パフォーマンス評価における多次元段階反応モデルの提案と評価

Proposal and Evaluation of Multidimensional Item Response Theory Model for Performance Assessment

八木嵩大[1*]　宇都雅輝[1]
Shudai Yagi[1]　Masaki Uto[1]

[1] 電気通信大学
[1] University of Electro-Communications

Abstract: Performance assessment has been attracted much attention in various assessment contexts as a method to measure higher abilities of examinees. A persistent difficulty of performance assessment is that ability measurement accuracy depends strongly on characteristics of raters such as severity and consistency. To resolve the problem, item response theory models that incorporate rater characteristic parameters have been proposed. Those models require the assumption of unidimensionality, which means that one latent ability is measured in a test. The assumption, however, might not be satisfied in performance assessment because multiple sub-abilities are often measured using a rubric with a set of assessment criteria. To solve the problem, this study proposes a new multi-dimensional item response theory model that incorporates rater characteristic parameters. Moreover, this study proposes the Markov chain Monte Carlo algorithm as a parameter estimation method for the proposed model, and demonstrates the effectiveness of the proposed model through simulation experiments and real data application.

1　はじめに

近年，大学入試や人事考課，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]．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（甘さ/厳しさなど）に依存する点が指摘されてきた [1]．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 [1, 2, 3, 4, 5]．これらのモデルでは評価者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，高精度な能力測定が実現できる [1, 3, 4, 5]．

これらの項目反応モデルは受験者の能力を一次元の値として推定する．しかし，パフォーマンス評価では，評価基準表を用いて複数次元の能力を測定するように設計されることがある [6, 7]．既存のモデルでは，このような多次元を仮定した能力測定はできない．一方，能力の多次元性を仮定した項目反応モデルとして，多次元項目反応モデルが知られている [8]．しかし，既存の多次元項目反応モデルは，評価者の特性を考慮した能力測定を行うことはできないため，パフォーマンス評価における能力測定精度が評価者特性に依存する問題が残る．

そこで，本研究では，評価者特性を考慮した多次元項目反応モデルを提案する．また，提案モデルのパラメータ推定法としてマルコフ連鎖モンテカルロ（MCMC）法を用いた手法を開発する．提案モデルの特徴は以下のとおりである．(1）能力尺度の適切な次元数をデータから推定できる．(2）測定している能力尺度を解釈できる．(3）評価者特性を考慮した多次元尺度での能力測定ができるため，従来の多次元項目反応モデルに比べ高精度な能力測定が可能である．

さらに，シミュレーション実験および実データ実験により提案モデルの有効性を示す．

2　提案モデル

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評価基準表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 $\mathcal{R} = \{1, \cdots, R\}$，評価基準表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 $\mathcal{K} = \{0, \cdots, K-1\}$ とおく．こ

*連絡先：電気通信大学情報理工学部 宇都研究室
〒 182-8585 東京都調布市調布ヶ丘 1-5-1.
E-mail: yagi@ai.lab.uec.ac.jp

こで，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．

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

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 $\mathbf{U}$ から，評価者の特性を考慮して多次元尺度で受験者の能力を推定できる項目反応モデルを提案する．項目反応モデルは，近代のテスト分野で広く実用・研究される潜在変数モデルである．提案モデルでは，受験者 i のパフォーマンスに関して評価者 r が評価項目 j について評点 k を与える確率 P_{ijrk} を次式で定義する．

$$P_{ijrk} = P^*_{ijrk} - P^*_{ijrk+1} \quad (2)$$

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

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．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．

提案モデルのパラメータ推定は，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 MCMC 法で行う．アルゴリズムの詳細は八木・宇都 [9] を参照されたい．

3　シミュレーション実験

3.1　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（2）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，先行研究 [10] 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 $j \in \mathcal{J}'$ の識別力パラメータを以下の値に設定した．

$$\begin{cases} \alpha_{jl} = 1.65 & j = J + l \\ \alpha_{jl} = 0.22 & j \neq J + l \end{cases} \quad (3)$$

   困難度パラメータは，カテゴリー数 $K = 2$ として，$\beta_{j1} = 0$ とした．

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 (2) 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 (2) 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二乗誤差（RMSE）を算出した．

6. 手順 (2)〜(5) を 10 回行い，RMSE の平均を算出した．

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．

実験結果を表 1 に示す．表 1 から，項目数や評価者数の増加に伴い，RMSE の値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., [11, 1, 12]）と一致した傾向を示している．また，次元数の増加により能力値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 [10] と一致した傾向となっている．

以上より，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．

3.2　情報量基準に基づく次元数推定の妥当性評価

ここでは，情報量基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 を情報量基準として用い，以下の実験を行なった．

表 1: シミュレーション実験における RMSE

		$J=5$			$J=10$			$J=15$		
	L	$R=5$	$R=10$	$R=15$	$R=5$	$R=10$	$R=15$	$R=5$	$R=10$	$R=15$
α	1	0.232	0.166	0.137	0.203	0.173	0.146	0.198	0.139	0.158
	2	0.343	0.271	0.248	0.332	0.288	0.248	0.352	0.329	0.238
	3	0.341	0.356	0.296	0.446	0.383	0.389	0.458	0.356	0.352
β	1	0.172	0.151	0.100	0.179	0.144	0.102	0.169	0.132	0.142
	2	0.180	0.128	0.113	0.154	0.128	0.112	0.177	0.135	0.130
	3	0.185	0.149	0.129	0.181	0.152	0.137	0.187	0.147	0.120
α_r	1	0.116	0.143	0.140	0.100	0.087	0.102	0.080	0.096	0.101
	2	0.125	0.127	0.128	0.087	0.105	0.097	0.118	0.081	0.088
	3	0.105	0.135	0.126	0.136	0.126	0.101	0.128	0.094	0.080
ϵ_r	1	0.214	0.198	0.209	0.180	0.225	0.183	0.144	0.166	0.233
	2	0.265	0.181	0.189	0.158	0.171	0.136	0.176	0.201	0.174
	3	0.224	0.177	0.182	0.272	0.163	0.156	0.143	0.164	0.165
θ	1	0.308	0.246	0.200	0.249	0.201	0.186	0.234	0.186	0.179
	2	0.448	0.342	0.293	0.357	0.273	0.220	0.315	0.280	0.202
	3	0.496	0.399	0.306	0.487	0.359	0.288	0.385	0.313	0.270

表 2: シミュレーション実験における次元数選択

		$R=5$						$R=10$						$R=15$					
		$J=5$		$J=10$		$J=15$		$J=5$		$J=10$		$J=15$		$J=5$		$J=10$		$J=15$	
L_t	L_e	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC
1	1	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
	2	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0
	3	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0
2	1	**1.2**	**1.4**	**1.2**	1.8	1.7	2.4	1.6	2.1	1.6	2.5	2.2	2.9	1.7	2.0	2.1	2.7	2.9	3.0
	2	1.8	1.7	1.8	**1.4**	**1.4**	**1.1**	**1.5**	**1.3**	**1.5**	**1.0**	**1.2**	**1.0**	**1.5**	**1.4**	**1.3**	**1.1**	**1.0**	**1.0**
	3	3.0	2.9	3.0	2.8	2.9	2.5	2.9	2.6	2.9	2.5	2.6	2.1	2.8	2.6	2.6	2.2	2.1	2.0
3	1	**1.3**	**1.6**	**1.4**	2.6	1.7	2.6	1.8	2.7	2.9	3.0	2.9	3.0	2.5	2.8	3.0	3.0	3.0	3.0
	2	1.7	**1.6**	1.7	**1.4**	**1.5**	**1.3**	**1.6**	**1.3**	**1.4**	1.6	1.6	1.9	**1.3**	**1.4**	**1.3**	1.8	1.9	2.0
	3	3.0	2.8	2.9	2.0	2.8	2.1	2.6	2.0	1.7	**1.4**	**1.5**	**1.1**	2.2	1.8	1.7	**1.2**	**1.1**	**1.0**

1. 真の次元数を L_t とし，モデルパラメータとデータ $\mathbf{U}$ を生成した．
2. 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なった．

上記の実験を 10 回繰り返し，順位の平均を算出した．また，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．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．

得られた結果を表 2 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．

真の次元数 $L_t = 1$ の場合は，すべての場合において正しい次元数 $L_e = 1$ を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ていることがわかる．以上から，情報量基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．

4 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，National Assessment of Educational Progress（NAEP）2007 [13] で出題された課題を日本語に翻訳したもので

表 3: 本実験で使用したルーブリック

	問題解決力		論理的思考力		
	項目 1:背景と問題（与えられたテーマから自分で問題を設定する）	**項目 2：主張と結論**（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く）	**項目 3：根拠と事実**（主張を支える根拠を述べ，根拠の真実性を立証する事実を明らかにする）	**項目 4：対立意見の検討**（自分の主張と対立する意見を取り上げ，それに対して論駁を行う）	**項目 5：全体構成**（問題の設定から結論にいたる過程を論理的に組み立て表現する）
$k=3$	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事実・データが示されている．	自分の主張と対立するいくつかの意見を取り上げ，それらすべてに対して論駁 (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事実・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事実･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

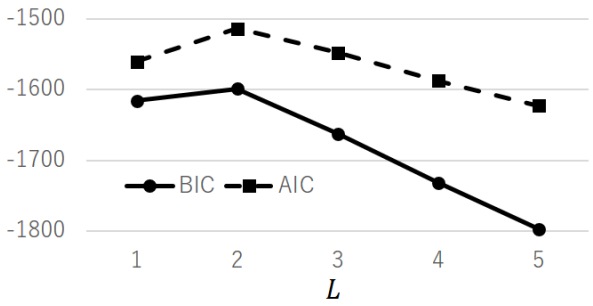

図 1: 実データにおける次元数選択

表 4: 項目パラメータ推定値

	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

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 [14] が開発した表 3 のルーブリックを用いて 4 段階で行わせた．表 3 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．

4.1 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 1 に示す．図 1 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基準値である．図 1 より，いづれの情報量基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．

4.2 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 章で述べたように，提案モデルでは，項目識別力に着目することで各尺度の意味を解釈できる．ここで，項目識別力の推定値を表 4 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．

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．

表 5: 評価者パラメータ推定値

	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

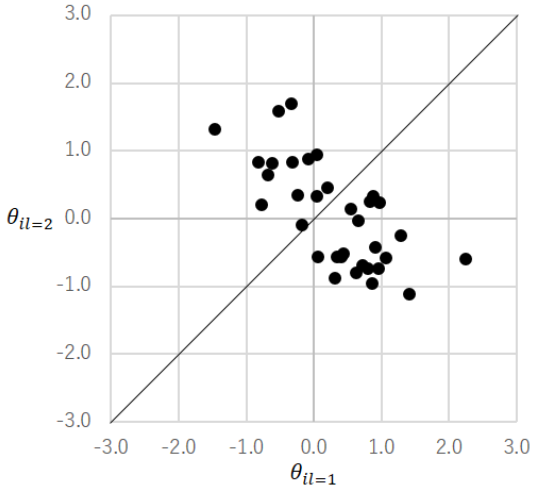

図 2: 能力推定値

4.3 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 4 に，評価者特性値を表 5 に示す．表 4 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 の「主張と結論」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 5 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，評価が相対的に厳しいが，特に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．

4.4 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 2 に示す．図 2 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，このような下位尺度ごとの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，このような多次元での能力推定が可能となる．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た能力測定を行うことができるため，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．

4.5 能力測定の精度評価

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の組を評価者群と呼び，評価者群を 60 組生成した．
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，各評価者群における能力パラメータを推定した．
4. 60 組の評価者群から任意の 2 組を選ぶ組み合わせの集合（$C_{60}^2 = 1770$ 通り）に対して，能力パラメータに関して RMSE を算出し，その平均を求めた．

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．

ここでは，提案モデルの能力測定精度を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なった．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なった．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 6: 能力測定精度の評価結果

	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	$t = 5.309$	$t = 24.919$	-
	$p < 0.01$	$p < 0.01$	-

表 6 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．表 6 から，提案モデルが従来の多次元段階反応モデルと比べて，優位に高い能力測定精度を示したことがわかる．これは，多次元段階反応モデルではデータを 2 相化する必要があるため，受験者に対する評点データが少なくなることが要因であると考えられる．また，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，能力推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能力測定精度を向上できたことを示している．

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．

5 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．さらに，情報量基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．

参考文献

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

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

[3] R.J. Patz and B.W. Junker, “Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses,” Journal of Educational and Behavioral Statistics, vol.24, pp.342–366, 1999.

[4] L.T. DeCarlo, “A model of rater behavior in essay grading based on signal detection theory,” Journal of Educational Measurement, vol.42, no.1, pp.53–76, 2005.

[5] M. Uto and M. Ueno, “Item response theory without restriction of equal interval scale for rater’s score,” Proc. International Conference on Artificial Intelligence in Education, pp.363–368, 2018.

[6] 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.

[7] 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.

[8] M.D. Reckase, Multidimensional Item Response Theory Models., Springer, 2009.

[9] 八木嵩大，宇都雅輝，“パフォーマンス評価における多次元尺度を構成する項目反応モデル，” 人工知能学会 先進的学習科学と工学研究会，vol.B5，no.01，pp.19–24，2018.

[10] M. Martin-Fernandez and J. Revuelta, “Bayesian estimation of multidimensional item response models. a comparison of analytic and simulation algorithms,” International Journal of Methodology and Experimental Psychology, vol.38, no.1, pp.25–55, 2017.

[11] 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.

[12] C.M. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics), Springer-Verlag, 2006.

[13] D. Salahu-Din, H. Persky, and J. Miller, “The nation’s report card: Writing 2007,” Technical report, National Center for Education Statistics, 2008.

[14] 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.

[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.