha_{JL}\}$，$\boldsymbol{\beta} = \{\beta_{11}, \cdots, \beta_{JK-1}\}$，$\boldsymbol{\alpha_r} = \{\alpha_1, \cdots, \alpha_R\}$，$\boldsymbol{\epsilon} = \{\epsilon_1, \cdots, \epsilon_R\}$ と表す．また，各パラメータの事前分布を $g(\boldsymbol{\theta}|\tau_\theta)$，$g(\boldsymbol{\alpha_j}|\tau_{\alpha_j})$，$g(\boldsymbol{\beta}|\tau_\beta)$，$g(\boldsymbol{\alpha_r}|\tau_{\alpha_r})$，$g(\boldsymbol{\epsilon}|\tau_\epsilon)$ とする．ただし，τ_θ，τ_{α_j}，τ_β，τ_{α_r}，τ_ϵ は各事前分布のパラメータ（ハイパーパラメータ）を表す．このとき，反応データ $\mathbf{U}$ を所与として，パラメータの事後分布は以下のように導かれる．

$$\begin{aligned} &g(\boldsymbol{\theta}, \boldsymbol{\alpha_j}, \boldsymbol{\beta}, \boldsymbol{\alpha_r}, \boldsymbol{\epsilon}, |\mathbf{U}) \propto L(\mathbf{U}|\boldsymbol{\theta}, \boldsymbol{\alpha_j}, \boldsymbol{\beta}, \boldsymbol{\alpha_r}, \boldsymbol{\epsilon}) \\ &g(\boldsymbol{\theta}|\tau_\theta) g(\boldsymbol{\alpha_j}|\tau_{\alpha_j}) g(\boldsymbol{\beta}|\tau_\beta) g(\boldsymbol{\alpha_r}|\tau_{\alpha_r}) g(\boldsymbol{\epsilon}|\tau_\epsilon) \end{aligned} \quad (6)$$

ここで，

$$L(\mathbf{U}|\boldsymbol{\theta}, \boldsymbol{\alpha_j}, \boldsymbol{\beta}, \boldsymbol{\alpha_r}, \boldsymbol{\epsilon}) = \prod_{i=1}^{I} \prod_{j=1}^{J} \prod_{r=1}^{R} \prod_{k=0}^{K-1} (P_{ijrk})^{z_{ijrk}} \quad (7)$$

$$z_{ijrk} = \begin{cases} 1 : x_{ijr} = k \text{ のとき} \\ 0 : \text{上記以外} \end{cases} \quad (8)$$

提案アルゴリズムでは，式 (6) の事後分布を MCMC により求める．ここで，$\boldsymbol{\lambda} = (\boldsymbol{\theta}, \boldsymbol{\alpha_j}, \boldsymbol{\beta}, \boldsymbol{\alpha_r}, \boldsymbol{\epsilon})$ とすると，アルゴリズムの大枠は，$\tau \in \boldsymbol{\lambda}$ を $\boldsymbol{\lambda}^{-\tau} = \boldsymbol{\lambda} \backslash \{\tau\}$ を所与とした完全条件付き事後分布からサンプリングすることを繰り返すというものである．ただし，項目反応理論においてはこれらの分布が解析的に求まらないため [35]，このサンプリングはメトロポリスヘイスティングを用いて行う．具体的には，$\tau \in \boldsymbol{\lambda}$ について，候補値 τ^* を提案分布 $N(\tau, \sigma_0^2)$ からサンプリングし，候補値 τ^* を次の採択確率で採択/棄却するという手順を，全てのパラメータについて行う．

$$a(\tau^*|\tau) = min\left(\prod_{l=1}^{L}\frac{L(\boldsymbol{U}|\tau^*,\boldsymbol{\lambda}^{-\tau})g(\tau^*)}{L(\boldsymbol{U}|\tau,\boldsymbol{\lambda}^{-\tau})g(\tau)},1\right)$$

このサンプリングを十分に繰り返し，得られたパラメータ・サンプルの平均値を EAP 推定値とする．ただし，分布が収束したと推測されるまでのバーンイン期間は，パラメータの初期値の影響が残るため推定に利用しない．本研究では，バーンイン期間は 30,000 とし，自己相関を考慮して 30,000 時点から 50,000 時点までのサンプルを 100 間隔で抽出して EAP 推定値を求める．また，各パラメータの事前分布については，先行研究 [8] と同様に，$\theta_{il} \sim N(0.0, 1.0^2)$，$\alpha_{jl} \sim LN(1.0, 0.5^2)$，$\beta_{j1} \sim N(-1.5, 1.0^2)$，$\beta_{j2} \sim N(0.0, 1.0^2)$，$\beta_{j3} \sim N(1.5, 1.0^2)$，$\alpha_r \sim LN(1.0, 0.5^2)$，$\epsilon_r \sim N(0.0, 1.0^2)$ とする．ここで，$N(\mu, \sigma^2)$ は平均 μ，標準偏差 σ の正規分布を，$LN(\mu', \sigma'^2)$ は平均 μ'，標準偏差 σ' の対数正規分布を表す．

5. シミュレーション実験

5.1 パラメータ推定精度

本節では，MCMC アルゴリズムによる提案モデルのパラメータ推定精度をシミュレーション実験により評価する．

ここで，l 次元目の識別力パラメータのベクトルを $\boldsymbol{\alpha}_l = \{\alpha_{jl}|j \in \mathcal{J}\}$，$l$ 次元目の能力ベクトルを $\boldsymbol{\theta}_l = \{\theta_{il}|i \in \mathcal{I}\}$ とするとき，提案モデルでは l 次元目のパラメータ（$\boldsymbol{\alpha}_l$，$\boldsymbol{\theta}_l$）と l' 次元目のパラメータ（$\boldsymbol{\alpha}_{l'}$，$\boldsymbol{\theta}_{l'}$）を入れ替えても式 (5) の反応確率は変化しないため，これらのパラメータ推定値は一意に定まらない．実データの分析においてはパラメータ推定後に各次元の解釈を行うためこの不定性は問題とならないが，本節で行うようなパラメータ・リカバリの精度評価ではこの不定性を解消しなければ適切に評価できない．そこで，ここでは，先行研究 [36] に基づき，識別力が極端な値となるダミー項目を用いて次元の識別性の問題を解消する．具体的には，ダミー項目 $\mathcal{J}' \in \{J+1, \cdots, J+L\}$ を用いて，以下の手順でパラメータ推定精度の評価を行った．

（1） ダミー項目 $j \in \mathcal{J}'$ はカテゴリー数 $K = 2$ とし，パラメータを以下の値に設定した．

$$\begin{cases}\alpha_{jl} = 1.65 & j = J+l \\ \alpha_{jl} = 0.22 & j \neq J+l\end{cases} \tag{9}$$

$$\beta_{jk} = 0 \tag{10}$$

（2） ダミー項目以外の項目 $j \in \mathcal{J}$ のパラメータと評価者パラメータ，受験者の能力値を **4.3** に示した分布に従ってランダムに生成した．

（3） 手順 (1) と手順 (2) で生成したパラメータを所与として，データ $\mathbf{U}$ を式 (5) に基づいて生成した．

（4） 生成したデータから MCMC を用いてパラメータ推定を行った．このとき，ダミー項目のパラメータは手順 (1) で生成した値を所与とした．また，ダミー項目のパラメータを所与とすることでモデルの識別性が保たれるため，本推定では式 (5) における $\alpha_{r=1} = 1$，$\epsilon_1 = 0$ の制約は適用しなかった．

（5） 得られたパラメータ推定値と手順 (1) で生成したパラメータ真値との平均平方 2 乗誤差（RMSE: Root Mean Square Error）とバイアスを算出した．

（6） 手順 (2)～(5) を 50 回行い，RMSE とバイアスの平均と標準偏差を算出した．

上記の実験を，評価項目数 $J = 5, 10, 15$，評価者数 $R = 5, 10, 15$，次元数 $L = 1, 2, 3$ のそれぞれの場合において行った．受験者数と評価カテゴリー数は，次章で行う実データ実験の設定に合わせて $I = 30$，$K = 4$ とした．

本実験で得られた RMSE の平均と標準偏差を表 2 に示す．表 2 から，項目数や評価者数の増加に伴い，能力値の RMSE が減少する傾向が読み取れる．これは，項目や評価者の増加により能力パラメータに対するデータ数が増加するためであり，先行研究 (e.g., [8], [16], [37]) と一致した傾向を示している．また，項目パラメータの RMSE は評価者の増加に伴って減少し，評価者パラメータの RMSE は項目の増加に伴って減少する傾向も確認できる．これは，評価者の増加に伴って項目パラメータに対するデータ数が増加し，項目の増加に伴って評価者パラメータに対するデータ数が増加するためであり，先行研究 (e.g., [8], [16], [37]) と一致した傾向となっている．なお，項目数が増加しても項目パラメータに対するデータ数は増加しないため，項目パラメータの RMSE は項目数を増やしても必ずしも減少しない点に注意されたい．評価者数と評価者パラメータの関係もこれと同様である．また，次元数の増加により能力値と項目識別力の推定精度が悪くなる傾向も読み取れる．これは，次元数が増加すると，データ数一定のまま能力値と項目識別力パラメータの数が増加するためであり，多次元項目反応モデルの先行研究 [36] と一致した傾向となっている．なお，表 2 では，少数ではあるが，次元数 $L = 3$ のときに評

表 2　パラメータ・リカバリ実験における RMSE の平均値と標準偏差（カッコ内）
Table 2　Average and standard deviation of RMSE values for parameter recovery experiment.

		$J=5$			$J=10$			$J=15$		
	L	$R=5$	$R=10$	$R=15$	$R=5$	$R=10$	$R=15$	$R=5$	$R=10$	$R=15$
α_{jl}	1	0.229	0.146	0.140	0.231	0.161	0.142	0.231	0.158	0.138
		(0.084)	(0.057)	(0.044)	(0.072)	(0.052)	(0.052)	(0.065)	(0.051)	(0.050)
	2	0.243	0.194	0.165	0.237	0.189	0.174	0.245	0.173	0.154
		(0.064)	(0.067)	(0.049)	(0.068)	(0.044)	(0.044)	(0.043)	(0.032)	(0.032)
	3	0.281	0.203	0.174	0.279	0.208	0.174	0.258	0.201	0.170
		(0.071)	(0.052)	(0.047)	(0.061)	(0.038)	(0.047)	(0.049)	(0.038)	(0.027)
β_{jk}	1	0.163	0.141	0.128	0.166	0.145	0.121	0.165	0.146	0.106
		(0.056)	(0.042)	(0.076)	(0.045)	(0.046)	(0.051)	(0.030)	(0.052)	(0.030)
	2	0.171	0.151	0.116	0.169	0.141	0.137	0.166	0.136	0.133
		(0.053)	(0.068)	(0.037)	(0.032)	(0.066)	(0.032)	(0.030)	(0.032)	(0.056)
	3	0.183	0.148	0.151	0.180	0.148	0.125	0.174	0.148	0.135
		(0.061)	(0.050)	(0.077)	(0.045)	(0.049)	(0.044)	(0.035)	(0.032)	(0.039)
α_r	1	0.140	0.128	0.130	0.107	0.103	0.102	0.089	0.093	0.087
		(0.052)	(0.045)	(0.054)	(0.052)	(0.031)	(0.033)	(0.037)	(0.039)	(0.032)
	2	0.124	0.124	0.130	0.118	0.105	0.099	0.093	0.083	0.095
		(0.045)	(0.058)	(0.033)	(0.055)	(0.033)	(0.037)	(0.049)	(0.036)	(0.047)
	3	0.136	0.131	0.116	0.094	0.100	0.097	0.086	0.084	0.087
		(0.041)	(0.050)	(0.035)	(0.042)	(0.040)	(0.027)	(0.032)	(0.027)	(0.030)
ϵ_r	1	0.204	0.211	0.195	0.184	0.184	0.171	0.170	0.177	0.152
		(0.070)	(0.078)	(0.053)	(0.098)	(0.078)	(0.070)	(0.097)	(0.088)	(0.060)
	2	0.215	0.200	0.195	0.177	0.185	0.182	0.165	0.156	0.141
		(0.097)	(0.067)	(0.058)	(0.082)	(0.078)	(0.063)	(0.078)	(0.059)	(0.050)
	3	0.192	0.196	0.188	0.163	0.162	0.164	0.167	0.165	0.143
		(0.089)	(0.075)	(0.051)	(0.067)	(0.051)	(0.047)	(0.075)	(0.065)	(0.054)
θ_{il}	1	0.329	0.240	0.222	0.253	0.198	0.175	0.235	0.179	0.145
		(0.075)	(0.057)	(0.074)	(0.065)	(0.057)	(0.046)	(0.058)	(0.053)	(0.040)
	2	0.439	0.333	0.275	0.371	0.284	0.261	0.337	0.258	0.214
		(0.064)	(0.055)	(0.042)	(0.076)	(0.056)	(0.059)	(0.057)	(0.046)	(0.050)
	3	0.469	0.371	0.314	0.444	0.311	0.274	0.385	0.287	0.243
		(0.062)	(0.046)	(0.037)	(0.087)	(0.044)	(0.039)	(0.049)	(0.038)	(0.032)

価者数（または項目数）の増加に伴って項目（または評価者）パラメータの RMSE が増加してしまうケースが見受けられる．次元数が 3 の場合には，能力値や項目識別力の推定誤差が比較的大きく，この誤差は他のパラメータの推定値にも反映される．この影響が，評価者数や項目数の増加によるパラメータ推定精度の改善を打ち消してしまったため，このような結果が得られたと考えらえる．

他方で，本実験で得られたバイアスの平均と標準偏差を表 3 に示す．表 3 から，いずれの条件においてもバイアスの平均は 0 に近い値を示しており，系統的な過大（または過少）推定の傾向も認められないことがわかる．

以上の結果から，MCMC アルゴリズムにより提案モデルのパラメータを適切に推定できることが確認できた．

5.2　情報量規準に基づく次元数推定の妥当性評価

ここでは，情報量規準を用いた次元数推定の妥当性を評価する．具体的には，BIC と AIC を情報量規準として用い，以下の実験を行った．

（1）　真の次元数を L_t とし，モデルパラメータを **4.3** で示した分布に従って生成した．

（2）　生成したパラメータを所与として，式 (5) に基づいてデータ $\mathbf{U}$ を生成した．

（3）　データ $\mathbf{U}$ を用いて次元数 $L_e=1,2,3$ を仮定して MCMC によるパラメータ推定を行い，情報量が高い次元数順に順位づけを行った．

（4）　上記の実験を 50 回繰り返し，順位の平均と標準偏差を算出した．

以上の実験は，項目数 $J=5,10,15$，評価者数 $R=5,10,15$，真の次元数 $L_t=1,2,3$ のそれぞれの場合において同様に行った．また，項目数と評価者数が多い場合を想定して，項目数 $J=20,25,30$，評価者数 $R=20,25,30$ の場合でも同様の実験を行った．受験者数とカテゴリー数は，前節の実験同様，$I=30$，$K=4$ に設定した．

表 3　パラメータ・リカバリ実験におけるバイアスの平均値と標準偏差（カッコ内）
Table 3　Average and standard deviation of bias values for parameter recovery experiment.

		$J=5$			$J=10$			$J=15$		
	L	$R=5$	$R=10$	$R=15$	$R=5$	$R=10$	$R=15$	$R=5$	$R=10$	$R=15$
α_{jl}	1	0.039	−0.001	−0.006	−0.005	−0.017	0.007	−0.026	−0.011	−0.015
		(0.127)	(0.091)	(0.090)	(0.136)	(0.091)	(0.101)	(0.122)	(0.087)	(0.088)
	2	−0.014	−0.036	0.003	−0.033	−0.015	0.003	0.002	−0.006	−0.002
		(0.096)	(0.102)	(0.092)	(0.088)	(0.082)	(0.084)	(0.071)	(0.065)	(0.074)
	3	−0.011	0.012	−0.017	0.002	−0.015	−0.000	−0.016	0.008	−0.007
		(0.104)	(0.087)	(0.072)	(0.073)	(0.073)	(0.067)	(0.079)	(0.076)	(0.057)
β_{jk}	1	0.007	0.007	0.004	0.006	0.022	0.013	0.009	0.021	0.012
		(0.049)	(0.045)	(0.041)	(0.045)	(0.046)	(0.042)	(0.039)	(0.048)	(0.025)
	2	0.004	0.026	0.006	−0.003	0.004	0.022	0.010	0.015	0.018
		(0.055)	(0.046)	(0.031)	(0.039)	(0.038)	(0.049)	(0.034)	(0.036)	(0.043)
	3	0.007	0.007	0.012	0.003	0.013	−0.001	0.008	0.019	0.019
		(0.055)	(0.046)	(0.057)	(0.044)	(0.050)	(0.040)	(0.042)	(0.044)	(0.040)
α_r	1	0.006	0.015	0.005	0.051	0.016	0.016	0.020	0.033	0.009
		(0.104)	(0.091)	(0.101)	(0.085)	(0.077)	(0.077)	(0.074)	(0.076)	(0.071)
	2	0.004	0.028	−0.010	0.034	0.021	0.013	0.026	0.017	0.026
		(0.090)	(0.088)	(0.080)	(0.104)	(0.078)	(0.072)	(0.084)	(0.064)	(0.080)
	3	0.034	0.014	0.020	−0.008	0.034	0.015	0.013	0.017	0.017
		(0.079)	(0.089)	(0.071)	(0.079)	(0.070)	(0.068)	(0.070)	(0.061)	(0.066)
ϵ_r	1	0.015	−0.029	0.037	−0.023	−0.034	−0.017	−0.014	−0.038	−0.031
		(0.147)	(0.162)	(0.112)	(0.174)	(0.153)	(0.134)	(0.169)	(0.170)	(0.120)
	2	0.009	0.016	−0.017	0.014	0.004	−0.007	0.001	−0.022	−0.008
		(0.176)	(0.139)	(0.119)	(0.156)	(0.156)	(0.135)	(0.152)	(0.124)	(0.097)
	3	−0.020	0.013	0.029	−0.017	0.010	0.006	−0.009	0.013	0.010
		(0.165)	(0.137)	(0.118)	(0.115)	(0.115)	(0.117)	(0.154)	(0.148)	(0.110)
θ_{il}	1	0.032	−0.007	0.035	−0.006	0.014	0.016	0.006	0.016	−0.004
		(0.134)	(0.107)	(0.116)	(0.104)	(0.111)	(0.093)	(0.128)	(0.106)	(0.074)
	2	0.016	0.035	−0.013	0.006	−0.002	0.024	0.009	0.002	0.015
		(0.097)	(0.090)	(0.071)	(0.084)	(0.076)	(0.097)	(0.079)	(0.071)	(0.073)
	3	−0.002	0.012	0.019	−0.003	0.011	−0.001	0.005	0.022	0.017
		(0.073)	(0.062)	(0.063)	(0.054)	(0.060)	(0.052)	(0.064)	(0.064)	(0.056)

表 4　次元数選択精度
Table 4　Accuracy of dimensionality determination.

		$R=5$						$R=10$						$R=15$					
		$J=5$		$J=10$		$J=15$		$J=5$		$J=10$		$J=15$		$J=5$		$J=10$		$J=15$	
L_t	L_e	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC
1	1	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.02**	**1.00**	**1.00**
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.14)	(0.00)	(0.00)
	2	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	1.98	2.00	2.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.14)	(0.00)	(0.00)
	3	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
2	1	**1.12**	**1.28**	**1.36**	1.90	1.80	2.50	**1.44**	1.84	2.22	2.82	2.42	2.88	**1.46**	1.98	2.24	2.76	2.80	2.98
		(0.32)	(0.53)	(0.56)	(0.88)	(0.72)	(0.67)	(0.61)	(0.70)	(0.70)	(0.43)	(0.67)	(0.38)	(0.67)	(0.73)	(0.74)	(0.43)	(0.40)	(0.14)
	2	1.88	1.76	1.68	**1.44**	**1.38**	**1.10**	1.62	**1.34**	**1.16**	**1.04**	**1.10**	**1.02**	1.64	**1.28**	**1.18**	**1.00**	**1.00**	**1.00**
		(0.32)	(0.43)	(0.47)	(0.50)	(0.49)	(0.30)	(0.49)	(0.47)	(0.37)	(0.20)	(0.30)	(0.14)	(0.48)	(0.45)	(0.38)	(0.00)	(0.00)	(0.00)
	3	3.00	2.96	2.96	2.66	2.82	2.40	2.94	2.82	2.62	2.14	2.48	2.10	2.90	2.74	2.58	2.24	2.20	2.02
		(0.00)	(0.20)	(0.20)	(0.47)	(0.38)	(0.49)	(0.24)	(0.38)	(0.49)	(0.40)	(0.50)	(0.30)	(0.30)	(0.44)	(0.49)	(0.43)	(0.40)	(0.14)
3	1	**1.12**	1.58	1.68	2.56	2.38	2.98	1.90	2.48	2.74	3.00	2.94	3.00	2.34	2.74	2.96	3.00	3.00	3.00
		(0.38)	(0.67)	(0.73)	(0.67)	(0.75)	(0.14)	(0.81)	(0.73)	(0.48)	(0.00)	(0.31)	(0.00)	(0.74)	(0.59)	(0.20)	(0.00)	(0.00)	(0.00)
	2	1.90	**1.52**	**1.56**	**1.36**	**1.42**	1.60	**1.40**	**1.36**	**1.34**	1.66	1.62	1.82	**1.24**	**1.28**	1.60	1.82	1.86	1.98
		(0.30)	(0.50)	(0.50)	(0.48)	(0.49)	(0.53)	(0.49)	(0.48)	(0.47)	(0.47)	(0.49)	(0.38)	(0.43)	(0.45)	(0.49)	(0.38)	(0.35)	(0.14)
	3	2.98	2.90	2.76	2.08	2.20	**1.42**	2.70	2.16	1.92	**1.34**	**1.44**	**1.18**	2.42	1.98	**1.44**	**1.18**	**1.14**	**1.02**
		(0.14)	(0.30)	(0.59)	(0.77)	(0.82)	(0.49)	(0.50)	(0.76)	(0.74)	(0.47)	(0.57)	(0.38)	(0.64)	(0.62)	(0.57)	(0.38)	(0.35)	(0.14)

表 5　データ数が多い場合の次元数選択精度
Table 5　Accuracy of dimensionality determination for larger data.

		$R=20$						$R=25$						$R=30$					
		$J=20$		$J=25$		$J=30$		$J=20$		$J=25$		$J=30$		$J=20$		$J=25$		$J=30$	
L_t	L_e	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC
1	1	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.06**	**1.02**	**1.04**	**1.00**	**1.04**	**1.02**	**1.04**	**1.00**	**1.02**	**1.00**	**1.04**
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.31)	(0.14)	(0.28)	(0.00)	(0.20)	(0.14)	(0.28)	(0.00)	(0.14)	(0.00)	(0.20)
	2	2.00	2.00	2.00	2.00	2.00	2.00	2.02	2.00	1.98	1.98	2.00	1.96	1.98	1.98	2.00	1.98	2.00	1.98
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.14)	(0.28)	(0.14)	(0.14)	(0.00)	(0.20)	(0.14)	(0.14)	(0.00)	(0.14)	(0.00)	(0.24)
	3	3.00	3.00	3.00	3.00	3.00	3.00	2.98	2.94	3.00	2.98	3.00	3.00	3.00	2.98	3.00	3.00	3.00	2.98
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.14)	(0.24)	(0.00)	(0.14)	(0.00)	(0.00)	(0.00)	(0.14)	(0.00)	(0.00)	(0.00)	(0.14)
2	1	3.00	3.00	3.00	3.00	3.00	3.00	2.98	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.14)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	2	**1.00**	**1.00**	**1.02**	**1.02**	**1.00**	**1.02**	**1.00**	**1.02**	**1.02**	**1.02**	**1.00**	**1.02**	**1.00**	**1.00**	**1.00**	**1.02**	**1.00**	**1.02**
		(0.00)	(0.00)	(0.14)	(0.14)	(0.00)	(0.14)	(0.00)	(0.14)	(0.14)	(0.14)	(0.00)	(0.14)	(0.00)	(0.00)	(0.00)	(0.14)	(0.00)	(0.14)
	3	2.00	2.00	1.98	1.98	2.00	1.98	2.02	1.98	1.98	1.98	2.00	1.98	2.00	2.00	2.00	1.98	2.00	1.98
		(0.00)	(0.00)	(0.14)	(0.14)	(0.00)	(0.14)	(0.14)	(0.14)	(0.14)	(0.14)	(0.00)	(0.14)	(0.00)	(0.00)	(0.14)	(0.00)	(0.14)	(0.00)
3	1	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	2	1.98	2.00	2.00	2.00	2.00	2.00	1.96	1.98	1.98	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00
		(0.14)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.20)	(0.14)	(0.14)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	3	**1.02**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.04**	**1.02**	**1.02**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
		(0.14)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.20)	(0.14)	(0.14)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)

得られた結果を表 4 と表 5 に示す．表中の値は，各条件下において，真の次元数が L_t のときに次元数 L_e を仮定して得られた情報量の順位の平均（カッコ内は標準偏差）を表す．順位の値が小さいほど，その次元数 L_e が最適値として多く選択されたことを意味する．

まず，真の次元数 $L_t=1$ のときの順位に着目すると，全ての条件において正しい次元数 $L_e=1$ を精度よく選択していることがわかる．真の次元数 $L_t=2$，$L_t=3$ のときには，評価者数や項目数が増加しデータ数が増加するほど，正しい次元数を精度よく選択できる傾向があることがわかる．また，AIC や BIC はデータ数が少ない場合には真モデルより単純なモデルを選択する傾向があることが知られており [38]，本実験でも，データ数が少ないときにはこの傾向が読み取れる．更に，表 5 より，データ数が十分に多いときには，漸近一致性を有する BIC [39] が漸近一致性をもたない AIC に比べて高精度に真の次元を推定していることがわかる．

以上から，情報量規準を用いた提案モデルの次元数選択が，理論通りに動作する妥当な方法であることが確認できた．

6. 実データ実験

本章では，実データ適用を通して，提案モデルの有効性を評価する．本研究では，実データを収集するために，34 名の大学生と大学院生にエッセイ課題を行わせ，各課題に対して提出された回答文を 10 名の評価者に採点させた．本実験で利用したエッセイ課題は，NAEP（National Assessment of Educational Progress）2007 [40] で出題された課題を日本語に翻訳したものであり，専門知識や特別な事前知識を必要としない内容である．また，評価者による採点は，松下ら [7] が開発した表 6 のルーブリックを用いて 4 段階で行わせた．表 6 のルーブリックは，評価項目 1 と 2 が「問題解決力」を，評価項目 3〜5 が「論理的思考力」を測定すると想定して開発されている．本研究では，このデータに対して提案モデルを適用する．

6.1 次元数の決定

本実験では，適切な次元数を決定するために，実データ $\mathbf{U}$ から次元数 $L=1,\cdots,5$ を仮定して BIC と AIC を算出した．結果を図 2 に示す．図 2 の横軸は次元数 L の値であり，縦軸は各次元を仮定したときの情報量規準値である．図 2 より，いずれの情報量規準を用いても最適な能力の次元数は $L=2$ となったことがわかる．これは，ルーブリック作成者の想定した尺度数と合致している．そこで，以降では，$L=2$ として提案モデルの適用を行う．

6.2 尺度の解釈

ここでは，$L=2$ の提案モデルで推定されたパラメータ値に基づき，各次元の尺度について解釈を行う．**4.** で述べたように，提案モデルでは，項目識別力と項目の内容に着目することで各尺度の意味を解釈できる．

表 6　実データ実験で使用したルーブリック
Table 6　Rubric used in actual data experiment.

	項目 1：背景と問題	項目 2：主張と結論	項目 3：根拠と事実	項目 4：対立意見の検討	項目 5：全体構成
$k=3$	与えられたテーマから問題を設定し，論ずる意義も含め，その問題を取り上げた理由や背景について述べている．	設定した問題に対し，展開してきた自分の主張を関連づけながら，結論を導いている．結論は一般論にとどまらず，独自性を有している．	自分の主張の根拠が述べられており，かつ根拠の真実性を立証する信頼できる複数のデータが示されている．	自分の主張と対立する幾つかの意見を取り上げ，それら全てに対して論駁 (問題点の指摘) を行っている．	問題の設定から結論にいたる論理的な組み立て，記述の順序，パラグラフの接続が整っている．概要は本文の内容を的確に要約している．
$k=2$	与えられたテーマから問題を設定し，その問題を取り上げた理由や背景について述べている．	設定した問題に対し，展開してきた自分の主張を関連づけながら，結論を導いている．	自分の主張の根拠が述べられており，かつ根拠の真実性を立証する信頼できるデータが少なくとも一つ示されている．	自分の主張と対立する少なくとも一つの意見を取り上げ，それに対して論駁 (問題点の指摘) を行っている．	問題の設定から結論にいたる論理的な組み立て，記述の順序，パラグラフの接続がおおむね整っている．
$k=1$	与えられたテーマから問題を設定しているが，その問題を取り上げた理由や背景の内容が不十分である．	結論は述べられているが，展開してきた自分の主張との関連づけが不十分である．	自分の主張の根拠は述べられているが，根拠の真実性を立証する信頼できるデータが明らかにされていない．	自分の主張と対立する意見を取り上げているが，それに対して論駁（問題点の指摘）がなされていない．	問題の設定から結論にいたるアウトラインはたどれるが，記述の順序やパラグラフの接続に難点のある箇所が散見される．
$k=0$	$k=1$ 未満の水準	$k=1$ 未満の水準	$k=1$ 未満の水準	$k=1$ 未満の水準	$k=1$ 未満の水準

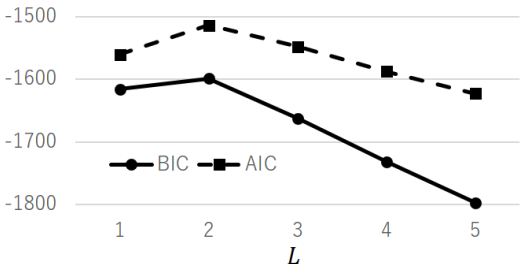


図 2　実データにおける次元数選択
Fig. 2　Dimensionality selection for actual data.

表 7　項目パラメータ推定値
Table 7　Item parameter estimates.

	項目 1	項目 2	項目 3	項目 4	項目 5
$\alpha_{jl=1}$	0.810	1.073	0.629	0.350	1.084
$\alpha_{jl=2}$	0.745	0.495	0.383	1.639	0.591
$\beta_{jk=1}$	−3.946	−3.884	−3.477	−1.342	−3.606
$\beta_{jk=2}$	−0.973	−1.009	−0.502	1.064	−0.875
$\beta_{jk=3}$	2.019	1.703	2.687	3.551	2.805

ここで，項目識別力の推定値を表 7 に示す．

まず，評価項目ごとに各次元の識別力を比較すると，評価項目 1，2，3，5 では次元 1 の識別力が相対的に大きく，評価項目 4 では次元 2 の識別力が大きく推定されている．これは，評価項目 1，2，3，5 と評価項目 4 がそれぞれ異なる能力尺度を測定していることを示唆している．ルーブリック作成者は，評価項目 1，2 と評価項目 3，4，5 が異なる尺度を構成していると想定していたが，本分析ではこの解釈とは異なる結果が得られたことがわかる．ルーブリックの内容を精査すると，評価項目 1，2，3，5 が自身の主張を正当化する論理構成力に重点をおくのに対し，評価項目 4 では他者の視点を想定した分析力が求められていると解釈できる．

表 8　評価者パラメータ推定値
Table 8　Rater parameter estimates.

	評価者 1	評価者 2	評価者 3	評価者 4	評価者 5
α_r	1.000	1.343	0.845	1.072	1.115
ϵ_r	0.000	−0.652	0.567	−1.327	−0.279
	評価者 6	評価者 7	評価者 8	評価者 9	評価者 10
α_r	1.059	1.079	1.649	1.033	1.883
ϵ_r	0.081	0.984	−0.006	0.013	1.112

以上のように，提案モデルでは，測定対象の能力尺度をデータに基づいて分析できることがわかる．

6.3　項目困難度と評価者特性

提案モデルでは，前節で説明した項目識別力に加えて，項目困難度と評価者の特性についても分析することができる．ここで，実データから推定された，項目困難度を表 7 に，評価者特性値を表 8 に示す．表 7 から，評価項目間で困難度に差異があることがわかる．例えば，評価項目 4 は β_{j1}，β_{j2} が他の項目より極端に高く，低得点を得にくい項目であることがわかる．反対に，評価項目 2 は β_{j3} が最も低く，最高点を得やすい項目であることがわかる．また，表 8 から，評価の厳しさや一貫性も評価者間で差異があることが確認できる．例えば，評価者 3 は一貫性が最も低いことから，評価のランダムネスが大きい評価者であると解釈できる．一貫性と厳しさが最も高い評価者 10 は，評

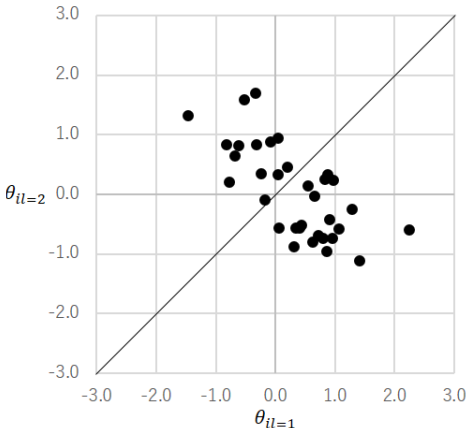


図 3　能力推定値
Fig. 3　Ability estimates.

価は相対的に厳しいものの，能力の高い受験者層を精度よく評価できる評価者であるといえる．また，評価の厳しさが最も小さい評価者 4 は，相対的に評価が甘い傾向があると解釈できる．

6.4　能力推定値

提案モデルでは，上述した評価者と評価項目の特性を考慮して，多次元尺度で受験者の能力を推定することができる．実データから推定された受験者の能力分布を図 3 に示す．図 3 は，横軸が 1 次元目の能力を，縦軸が 2 次元目の能力を表している．各プロットが個々の受験者を表す．能力の一次元性を仮定したモデルでは，このような多次元尺度での能力推定は実現できないが，提案モデルでは能力の多次元を導入したことによりこれが可能となる．また，提案モデルは，従来の多次元段階反応モデルとは異なり，評価者の特性を考慮して能力を推定できるため，より高精度な能力測定が実現できると期待される．そこで，次節では，提案モデルにより，能力測定の精度が向上するかを評価する．

6.5　能力測定の精度評価

ここでは，評価者の特性を考慮したことによる能力測定精度の改善について評価するために，能力測定の精度を，異なる評価者群から推定された能力値の安定性としてみなして評価を行う [15]．具体的には，同一の受験者群に対して，ある評価者群 A を用いて得られた能力推定値が，異なる評価者群 B から得られた能力推定値と近ければ，能力測定の精度が高いと解釈する．この考え方に基づき，以下の手順で精度を評価した．

（1）　実データを用いてパラメータを推定した．

（2）　評価者 10 人からランダムに 5 人選択して作成した評価者群を 60 群生成した．

（3）　手順 (1) で推定した項目パラメータ，評価者パラメータを所与とし，各評価者群の評点データから能力パラメータを推定した．

（4）　全ての評価者群間で能力推定値の RMSE を算出し，その平均値と標準偏差を求めた．

上記の実験では，RMSE が小さいほど，評価者の変化による能力推定値の変動が小さいことを表し，能力測定精度が高いことを意味する．

本実験では，提案モデルの能力測定精度を **3.3** で紹介した従来の多次元段階反応モデルと比較する．ただし，従来の多次元段階反応モデルでは 3 相データを直接には扱えないため，評価者得点の最頻値を用いて「受験者」×「評価項目」の 2 相データに変換して適用を行った．ただし，この方法との比較のみでは，精度の変化が 2 相データ化によるものか，評価者特性を考慮したことによるものかを明確には区別できない．そこで，3 相データを適用しつつ評価者特性の有無の影響を分析するために，提案モデルにおける評価者パラメータを $\alpha_r = 1$，$\epsilon_r = 0$，$\forall r$ とした場合についても精度の評価を行った．また，本実験では，各手法によって得られる RMSE の平均値の優位差を評価するために，Tukey 法による多重比較を行った．

表 9 に実験結果を示す．表では，「従来モデル」が多次元段階反応モデルの結果を表し，「評価者母数固定モデル」が評価者パラメータを固定した提案モデルの結果を表す．また，μ は RMSE の平均値，σ はその標準偏差，t は検定統計量を表す．提案モデルを，評価者パラメータを一定にした提案モデルと比較すると，提案モデルが優位に高い能力測定精度を示したことがわかる．これは，能力測定精度が評価者特性に依存することを意味しており，評価者特性を考慮した能力推定によりこの精度を向上できたことを示している．また，従来の多次元段階反応モデルは，他のモデルと比べて著しく能力測定精度が低いことがわかる．これは，

表 9　能力測定精度の評価結果
Table 9　Accuracy of ability measurement.

	提案モデル	従来モデル	評価者母数固定モデル
	$\mu = 0.432$	$\mu = 0.514$	$\mu = 0.446$
	$\sigma = 0.118$	$\sigma = 0.088$	$\sigma = 0.134$
従来モデル	$t = 30.227$	-	-
	$p < 0.01$	-	-
評価者母数固定モデル	$t = 5.309$	$t = 24.919$	-
	$p < 0.01$	$p < 0.01$	-

多次元段階反応モデルでは評価者特性を考慮できないことに加え，データの 2 相化により受験者に対する評点データが減少するためと考えられる．

以上の実験から，提案モデルが能力測定の能力測定精度向上に有効であることが確認できた．

7. むすび

本研究では，パフォーマンス評価において，評価者の特性を考慮して多次元尺度で受験者の能力を測定できる新たな項目反応モデルを提案した．提案モデルは，既存の多値型多次元項目反応モデルに対して，評価者の特性を表すパラメータを付与したモデルとして定式化した．また，提案モデルのパラメータ推定手法として，MCMC アルゴリズムを用いたアルゴリズムを提案し，シミュレーション実験によりアルゴリズムの妥当性を示した．更に，情報量規準に基づくモデル選択のアプローチを提案モデルに適用することで，能力尺度の最適な次元数を推定できることを，シミュレーション実験により示した．実データ実験では，モデルのパラメータ推定値に基づいて各次元の能力尺度の意味を解釈できることを示した．また，提案モデルが評価者特性を考慮した高精度な能力測定を実現できることを，従来モデルとの比較により示した．

今後は，より多様なデータに適用して提案モデルの有効性を検証していきたい．また，本研究では，受験者は一つの課題を与えられると仮定したが，実際には複数の課題を与えることが多いため，今後は提案モデルに課題の特性パラメータを付与した 4 相モデルへの拡張についても検討したい．

謝辞　本研究は JSPS 科研費 17H04726，17K20024 の助成を受けたものです．

文　献

[1] R. Schendel and A. Tolmie, “Assessment techniques and students’ higher-order thinking skills,” Assessment & Evaluation in Higher Education, vol.42, no.5, pp.673–689, 2017.

[2] Y. Abosalem, “Beyond translation: Adapting a performance-task-based assessment of critical thinking ability for use in Rwanda,” Int. J. Secondary Education, vol.4, no.1, pp.1–11, 2016.

[3] Y. Rosen and M. Tager, “Making student thinking visible through a concept map in computer-based assessment of critical thinking,” J. Educational Computing Research, vol.50, no.2, pp.249–270, 2014.

[4] O.L. Liu, L. Frankel, and K.C. Roohr, “Assessing critical thinking in higher education: Current state and directions for next-generation assessment,” ETS Research Report Series, vol.2014, no.1, pp.1–23, 2014.

[5] H.J. Bernardin, S. Thomason, M.R. Buckley, and J.S. Kane, “Rater rating-level bias and accuracy in performance appraisals: The impact of rater personality, performance management competence, and rater accountability,” Human Resource Management, vol.55, no.2, pp.321–340, 2016.

[6] 宇都雅輝，植野真臣，“パフォーマンス評価のため項目反応モデルの比較と展望，” 日本テスト学会誌，vol.12, no.1, pp.55–75, 2016.

[7] 松下佳代，小野和宏，高橋雄介，“レポート評価におけるルーブリックの開発とその信頼性の検討，” 大学教育学会誌，vol.35, no.1, pp.107–115, 2013.

[8] M. Uto and M. Ueno, “Item response theory for peer assessment,” IEEE Trans. Learning Technologies, vol.9, no.2, pp.157–170, 2016.

[9] N.L.A. Kassim, “Judging behaviour and rater errors: An application of the many-facet Rasch model,” GEMA Online Journal of Language Studies, vol.11, no.3, pp.179–197, 2011.

[10] C.M. Myford and E.W. Wolfe, “Detecting and measuring rater effects using many-facet Rasch measurement: Part I,” J. Applied Measurement, vol.4, pp.386–422, 2003.

[11] T. Eckes, “Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis,” Language Assessment Quarterly, vol.2, no.3, pp.197–221, 2005.

[12] 宇都雅輝，植野真臣，“ピアアセスメントにおける異質評価者に頑健な項目反応理論，” 信学論（D），vol.J101-D, no.1, pp.211–224, Jan. 2018.

[13] T. Eckes, Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluating Rater-Mediated Assessments, Peter Lang Pub., 2015.

[14] 宇佐美慧，“採点者側と受験者側のバイアス要因の影響を同時に評価する多値型項目反応モデル：MCMC アルゴリズムに基づく推定，” 教育心理学研究，vol.58, no.2, pp.163–175, 2010.

[15] 宇佐美慧，“論述式テストの運用における測定論的問題とその対処，” 日本テスト学会誌，vol.9, no.1, pp.145–164, 2013.

[16] M. Uto and M. Ueno, “Empirical comparison of item response theory models with rater’s parameters,” Heliyon, Elsevier, vol.4, no.5, pp.1–32, 2018.

[17] J.M. Linacre, Many-faceted Rasch Measurement, MESA Press, 1989.

[18] R.J. Patz, B.W. Junker, M.S. Johnson, and L.T. Mariano, “The hierarchical rater model for rated test items and its application to large-scale educational assessment data,” J. Educational and Behavioral Statistics, vol.27, no.4, pp.341–366, 1999.

[19] 宇都雅輝，植野真臣，“ピアアセスメントの低次評価者母数をもつ項目反応理論，” 信学論（D），vol.J98-D, no.1, pp.3–16, Jan. 2015.

[20] 植野真臣，ソンムァンポクポン，岡本敏雄，永岡慶三，“ピアアセスメントにおける評価者特性を考慮した項目反応理論，” 信学論（D），vol.J91-D, no.2, pp.377–388, Feb. 2008.
[21] 鈴木雅之，“ルーブリックの提示による評価基準・評価目的の教示が学習者に及ぼす影響，” 教育心理学研究，vol.59, no.2, pp.131–143, 2011.
[22] 中嶋一恵，浦川末子，白石景一，下釜綾子，永野　司，中村浩美，中島健一郎，滝川由香里，本村弥寿子，“ルーブリックを使用した学外実習評価基準の作成について，” 長崎女子短期大学紀要，2014.
[23] 孫　媛，“多次元データに対する項目反応モデル，” 学術情報センター紀要，vol.9, pp.103–111, 1997.
[24] L.R. Hutten, Some Empirical Evidence for Latent Trait Model Selection, ERIC Clearinghouse, 1980.
[25] E. Muraki and J.E. Carlson, “Full-information factor analysis for polytomous item responses,” Applied Psychological Measurement, vol.19, no.1, pp.73–90, 1995.
[26] M.D. Reckase, Multidimensional Item Response Theory Models, Springer, 2009.
[27] F.M. Lord, Applications of item response theory to practical testing problems, Erlbaum Associates, 1980.
[28] F. Samejima, “Estimation of latent ability using a response pattern of graded scores,” Psychometrika Monography, vol.17, pp.1–100, 1969.
[29] E. Muraki, “A generalized partial credit model: Application of an EM algorithm,” Applied Psychological Measurement, vol.16, no.2, pp.159–176, 1992.
[30] L.R. Fabrigar, D.T. Wegener, R.C. MacCallum, and E.J. Strahan, “Evaluating the use of exploratory factor analysis in psychological research,” Psychological Methods, vol.4, no.3, pp.272–299, 1999.
[31] G. Schwarz, “Estimating the dimensions of a model,” Annals of Statistics, vol.6, pp.461–464, 1978.
[32] H. Akaike, “A new look at the statistical model identification,” IEEE Trans. Autom. Control, vol.19, pp.716–723, 1974.
[33] F.B. Baker and S.H. Kim, Item Response Theory: Parameter Estimation Techniques, Statistics, textbooks and monographs, Marcel Dekker, 2004.
[34] J.-P. Fox, Bayesian item response modeling: Theory and applications, Springer, 2010.
[35] R.J. Patz and B.W. Junker, “Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses,” J. Educational and Behavioral Statistics, vol.24, pp.342–366, 1999.
[36] M. Martin-Fernandez and J. Revuelta, “Bayesian estimation of multidimensional item response models. a comparison of analytic and simulation algorithms,” Int. J. Methodology and Experimental Psychology, vol.38, no.1, pp.25–55, 2017.
[37] C.M. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics), Springer-Verlag, 2006.
[38] 植野真臣，“ベイジアンネットワークの統計的学習，” 人工知能学会誌，vol.25, no.6, pp.803–810, 2010.
[39] 渡辺澄夫，ベイズ統計の理論と方法，コロナ社，2012.
[40] D. Salahu-Din, H. Persky, and J. Miller, “The nation’s report card: Writing 2007,” Technical report, National Center for Education Statistics, 2008.

（2019 年 2 月 17 日受付，4 月 23 日再受付，
5 月 31 日早期公開）

八木　嵩大

2019 年電気通信大学情報理工学部卒．e テスティング，項目反応理論の研究に従事．日本テスト学会第 16 回大会において大会奨励賞受賞．平成 30 年度電気通信大学目黒会賞受賞．

宇都　雅輝　（正員）

2013 年電気通信大学大学院情報システム学研究科博士後期課程了．博士 (工学)．長岡技術科学大学を経て，2015 年より電気通信大学助教に着任，現在に至る．e テスティング，e ラーニング，人工知能，ベイズ統計，自然言語処理などの研究に従事．