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

著者	宇都 雅輝, 植野 真臣
雑誌名	電子情報通信学会論文誌. D, 情報・システム
巻	J101-D
号	1
ページ	211-224
発行年	2018-01-01
URL	http://id.nii.ac.jp/1438/00008840/

doi: 10.14923/transinfj.2017JDP7055

論　文

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

宇都　雅輝[†a)]　　植野　真臣[†]

Robust Item Response Theory Model for Aberrant Raters in Peer Assessment

Masaki UTO[†a)] and Maomi UENO[†]

あらまし　近年，MOOCs に代表される大規模 e ラーニングの普及に伴い，ピアアセスメントを学習者の能力測定に用いるニーズが高まっている．一方で，ピアアセスメントによる能力測定の課題として，その測定精度が評価者の特性に強く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．しかし，既存モデルでは，評価基準が他の評価者と極端に異なる “異質評価者” の特性を必ずしも表現できないため，異質評価者が存在する可能性があるピアアセスメントに適用したとき能力測定精度が低下する問題が残る．この問題を解決するために，本論文では，1）評価の厳しさ，2）一貫性，3）尺度範囲の制限，に対応する評価者特性パラメータを付与した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とおりである．1）評価者の特性を柔軟に表現できるため，異質評価者の採点データに対するモデルのあてはまりを改善できる．2）異質評価者の影響を正確に能力測定値に反映できるため，異質評価者が存在するピアアセスメントにおいて，既存モデルより高精度な能力測定が期待できる．本論文では，シミュレーション実験と実データ実験から提案モデルの有効性を示す．

キーワード　項目反応理論，ピアアセスメント，評価者特性，異質評価者，能力測定．

1. まえがき

近年，社会構成主義に基づく学習評価法として，学習者同士による相互評価法を表すピアアセスメント [1] が注目されている．学習場面におけるピアアセスメントの利用には次のような利点がある [2]～[7]．（1）学習者に評価者の役割を与えることで学習モチベーションが向上する．（2）他者からのフィードバックにより学習者の内省が促進される．（3）同等の立場である学習者からのフィードバックは理解しやすい．（4）他者の成果物からの学びが促される．（5）学習者同士で評価を行うため，教師の負担が軽減され，学習者数が多い場合でも評価を実施できる．（6）成人学生の場合，教師一人による採点より，複数名の学習者による採点の方が信頼性が高い．

以上のような利点を有することから，ピアアセスメントは様々な学習・評価場面で活用されてきた [6]～[14]．具体的には，学習者同士でフィードバックを与え合わせることによる学習支援ツールとして利用されることが一般的であった（e.g., [1], [4], [6], [12]～[14]）．一方，近年では，Massive Open Online Courses（MOOCs）に代表される大規模 e ラーニングの普及に伴い，ピアアセスメントを学習者の能力測定に用いるニーズが高まっている [7], [15]．学習者数が大幅に増加すると，少数の教師が全ての学習者を評価することは困難になる．しかし，ピアアセスメントでは，各学習者に数名の評価者を割り当てることで，教師や学習者の負担を大幅に増加させることなく評価を実施できる [7], [15], [16]．また，社会構成主義の考え方に基づくと，能力とは，同一コミュニティのメンバが判断するものと解釈できるため [17]，ピアアセスメントによる能力測定は妥当であるといえる．以上から，本研究では，ピアアセスメントを学習者の能力測定に用いる場合に着目する．

ピアアセスメントに基づく能力測定の課題として，学習者に与えられる評価点が評価者の特性（例えば，評価の厳しさなど）に強く依存し，これが能力測定精度の低下を引き起こす問題が指摘されてきた [2], [6], [16], [18]～[24]．この問題を解決する手法の一つとして，数理モデルを用いたテスト理論の一つであ

[†] 電気通信大学大学院情報理工学研究科，調布市
Graduate School of Informatics and Engineering, The University of Electro-Communications, 1–5–1 Chofugaoka, Chofu-shi, 182–8585 Japan
a) E-mail: uto@ai.lab.uec.ac.jp
DOI:10.14923/transinfj.2017JDP7055

る項目反応理論（Item Response Theory: IRT）[25] に対し，評価者特性を表すパラメータを付与したモデルが近年多数提案されている [2], [3], [26]〜[31].

このような IRT モデルとして最も代表的なモデルが多相ラッシュモデル [29] である．多相ラッシュモデルには幾つかのバリエーションが知られているが，最も基礎的なモデルはラッシュ型 IRT モデル（e.g., ラッシュモデル [32]，評定尺度モデル [33]，部分採点モデル [34]）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 [22], [23], [30]．多相ラッシュモデルは，全ての課題について識別力が一定であることを仮定するが，現実にはこの仮定は成り立たないことが多い [3], [19], [20], [27], [28], [31]．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を表現できる一般化部分採点モデル [35] や段階反応モデル [36] に対して評価者パラメータを付与したモデルが提案されてきた．例えば，Patz and Junker [28] は，一般化部分採点モデルに評価者の厳しさパラメータを付与したモデルを提案している．また，宇佐美 [31] は，課題間での識別力の差異に加え，評価者内/評価者間における評価の一貫性にも差異があることを指摘し，評価者一貫性パラメータを加えた一般化部分採点モデルを提案している．他方，段階反応モデルの拡張として，Ueno and Okamoto [3] は，各評価カテゴリーに対する評価者の厳しさを表すパラメータを付与したモデルを提案している．更に，Uto and Ueno [2] は，宇佐美 [31] と同様に，能力測定精度が評価者一貫性にも依存することを指摘し，一貫性パラメータを付与した段階反応モデルを提案している．

これらの IRT モデルでは評価者特性を考慮して学習者の能力を推定できるため，素点の合計や平均といった単純な得点化法に比べて高精度な能力測定が可能である [2], [19], [37]．しかし，既存モデルでは，評価基準が他の評価者と極端に異なる “異質評価者” の特性を必ずしも十分に表現できないため，異質評価者が存在する可能性があるピアアセスメントに適用したとき能力測定精度が低下する問題が残る [19], [38]．ピアアセスメントでは，様々なバックグラウンドの学習者が評価者として参入するため，異質な特性の評価者が混在する可能性は高いと予測できる．

この問題を解決するために，本研究では，以下の三つの特性に対応する評価者パラメータを付与した新たな IRT モデルを提案する．

（1） 評価の厳しさ（甘さ）：全体として低い（または高い）評価を与える傾向の程度．

（2） 評価の一貫性：測定対象の能力真値に対し，評価者が与える評価点が一致・一貫している程度．

（3） 尺度範囲の制限：特定の評価カテゴリーを過剰に使用する（または特定の評価カテゴリーの使用を避ける）傾向の程度．

提案モデルの利点は以下のとおりである．

（1） 評価者の特性を柔軟に表現できるため，異質評価者の採点データに対するモデルのあてはまりを改善できる．

（2） 異質評価者の影響を正確に能力測定値に反映できるため，異質評価者が存在するピアアセスメントにおいて，既存モデルより高精度な能力測定が期待できる．

本研究では，シミュレーション実験と実データ実験を通して，提案手法の有効性を評価する．

2. ピアアセスメントデータ

本研究では，ピアアセスメントデータとして，課題 $i \in \{1, \cdots, I\}$ における学習者 $j \in \{1, \cdots, J\}$ の成果物に評価者 $r \in \{1, \cdots, J\}$ が与える評価カテゴリー $k \in \{1, \cdots, K\}$ で構成される評価データを想定する．ここで，u_{ijr} を課題 i における学習者 j の成果物に対する評価者 r の評価カテゴリーとすると，評価データ $\boldsymbol{U}$ は以下のように定義できる．

$$\boldsymbol{U} = \{u_{ijr} | u_{ijr} \in \{-1, 1, \cdots, K\}, \forall i, \forall j, \forall r\}$$

ここで，$u_{ijr} = -1$ は欠測データを表す．

本研究では，上記のピアアセスメントデータ $\boldsymbol{U}$ に項目反応理論を適用し，学習者の能力を高精度に測定することを目指す．

3. 項目反応理論

項目反応理論（Item Response Theory: IRT）は数理モデルを用いたテスト理論の一つである [25]．IRT の利用には次のような利点がある [2], [3]．(1) 能力測定精度の低い異質項目の影響を小さくして学習者の能力を推定できる．(2) 異なる項目への学習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．

IRT は適応型テストや等質テスト自動構成といった現代のテスト運用の基礎を成す理論であり，TOEFL [39] や IT パスポート試験 [40]，医療系大学間共用試験 [41] などの国内外の大規模試験を含め，様々な評価場面で

広く実用化されている．

これまで，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．一方で，近年では，論述式・記述式テストのような多段階カテゴリー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 [20], [42]．本研究で扱う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，段階反応モデル（Graded Response Model: GRM）[36] や一般化部分採点モデル（Generalized Partial Credit Model: GPCM）[35] が知られている．本研究では，これらの多値型 IRT モデルを基礎モデルとして用いるため，以降ではこれらのモデルについて詳述する．

3.1 段階反応モデル

GRM は，Samejima [36] 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 が評価カテゴリー k を得る確率を次式で定義する．

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

$$\begin{cases} P^*_{ijk} = [1 + \exp(-\alpha_i(\theta_j - b_{ik}))]^{-1} \\ P^*_{ij0} = 1,\ P^*_{ijK} = 0. \end{cases} \tag{2}$$

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 より大きいカテゴリーを得る困難度を表すパラメータである．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．

3.2 一般化部分採点モデル

GPCM では同様の確率 P_{ijk} を次式で定義する．

$$P_{ijk} = \frac{\exp\sum_{m=1}^{k}[\alpha_i(\theta_j - \beta_{im})]}{\sum_{l=1}^{K}\exp\sum_{m=1}^{l}[\alpha_i(\theta_j - \beta_{im})]} \tag{3}$$

ここで，β_{ik} は，課題 i においてカテゴリー $k-1$ からカテゴリー k に遷移する困難度を表し，ステップパラメータと呼ばれる．GPCM では，モデルの識別性のために $\beta_{i1} = 0; \forall i$ と制約する．

GPCM は，ステップパラメータ β_{ik} を $\beta_i + d_{ik}$ と分解し，次のように表す場合もある．

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

ここで，β_i は課題 i の困難度を表す位置パラメタ，d_{ik} は課題 i のカテゴリー k に対する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，$d_{i1} = 0$, $\sum_{k=2}^{K} d_{ik} = 0 : \forall i$ と制約される．

GPCM は，評定尺度モデル（Rating Scale Model: RSM）[33] や部分採点モデル（Partial Credit Model: PCM）[34] 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 β_{ik} を $\beta_i + d_k$ と分解したモデルとして定義される．ただし，d_k はカテゴリー $k-1$ から k に遷移する困難度を表すステップパラメータである．

3.3 項目特性パラメータの解釈

本節では，上述の多値型 IRT モデルの中で最も多くのパラメータで定義される式 (4) の GPCM に基づき，項目特性パラメータの解釈を説明する．このために，異なる特性パラメータをもつ三つの課題に対する GPCM の反応曲線を図 1 に示した．カテゴリー数は $K = 5$ とした．図 1 の横軸は学習者の能力 θ，縦軸は各カテゴリーへの反応確率 P_{ijk} を示す．

図 1 より，課題 1 に比べて困難度 β_i が高い課題 2 では，反応曲線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，課題 2 で得られる評価が課題 1 に比べて平均的に低くなることを意味している．

他方，識別力 α_i が低い課題 3 では，識別力が高い課題（課題 1，2）に比べ，能力値を所与としたときの各カテゴリーへの反応確率の差異が小さくなっている．これは，識別力が低い課題では，能力真値に一貫・一致した評価が得られにくい，すなわち能力真値と観測評点の相関が低くなる傾向があることを表す．

また，ステップパラメータ $d_{ik}; k \geq 2$ については，隣接するカテゴリー間の差異 $d_{ik+1} - d_{ik}$ が大きくなるほど，カテゴリー k に対する反応確率が増加する．図 1 の例では，どの課題も $d_{i3} - d_{i2}$ が大きいため，評価カテゴリー 2 への反応確率が能力尺度の広い範囲で他カテゴリーより高くなっており，カテゴリー 2 が全体として得られやすいという性質が表現されている．

3.4 GRM と GPCM の比較

GRM と GPCM はどちらもリッカート型多値データに適用でき，パラメータ数も同数で定義できる．モデルの本質的な差異は，評価カテゴリーに対する反応プロセスにどのような仮定を置くかにある [43], [44]．

GPCM は，カテゴリー $k-1$ に対するカテゴリー k への相対的な反応のし易さ P_{ijk}/P_{ijk-1} がロジスティックモデルに従うと仮定することで導出される．すなわち，GPCM では，カテゴリー k に反応するか否かが，隣接カテゴリー $k-1$ との比較から決定されると仮定している．一方で，GRM は，各カテゴリー

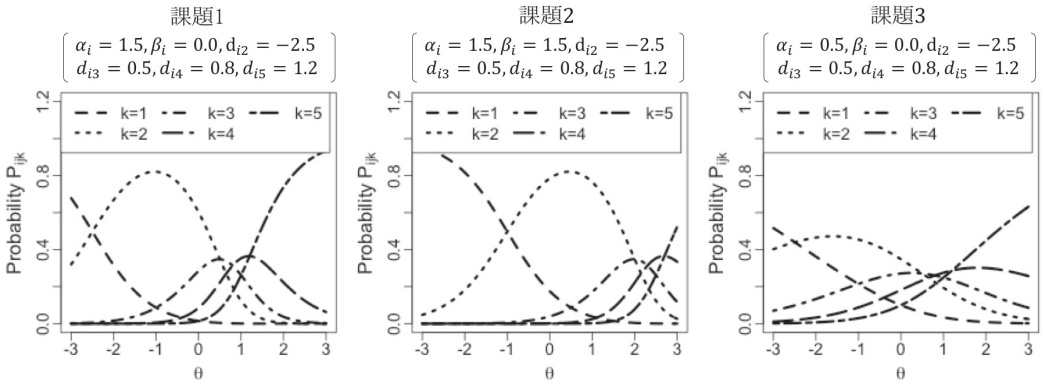

図 1　一般化部分採点モデルの項目特性曲線
Fig. 1　Item response curves of generalized partial credit model.

k について，カテゴリー k 以上に反応する確率 P^*_{ijk} がロジスティックモデルに従うと考える．すなわち，GRM では，各カテゴリーに対する絶対的な基準が存在することが仮定されている．

以上のような差異があるものの，分析対象データに対してどちらのモデル化が適切であるかを事前に決定することは一般に困難である [44]〜[47]．実データを用いてモデル比較を行った先行研究においても，扱うデータによって適切なモデルが異なることを報告している [45]〜[47]．そのため，実践的には，モデルの選択は恣意的になされることが多い [47]．一方で，Sung and Kang [46] は，GRM に従うデータに GPCM がよく当てはまるのに対し，GPCM に従うデータに対して GRM は必ずしも当てはまらないことを報告している．この結果は，事前情報がない場合には，GPCM を採用した方が高い性能が得られる可能性が高いことを示唆する．そこで，本研究で開発するモデルでは GPCM を基礎モデルとして採用する．

4. 評価者特性パラメータを付与した項目反応モデル

3. で紹介した多値型 IRT モデルは，学習者 × 課題の 2 相データへの適用を想定している．一方で，**2.** で述べたように，本論で扱うピアアセスメントデータは学習者 × 課題 × 評価者の三相データとなる．このような三相データに対して，これまで紹介した多値型 IRT モデルを直接適用することはできない [2], [3]．

この問題を解決するアプローチとして，評価者特性を表すパラメータを加えた IRT モデルが多数提案されてきた [2], [3], [26]〜[29], [31]．以降では，これらの既存の IRT モデルを紹介する．

4.1 多相ラッシュモデル

評価者パラメータを付与した IRT モデルとして最も一般的なモデルは，Linacre [29] が提案した多相ラッシュモデル（Many-Facet Rasch Models: MFRMs）である．多相ラッシュモデルは，ラッシュ型モデル（ラッシュモデル [32]，RSM，PCM など）に課題と学習者以外の要因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．多相ラッシュモデルには多数のバリエーションが存在するが [22], [23], [30]，最も基礎的なモデルは Common Step モデルとして知られている．Common Step MFRM では，課題 i における学習者 j の成果物に評価者 r が評価カテゴリー k を与える確率 P_{ijrk} を次式で与える．

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

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

多相ラッシュモデルは，全ての課題について識別力が一定と仮定する．しかし，現実にはこの仮定は成り立たないことが多い [3], [19], [20], [27], [28], [31]．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を考慮できる GPCM や GRM に対して評価者パラメータを付与したモデルが提案されてきた．

4.2 評価者パラメータを付与した GPCM

Patz and Junker [28] は，GPCM に評価者の厳しさを表すパラメータを付与した以下のモデルを提案し

ている．

$$P_{ijrk} = \frac{\exp\sum_{m=1}^{k}\left[\alpha_i(\theta_j - \beta_{im} - \rho_{ir})\right]}{\sum_{l=1}^{K}\exp\sum_{m=1}^{l}\left[\alpha_i(\theta_j - \beta_{im} - \rho_{ir})\right]}, \tag{6}$$

ここで，ρ_{ir} は課題 i における評価者 r の厳しさを表す．ここでは，モデルの識別性のために $\beta_{i1} = 0$, $\rho_{i1} = 0; \forall i$ が仮定される．

また，宇佐美 [31] は，課題間での識別力の差異に加え，評価者内・評価者間における評価の一貫性にも差異があることを指摘し，GPCM に評価の一貫性パラメータを加えた以下のモデルを提案している．

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

ここで，α_r は評価者 r の評価の一貫性，d_r は評価者 r の評価点の分散を表す．パラメータの識別性のために，$\Pi_r\alpha_r = 1$，$\sum_r \beta_r = 0$，$\Pi_r d_r = 1$，$d_{i1} = 0$ を仮定する．

4.3 評価者パラメータを付与した GRM

GRM に評価者パラメータを付与したモデルとして，Ueno and Okamoto [3] は以下の IRT モデルを提案している．

$$P_{ijrk} = P^*_{ijrk-1} - P^*_{ijrk}, \tag{8}$$
$$\begin{cases} P^*_{ijrk} = [1 + \exp(-\alpha_i(\theta_j - b_i - \varepsilon_{rk}))]^{-1}, \\ P^*_{ijr0} = 1,\ P^*_{ijrK} = 0. \end{cases}$$

ここで，b_i は課題 i の困難度を表し，ε_{rk} は評価カテゴリー k に対する評価者 r の厳しさを表す．ε_{rk} には順序制約 $\varepsilon_{r1} < \varepsilon_{r2} < \cdots < \varepsilon_{rK-1}$ を仮定する．モデルの識別性のために $\varepsilon_{11} = -2.0$ と制約する．

また，Uto and Ueno [2] は，宇佐美 [31] と同様，能力測定精度が評価者の一貫性にも依存することを指摘し，評価者の一貫性パラメータを付与した次の GRM を提案している．

$$P_{ijrk} = P^*_{ijrk-1} - P^*_{ijrk}, \tag{9}$$
$$\begin{cases} P^*_{ijrk} = [1 + \exp(-\alpha_i\alpha_r(\theta_j - b_{ik} - \varepsilon_r))]^{-1}, \\ P^*_{ijr0} = 1,\ P^*_{ijrK} = 0. \end{cases}$$

ここで，ε_r は評価者 r の厳しさを表す．パラメータの識別性のために $\alpha_{r=1} = 1$，$\varepsilon_1 = 0$ を仮定する．

4.4 既存モデルの問題点

上記の IRT モデルでは評価者特性を考慮して学習者の能力を推定するため，素点の合計や平均といった単純な得点化法に比べて高精度な能力測定が可能である [2], [19], [37]．一方で，既存モデルでは，採点基準が他の評価者と極端に異なる “異質評価者” の特性を必ずしも十分に表現できないため，異質評価者が存在する可能性があるピアアセスメントに適用すると能力測定精度が低下する問題が残る [19], [38]．

具体的には，多相ラッシュモデルや Patz and Junker [28]，Ueno and Okamoto [3] のモデルでは，全ての評価者が等しく一貫した基準で採点を行うと仮定するため，一貫性が極端に低い異質評価者が存在した場合にも，その評価結果を理想的な評価者と同等の重みで能力推定値に反映してしまい，能力測定誤差の増加を引き起こす．他方，Ueno and Okamoto [3] 以外のモデルでは，各評価カテゴリーに対する反応傾向が評価者間で共通であると仮定するため，評価カテゴリーの利用傾向が異質な評価者の反応分布が実態と大きく乖離して推定されてしまい，これが測定誤差の増加要因となる．

ピアアセスメントでは，様々なバックグラウンドをもつ学習者が評価者の役割を担うため，このような異質評価者が混在する可能性は高いと予測できる．

5. 異質評価者に頑健な項目反応モデル

4.4 で述べた問題を解決するために，本研究では，1）評価の厳しさ，2）評価者一貫性，3）尺度範囲の制限，に対応する評価者特性パラメータを付与した以下の GPCM を提案する．

$$P_{ijrk} = \frac{\exp\sum_{m=1}^{k}\left[\alpha_r\alpha_i(\theta_j - \beta_i - \beta_r - d_{rm})\right]}{\sum_{l=1}^{K}\exp\sum_{m=1}^{l}\left[\alpha_r\alpha_i(\theta_j - \beta_i - \beta_r - d_{rm})\right]} \tag{10}$$

ただし，パラメータの識別性のために，$\alpha_{r=1} = 1$，$\beta_{r=1} = 0$，$d_{r1} = 0$，$\sum_{k=2}^{K} d_{rk} = 0$ を仮定する．

提案モデルでは，α_r が一貫性を，β_r が評価の厳しさを，d_{rk} が尺度範囲の制限を表現する．このことを示すために，表 1 のパラメータをもつ 4 名の評価者に対する提案モデルの項目反応曲線を図 2 に示す．図 2 では，横軸が能力値 θ，縦軸が反応確率 P_{ijrk} を表す．

図 2 から，評価の一貫性 α_r が低い Rater2 の分布では，一貫性が高い Rater1 の分布に比べて，評価カ

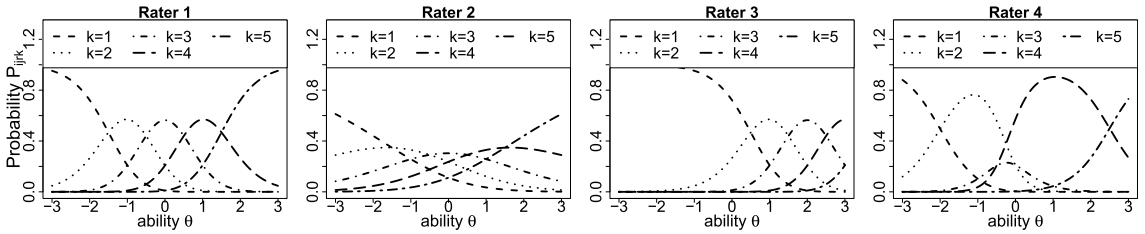

図 2　特性の異なる 4 名の評価者に対する提案モデルの反応曲線
Fig. 2　Item response curves of the proposed model for different four raters.

表 1　図 2 の評価者の特性パラメータ
Table 1　Rater parameters used in Fig. 2.

	α_r	β_r	d_{rk}
Rater 1	2.0	0.0	{0.0, −1.5, −0.5, 0.5, 1.5}
Rater 2	0.5	0.0	{0.0, −1.5, −0.5, 0.5, 1.5}
Rater 3	2.0	2.0	{0.0, −1.5, −0.5, 0.5, 1.5}
Rater 4	2.0	0.0	{0.0, −2.0, 0.0, −0.5, 2.5}

テゴリー間での反応確率の差異が小さくなっていることがわかる．これは，Rater2 は，同一の能力の学習者に対しても評価がバラつきやすく，能力真値と一致・一貫した評価が付与されにくいことを表現している．すなわち，α_r は能力真値と観測評点の相関に対応しており，相関が低いほどこの値が低く推定される傾向がある．

他方，評価の厳しさ β_r が大きい Rater3 では，Rater1 に比べ，反応曲線全体が右に移動していることがわかる．これは，Rater3 が与える評価が Rater1 に比べて平均的に低い，すなわち評価が厳しいことを意味する．

また，ステップパラメータ d_{rk} の値を，隣接する評価カテゴリーとの差異 $d_{rk+1} - d_{rk}$ が大きくなるように定めると，評価カテゴリー k への反応確率が高くなる．これは，評価者 r の使用する評価カテゴリーが $d_{rk+1} - d_{rk}$ の大きいカテゴリーのみに制限されることを意味し，尺度範囲の制限の特性を反映していると解釈できる．例えば，図 2 の Rater1 と Rater4 を比較すると，Rater4 では $d_{r3} - d_{r2}$ と $d_{r5} - d_{r4}$ が Rater1 より大きいため，カテゴリー 2 と 4 への反応確率が Rater1 の反応確率より高くなっている．θ は平均 0.0 の標準正規分布に従うので，カテゴリー 2 と 4 のみに反応が集中する傾向となる．

既存モデルと比較した提案モデルの特徴は次のとおりである．

（1） 多相ラッシュモデル（式 (5)），Patz and Junker（式 (6)），Ueno and Okamoto（式 (8)）のモデルと比較すると，提案モデルでは評価者一貫性パラメータ α_r を採用している点が特徴である．**4.4** で述べたように，一貫性パラメータをもたないモデルでは，一貫性が低い異質評価者と一貫性が高い理想的な評価者の評価結果を同等の重みで能力推定値に反映してしまい，これが能力測定精度の低下要因となる．一方で，提案モデルで採用した評価者一貫性パラメータ α_r は，一貫性の低い評価者の影響を小さくする働きをするため [2], [31]，提案モデルでは，このような評価者の存在に起因する能力測定精度の低下を回避できる．

（2） 多相ラッシュモデル（式 (5)），Patz and Junker（式 (6)），宇佐美（式 (7)），Uto and Ueno（式 (9)）のモデルと比較すると，提案モデルでは尺度範囲の制限の特性を表現できる点が異なる．**4.4** で述べたように，これらの既存モデルでは，評価者間で評価カテゴリーの利用傾向が一致していると仮定するため，尺度範囲の制限が強い評価者の反応分布が実態と乖離して表現され，これが能力測定精度の低下要因となる．一方で上述のとおり，提案モデルではステップパラメータ d_{rk} により評価者の尺度範囲の制限を柔軟に表現できるため，この特性に関して異質性の高い評価者が存在しても能力測定精度の低下を回避できる．

以上のように提案モデルは，異質評価者の特性を既存モデルより柔軟に捉えることができるため，異質評価者の採点データに対する当てはまりを改善できると期待できる．更に，異質評価者の特性を正確に能力測定値に反映できるため，異質評価者数が存在するピアアセスメントに適用しても高い能力測定精度を維持できると考えられる．

6. シミュレーション実験

本章では，シミュレーション実験によるモデル比較を通して提案モデルの有効性を評価する．以降では，簡単のために，式 (5) のモデルを *MFRM*，式 (6) のモ

デルを *Patz1999*，式 (7) のモデルを *Usami2010*，式 (8) のモデルを *Ueno2008*，式 (9) のモデルを *Uto2016* とそれぞれ呼称する．

6.1 情報量基準に基づくモデル比較

提案モデルは，1）評価の厳しさ，2）評価者一貫性，3）尺度範囲の制限，に関する異質性を既存モデルより柔軟に表現できる．したがって，これらの異質性を有する評価者が多数存在する場合，評価データに対するモデルの当てはまりが既存モデルより優れると期待できる．この点を明らかにするために，本節では，異質評価者に対応したシミュレーションデータを用いて，情報量基準に基づくモデル比較を行う．

実験手順は以下のとおりである．

（1） $J = 30$，$I = 4$，$K = 5$ として，提案モデル，MFRM，Patz1999，Usami2010，Ueno2008，Uto2016 のパラメータ真値を表 2 の分布に従いランダムに生成し，それらを所与として各モデルから評価データ $\boldsymbol{U}$ をサンプリングした．

（2） 各評価者 r のデータ $\boldsymbol{U}_r = \{u_{ijr} | \forall i, j\} \subset \boldsymbol{U}$ を表 3 に従って変形し，異質評価者の反応を模した評価データ $\boldsymbol{U}' = \{\boldsymbol{U}'_1, \cdots, \boldsymbol{U}'_R\}$ を生成した．このとき，r 番目の評価者データには表 3 内の $r\%4+1$ 行目の規則を適用した．ここで % は剰余を表す．表 3 では，規則（A）が評価者一貫性に関する，規則（B）と（C）が尺度範囲の制限に関する，規則（D）が評価の厳しさに関する異質評価者のデータ生成規則を表す．尺度範囲の制限に関する異質性には複数のパターンが考えられるが，ここでは，単一の評点に偏る場合（規則（B））と二つの評点のみに偏る場合（規則（C））の 2 パターンを想定した．単一評点の過剰使用は判断に自信がない評価者に，特定の二つの評点に偏る傾向は上位・下位程度の大まかな区別しか行わない評価者に多くみられ，尺度範囲の制限の代表例として知られる [22], [24]．三つ以上の評点に偏る場合は異質性が低いと考え，異質評価者の生成規則には加えていない．

（3） 変換前の各評価データ $\boldsymbol{U}$ に対して，提案モデル，MFRM，Patz1999，Usami2010，Ueno2008，Uto2016 を仮定してパラメータ推定を行った．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マルコフ連鎖モンテカルロ法（Markov chain Monte Carlo: MCMC）[2], [28], [31] による期待事後確率推定（Expected a Posteriori: EAP）法で行った．パラメータの事前分布には表 2 を用いた．

（4） 手順（3）で推定した能力と課題パラメータの値を所与として，異質評価者データ $\boldsymbol{U}'$ から各モデルの評価者特性パラメータを MCMC で推定した．

（5） 手順（4）の推定結果を元に情報量基準によるモデル選択を行い，各モデルの順位付けを行った．情報量基準には，モデル選択問題において広く利用される Akaike Information Criterion（AIC）[48]，Widely Applicable Information Criterion（WAIC）[49]，Bayesian Information Criterion（BIC）[50]，周辺ゆう度（ML）を用いた．ただし，ML はゆう度関数を全てのパラメータで周辺化する必要があり厳密には計算できないため，MCMC サンプルを用いた近似手法 [51] により求めた．ここで，AIC，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

表 2　パラメータ分布
Table 2　Parameter distributions.

$\log \alpha_i \sim N(0.1, 0.4)$, $\log \alpha_r \sim N(0.0, 0.5)$
β_i, β_r, β_{ik}, ε_r, $\rho_{ir} \sim N(0.0, 1.0)$
d_{ik}, d_{rk}, d_k, b_i, d_r, $\theta_j \sim N(0.0, 1.0)$
$b_{ik}, \varepsilon_{rk} \sim MN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
$\boldsymbol{\mu} = \{-1.50, -0.75, 0.75, 1.50\}$,
$\boldsymbol{\Sigma} = \begin{pmatrix} 0.25 & 0.16 & 0.16 & 0.16 \\ 0.16 & 0.25 & 0.16 & 0.16 \\ 0.16 & 0.16 & 0.25 & 0.16 \\ 0.16 & 0.16 & 0.16 & 0.25 \end{pmatrix}$

表 3　シミュレーション実験で利用する異質評価者データの生成規則
Table 3　Generation rules of aberrant raters' data used for simulation experiments.

異質特性	異質評価者データ $\boldsymbol{U}'_r$ の生成方法
（A） 一貫性が低い（評価のランダムネスが強い）	評価データ $\boldsymbol{U}_r$ の 70%をランダムに発生させたカテゴリーに置換したデータを $\boldsymbol{U}'_r$ とする．
（B） 尺度範囲の制限が極端に強い（単一の評点に反応が偏る）	ランダムに一つのカテゴリー $k' \in \{1, \cdots, K\}$ を選択し，評価データ $\boldsymbol{U}_r$ からランダムに選択した 70%のデータを k' に置換したデータを $\boldsymbol{U}'_r$ とする．
（C） 尺度範囲の制限が強い（二つの評点に反応が偏る）	ランダムに二つのカテゴリー k'', k'''（ただし，$k'' > k'''$）を選択し，評価データ $\boldsymbol{U}_r$ からランダムに選択した 70%のデータに対して，データの平均値より大きい場合には k'' に，そうでない場合には k''' に置換したデータを $\boldsymbol{U}'_r$ とする．
（D） 厳しい，または甘い（全体的に低い，または高い評点に偏る）	ランダムに選択したカテゴリー $k'''' \in \{-K+1, \cdots, -1, 1, \cdots, K-1\}$ を所与とし，評価データ $\boldsymbol{U}_r$ からランダムに選択した 70%のデータを $x_{ijr} + k''''$ としたデータを $\boldsymbol{U}'_r$ とする．ただし，変換後のカテゴリーが 1 を下回る場合には 1，5 を超える場合には K とした．

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．一方で，BIC と ML は一致性をもつモデル選択基準であり，漸近的に真のモデルが選択される．

（6） 以上の手順を 10 回繰り返し，各情報量基準で推定された各モデルの順位の平均と標準偏差を求めた．また，比較のために，評価データ $\boldsymbol{U}$ に対しても同様のモデル比較実験を行った．

実験結果を表 4 に示す．表 4 より，異質評価者を含まないデータ $\boldsymbol{U}$（以降，通常評価者データと呼ぶ）に対しては，データ発生に用いたモデルが全ての情報量基準で最適なモデルとして選択されていることがわかる．一方で，異質評価者のデータに対しては，データ発生元のモデルにかかわらず，提案モデルが全ての情報量基準で最適なモデルとして推定された．これは，

表 4 シミュレーション実験における情報量基準を用いたモデル比較
Table 4 Model comparison using information criteria in simulation experiments.

		データ発生モデル	提案モデル	MFRM	Patz1999	Usami2010	Ueno2008	Uto2016
異質評価者	AIC	提案モデル	**1.00(0.00)**	5.80(0.16)	2.80(0.16)	5.20(0.16)	2.50(0.65)	3.70(0.41)
		MFRM	**1.00(0.00)**	6.00(0.00)	2.40(0.24)	5.00(0.00)	2.70(0.41)	3.90(0.09)
		Patz1999	**1.30(0.21)**	5.00(0.00)	3.40(0.24)	6.00(0.00)	1.70(0.21)	3.60(0.24)
		Usami2010	**1.00(0.00)**	5.50(0.25)	2.60(0.24)	5.50(0.25)	2.60(0.64)	3.80(0.16)
		Ueno2008	**1.00(0.00)**	5.40(0.24)	2.90(0.29)	5.60(0.24)	2.30(0.41)	3.80(0.16)
		Uto2016	**1.00(0.00)**	5.00(0.00)	3.20(0.16)	6.00(0.00)	2.00(0.00)	3.80(0.16)
	WAIC	提案モデル	**1.00(0.00)**	5.70(0.21)	5.30(0.21)	2.70(0.21)	2.50(0.65)	3.80(0.16)
		MFRM	**1.00(0.00)**	6.00(0.00)	5.00(0.00)	2.10(0.09)	3.00(0.20)	3.90(0.09)
		Patz1999	**1.40(0.24)**	5.00(0.00)	6.00(0.00)	3.30(0.21)	1.60(0.24)	3.70(0.21)
		Usami2010	**1.00(0.00)**	5.50(0.25)	5.50(0.25)	2.60(0.24)	2.60(0.64)	3.80(0.16)
		Ueno2008	**1.00(0.00)**	5.30(0.21)	5.70(0.21)	2.90(0.29)	2.30(0.41)	3.80(0.16)
		Uto2016	**1.00(0.00)**	5.00(0.00)	6.00(0.00)	3.00(0.00)	2.00(0.00)	4.00(0.00)
	BIC	提案モデル	**1.00(0.00)**	5.40(0.24)	5.60(0.24)	2.80(0.36)	2.60(0.64)	3.60(0.44)
		MFRM	**1.00(0.00)**	6.00(0.00)	5.00(0.00)	2.20(0.16)	3.10(0.49)	3.70(0.21)
		Patz1999	**1.30(0.21)**	5.00(0.00)	6.00(0.00)	3.50(0.25)	1.70(0.21)	3.50(0.25)
		Usami2010	**1.00(0.00)**	5.40(0.24)	5.60(0.24)	2.60(0.24)	2.60(0.64)	3.80(0.16)
		Ueno2008	**1.00(0.00)**	5.10(0.09)	5.90(0.09)	3.00(0.40)	2.30(0.41)	3.70(0.21)
		Uto2016	**1.10(0.09)**	5.00(0.00)	6.00(0.00)	3.20(0.16)	1.90(0.09)	3.80(0.16)
	ML	提案モデル	**1.00(0.00)**	5.90(0.09)	5.10(0.09)	2.80(0.16)	2.30(0.41)	3.90(0.09)
		MFRM	**1.00(0.00)**	6.00(0.00)	5.00(0.00)	2.40(0.24)	2.70(0.41)	3.90(0.09)
		Patz1999	**1.30(0.21)**	5.00(0.00)	6.00(0.00)	3.30(0.21)	1.70(0.21)	3.70(0.21)
		Usami2010	**1.00(0.00)**	5.80(0.16)	5.20(0.16)	2.60(0.24)	2.60(0.64)	3.80(0.16)
		Ueno2008	**1.00(0.00)**	5.50(0.25)	5.50(0.25)	2.90(0.29)	2.30(0.41)	3.80(0.16)
		Uto2016	**1.00(0.00)**	5.00(0.00)	6.00(0.00)	3.00(0.00)	2.00(0.00)	4.00(0.00)
通常評価者	AIC	提案モデル	**1.10(0.09)**	5.00(0.40)	5.60(0.44)	2.00(0.20)	4.10(1.09)	3.20(0.16)
		MFRM	3.30(0.41)	**1.00(0.00)**	4.70(0.81)	3.50(2.45)	5.30(1.01)	3.20(1.76)
		Patz1999	4.70(0.61)	4.90(0.89)	**1.00(0.00)**	2.60(0.44)	5.30(0.61)	2.50(0.25)
		Usami2010	1.90(0.09)	5.90(0.09)	4.60(0.44)	**1.10(0.09)**	4.00(1.00)	3.50(0.25)
		Ueno2008	2.40(0.84)	6.00(0.00)	4.80(0.16)	3.80(0.36)	**1.00(0.00)**	3.00(0.40)
		Uto2016	3.30(1.41)	5.90(0.09)	4.40(0.44)	2.30(0.21)	4.10(0.69)	**1.00(0.00)**
	WAIC	提案モデル	**1.00(0.00)**	5.50(0.25)	5.30(0.41)	2.10(0.09)	3.80(0.96)	3.30(0.21)
		MFRM	3.20(0.56)	**1.10(0.09)**	5.10(0.49)	3.60(2.84)	5.00(1.20)	3.00(1.40)
		Patz1999	4.00(0.60)	5.70(0.21)	**1.00(0.00)**	2.40(0.24)	5.10(0.49)	2.80(0.36)
		Usami2010	1.90(0.09)	6.00(0.00)	4.40(0.64)	**1.10(0.09)**	3.80(0.56)	3.80(0.56)
		Ueno2008	2.40(0.84)	6.00(0.00)	4.80(0.16)	3.70(0.41)	**1.00(0.00)**	3.10(0.49)
		Uto2016	3.30(1.41)	6.00(0.00)	4.60(0.24)	2.30(0.21)	3.80(0.36)	**1.00(0.00)**
	BIC	提案モデル	**1.40(0.24)**	4.30(0.21)	5.90(0.09)	1.90(0.89)	4.50(1.25)	3.00(0.20)
		MFRM	5.10(0.49)	**1.00(0.00)**	4.30(0.61)	2.90(1.29)	5.00(1.00)	2.70(1.41)
		Patz1999	5.00(0.60)	4.40(1.44)	**1.00(0.00)**	2.80(0.56)	5.30(0.61)	2.50(0.25)
		Usami2010	2.10(0.09)	5.70(0.41)	4.90(0.29)	**1.00(0.00)**	4.00(1.00)	3.30(0.41)
		Ueno2008	2.60(1.04)	5.20(0.76)	5.40(0.44)	4.10(0.49)	**1.00(0.00)**	2.70(0.21)
		Uto2016	3.70(0.81)	5.70(0.21)	4.50(1.05)	2.00(0.00)	4.10(0.69)	**1.00(0.00)**
	ML	提案モデル	**1.00(0.00)**	5.70(0.21)	5.10(0.29)	2.10(0.09)	3.70(1.01)	3.40(0.24)
		MFRM	2.00(0.40)	**1.80(0.76)**	4.30(0.61)	3.70(3.61)	5.10(1.09)	4.10(2.29)
		Patz1999	3.80(0.56)	6.00(0.00)	**1.00(0.00)**	2.40(0.24)	4.90(0.09)	2.90(0.49)
		Usami2010	1.90(0.09)	6.00(0.00)	4.20(0.56)	**1.10(0.09)**	3.70(0.41)	4.10(0.89)
		Ueno2008	2.30(0.81)	6.00(0.00)	4.70(0.21)	3.70(0.81)	**1.00(0.00)**	3.30(0.21)
		Uto2016	3.20(1.56)	6.00(0.00)	4.50(0.45)	2.40(0.24)	3.90(0.29)	**1.00(0.00)**

表 3 に挙げた異質評価者の特性を提案モデルが最も柔軟に表現できたためと解釈できる．以上から，表 3 のような異質評価者が多数存在する場合，真のモデルにかかわらず，提案モデルが最適なモデルとみなせることが示せた．

6.2 能力測定精度の比較

一般にデータへの当てはまりが良い IRT モデルは，能力測定のバイアス要因を高精度に取り除けるため，高い能力測定精度が期待できる [2], [19]．そこで，本節では，能力測定精度の観点でモデルの比較を行い，提案モデルの有効性を評価する．

6.2.1 相関に基づく評価

ここでは，**6.1** の実験と同様に各モデルからデータを発生させ，それらのデータから各モデルで能力推定を行い，推定値と能力真値とを比較する実験を行う．このような推定誤差の評価には，推定値と真値の平均平方 2 乗誤差（Root mean square error: RMSE）が広く用いられる [2], [19], [31]．しかし，本実験において RMSE を用いると，真のモデルと推定モデルが異なる場合の誤差が，モデルが一致している場合より過大に推定され，これらの場合についての精度を比較できない．そこで，本実験では，真の能力値と能力推定値との相関を用いて能力測定精度を評価する．モデルの不一致により RMSE が過大推定される場合でも，相関が高ければ，学習者の相対的な順序を正しく推定できており，妥当な能力測定がなされたとみなせる．

本実験の手順は以下のとおりである．

（1） **6.1** の手順（1）（2）と同様に，各 IRT モデルのパラメータ真値とそれに従う評価データ $\boldsymbol{U}$ を発生させ，各データ $\boldsymbol{U}$ に対応する異質評価者データ $\boldsymbol{U}'$ を生成した．

（2） **6.1** の手順（3）と同様に，各評価データ $\boldsymbol{U}$ から MCMC により各モデルのパラメータを推定した．

（3） **6.1** の手順（4）と同様に，評価データ $\boldsymbol{U}$ から推定された学習者の能力と課題パラメータを所与として，異質評価者データ $\boldsymbol{U}'$ から各モデルの評価者特性パラメータを推定した．

（4） 評価データ $\boldsymbol{U}$ と異質評価者データ $\boldsymbol{U}'$ のそれぞれについて，各学習者の能力を各評価者のデータ $\boldsymbol{U}_r$，$\boldsymbol{U}'_r$ から推定した．ただし，評価データ $\boldsymbol{U}$ からの推定では手順（2）で推定された課題と評価者パラメータを所与とし，異質評価者データ $\boldsymbol{U}'$ からの推定では手順（2）で推定された課題パラメータと手順（3）で推定された評価者パラメータを所与とした．

（5） 各評価者のデータから推定された能力値と能力真値との相関を求めた．

（6） 以上の手順を 10 回繰り返し，相関の平均と標準偏差を求めた．

実験結果を表 5 に示す．表 5 より，通常評価者のデータでは，MFRM が真のモデルの場合を除き，真のモデルを用いて推定したときに最も高い相関を示したことがわかる．真のモデルが MFRM の場合には，提案モデルが最も相関が高かったが，実際には全てのモデル間における差異が 0.01 未満と非常に小さいため，モデル間で優位な差異はないと解釈できる．MFRM は最も単純なモデルであり，他の全てのモデルの下位モデルとみなせる．そのため，他のモデルでも MFRM のデータ生成過程を再現でき，結果としてモデル間の差異が失われたと解釈できる．

他方，異質評価者のデータに対しては，真のモデルにかかわらず，提案モデルが最も相関が高かったことがわかる．前節で示したとおり，既存モデルでは，異質評価者の特性を必ずしも捉えることができないため，能力測定精度が大きく低下したと考えられる．これに

表 5 能力真値と推定値の相関
Table 5 Correlations between true ability and ability estimates.

	データ発生モデル	提案モデル	MFRM	Patz1999	Usami2010	Ueno2008	Uto2016
異質評価者	提案モデル	**0.556(0.123)**	0.474(0.143)	0.479(0.145)	0.533(0.127)	0.502(0.136)	0.532(0.123)
	MFRM	**0.573(0.105)**	0.497(0.131)	0.486(0.154)	0.545(0.123)	0.512(0.133)	0.555(0.111)
	Patz1999	**0.577(0.119)**	0.516(0.147)	0.540(0.130)	0.553(0.128)	0.521(0.128)	0.556(0.114)
	Usami2010	**0.556(0.121)**	0.440(0.164)	0.453(0.159)	0.539(0.149)	0.471(0.150)	0.514(0.134)
	Ueno2008	**0.493(0.145)**	0.451(0.145)	0.433(0.142)	0.476(0.151)	0.457(0.130)	0.480(0.151)
	Uto2016	**0.499(0.151)**	0.436(0.164)	0.438(0.129)	0.490(0.136)	0.454(0.141)	0.454(0.139)
通常評価者	提案モデル	**0.864(0.037)**	0.831(0.055)	0.841(0.053)	0.855(0.044)	0.840(0.048)	0.848(0.045)
	MFRM	**0.846(0.044)**	0.841(0.053)	0.840(0.048)	0.834(0.054)	0.840(0.043)	0.833(0.050)
	Patz1999	0.837(0.055)	0.814(0.064)	**0.886(0.038)**	0.843(0.048)	0.815(0.060)	0.830(0.051)
	Usami2010	0.834(0.076)	0.796(0.083)	0.816(0.067)	**0.843(0.066)**	0.815(0.078)	0.809(0.077)
	Ueno2008	0.800(0.081)	0.771(0.084)	0.795(0.072)	0.795(0.069)	**0.802(0.070)**	0.796(0.075)
	Uto2016	0.801(0.069)	0.710(0.114)	0.744(0.095)	0.787(0.085)	0.764(0.103)	**0.810(0.075)**

対し，提案モデルでは，その特性を柔軟に捉えて能力測定値に反映できたため，測定精度の低下を緩慢にできたと解釈できる．

6.2.2 推定誤差に基づく評価

6.2.1 の実験では，真のモデルと推定モデルが異なる場合について能力測定精度を比較するために，相関を利用した．本項では，真のモデルと推定モデルが同じ場合を対象として，RMSE による測定誤差の評価を行う．本実験の手順は以下のとおりである．

（1） **6.1** の手順（1）（2）と同様に，各 IRT モデルのパラメータ真値と評価データ $\boldsymbol{U}$ を発生させ，各データ $\boldsymbol{U}$ に対応する異質評価者データ $\boldsymbol{U}'$ を生成した．

（2） 学習者の能力と課題パラメータの真値を所与として，異質評価者データ $\boldsymbol{U}'$ から評価者パラメータを推定した．

（3） 評価データ $\boldsymbol{U}$ と異質評価者データ $\boldsymbol{U}'$ のそれぞれから学習者の能力を推定し，能力推定値と能力真値との RMSE を求めた．このとき，評価データ $\boldsymbol{U}$ からの推定では，課題と評価者パラメータの真値を所与とした．異質評価者データ $\boldsymbol{U}'$ からの推定では，課題パラメータの真値と手順（2）で推定した評価者パラメータを所与とした．

（4） 以上の手順を 10 回繰り返し，RMSE の平均と標準偏差を求めた．

実験結果を表 6 に示す．表 6 より，通常評価者データでは，モデル間の RMSE の差異が 0.02 以下と微小であり，どのモデルも同程度の精度で推定が行われたといえる．一方で，異質評価者のデータに対しては，提案モデルの RMSE が最小となったことが確認できる．

6.3 考　　察

以上のシミュレーション実験から，表 3 のような異質評価者が多数存在する場合，真のデータ発生源によらず，提案モデルがデータに最もよく当てはまり，能力測定精度も最も高くなることが示せた．ただし，表 3 に挙げた異質評価者の一部しか存在しないような場合には，必ずしも提案モデルが最適となる保証はない．一般に，データを適切に表現できるモデルが複数存在する場合，パラメータ数の少ないモデルの方が優れた性能を示す可能性が高い [2], [19], [26]．これは，パラメータ数が少ないモデルの方が，パラメータ数に対するデータ数が多くなり，パラメータの推定精度が向上するためである [2], [19], [26]．したがって，例えば，表 3 の異質評価者（A）のような一貫性の低い評価者が存在しない場合には，評価者の厳しさと尺度範囲の制限の特性を表現でき，提案モデルよりパラメータ数が少ない Ueno2008 の採用が適切であると考えられる．また，異質評価者（B），（C）のような尺度範囲の制限が強い評価者が存在しない場合には，評価者一貫性と厳しさの特性を表現でき，提案モデルより少数のパラメータで記述される Uto2016 や Usami2010 の方が優れると予想される．しかし，**1.** や **4.4** で述べたように，学習者同士のピアアセスメントでは様々な異質特性をもつ評価者が存在する可能性が高く，そのような場合には提案モデルが既存モデルより優れた性能を示すことが本章の実験から示された．次章では，現実の学習者同士のピアアセスメントにおいて，表 3 の異質特性をもつ評価者が存在し，提案モデルが有効であることを実データにより示す．

表 6　シミュレーション実験による能力測定誤差（RMSE）
Table 6　Ability measurement errors (RMSE) evaluated using simulation data.

	異質評価者 平均（標準偏差）	通常評価者 平均（標準偏差）
提案モデル	**0.205** (0.175)	0.080 (0.064)
MFRM	0.413 (0.324)	0.088 (0.067)
Patz1999	0.473 (0.379)	**0.078** (0.061)
Usami2010	0.254 (0.226)	0.082 (0.068)
Ueno2008	0.466 (0.365)	0.095 (0.066)
Uto2016	0.436 (0.348)	0.087 (0.068)

7. 実データ適用

本章では，被験者実験により収集した実際のピアアセスメントデータを用いて，提案手法の有効性を評価する．

7.1 実 デ ー タ

本研究で行った被験者実験では，まず，30 名の大学生と大学院生に対して，四つのライティング課題を行わせた．ライティング課題は，National Assessment of Educational Progress（NAEP）の 2002 年 [52] と 2007 年 [53] 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を選択した．各課題に対して提出された成果物を，30 名の被験者に評価させた．評価は NAEP grade 12 [53] に基づいて作成した 5 段階カテゴリーの評価基準を用いて行わせた．被験者にはレクチャーなどを与えずに評価を行わせており，専門分野や学年も異なることから，異質な特性の評価者が混在している可能性は高いといえる．

表 7　実データに対する情報量基準値
Table 7　Information criteria calculated from actual data.

	AIC	WAIC	BIC	ML
提案モデル	**−4431.53**	**−4396.07**	−4561.84	**−4324.23**
MFRM	−4650.06	−4646.46	−4696.30	−4615.25
Patz1999	−4662.97	−4646.08	−4776.47	−4575.41
Usami2010	−4458.22	−4434.58	−4554.20	−4377.75
Ueno2008	−4541.02	−4504.17	−4651.02	−4445.21
Uto2016	−4442.92	−4434.82	**−4518.58**	−4385.57

以上の被験者実験から得られたピアアセスメントデータを用いて提案モデルの有効性を評価した．

7.2　情報量基準によるモデル比較

本節では，情報量基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，上記のピアアセスメントデータから 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，AIC, WAIC，BIC，ML を推定した．

実験結果を表 7 に示す．表 7 から，BIC を除く全ての情報量基準において，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．BIC は，ML の漸近近似であるにもかかわらず，ML とは異なるモデルを選択している．BIC は，パラメータ数に対してデータ数が少ない場合，真のモデルよりパラメータ数が少ないモデルを選択する傾向がある [54], [55]．本実験でも，MFRM に次いでパラメータ数が少ない Uto2016 が選択されており，この解釈と一致する．MFRM はパラメータ数は最小であるもの，評価者の厳しさしか考慮しておらず，データの表現力が極めて低いため BIC でも選択されなかったと考えられる．ML では提案モデルが選択されているため，データ数が増加すれば，BIC を用いても提案モデルが選択されると予測できる．

以上から，実データに対して提案モデルが最適なモデルであることが確認できた．

7.3　評価者特性の分析

提案モデルが実データに適合した要因として，提案モデルで導入した（1）評価の厳しさ，（2）評価者一貫性，（3）尺度範囲の制限の特性について異質性の高い評価者が多数存在していたことが考えられる．このことを確認するために，提案モデルを用いて推定した各評価者の特性パラメータ値を表 8 に示した．表 8 より次の点が確認できる．

（1）　評価者 2，24，26 など，一貫性 α_r が極端に低い評価者の存在が複数確認できる．

（2）　評価者 7，9，26 のように評価の厳しさ β_r が非常に高い，極端に厳しい評価者の存在も確認できる．

（3）　尺度範囲の制限の特性については複数の異質パターンが確認できる．例えば，評価者 2 や 28 は $d_{r5} - d_{r4}$ のみが大きいことから，評点 4 のみに評価が集中する傾向があると解釈できる．また，評価者 1，4，9，18 は $d_{r3} - d_{r2}$ と $d_{r5} - d_{r4}$ が相対的に大きく，評点 2 と 4 を過剰利用する傾向が読み取れる．

表 8　実データから推定された評価者パラメータ
Table 8　Rater parameters estimated from actual data.

評価者	α_r	β_r	d_{r2}	d_{r3}	d_{r4}	d_{r5}
1	1.000	0.000	−1.169	−0.154	0.152	1.171
2	0.638	0.132	−0.383	−0.460	−0.163	1.007
3	1.267	0.393	−0.991	−0.308	0.477	0.822
4	1.115	0.025	−1.695	−0.416	0.051	2.059
5	0.963	−0.334	−1.740	−0.372	0.740	1.372
6	0.928	−0.078	−1.774	−0.145	0.386	1.532
7	0.746	0.856	−0.357	−0.546	0.882	0.022
8	1.809	0.301	−1.511	−0.680	0.701	1.489
9	1.091	0.793	−1.857	−0.034	0.414	1.477
10	0.797	−0.111	−0.445	−0.089	0.133	0.401
11	1.137	−0.262	−1.645	−0.584	0.626	1.602
12	1.029	−0.182	−1.780	−0.651	0.603	1.828
13	0.858	0.648	−1.171	−0.129	0.694	0.606
14	0.881	0.235	−1.935	−0.017	0.595	1.358
15	1.374	−0.128	−1.480	−0.897	0.618	1.759
16	1.249	0.148	−0.111	−1.637	−0.295	2.043
17	1.261	−0.413	−1.231	−0.846	0.567	1.509
18	1.670	0.206	−1.307	−0.299	0.393	1.213
19	1.770	0.455	−2.278	−0.459	1.829	0.908
20	1.261	0.698	−1.506	−0.599	0.340	1.764
21	0.745	0.004	−1.137	0.083	0.623	0.431
22	1.354	0.249	−2.051	−0.308	0.755	1.604
23	1.153	0.188	−1.493	−1.501	0.927	2.068
24	0.568	0.231	−1.376	−0.458	0.792	1.042
25	0.829	−0.126	−0.536	0.030	0.236	0.270
26	0.571	0.773	−1.027	0.106	0.268	0.653
27	0.920	−0.079	−0.941	0.130	−0.374	1.185
28	0.855	−0.397	−0.589	−0.943	−0.441	1.973
29	1.338	0.118	−1.423	−0.253	0.494	1.182
30	0.834	−0.285	−1.741	0.715	−0.067	1.092

以上から，実データにおいて，提案モデルで採用した三つの評価者特性について異質性が高い評価者が多数程度が存在したことがわかる．既存モデルではこれらの特性を必ずしも表現できないのに対し，提案モデルではこれらの特性を柔軟に反映できるため，前節の実験において実データへの当てはまりが最も高かったと解釈できる．

7.4　能力推定精度の評価

本節では，実データを用いて能力測定精度の評価を行う．実験手順は以下のとおりである．

（1）　提案モデル, MFRM, Patz1999, Usami2010, Ueno2008，Uto2016 について，実データを用いて

MCMC によるパラメータ推定を行った．事前分布には表 2 を用いた．

（2） 推定された課題パラメータと評価者パラメータを所与として，個々の評価者のデータ $\boldsymbol{U}_r$ を用いて各学習者の能力を推定した．

（3） 評価者ごとに，手順（2）で推定された能力値と手順（1）で完全データから推定した能力値とのRMSE を計算した．

比較のために，評価点の平均値を能力推定値とする手法についても同様の実験を行った．

実験結果を図 3 に示す．図 3 より，ほぼ全ての評価者において，提案モデルを用いたときの能力測定誤差が最小となったことが確認できる．提案モデルが既存モデルに劣る場合でも RMSE 最小値と提案モデルの RMSE との差異は，その他のモデル間や評価者間での差に比べて微小であることから，提案モデルはどのような評価者に対しても既存モデルと同程度以上の能力測定精度を達成できたと解釈できる．

また，各モデルにおける RMSE の平均値と標準偏差を表 9 に示した．表 9 から，提案モデルを用いたときに RMSE の平均値が最小となったことが確認できる．そこで，RMSE の平均値の差異が優位であるかを確認するために平均値の差の検定を行った．まず，等分散性の検定を行ったところ等分散性が認められたため（検定統計量 $= 63.093$，$p < 0.001$），ANOVA の方法で平均値の差の検定を行った．分散分析の結果，モデル間の平均値に有意差が検出された（検定統計量 $= 14.635$，$p < 0.01$）．そこで，提案モデルにおける RMSE の平均値が既存モデルと比べて優位に小さいことを確認するために，Dunnet の手法により提案モデルと既存モデルとの多重比較を行った．結果を表 9 の「検定統計量」の列に示す．表 9 より，提案モデルの能力測定誤差が既存モデルに比べて優位に小さいことが確認できた．

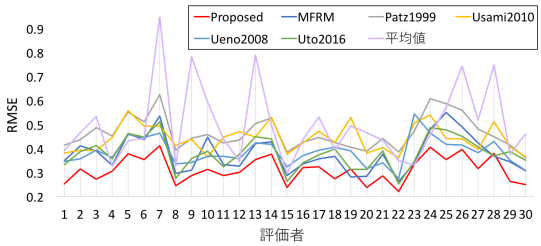

図 3 評価者ごとの能力測定精度
Fig. 3 Ability measurement accuracy for each rater.

表 9 実データによる能力推定精度の評価結果
Table 9 Ability measurement accuracies evaluated using actual data.

	平均	標準偏差	検定統計量（p 値）
提案モデル	**0.313**	0.053	-
MFRM	0.379	0.075	3.025 ($p < 0.05$)
Patz1999	0.464	0.067	6.995 ($p < 0.01$)
Usami2010	0.443	0.057	5.987 ($p < 0.01$)
Ueno2008	0.386	0.055	3.385 ($p < 0.01$)
Uto2016	0.382	0.065	3.192 ($p < 0.01$)
平均値	0.499	0.157	7.709 ($p < 0.01$)

本章で示した実データ実験から，異質評価者が混在すると想定できる実際のピアアセスメントにおいて，提案モデルがデータに最もよく当てはまり，能力測定精度も優位に改善できることが示せた．

8. むすび

本論文では，異質評価者が多数存在するピアアセスメントにおいて，学習者の能力測定精度を改善するために，異質評価者に頑健な IRT モデルを提案した．具体的には，（1）評価の厳しさ，（2）一貫性，（3）尺度範囲の制限，の三つの評価者特性パラメータを付与した一般化部分採点モデルを提案した．また，提案モデルの利点として次の 2 点に言及した．1）評価者の特性を既存モデルより柔軟に捉えることができるため，異質評価者の採点データに対する当てはまりを改善できる．2）異質評価者の特性を正確に能力測定値に反映できるため，異質評価者数が存在するピアアセスメントに適用しても高い能力測定精度を維持できる．

本論文では，シミュレーション実験から，異質評価者が多数存在する場合，真のモデルによらず，提案モデルがデータに最もよく当てはまり，能力測定精度も最大となることを示した．更に，異質な評価者が混在しやすい場面を想定して収集した実際のピアアセスメントデータを用いた実験を行った．実験の結果，シミュレーション実験同様，提案モデルが実データに最もよく当てはまり，能力測定精度も既存モデルより有意に高かったことを示した．

本研究では，能力測定精度のみに着目したが，提案モデルは多様な評価者特性値を与えるため，それらを評価能力の測定や異質評価者のトレーニングなどに応用することも可能である．このような応用を今後の課題としたい．

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

文　献

[1] K.J. Topping, E.F. Smith, I. Swanson, and A. Elliot, “Formative peer assessment of academic writing between postgraduate students,” Assessment & Evaluation in Higher Education, vol.25, no.2, pp.149–169, 2000.

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

[3] M. Ueno and T. Okamoto, “Item response theory for peer assessment,” Proc. IEEE International Conference on Advanced Learning Technologies, pp.554–558, 2008.

[4] S. Bostock, “Student peer assessment,” Higher Education Academy Articles, 2001.

[5] R.L. Weaver and H.W. Cotrell, “Peer evaluation: A case study,” Innovative Higher Education, vol.11, no.1, pp.25–39, 1986.

[6] H. Suen, “Peer assessment for massive open online courses (MOOCs),” The International Review of Research in Open and Distributed Learning, vol.15, no.3, pp.313–327, 2014.

[7] N.B. Shah, J. Bradley, S. Balakrishnan, A. Parekh, K. Ramchandran, and M.J. Wainwright, “Some scaling laws for MOOC assessments,” ACM KDD Workshop on Data Mining for Educational Assessment and Feedback, 2014.

[8] M. Ueno, “Data mining and text mining technologies for collaborative learning in an ILMS “samurai”,” Proc. IEEE International Conference on Advanced Learning Technologies, pp.1052–1053, 2004.

[9] S.S. Lin, E.Z.F. Liu, and S.M. Yuan, “Web-based peer assessment: Feedback for students with various thinking-styles,” J. Computer Assisted Learning, vol.17, no.4, pp.420–432, 2001.

[10] S. Trahasch, “From peer assessment towards collaborative learning,” Proc. ASEE/IEEE Frontiers in Education Conference, vol.2, pp.16–20, 2004.

[11] P. Davies, “Review in computerized peer-assessment. Will it affect student marking consistency?,” Proc. 11th CAA International Computer Assisted Conference, pp.143–151, 2007.

[12] Y.T. Sung, K.E. Chang, S.K. Chiou, and H.T. Hou, “The design and application of a web-based self- and peer-assessment system,” Computers & Education, vol.45, no.2, pp.187–202, 2005.

[13] J. Sitthiworachart and M. Joy, “Effective peer assessment for learning computer programming,” Proc. Ninth Annual SIGCSE Conference on Innovation and Technology in Computer Science Education, pp.122–126, 2004.

[14] K. Cho and C.D. Schunn, “Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system,” Computers & Education, vol.48, no.3, pp.409–426, 2007.

[15] L. Moccozet and C. Tardy, “An assessment for learning framework with peer assessment of group works,” Proc. International Conference on Information Technology Based Higher Education and Training, pp.1–5, 2015.

[16] C. Piech, J. Huang, Z. Chen, C. Do, A. Ng, and D. Koller, “Tuned models of peer assessment in MOOCs,” Proc. Sixth International Conference of MIT’s Learning International Networks Consortium, pp.153–160, 2013.

[17] J. Lave and E. Wenger, Situated Learning – Legitimate Peripheral Participation, ed. R. Pea and J.S. Brown, Cambridge University Press, New York, Port Chester, Melbourne, Sydney, 1991.

[18] S.J. Lurie, A.C. Nofziger, S. Meldrum, C. Mooney, and R.M. Epstein, “Effects of rater selection on peer assessment among medical students,” Medical Education, vol.40, no.11, pp.1088–1097, 2006.

[19] 宇都雅輝，植野真臣，“パフォーマンス評価のための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.

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

[21] F.E. Saal, R.G. Downey, and M.A. Lahey, “Rating the ratings: Assessing the psychometric quality of rating data,” Psychological Bulletin, vol.88, no.2, pp.413–428, 1980.

[22] 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.

[23] C.M. Myford and E.W. Wolfe, “Detecting and measuring rater effects using many-facet Rasch measurement: Part II,” J. Applied Measurement, vol.5, pp.189–227, 2004.

[24] 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.

[25] F.M. Lord, Applications of item response theory to practical testing problems, Erlbaum Associates, 1980.

[26] 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.

[27] 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.

[28] 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, no.4, pp.342–366, 1999.

[29] J.M. Linacre, Many-faceted Rasch Measurement, MESA Press, 1989.
[30] T. Eckes, Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluating Rater-Mediated Assessments, Peter Lang Pub. Inc., 2015.
[31] 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.
[32] G. Rasch, Probabilistic models for some intelligence and attainment tests, The University of Chicago Press, 1980.
[33] D. Andrich, “A rating formulation for ordered response categories,” Psychometrika, vol.43, no.4, pp.561–573, 1978.
[34] G. Masters, “A Rasch model for partial credit scoring,” Psychometrika, vol.47, no.2, pp.149–174, 1982.
[35] E. Muraki, “A generalized partial credit model,” Handbook of Modern Item Response Theory, ed. W.J. van der Linden and R.K. Hambleton, pp.153–164, Springer, 1997.
[36] F. Samejima, “Estimation of latent ability using a response pattern of graded scores,” Psychometrika Monography, vol.17, pp.1–100, 1969.
[37] T. Nguyen, M. Uto, Y. Abe, and M. Ueno, “Reliable peer assessment for team project based learning using item response theory,” Proc. International Conference on Computers in Education, pp.144–153, 2015.
[38] 宇都雅輝，植野真臣，“評価者と課題の多様な特性を考慮した項目反応モデル，” 日本テスト学会第 14 回大会発表論文抄録集，pp.62–63, 2016.
[39] Educational Testing Service, “The TOEFL Test,” https://www.ets.org/toefl/
[40] 独立行政法人情報処理推進機構，“IT パスポート試験，” https://www3.jitec.ipa.go.jp/JitesCbt/
[41] 公益社団法人医療系大学間共用試験実施評価機構，“臨床実習開始前の「共用試験」第 13 版（平成 27 年度），” http://www.cato.umin.jp/e-book/13/index.html
[42] M. Matteucci and L. Stracqualursi, “Student assessment via graded response model,” Statistica, vol.66, pp.435–447, 2006.
[43] E. Muraki, “A generalized partial credit model: Application of an EM algorithm,” Applied Psychological Measurement, vol.16, no.2, pp.159–176, 1992.
[44] S.P. Reise and D.A. Revicki, Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment (Multivariate Applications Series), Routledge, Taylor & Francis Group, 2014.
[45] M.L. Nering and R. Ostini, Handbook of Polytomous Item Response Theory Models, Routledge, Taylor & Francis Group, 2010.
[46] H.J. Sung and T. Kang, “Choosing a polytomous IRT model using Bayesian model selection methods,” National Council on Measurement in Education Annual Meeting, pp.1–36, 2006.
[47] O. Naumenko, “Comparison of various polytomous item response theory modeling approaches for task-based simulation CPA exam data,” Technical Report, CPA Examination Technical Reports, 2014.
[48] H. Akaike, “A new look at the statistical model identification,” IEEE Trans. Autom. Control, vol.19, pp.716–723, 1974.
[49] S. Watanabe, “Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory,” J. Machine Learning Research, vol.11, pp.3571–3594, 2010.
[50] G. Schwarz, “Estimating the dimensions of a model,” Annals of Statistics, vol.6, pp.461–464, 1978.
[51] M. Newton and A.E. Raftery, “Approximate Bayesian inference by the weighted likelihood bootstrap,” J. Royal Statistical Society. Series B: Methodological, vol.56, no.1, pp.3–48, 1994.
[52] H. Persky, M. Daane, and Y. Jin, “The nation’s report card: Writing 2002,” Technical Report, National Center for Education Statistics, 2003.
[53] D. Salahu-Din, H. Persky, and J. Miller, “The nation’s report card: Writing 2007,” Technical Report, National Center for Education Statistics, 2008.
[54] R. Eggeling, T. Roos, P. Myllymäki, and I. Grosse, “Robust learning of inhomogeneous PMMs,” Proc. 17th International Conference on Artificial Intelligence and Statistics (AISTATS), pp.229–237, 2014.
[55] M. Drton, “A Bayesian information criterion for singular models,” J. Royal Statistical Society: Series B (Statistical Methodology), vol.79, no.2, pp.1–38, 2017.

（平成 29 年 5 月 27 日受付，8 月 3 日再受付，
9 月 8 日早期公開）

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修了．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．

植野　真臣　（正員）

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．