THE IEICE TRANSACTIONS ON INFORMATION AND SYSTEMS (JAPANESE EDITION)

電子情報通信学会
論文誌 D 情報・システム

DOI:10.14923/transinfj.2019JDP7065
早期公開日:2020/01/15

情報・システムソサイエティ
一般社団法人 電子情報通信学会
THE INFORMATION AND SYSTEMS SOCIETY
THE INSTITUTE OF ELECTRONICS, INFORMATION AND COMMUNICATION ENGINEERS

論　文

ルーブリック評価における項目反応理論

宇都雅輝[†a)]　　植野真臣[†b)]

Item response theory for rubric-based assessment

Masaki UTO[†a)] and Maomi UENO[†b)]

あらまし　近年，学習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ルーブリック評価が注目されている．ルーブリックは評価者の主観による評価基準をより客観的にするためのツールであるが，それでも評価がパフォーマンス課題や評価者，ルーブリックの評価観点の特性に依存してしまうことが指摘されてきた．この問題を解決する手法の一つとして，これらの特性を考慮して学習者の能力を測定できる項目反応モデルが近年多数提案されている．しかし，既存モデルは学習者・課題・評価者・評価観点で構成される 4 相の評価データに直接には適用できず，課題・評価者・評価観点の特性を同時に考慮した能力測定は実現できない．また，ルーブリック評価の評点は一般に段階カテゴリとして与えられ，各カテゴリに対する評価基準は評価者と評価観点の特性に依存する．しかし，既存モデルでは評価基準は評価者と評価観点のいずれか一方にのみ依存すると仮定している．以上の問題を解決するために，本論文では，評価観点と評価者の評価基準を考慮して，ルーブリック評価の 4 相データから学習者の能力を測定できる新たな項目反応モデルを提案する．また，シミュレーション実験と実データ実験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，ルーブリック，項目反応理論，教育評価，評価者バイアス

1. ま え が き

近年，学習評価場面において，論理的・批判的思考力や表現力といった学習者の真正な能力を測定するニーズが高まっており，このような能力を測定する手法の一つとしてルーブリック評価が注目されている [1]〜[5]．ルーブリック評価とは，現実的な課題に対する学習者のパフォーマンスを，ルーブリックと呼ばれる評価基準表を用いて評価者が採点する方法であり，記述・論述式試験やレポート課題，グループディスカッションやプレゼンテーション課題などの形式で利用されてきた．ルーブリックを利用する利点としては，測定対象の能力を明確化できることや，評価者の主観的な評価を客観的にさせることなどが挙げられる [4], [5]．

しかし，それでもルーブリック評価では，学習者の能力測定精度がパフォーマンス課題や評価者，ルーブリックの評価観点の特性に依存してしまうことが指摘されてきた [6]〜[13]．この問題を解決する手法の一つとして，これら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 [8]〜[11]．具体的には，課題と評価者の特性パラメータを付与したモデル [14]〜[19] や，評価者とルーブリックの評価観点の特性を考慮したモデル [12], [13] が提案されてきた．これらの項目反応モデルは，素点平均などの単純な手法と比べて高精度な能力測定が実現できる [10], [15], [16]．しかし，既存モデルをルーブリック評価に適用する場合，次の問題が残る．

（1）　ルーブリック評価で得られるデータは学習者 × 課題 × 評価者 × 評価観点の 4 相データとなる．しかし，既存モデルは学習者 × 課題 × 評価者，または学習者 × 評価者 × 評価観点の 3 相データへの適用を仮定しているため，ルーブリック評価の 4 相データに直接には適用できず，課題・評価者・評価観点の特性を同時に考慮した能力測定は実現できない．

（2）　ルーブリック評価の評点は，一般に順序尺度に従う段階カテゴリとして与えられる．各カテゴリに対する評価基準はルーブリックの評価観点ごとに定義され，理想的には評価観点の特性のみで決まる．しかし，現実には評価者ごとに評価基準の解釈が異なるこ

† 電気通信大学，調布市
University of Electro-Communications, Chofugaoka 1-5-1, Chofu-shi, Tokyo, 182–8585 Japan
a) E-mail: uto@ai.lab.uec.ac.jp
b) E-mail: ueno@ai.is.uec.ac.jp

とが多いため [4]〜[6]，各カテゴリに対する評価基準は評価観点だけでなく評価者の特性にも依存する．これに対し，既存モデルでは，評価基準は評価者と評価観点のいずれか一方にのみ依存すると仮定している．

以上の問題を解決するために，本研究では，ルーブリック評価の 4 相データに適用でき，評価観点と評価者の評価基準を考慮できる新たな項目反応モデルを提案する．提案モデルの利点は次の通りである．

（1） 4 相データから課題・評価者・評価観点の特性を同時に考慮して学習者の能力を測定できるため，従来モデルと比べて高精度な能力測定が期待できる．

（2） 評価観点だけでなく評価者の評価基準も考慮できるためデータへの当てはまりが改善され，能力測定精度が向上すると期待できる．

本論文では，シミュレーション実験と実データ実験を通して，提案モデルの有効性を評価する．

2. ルーブリック評価データ

本研究では，最も一般化されたルーブリック評価の状況として，複数の課題に対する学習者のパフォーマンスを，複数の評価者がルーブリックを用いて複数の評価観点に基づいて採点する場合を想定する．ルーブリックとは，パフォーマンスの質を評価するために用いられる評価基準表のことであり， 一つ以上の評価観点とそれについての数値的な尺度および尺度の中身を説明する記述語から構成される [3]．一般に尺度には順序尺度が用いられ，段階評価カテゴリで評点が与えられる [2]〜[5]．例として，松下ら [3] が開発したライティング評価のためのルーブリックを表 1 に示す．この例では 5 つの評価観点について，それぞれ 4 段階の評価カテゴリが定義されている．

ここで，I, J, R, C, K をそれぞれ課題数，学習者数，評価者数，評価観点数，評価カテゴリ数とすると，表 1 のようなルーブリックを用いた評価データ $\mathbf{X}$ は，課題 $i \in \mathcal{I} = \{1, \ldots, I\}$ における学習者 $j \in \mathcal{J} = \{1, \ldots, J\}$ のパフォーマンスに対し，評価者 $r \in \mathcal{R} = \{1, \ldots, R\}$ が評価観点 $c \in \mathcal{C} = \{1, \ldots, C\}$ に基づいて与える評点 $x_{ijrc} \in \mathcal{K} = \{1, \ldots, K\}$ の集合として以下で定義できる．

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

ここで，$x_{ijrc} = -1$ は欠測データを表す．

3. 項目反応理論

本研究の目的は，前節で定義したルーブリック評価データ $\mathbf{X}$ から，課題・評価者・ルーブリックの評価観点の特性を考慮した高精度な能力測定を行うことにある．このような能力測定を行うために，本研究では，項目反応理論（Item response theory: IRT）[20] を利用する．なお，本研究では測定対象の能力に一次元性を仮定する．

IRT 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．IRT は，正誤判定問題や多肢選択式問題などの 2 値の正誤データを扱うテストに対して広く適用されてきた．また，近年では，論述式・記述式テストのような多段階カテゴリを用いた評価データに対し，多値型 IRT モデルを適用する研究も進められている [21], [22]．本研究で扱うようなリッカート型データに適用できる代表的な多値型 IRT モデルとしては，段階反応モデル（Graded Response Model: GRM）[23] や一般化部分採点モデル（Generalized Partial Credit Model: GPCM）[24] が知られている．

3.1 段階反応モデル

GRM は，Samejima [23] が考案した多値型 IRT モデルであり，課題 i において学習者 j が評点 k を得る確率 P_{ijk} を次式で定義する．

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

P^*_{ijk} は課題 i において学習者 j が k より大きい評点を得る確率を表し，次式で定義される．

$$\begin{cases} P^*_{ijk} = [1 + \exp(-D\alpha_i(\theta_j - b_{ik}))]^{-1} \\ P^*_{ij0} = 1,\ P^*_{ijK} = 0. \end{cases} \quad (3)$$

ここで，θ_j は学習者 j の能力，α_i は課題 i の識別力，b_{ik} は課題 i において k より大きい評点を得る困難度を表す．困難度パラメータ b_{ik} には順序制約 $b_{i1} < b_{i2} < \cdots < b_{iK-1}$ が課される．定数 D はロジスティック関数を累積正規分布関数に近似するための定数であり，一般に 1.7 が利用される．

3.2 一般化部分採点モデル

GPCM では反応確率 P_{ijk} を次式で定義する．

$$P_{ijk} = \frac{\exp \sum_{m=1}^{k} [D\alpha_i(\theta_j - \beta_i - d_{im})]}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} [D\alpha_i(\theta_j - \beta_i - d_{im})]} \quad (4)$$

表 1: ライティング評価ルーブリック

	観点 1：背景と問題	観点 2：主張と結論	観点 3：根拠と事実	観点 4：対立意見の検討	観点 5：全体構成
$k=4$	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．
$k=3$	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．
$k=2$	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．
$k=1$	$k=2$ 未満の水準	$k=2$ 未満の水準	$k=2$ 未満の水準	$k=2$ 未満の水準	$k=2$ 未満の水準

ここで，β_i は課題 i の困難度を表す位置パラメータであり，d_{ik} は課題 i において評点 k を得る困難度を表すステップパラメータである．ただし，モデルの識別性のために，$d_{i1}=0$，$\sum_{k=2}^{K} d_{ik}=0:\forall i$ と制約する．

GPCM は，評定尺度モデル（Rating Scale Model: RSM) [25] や部分採点モデル（Partial Credit Model: PCM) [26] などの複数の多値型 IRT モデルの一般形となっている．PCM は GPCM において $\alpha_i=1.0;\forall i$ と制約したモデル，RSM は PCM において $d_{ik}=d_k;\forall i$ と制約したモデルとして定義される．ただし，d_k は評点 k を得る困難度を表すパラメータである．

4. 評価者特性を考慮した項目反応モデル

3. で紹介した多値型 IRT モデルは，課題における学習者の評点で構成される学習者 × 課題の二相データに適用される．一方で，本研究で扱うようなパフォーマンス課題に対する評価では，個々の対象を複数の評価者で採点することが一般的であり，評価データは学習者 × 課題 × 評価者の三相データとなる．上記の多値型 IRT モデルは，このような三相データに対して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 IRT モデルが近年多数提案されている [8]〜[11].

4.1 課題と評価者の特性を考慮した IRT モデル

評価者パラメータを付与した代表的な IRT モデルとして，多相ラッシュモデル（MFRM: Many-Facet Rasch Model）[14] が知られている．MFRM にはいくつかのバリエーションが存在するが [8], [9]，一般には RSM に評価者の厳しさを表すパラメータを付与したモデルとして定式化される．このモデルでは，課題 i における学習者 j のパフォーマンスに評価者 r が評点 k を与える確率 P_{ijrk} を次式で定義する．

$$P_{ijrk}=\frac{\exp\sum_{m=1}^{k}\left[D(\theta_j-\beta_i-\beta_r-d_m)\right]}{\sum_{l=1}^{K}\exp\sum_{m=1}^{l}\left[D(\theta_j-\beta_i-\beta_r-d_m)\right]}, \tag{5}$$

ここで，β_r は評価者 r の厳しさを表すパラメータである．モデルの識別性のために $\sum_{r=1}^{R}\beta_r=0$，$d_1=0$, $\sum_{k=2}^{K} d_k=0$ を仮定する．

MFRM では，1）全ての課題について識別力が一定であること，2）全ての評価者が同等の一貫性を有すること，が仮定される．しかし，現実にはこれらの仮定は成り立たないことが多い [10], [27], [28]．そこで，この制約を緩めたモデルとして，課題間での識別力の差異と評価者間の一貫性の差異を考慮できるモデルが提案されている．

課題識別力と評価者一貫性を考慮した最先端モデルの一つが Uto and Ueno のモデル [15] である．このモデルは GRM の拡張モデルとして定式化され，反応確率 P_{ijrk} を次式で定義する．

$$P_{ijrk}=P^{*}_{ijrk-1}-P^{*}_{ijrk}, \tag{6}$$

P^{*}_{ijrk} は課題 i における学習者 j のパフォーマンスに評価者 r が k より大きい評点を与える確率を表し，次式で定義される．

$$\begin{cases} P^*_{ijrk} = [1 + \exp(-D\alpha_i\alpha_r(\theta_j - b_{ik} - \varepsilon_r))]^{-1}, \\ P^*_{ijr0} = 1,\ P^*_{ijrK} = 0. \end{cases}$$

ここでは，α_r が評価者 r の一貫性を，ϵ_r は評価者 r の厳しさを表す．モデルの識別性のために $\prod_{r=1}^{R} \alpha_r = 1$，$\sum_{r=1}^{R} \varepsilon_r = 0$ を仮定する．

また，課題識別力と評価者一貫性を考慮した GPCM も提案されている [16]．このモデルでは反応確率 P_{ijrk} を次式で定義する．

$$P_{ijrk} = \frac{\exp \sum_{m=1}^{k} [D\alpha_i\alpha_r(\theta_j - \beta_i - \beta_r - d_{rm})]}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} [D\alpha_i\alpha_r(\theta_j - \beta_i - \beta_r - d_{rm})]} \tag{7}$$

ここで，d_{rk} は評価者 r の評点 k に対する厳しさを表すステップパラメータである．モデルの識別性のために，$\prod_{r=1}^{R} \alpha_r = 1$，$\sum_{r=1}^{R} \beta_r = 0$，および $d_{r1} = 0$，$\sum_{k=2}^{K} d_{rk} = 0 : \forall r$ を仮定する．

式（6）と式（7）のモデルの本質的な差異は，各評価カテゴリに対する評価基準が課題と評価者のどちらに依存すると仮定するかにある．式（6）のモデルでは課題パラメータ b_{ik} が，式（7）のモデルでは評価者パラメータ d_{rk} がカテゴリ k の基準を定めている．

4.2 評価者とルーブリックの特性を考慮した IRT モデル

前節で紹介したモデルでは，課題と評価者の特性を考慮した能力測定を行うことができる．他方で，ルーブリック評価では一般に複数の評価観点に基づいて採点を行うため，能力測定精度は課題や評価者の特性だけでなく，ルーブリックの評価観点の特性にも依存する．評価観点の特性を考慮した IRT モデルとして，八木・宇都 [12] は，Uto and Ueno のモデル [15] の課題パラメータを評価観点の特性パラメータとみなし，能力尺度を多次元に拡張した IRT モデルを提案している．能力の 1 次元性を仮定した場合，このモデルは，学習者 j のパフォーマンスに対して評価者 r が評価観点 c について評点 k を与える確率 P_{jrck} を次式で定義する．

$$P_{jrck} = P^*_{jrck-1} - P^*_{jrck}, \tag{8}$$

P^*_{jrck} は，学習者 j のパフォーマンスに対して評価者 r が評価観点 c について k より大きい評点を与える確率を表し，次式で定義される．

$$\begin{cases} P^*_{jrck} = [1 + \exp(-D\alpha_c\alpha_r(\theta_j - b_{ck} - \varepsilon_r))]^{-1} \\ P^*_{jrc0} = 1,\ \ P^*_{jrcK} = 0 \end{cases}$$

ここで，α_c は評価観点 c の識別力を表し，b_{ck}（$b_{c1} < b_{c2} < \cdots < b_{cK-1}$）は評価観点 c において評点 k より大きい評点を得る困難度を表す．モデルの識別性のために $\prod_{r=1}^{R} \alpha_r = 1$，$\sum_{r=1}^{R} \varepsilon_r = 0$ を仮定する．このモデルでは，各評価カテゴリに対する評価基準は b_{ck} により表現されており，評価観点に依存して決まると仮定している．

評価者と評価観点の特性を考慮したモデルとしては，Hua and Wind [13] のモデルも知られている．このモデルは，MFRM の課題パラメータを評価観点パラメータとみなしたモデルであり，式（8）の下位モデルとみなせる．

4.3 既存モデルの問題点

上述した IRT モデルを利用することで，素点平均などの単純な手法と比べて高精度な能力測定が実現できる．しかし，既存モデルを 2. で定義したルーブリック評価データに適用する場合，以下の問題が残る．

（1） ルーブリック評価データは学習者 × 課題 × 評価者 × 評価観点の 4 相データとなるが，既存モデルは 3 相データへの適用のみを想定している．したがって，ルーブリック評価の 4 相データには直接には適用できず，課題・評価者・評価観点の特性を同時に考慮した能力測定も実現できない．

（2） 既存モデルでは，各評価カテゴリに対する評価基準が課題・評価者・評価観点のいずれか一つの要因のみに依存すると仮定する．各カテゴリに対する評価基準はルーブリックの評価観点ごとに定義され，理想的には評価観点の特性のみで決まると仮定できる．しかし，現実には評価者ごとに評価基準の解釈が異なることが多いため [4]〜[6]，各カテゴリに対する評価基準は評価観点だけでなく評価者の特性にも依存する．

以上の問題を解決するために，本研究では，ルーブリック評価の 4 相データに適用でき，評価観点と評価者の評価基準を考慮できる IRT モデルを提案する．

5. 提案モデル

提案モデルでは，課題 i に対する学習者 j のパフォーマンスに，評価者 r が評価観点 c に基づいて評点 k を与える確率 P_{ijrck} を次式で定義する．

$$P_{ijrck} =$$

$$\frac{\exp\sum_{m=1}^{k}\left[D\alpha_i\alpha_r\alpha_c(\theta_j-\beta_i-\beta_r-\beta_c-\tau_r d_{cm})\right]}{\sum_{l=1}^{K}\exp\sum_{m=1}^{l}\left[D\alpha_i\alpha_r\alpha_c(\theta_j-\beta_i-\beta_r-\beta_c-\tau_r d_{cm})\right]} \tag{9}$$

ここで，β_c は評価観点 c の困難度を表すパラメータである．d_{ck} は評価観点 c において評点 k を得る困難度を表すステップパラメータであり，各カテゴリに対する評価基準を表現する．また，τ_r は評価者 r が評価カテゴリの適用範囲をどの程度広く解釈しているかを表すパラメータである．

提案モデルでは，各カテゴリ k に対する評価基準を評価者パラメータ τ_r と評価観点パラメータ d_{ck} の積 $\tau_r d_{ck}$ で表現している点に特徴がある．これにより評価観点と評価者の双方の評価基準を考慮できるため，従来モデルと比べてデータへの当てはまりが改善され，能力測定精度が向上すると期待できる．

5.1 モデルの識別性

提案モデルは，パラメータの値を一意に決定できない識別不能の問題を有する．IRT では，能力値 θ_j が標準正規分布に従うと仮定することが一般的であり，これにより θ_j は識別可能となる [29]．また，ステップパラメータ d_{ck} は，GPCM やその拡張モデルと同様に，$d_{c1}=0$，$\sum_{k=2}^{K} d_{ck}=0:\forall c$ と制約することで識別可能となる．しかし，提案モデルでは，これらの制約を課しても，$\alpha_i\alpha_r\alpha_c$ と $\tau_r d_{ck}$，$-\beta_i-\beta_r-\beta_c$ の各項において識別不能の問題が残る．

$-\beta_i-\beta_r-\beta_c$ の項については，例えば，任意の定数 h を用いて β_i と β_r を β_i+h，β_r-h と線形変換しても反応確率が不変であることから，識別不能であることがわかる．このような識別性の問題は，パラメータの平均値に制約を課すことで解消できる [29], [30]．ここでは，β_i，β_r，β_c の 3 つのパラメータのうち 2 つを制約すれば識別可能となるため，本研究では，$\sum_{i=1}^{I}\beta_i=0$，$\sum_{c=1}^{C}\beta_c=0$ と制約する．

$\alpha_i\alpha_r\alpha_c$ と $\tau_r d_{ck}$ の項については，例えば，α_i と α_r を任意の定数 h を用いて $\alpha_i h$，$\frac{\alpha_r}{h}$ としても反応確率が不変であることから識別不能であることがわかる．このような識別性の問題は，パラメータの積に制約を与えることで解消できる [29], [30]．$\alpha_i\alpha_r\alpha_c$ の項については，2 つのパラメータを制約すれば識別可能となるため，ここでは，$\prod_{i=1}^{I}\alpha_i=1$，$\prod_{c=1}^{C}\alpha_c=1$ と制約する．$\tau_r d_{ck}$ については，一方のパラメータに制約を課せばよいため，$\prod_{r=1}^{R}\tau_r=1$ と制約する．

5.2 パラメータの解釈

本節では，提案モデルの評価観点パラメータと評価者パラメータの解釈について説明する．このために，評価カテゴリ数 $K=4$ において，表 2 のパラメータを所与としたときの，提案モデルの項目反応曲線（ICC: Item Characteristic Curve）を図 1 に示す．なお，表 2 では，パラメータの意味が理解しやすい例を示すために，モデルの識別性の条件式を必ずしも満たさないパラメータ値を用いているが，条件式を満たす値でも解釈は同様である．各図は，横軸が学習者の能力 θ_j を表し，縦軸が各評点への反応確率 P_{ijrck} を表す．図 1 から，いずれの ICC においても，能力が低いほど低い評点を得る確率が高くなり，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．

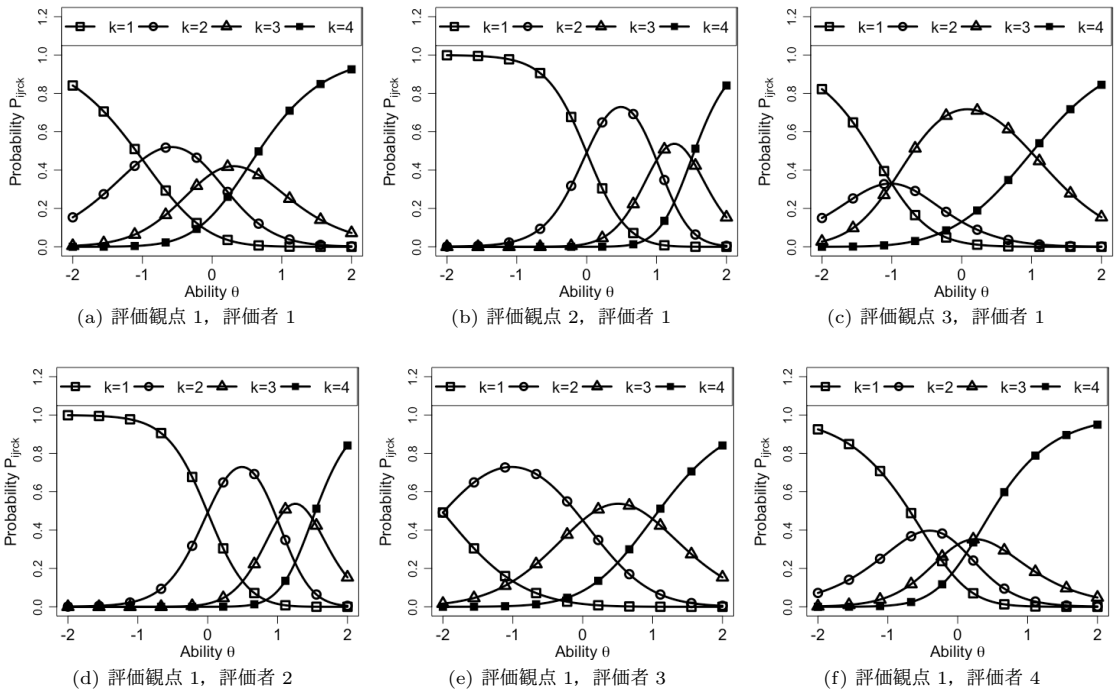
表 2: 図 1 で使用したパラメータ

	α_c	β_c	d_{c1}	d_{c2}	d_{c3}	d_{c4}
評価観点 1	1.0	0.0	0.0	-1.0	0.0	0.5
評価観点 2	2.0	1.0	0.0	-1.0	0.0	0.5
評価観点 3	1.0	0.0	0.0	-1.0	-1.0	1.0
	α_r	β_r	τ_r			
評価者 1	1.0	0.0	1.0			
評価者 2	2.0	1.0	1.0			
評価者 3	1.0	0.0	2.0			
評価者 4	1.0	0.0	0.5			

ここで，図 1 の (a)〜(c) は評価者パラメータ一定のもとで評価観点パラメータを変更した場合に対応し，(a) と (d)〜(f) は評価観点パラメータ一定のもとで評価者パラメータを変更した場合に対応する．まず，評価観点特性の解釈を説明するために，(a) を基準に (b) と (c) を比較する．

(b) は (a) から評価観点の識別力と困難度のパラメータ値を大きくした場合の ICC である．(b) の ICC では，能力値が変化したときの反応確率の変動が大きくなっていることがわかる．これは，識別力の高い評価観点では，能力値に応じた評点が与えられやすく，同等の能力の学習者には同一の評点が安定して与えられることを表現している．また，(b) では ICC が全体として右に移動していることが確認できる．これは，困難度の高い評価観点では，高い評点を得るためにより高い能力が必要であることを表現している．

(c) は (a) から，ステップパラメータ d_{ck} の値を変化させた場合の ICC である．このパラメータは，隣接する値 $d_{ck+1}-d_{ck}$ の差が大きくなるほど，評点 k と評点 $k+1$ の基準の乖離が大きいことを意味する．ICC 上では，評点 k への反応確率を能力尺度の広い範囲で高くすることでこの特性が表現される．例えば，$d_{c4}-d_{c3}$ が大きく，$d_{c3}-d_{c2}$ が小さい (c) の ICC は，



(a) 評価観点 1，評価者 1　(b) 評価観点 2，評価者 1　(c) 評価観点 3，評価者 1

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

図 1: 表 2 のパラメータを適用した場合の項目反応曲線

(a) のそれと比べて，評点 3 への反応確率が高くなる能力値の範囲を広く，評点 2 の確率が高くなる範囲を狭く表現している．提案モデルでは，このように各カテゴリ k に対する評価基準を評価観点ごとに表現する．

次に，評価者特性の解釈について説明するために，(a) を基準に (d)〜(f) を比較する．

(d) は (a) から評価者の一貫性と厳しさのパラメータ値を大きくした場合の ICC である．(d) の ICC では，能力の変動に対する反応確率の変化が大きくなっている．これは，一貫性の高い評価者は，学習者の能力と相関した評点を与えるとともに，同等の能力の学習者に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．また，(d) の分布は全体として右に移動しており，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が表現されている．

(e) と (f) は，パラメータ τ_r が (a) と比べて大きい場合と小さい場合に対応する．τ_r が大きいほど，K 段階カテゴリの平均値 $\frac{K}{2}$ 付近の評価カテゴリへの反応確率が，能力尺度の広い範囲で高く表現される．例えば，τ_r が大きい (e) の ICC では，τ_r が小さい (f) の ICC と比べて，4 段階カテゴリの平均値付近にあたる評点 2 と 3 への反応確率が能力尺度の広い範囲で高くなっている．これは (e) の評価者が評点 2 と 3 の適用範囲を (f) の評価者よりも広く解釈していることを表現している．提案モデルでは，このように評価カテゴリに対する評価者ごとの基準の差異を表現する．

課題の困難度と識別力については，評価観点の困難度と識別力と類似した解釈が可能である．課題特性の詳細な解釈については，関連論文 [10], [11] が詳しい．

5.3 パラメータ推定手法

本節では提案モデルのパラメータ推定法について述べる．IRT の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尤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 [31]．一方で，本研究で扱うような複雑な IRT 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）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知られている [15], [29]．IRT 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（Gibbs/MH）[15], [27], [28] が利用されてきた．この

表 3: パラメータ・リカバリ実験の結果

				RMSE										平均バイアス									
J	I	R	C	θ_j	α_i	α_r	α_c	β_i	β_r	β_c	τ_r	d_{ck}	$Avg.$	θ_j	α_i	α_r	α_c	β_i	β_r	β_c	τ_r	d_{ck}	$Avg.$
30	3	5	5	0.22	0.04	0.18	0.12	0.08	0.08	0.06	0.09	0.23	0.12	0.01	0.00	0.11	0.00	0.00	0.01	0.00	0.00	0.00	0.01
			10	0.15	0.04	0.12	0.10	0.10	0.05	0.05	0.03	0.15	0.09	0.00	0.00	0.10	-0.01	0.00	-0.01	0.00	0.01	0.00	0.01
		10	5	0.12	0.02	0.21	0.06	0.01	0.10	0.03	0.14	0.14	0.09	-0.02	0.00	0.10	0.01	0.00	-0.01	0.00	0.02	0.00	0.01
			10	0.10	0.02	0.09	0.07	0.08	0.06	0.05	0.11	0.12	0.08	-0.01	0.00	0.07	-0.01	0.00	-0.01	0.00	0.00	0.00	0.00
	5	5	5	0.10	0.09	0.05	0.09	0.03	0.03	0.01	0.13	0.15	0.08	0.00	0.01	0.00	0.02	0.00	-0.01	0.00	0.02	0.00	0.00
			10	0.11	0.07	0.09	0.09	0.06	0.12	0.04	0.06	0.15	0.09	0.03	0.00	0.09	0.00	0.00	0.02	0.00	0.00	0.00	0.02
		10	5	0.09	0.03	0.07	0.03	0.02	0.07	0.03	0.05	0.05	0.05	0.01	0.00	-0.03	0.00	0.00	0.02	0.00	0.02	0.00	0.00
			10	0.10	0.03	0.06	0.03	0.04	0.09	0.01	0.05	0.07	0.05	0.03	0.00	0.06	0.00	0.00	0.02	0.00	0.00	0.00	0.01
50	3	5	5	0.14	0.02	0.08	0.12	0.09	0.09	0.01	0.08	0.21	0.09	0.00	0.00	0.02	0.02	0.00	0.00	0.00	0.00	0.00	0.00
			10	0.14	0.05	0.09	0.12	0.09	0.06	0.04	0.10	0.13	0.09	0.00	0.00	0.09	-0.02	0.00	0.01	0.00	0.02	0.00	0.01
		10	5	0.10	0.03	0.13	0.06	0.06	0.09	0.04	0.10	0.11	0.08	0.01	-0.01	0.05	0.00	0.00	-0.01	0.00	0.02	0.00	0.01
			10	0.10	0.03	0.09	0.03	0.08	0.07	0.03	0.05	0.10	0.06	-0.01	0.00	0.09	0.00	0.00	-0.01	0.00	0.02	0.00	0.01
	5	5	5	0.12	0.06	0.13	0.04	0.09	0.10	0.04	0.10	0.11	0.09	0.01	0.00	-0.12	0.00	0.00	0.00	0.00	0.01	0.00	-0.01
			10	0.13	0.04	0.11	0.06	0.02	0.03	0.03	0.06	0.07	0.06	-0.03	-0.01	-0.09	0.01	0.00	0.00	0.00	0.03	0.00	-0.01
		10	5	0.12	0.01	0.08	0.04	0.07	0.08	0.01	0.07	0.08	0.06	0.01	0.00	0.07	0.00	0.00	-0.01	0.00	0.00	0.00	0.01
			10	0.04	0.05	0.05	0.04	0.02	0.03	0.02	0.03	0.08	0.04	0.01	-0.01	-0.02	0.00	0.00	0.02	0.00	0.00	0.00	0.00

アルゴリズムは，単純で実装が容易である反面，目標分布への収束が遅いという問題がある [32], [33].

Gibbs/MH より効率の良い MCMC アルゴリズムとして，ハミルトニアンモンテカルロ法（HMC）が知られている [34]．HMC では，ステップサイズとシミュレーション長という二つの決定変数を適切に選択することで，自己相関の低い良質なサンプルを得ることができ，高速に目標分布に収束することが知られている [32], [35]．近年では，HMC の決定変数を，サンプリングの過程で最適化できる No-U-Trun Sampler（NUT）[32] と呼ばれる手法が提案されている．NUT による MCMC は，Stan [36] と呼ばれるライブラリの整備により，様々な数理モデルに容易に適用できるようになったため，IRT を含む様々な統計・機械学習モデルの推定に近年広く利用されている [37]〜[39].

以上より，本研究では提案モデルのパラメータ推定手法として Stan を用いた NUT による MCMC 法を用いる．実装は RStan [40] を用いて行なった．提案モデルの Stan コードは付録に示した．パラメータの事前分布は θ_j, β_i, β_r, β_c, d_{ck}, $\log \alpha_i$, $\log \alpha_r$, $\log \alpha_c$, $\log \tau_r \sim N(0.0, 1.0^2)$ とした．ここで，$N(\mu, \sigma^2)$ は平均 μ，標準偏差 σ の正規分布を表す．本研究では，MCMC のバーンイン期間は 500 とし，$500 \sim 1,000$ 時点までの 500 サンプルを用いる．独立した MCMC を 3 チェイン実行し，得られたサンプルの期待値として EAP 推定値を求める．

5.4 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．実験手順は以下の通りである．

（1） モデルパラメータの真値を 5.3 節に示した分布に従ってランダムに生成した．

（2） 手順 (1) で生成したパラメータを所与として，データ $\mathbf{X}$ を式 (9) の提案モデルから生成した．

（3） 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．

（4） 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: Root Mean Square Error）とバイアスを算出した．

（5） 以上を 30 回行い，RMSE とバイアスの平均と標準偏差を求めた．

上記の実験を，学習者数 $J = 30, 50$, 課題数 $I = 3, 5$, 評価者数 $R = 5, 10$，評価観点数 $C = 5, 10$ の場合において行った．カテゴリ数は $K = 4$ とした．これらの実験条件は次章で行う実データ実験の規模と同程度となるように選定した．

実験結果を表 3 に示す．表 3 から，全パラメータの RMSE の平均値（$Avg.$ 列）は 0.1 程度となり，パラメータ別の最大値でも 0.2 程度にとどまっていることがわかる．誤差 0.1 や 0.2 という値は，標準正規分布に従うサンプルの 99.73% が含まれる範囲（$-3 \sim 3$）の 1.7% と 3.3% に相当し，十分に小さい値と解釈できる．また，関連研究 (e.g., [11], [12], [15]) と同様に，学習者数・課題数・評価者数・評価観点数の増加に伴い推定精度が改善する傾向も読み取れる．バイアスの平均については，いずれのパラメータも 0 に非常に近

表 4: パラメータ推定例

評価観点	$c=1$	$c=2$	$c=3$	$c=4$	$c=5$
α_c	1.130	0.916	1.210	0.810	0.986
β_c	-0.541	-0.143	0.052	0.750	-0.117
d_{c2}	-2.336	-1.383	-2.041	-1.496	-2.408
d_{c3}	-0.048	-0.438	0.041	-0.094	-0.383
d_{c4}	2.385	1.821	2.001	1.591	2.791
評価者	$r=1$	$r=2$	$r=3$	$r=4$	$r=5$
α_r	0.821	0.469	0.698	0.597	0.816
β_r	-0.489	-1.211	-0.228	-1.498	-0.232
τ_r	0.882	1.520	0.459	1.033	0.578
評価者	$r=6$	$r=7$	$r=8$	$r=9$	$r=10$
α_r	0.752	0.847	0.992	0.254	0.780
β_r	-0.295	0.053	-0.315	-0.236	1.360
τ_r	0.815	0.658	0.731	2.020	3.438
課題	$i=1$	$i=2$	$i=3$	$i=4$	
α_i	0.841	1.018	1.086	1.075	
β_i	-0.031	-0.054	0.075	0.009	

い値を示しており，系統的な過大（または過少）推定の傾向もないことが確認できる．また，MCMC の収束を示す Gelman-Rubin の収束判定指標 $\hat{R}$ [41], [42] を確認したところ，すべての場合で一般的な収束基準値である 1.1 を下回っていた．

以上の結果から，MCMC 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．

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 4 つの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テーマに対する自身の意見を述べる内容であり，専門知識や客観データは必要としない．また，評価者による採点は，松下ら [3] が開発した表 1 のルーブリックを用いて 4 段階で行わせた．

本研究では，この実データに提案モデルを適用し，モデルの有効性を評価する．ここで，表 4 に実データから求めた提案モデルのパラメータ推定例を示す．表 4 から，課題・評価者・評価観点についてそれぞれ特性差があることが読み取れる．また，評価観点間で d_{ck} の傾向が異なり，評価者間で τ_r が異なることも確認できる．これは各カテゴリ k に対する評価基準が評価者と評価観点のそれぞれに依存していることを示唆する．

6.1 比較モデル

以降では，提案モデルの性能を評価するために，4. で紹介した最先端モデルとの性能比較を行う．以降では簡単のために，式 (6)，式 (7)，式 (8) のモデルをそれぞれ，r-GRM，r-GPCM，r-MGRM と呼ぶ．ただし，これらの既存モデルは 4 相データに直接には適用できないため，本実験では次の二つの方法で実データに適用する．

（1） **4 相データを 3 相データに変換**

一つ目の方法は，従来モデルに適用できるように 4 相データを 3 相データに変換する方法である．具体的には，ルーブリック評価の 4 相データ $\mathbf{X}$ を，学習者 × 課題 × 評価者の 3 相データ $\mathbf{X}' = \{mode(\{x_{ijrc}|c \in \mathcal{C}\})|i \in \mathcal{I}, j \in \mathcal{J}, r \in \mathcal{R}\}$ に変換して r-GRM と r-GPCM を適用する．ここで，$mode(\mathbf{S})$ は集合 $\mathbf{S}$ の最頻値を返す関数とする．同様に，4 相データ $\mathbf{X}$ を学習者 × 評価者 × 評価観点の 3 相データ $\mathbf{X}'' = \{mode(\{x_{ijrc}|i \in \mathcal{I}\})|j \in \mathcal{J}, r \in \mathcal{R}, c \in \mathcal{C}\}$ に変換して r-MGRM を適用する．

（2） **4 相データに適用できるモデル定義に拡張**

二つ目の方法は，4 相データ $x_{ijrc} \in \mathbf{X}$ に対して反応確率が定義されるようにモデル式を変更する方法である．具体的には，r-GRM の式（6）と r-GPCM の式（7）の左辺を P_{ijrk} から P_{ijrck} に変更し，r-MGRM の式（8）の左辺を P_{jrck} から P_{ijrck} に変更する．これは，r-GRM と r-GPCM では，評価観点 c にかからわず，式（6）右辺および式（7）右辺で反応確率を計算し，r-MGRM ではいずれの課題 i についても，式（8）右辺で反応確率を計算することを意味する．

また，提案モデルで新たに導入した評価者パラメータ τ_r の効果を確認するために，提案モデルから τ_r を取り除いたモデルとの比較も行う．以降ではこのモデルを「提案 w/o τ_r」と呼ぶ．

6.2 情報量規準によるモデル比較

本節では，情報量規準に基づくモデル比較により提案モデルの性能を評価する．ここでは，MCMC により各モデルのパラメータを推定し，得られた推定値を用いて情報量規準を求めた．情報量規準には MCMC のパラメータサンプルから算出できる Widely Applicable Information Criterion (WAIC) [43] と近似対数周辺尤度 (ML) [44] を用いた．ここで，WAIC は汎化誤差を最小化するモデルを選択する手法であり，選択されたモデルは将来のデータの予測に優れたモデルと解釈できる．ML は一致性を持つモデル選択規準であり，漸近的に真のモデルが選択される．なお，WAIC と解釈を合わせるために，ML の値には -2 をかけた値

表 5: 情報量規準と能力測定精度

			3 相データ			4 相データ					
			r-GRM	r-GPCM	r-MGRM	r-GRM	r-GPCM	r-MGRM	提案 w/o τ_r	提案モデル	素点平均
情報量規準	WAIC		2278.18	2255.39	2837.81	14112.59	13982.13	13279.27	13286.24	**13033.21**	-
	ML		2211.41	2183.64	2766.27	14044.36	13907.34	13208.19	13207.10	**12943.06**	-
能力測定精度	平均		0.350	0.362	0.317	0.369	0.396	0.395	0.390	**0.450**	0.347
	標準偏差		0.137	0.144	0.157	0.145	0.136	0.136	0.135	0.130	0.147
	3 相データ	r-GPCM	$p < .001$	-	-	-	-	-	-	-	-
		r-MGRM	$p < .001$	$p < .001$	-	-	-	-	-	-	-
	4 相データ	r-GRM	$p < .001$	$p = .141$	$p < .001$	-	-	-	-	-	-
		r-GPCM	$p < .001$	$p < .001$	$p < .001$	$p < .001$	-	-	-	-	-
		r-MGRM	$p < .001$	$p < .001$	$p < .001$	$p < .001$	$p = .999$	-	-	-	-
		提案 w/o τ_r	$p < .001$	$p < .001$	$p < .001$	$p < .001$	$p = .501$	$p = .869$	-	-	-
		提案モデル	$p < .001$	$p < .001$	$p < .001$	$p < .001$	$p < .001$	$p < .001$	$p < .001$	-	-
		素点平均	$p = .993$	$p < .001$	$p < .001$	$p < .001$	$p < .001$	$p < .001$	$p < .001$	$p < .001$	-

を出力した．以上の情報量規準は値が小さいモデルほど，適切なモデルであることを意味する．

実験結果を表 5 に示す．表では，「3 相データ」の列が実データを 3 相データに変換して既存モデルを適用した場合の結果を表し，「4 相データ」の列は 4 相データに適用可能な定義に拡張した既存モデルと提案モデルの結果を表す．情報量規準値は同一のデータセットに適用した場合にのみモデル比較が可能となるため，3 相データ適用では r-GRM と r-GPCM のみ比較可能であり，4 相データ適用時と 3 相データ適用時の結果は比較できないことに注意してほしい．

表 5 から，4 相データに適用した場合，どちらの情報量規準においても，提案モデルが最適モデルとして選択されたことが確認できる．従来モデルについて比較すると，r-MGRM，r-GPCM，r-GRM の順で性能が高いことが確認できる．4. で述べたように，r-MGRM，r-GPCM，r-GRM の主な違いは評価カテゴリの評価基準が課題・評価者・評価観点のどの要因に依存すると仮定するかにあり，この実験結果は，評価基準は評価観点と評価者に強く依存し，課題への依存は小さいことを示唆している．3 相データ適用においても，r-GPCM が r-GRM より高い性能を示しており，この結果と一致している．また，評価者の評価基準パラメータを無視した「提案 w/o τ_r」の性能が提案モデルと比べて低いことから，評価観点と評価者の両方の特性を考慮して評価基準を表現することが重要であるとわかる．

以上より，提案モデルでは，評価観点と評価者の両方について評価基準を表現できるため，データへの当てはまりが最も高くなったと解釈できる．

6.3 能力測定の精度評価

ここでは，提案モデルの能力測定精度について評価する．本論文では能力測定精度を，先行研究 (e.g., [10], [11], [15], [16]) と同様に，「評価者や課題・評価観点が変化したとき，どの程度安定して能力値を推定できるか」として評価する．具体的には以下の実験で能力測定精度を求めた．

（1） 実データを用いて MCMC によりモデルパラメータを推定した．

（2） 各学習者に与えられた評点データの 95%をランダムに欠測させたデータを 100 パタン作成した．これは，各学習者に対する評価者・課題・評価観点をランダムに変更することに対応する．

（3） 手順 (1) で推定した課題・評価者・評価観点パラメータを所与として，各欠測データから学習者の能力を推定した．この手順は，各学習者の能力値を，評価者・課題・評価観点を変更しながら推定することに対応する．

（4） n 番目の欠測データから推定された能力値 $\boldsymbol{\theta}_n$ と n' 番目の欠測データから推定された能力値 $\boldsymbol{\theta}_{n'}$ との相関係数 $Cor(\boldsymbol{\theta}_n, \boldsymbol{\theta}_{n'})$ を $n \in \{1, \ldots, 100\}$, $n' \in \{n+1, \ldots, 100\}$ の全ての組み合わせについて求め，相関の平均と標準偏差を算出した．この相関は，評価者・課題・評価観点が変化しても安定して能力を推定できるほど高い値を示すため，本実験ではこれを能力測定精度の指標と解釈する．

本実験では，比較のために，素点の平均値を能力推定値とみなした場合についても同様の実験を行なった．また，相関係数の平均値にモデル間で有意な差があるかを確認するために，Tukey 法による多重比較を行った．

実験結果を表 5 に示す．まず既存モデルについて，3 相データに適用した場合と 4 相データに適用した場合を比較すると，すべての場合で 4 相データ適用時に精度が向上していることが確認できる．特に r-MGRM ではその効果が顕著であることが読み取れる．これは，4 相データを 3 相データにする際に，r-GRM と r-GPCM では評価観点相の削減によりデータ数が 1/5 になるのに対し，r-MGRM では評価者相の削減によりデータ数が 1/10 になるため，3 相データと 4 相データでの評点データ数の差異が大きかったことが理由と考えられる．この結果は，4 相データを用いて能力を測定することの有効性を示している．

次に 4 相データに適用した場合で比較を行うと，提案モデルが最も高い能力測定精度を示したことがわかる．また，従来モデルについて比較すると，r-MGRM と r-GPCM の精度が同程度であり，r-GRM は精度が悪い．6. 2 でも議論したように，これらのモデルの主な違いは各評価カテゴリの評価基準が課題・評価者・評価観点のどの要因に依存すると仮定するかであり，この結果は，評価観点と評価者について評価基準の差異を考慮することで能力測定精度が改善されることを示している．また，前節の実験と同様に，「提案 w/o τ_r」は提案モデルと比べて性能が低いことが確認できる．このことは，評価観点と評価者の両方の評価基準を考慮することが能力測定精度の改善に有効であることを意味する．

以上の実験から，提案モデルでは，評価観点と評価者について評価基準の差異を表現でき，4 相データから学習者の能力を測定できるため，能力測定の精度を向上できたことが確認できた．

7. むすび

本研究では，ルーブリック評価で得られる学習者 × 課題 × 評価者 × 評価観点の 4 相データから，評価観点と評価者の評価基準を考慮して学習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．また，提案モデルのパラメータ推定手法として，Stan を用いた No-U-turn sampler による MCMC アルゴリズムを提案し，シミュレーション実験によりパラメータ推定の妥当性を示した．さらに，実データを用いた実験では，提案モデルの特徴である，1）4 相データから能力を測定できること，2）評価観点と評価者の評価基準を考慮できることと，の二点がデータ適合と能力測定精度の向上に有効であることを示した．

今後は，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，本モデルで推定されるパラメータを分析することで，ルーブリック自体の分析・評価を行うこともできる．本モデルをルーブリックの作成・改善に活用することで，より良いルーブリックの開発を行いたい．さらに，本研究では測定対象の能力尺度に 1 次元性を仮定したが，今後は多次元尺度への拡張を行い，幅広い活用方法を検討したい．

謝辞

本研究は JSPS 科研費 17H04726，17K20024，19H05663，19K21751 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] O. Zlatkin-Troitschanskaia, R.J. Shavelson, S. Schmidt, and K. Beck, “On the complementarity of holistic and analytic approaches to performance assessment scoring,” British Journal of Educational Psychology, 2019.

[3] 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.

[4] 尚徳西谷，“文章力養成のためのルーブリック活用の教育的意義の検討–授業実践から見る教育手法–,” 京都大学高等教育研究，vol.23，pp.25–35，2017.

[5] S.M. Brookhart and F. Chen, “The quality and effectiveness of descriptive rubrics,” Educational Review, vol.67, no.3, pp.343–368, 2015.

[6] A.A. Rahman, J. Ahmad, R.M. Yasin, and N.M. Hanafi, “Investigating central tendency in competency assessment of design electronic circuit: Analysis using many facet Rasch measurement (MFRM),” International Journal of Information and Education Technology, vol.7, no.7, pp.525–528, 2017.

[7] 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.

[8] 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.

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

[10] 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.

[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] 八木嵩大，宇都雅輝，"パフォーマンス評価における多次元項目反応モデル，" 電子情報通信学会論文誌 D，vol.102，no.10，pp.708–720，2019.
[13] C. Hua and S.A. Wind, "Exploring the psychometric properties of the mind-map scoring rubric," Behaviormetrika, vol.46, no.1, pp.73–99, 2019.
[14] J.M. Linacre, Many-faceted Rasch Measurement, MESA Press, 1989.
[15] M. Uto and M. Ueno, "Item response theory for peer assessment," IEEE Transactions on Learning Technologies, vol.9, no.2, pp.157–170, 2016.
[16] 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，" 電子情報通信学会論文誌.D，vol.101，no.1，pp.211–224，2018.
[17] 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.
[18] 宇都雅輝，"論述式試験における評点データと文章情報を活用した項目反応トピックモデル，" 電子情報通信学会論文誌 D，vol.102，no.8，pp.553–566，2019.
[19] M. Uto, "Rater-effect IRT model integrating supervised LDA for accurate measurement of essay writing ability," Proc. International Conference on Artificial Intelligence in Education, pp.494–506, 2019.
[20] F.M. Lord, Applications of item response theory to practical testing problems, Erlbaum Associates, 1980.
[21] M. Matteucci and L. Stracqualursi, "Student assessment via graded response model," Statistica, vol.66, pp.435–447, 2006.
[22] 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.
[23] F. Samejima, "Estimation of latent ability using a response pattern of graded scores," Psychometrika Monography, vol.17, pp.1–100, 1969.
[24] E. Muraki, "A generalized partial credit model," Handbook of Modern Item Response Theory, eds. by W.J. van derLinden and R.K. Hambleton, pp.153–164, Springer, 1997.
[25] D. Andrich, "A rating formulation for ordered response categories," Psychometrika, vol.43, no.4, pp.561–573, 1978.
[26] G. Masters, "A Rasch model for partial credit scoring," Psychometrika, vol.47, no.2, pp.149–174, 1982.
[27] 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.
[28] 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.
[29] J.-P. Fox, Bayesian item response modeling: Theory and applications, Springer, 2010.
[30] M. Uto, D. Nguyen, and M. Ueno, "Group optimization to maximize peer assessment accuracy using item response theory and integer programming," IEEE Transactions on Learning Technologies (in press).
[31] F.B. Baker and S.H. Kim, Item Response Theory: Parameter Estimation Techniques, Statistics, textbooks and monographs, Marcel Dekker, 2004.
[32] M.D. Hoffman and A. Gelman, "The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo," Journal of Machine Learning Research, vol.15, pp.1593–1623, 2014.
[33] M. Girolami and B. Calderhead, "Riemann manifold Langevin and Hamiltonian Monte Carlo methods," Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol.73, no.2, pp.123–214, 2011.
[34] S. Brooks, A. Gelman, G. Jones, and X.L. Meng, Handbook of Markov Chain Monte Carlo, Chapman & Hall/ CRC Handbooks of Modern Statistical Methods, CRC Press, 2011.
[35] R.M. Neal, "MCMC using Hamiltonian dynamics," Handbook of Markov Chain Monte Carlo, vol.54, pp.113–162, 2010.
[36] B. Carpenter, A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell, "Stan: A probabilistic programming language," Journal of Statistical Software, Articles, vol.76, no.1, pp.1–32, 2017.
[37] Y. Luo and H. Jiao, "Using the Stan program for Bayesian item response theory," Educational and Psychological Measurement, vol.78, no.3, pp.384–408, 2018.
[38] Z. Jiang and R. Carter, "Using Hamiltonian Monte Carlo to estimate the log-linear cognitive diagnosis model via Stan," Behavior Research Methods, vol.51, no.2, pp.651–662, 2019.
[39] 松浦健太郎，Stan と R でベイズ統計モデリング，共立出版，2016.
[40] Stan Development Team, "RStan: the R interface to stan. R package version 2.17.3.," http://mc-stan.org, 2018.
[41] A. Gelman and D.B. Rubin, "Inference from iterative simulation using multiple sequences," Statist. Sci., vol.7, no.4, pp.457–472, 1992.
[42] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin, Bayesian Data Analysis, Third Edition, Chapman & Hall/CRC Texts in Statistical Science, Taylor & Francis, 2013.
[43] S. Watanabe, "Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory," Journal of Machine

Learning Research, pp.3571–3594, 2010.

[44] M. Newton and A.E. Raftery, “Approximate Bayesian inference by the weighted likelihood bootstrap,” Journal of the Royal Statistical Society. Series B: Methodological, vol.56, no.1, pp.3–48, 1994.

付　　録

```
data{
  int <lower=0> J;
  int <lower=0> I;
  int <lower=0> R;
  int <lower=0> C;
  int <lower=2> K;
  int <lower=0> N;//n_samples
  int <lower=1, upper=J> ExamineeID [N];
  int <lower=1, upper=I> ItemID [N];
  int <lower=1, upper=R> RaterID [N];
  int <lower=1, upper=C> RubricID [N];
  int <lower=1, upper=K> X [N];
}
transformed data{
  vector[K] c = cumulative_sum(rep_vector(1, K)) - 1;
}
parameters {
  vector[J] theta;
  real<lower=0> alpha_i [I-1];
  real<lower=0> alpha_r [R];
  real<lower=0> alpha_c [C-1];
  vector[I-1] beta_i;
  vector[R] beta_r;
  vector[C-1] beta_c;
  vector[K-2] beta_ck [C];
  real<lower=0> tau_r [R-1];
}
transformed parameters{
  real<lower=0> trans_alpha_i[I];
  real<lower=0> trans_alpha_c[C];
  vector[I] trans_beta_i;
  vector[C] trans_beta_c;
  real<lower=0> trans_tau_r[R];
  vector[K-1] category_est[C];
  vector[K] category_prm[C, R];
  trans_alpha_i[1] = 1.0 / prod(alpha_i);
  trans_alpha_c[1] = 1.0 / prod(alpha_c);
  trans_beta_i[1] = -1*sum(beta_i);
  trans_beta_c[1] = -1*sum(beta_c);
  trans_tau_r[1] = 1.0 / prod(tau_r);
  trans_alpha_i[2:I] = alpha_i;
  trans_alpha_c[2:C] = alpha_c;
  trans_beta_i[2:I] = beta_i;
  trans_beta_c[2:C] = beta_c;
  trans_tau_r[2:R] = tau_r;
  for(p in 1:C){
    category_est[p, 1:(K-2)] = beta_ck [p];
    category_est[p, K-1] = -1*sum(beta_ck [p]);
    for(r in 1:R){
      category_prm[p, r] = cumulative_sum(append_row
          (0, trans_tau_r[r]*category_est[p]));
    }
  }
}
model{
  trans_alpha_i ~ lognormal(0, 1);
  alpha_r ~ lognormal(0, 1);
  trans_alpha_c ~ lognormal(0, 1);
  trans_beta_i ~ normal(0, 1);
  beta_r ~ normal(0, 1);
  trans_beta_c ~ normal(0, 1);
  trans_tau_r ~ lognormal(0, 1);
  theta ~ normal(0, 1);
  for (p in 1:C) category_est [p,] ~ normal(0, 1);
  for (n in 1:N){
    X[n] ~ categorical_logit(1.7 *trans_alpha_i[ItemID[n
        ]]*alpha_r[RaterID[n]]*trans_alpha_c[RubricID[
        n]]*(c*(theta[ExamineeID[n]]-trans_beta_i[
        ItemID[n]]-beta_r[RaterID[n]]-trans_beta_c[
        RubricID[n]])-category_prm[RubricID[n],
        RaterID[n]]));
  }
}
generated quantities {
  vector[N] log_lik;
  for (n in 1:N){
    log_lik[n] = categorical_logit_log(X[n], 1.7 *
        trans_alpha_i[ItemID[n]]*alpha_r[RaterID[n]]*
        trans_alpha_c[RubricID[n]]*(c*(theta[
        ExamineeID[n]]-trans_beta_i[ItemID[n]]-
        beta_r[RaterID[n]]-trans_beta_c[RubricID[n]])
        -category_prm[RubricID[n], RaterID[n]]));
  }
}
```

（平成 xx 年 xx 月 xx 日受付）

宇都雅輝

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修了．博士 (工学)．長岡技術科学大学特任助教を経て，2015 年に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．

植野真臣

1994 年東京工業大学大学院総合理工学研究科修了．博士（工学），東京工業大学，千葉大学，長岡技術科学大学を経て，2006 年より電気通信大学勤務，同大学教授に着任，現在に至る．人工知能，e テスティング，e ラーニング，ベイズ統計，ベイジアンネットワークなどの研究に従事．

Abstract Rubric-based assessment has been attracted much attention as a method to measure higher-order abilities of learners. A persistent difficulty of the assessment is that ability measurement accuracy depends on characteristics of raters, performance tasks, and assessment criteria of a rubric. To resolve the problem, item response theory models that incorporate their characteristic parameters have been proposed. However, conventional models cannot estimate learner ability considering characteristics of all the three factors from rubric-based four-way rating data. In addition, they assume that a rating scale is determined by one of the three factors although it generally depends on both raters and assessment criteria. To resolve the problem, this study proposes a new item response theory model that incorporates rating scale parameters for raters and assessment criteria to estimate learner ability from the four-way rating data. This study demonstrates the effectiveness of the proposed model through simulation experiments and real data application.

Key words Performance assessment, rubric, item response theory, educational measurement, rater bias