2015 年度 日本テスト学会誌　Vol.12, No.1 55 - 75

展望論文

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

A Review of Item Response Models for Performance Assessment

宇都 雅輝 [1], 植野 真臣 [1]

Masaki Uto [1], Maomi Ueno [1]

[1] 電気通信大学

[1] The University of Electro-Communications

Correspondence concerning this article should be sent to: Masaki Uto, The University of Electro-Communications, 1-5-1 Chofugaoka, Chofu, Tokyo 182-8585.　E-mail: uto@is.uec.ac.jp

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

宇都 雅輝[1]，植野 真臣[1]

[1]電気通信大学

近年，大学入試や人事考課，学習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．パフォーマンス評価では，受験者を複数の課題に取り組ませ，その過程や成果物を複数の評価者で採点することが一般的である．しかし，このようなパフォーマンス評価では，評価の信頼性が課題や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，評価者と課題の特性を考慮して受験者の能力を推定できる項目反応モデルが多数提案され，その有効性が報告されてきた．これらの項目反応モデルでは，考慮される評価者や課題の特性がモデルごとに異なるため，評価場面の特性に合った適切なモデルを選択することが信頼性の改善において重要となる．そこで，本論文では，評価者と課題の特性を考慮した既存の項目反応モデルについて，それらの特徴を比較し整理する．さらに，その結果に基づき，評価場面の特徴に応じた適切なモデルの選択法について述べる．また，本論文では，実際のパフォーマンス評価への適用例を通して，これらの項目反応モデルが評価の信頼性改善に与える影響について評価する．

キーワード：項目反応理論，パフォーマンス評価，信頼性，評価者特性，多相データ

A Review of Item Response Models for Performance Assessment

Masaki Uto[1], Maomi Ueno[1]

[1] The University of Electro-Communications

Performance assessment has been attracted much attention in various assessment fields, such as entrance exam, employee evaluation and educational assessment. Performance assessment enables to assess examinees' practical and higher order skills, which are difficult to be assessed by traditional paper tests. In typical performance assessment, examinee's performances for multiple tasks are evaluated by multiple raters. However, it has been pointed out that reliability of such performance assessment strongly depends on characteristics of raters and tasks. As a method to improve the reliability, item response models which incorporate rater and task characteristic parameters has been proposed. Earlier studies reported that the models could improve the reliability of performance assessment because they can estimate ability of examinees considering characteristics of raters and tasks. When applying them to actual performance assessments, the selection of an optimal model for the assessment situation is important. Therefore, this paper reviews previous item response models that incorporate rater and task characteristic parameters and explains those characteristics. Furthermore, the paper proposes an approach to select an optimal model for assessment situations. Moreover, the paper demonstrates the effectiveness of the models through a real data application.

Keywords： Item response theory, performance assessment, reliability, rater characteristics, multi-way data

1. はじめに

近年，入学試験や入社試験を始めとする様々な評価場面において，論理的思考力や問題解決力，主体性といった高次の能力を測定するニーズが高まっており，これを実現する手法の一つとしてパフォーマンス評価が注目されている（植野・荘島, 2010; 宇佐美, 2015）．パフォーマンス評価は，受験者に課題を与え，その活動過程や成果物を直接評価する評価法であり（松下, 2012），これまでにも様々な評価場面で活用されてきた．例えば，大学入試における論述式テストや外国語試験におけるスピーキング・リスニングテスト，入社試験における面接やグループディスカッション，学習場面におけるレポートや学習プロセスの評価などが挙げられる．さらに，将来の導入が検討されている「大学入学希望者学力評価テスト（仮称）」では，思考力，判断力，表現力，主体性，協調性といった多面的・総合的な能力の測定を目指し，記述式テストやスピーキングテストなどのパフォーマンス評価の採用が検討されている（高大接続システム改革会議, 2015; 文部科学省中央教育審議会, 2014）．以上のように，パフォーマンス評価の重要性は今後ますます増加すると考えられる．

パフォーマンス評価は，課題に対する各受験者のパフォーマンスを，複数の評価者が評価する形式で行われることが一般的である．しかし，このようなパフォーマンス評価では，得られる評点が課題や評価者の特性に強く依存することが知られており，これにより評価の信頼性が低下する問題が指摘されてきた（De Gruijter, 1984; Lurie, Nofziger, Meldrum, Mooney, & Epstein, 2006; 宇佐美, 2013a; Uto & Ueno, 2015）．具体的には，課題の困難度や識別力，評価者の甘さ・厳しさ，一貫性，中心化傾向，尺度範囲の制限などの特性が，評価の信頼性低下を引き起こすバイアス要因として知られている（例えば, DeCarlo, Kim, & Johnson, 2011; Lu & Wang, 2006; Myford & Wolfe, 2003; Saal, Downey, & Lahey, 1980; 宇佐美, 2013a, 2013b; Uto & Ueno, 2015）．したがって，パフォーマンス評価では，このような評価者と課題のバイアスをできる限り取り除いて関心下の能力を測定することで，評価の信頼性を改善することが重要な課題となる．

平井（2007）や宇佐美（2013a, 2015）は，このような信頼性の高い能力測定を実現するためには，1) 十分な数の課題と評価者を用意すること，2) 明確な評価基準を作成すること，3) 評価者のトレーニングを十分に行うこと，4) 関心下の能力がパフォーマンスに十分に反映されるような課題内容とすること，などが重要であると指摘している．実際には，これらのアプローチに基づき，評価場面に応じた適切な評価環境をデザインすることが重要となる．例えば，測定対象の能力を細分化し，それらの能力を詳細な評価基準を用いて評価する分析的評価などでは，比較的客観的な評価基準を作成できると考えられる．このような場合には，評価基準を順守するように評価者をトレーニングすることで評価者特性に起因する評点の分散を十分に小さくでき，評価の信頼性を改善できる．一方で，概略評価や総合評価によって複合的な能力の測定を試みる場合や評価者の主観的な判断を要する評価基準を用いる場合などには，評価者のトレーニングのみでは評価者特性の影響を十分に取り除くことは困難といえる．このような場合には，評価者数や課題数を増やすことで，特定の評価者や課題が評価結果に与える影響を小さくすることが効果的と考えられる．しかし，実際の評価場面では，人的・時間的・経済的な制約から，十分な評価者数，課題数を用意することは難しい場合も多い．また，評価者トレーニングによって評価者特性を均一化することは短時間のトレーニングでは困難であることが知られている（平井, 2007）．したがって，現実のパフォーマンス評価では，評価者や課題特性による影響が残ることが多いといえる．

このような問題を解決する手法の一つとして，評価者や課題の特性を考慮して受験者の能力を推定することで，評価の信頼性改善を目指す手法が提案されてきた．具体的には，テスト理論の一つである項目反応理論の拡張モデルとして，評価者と課題の特性パラメータを付与した項目反応モデルが多数提案されてきた（例えば，DeCarlo, Kim, & Johnson, 2011; Linacre, 1989; Lu & Wang, 2006; Patz & Junker, 1999; Patz, Junker, Johnson, & Mariano, 2002; Ueno & Okamoto, 2008; 宇佐美, 2010; Uto & Ueno, 2015）．これらの項目反応モデルでは，評価者と課題のバイアスを補正して受験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で得られた評点よりも高い信頼性を示すことが報告されている（Nguyen, Uto, Abe, & Ueno, 2015; 宇都・植野, 2015; Uto & Ueno, 2015）．したがって，これらの項目反応モデルは，評価環境のデザインによって評価者や課題のバイアスを十分に取り除くことが難しい場合に評価の信頼性を

改善できる有効な手法と解釈できる．

評価者と課題の特性パラメータを付与した既存の項目反応モデルでは，採用している評価者・課題パラメータがモデルごとに異なるため，実際の評価場面に適用する場合，評価者数や課題数，想定される評価者・課題の特性に応じた適切なモデルの選択が重要となる．そこで，本論文では評価者と課題の特性を考慮した既存の項目反応モデルについて，それらの特徴を比較し整理する．さらに，その結果に基づき，評価場面の特徴に応じた適切なモデルの選択法について述べる．また，本論文では，実際のパフォーマンス評価への適用例を通して，これらの項目反応モデルが評価の信頼性改善に与える影響について評価する．

2. パフォーマンス評価における信頼性

評価の信頼性とは，測定結果が受験者の真の能力を反映している程度を表す概念である（Kim, 2012; 日本テスト学会, 2007; 宇佐美, 2013a）．受験者の能力を異なる課題や評価者によって評価したとき，能力測定結果が安定しているほど信頼性が高いと解釈される．

1節で述べたように，パフォーマンス評価では，個々の評点が課題や評価者の特性に強く依存することが知られている．

信頼性に影響を与える評価者特性としては，以下の特性が知られている（例えば，平井, 2006, 2007; Muraki, Hombo, & Lee, 2000; Myford & Wolfe, 2003; Saal, Downey, & Lahey, 1980; 宇佐美, 2013a, 2013b; Uto & Ueno, 2015; Wang & Yao, 2013）．

1. 評価の甘さ・厳しさ（Leniency / Severity）：評価の甘さは，すべての受験者に対して高い評点を与える傾向を表す．評価の厳しさは，全ての受験者に対して低い評点を与える傾向を表す．
2. 評価の一貫性（Consistency）：評価者内・評価者間で評価基準が一貫・一致している程度を表す．評価の一貫性が低い場合，類似したパフォーマンスに対しても評点がばらつき，安定した評価結果を得ることが困難になる．
3. 中心化傾向（Central Tendency）：評点が評価尺度の平均値付近に集中する傾向を表す．この特性が強い評価者は，能力の微小な差異を識別していないと解釈できる．
4. 尺度範囲の制限（Restriction of Range）：特定の評点に評価が集中する傾向を表す．中心化傾向は，尺度範囲の制限の特殊形であるが，これら二つの概念は区別して解釈することが一般的である（Kassim, 2011; Myford & Wolfe, 2003; Saal, Downey, & Lahey, 1980）．

一方で，信頼性に影響を与える課題特性としては，以下が想定されることが多い（DeCarlo, Kim, & Johnson, 2011; Lu & Wang, 2006; Patz & Junker, 1999; Ueno & Okamoto, 2008; 宇佐美, 2010, 2013b; Uto & Ueno, 2015）．

1. 困難度（Difficulty）：得られる評点が全体として低くなる特性を表す．
2. 識別力（Discrimination）：関心下の能力がその課題に対するパフォーマンスに反映される度合いを表す．識別力が高い課題では，測定対象の能力がパフォーマンスに適切に反映されるため，当該能力の高い受験者は高い評点を，低い評価者は低い評点を一貫して得やすくなると解釈できる．

以上のような評価者と課題の特性を考慮して受験者の能力を推定することで，評価の信頼性改善を目指すアプローチとして，評価者と課題の特性パラメータを付与した項目反応モデルが多数提案されてきた（例えば，DeCarlo, Kim, & Johnson, 2011; Linacre, 1989; Lu & Wang, 2006; Patz & Junker, 1999; Patz et al., 2002; Ueno & Okamoto, 2008; 宇佐美, 2010; Uto & Ueno, 2015）．これらの項目反応モデルは，受験者とテスト項目の2相データを扱う一般的な項目反応モデルにおいて，テスト項目を課題とみなし，評価者特性パラメータを付与したモデルとして定式化される．そこで，次節では，まず，一般的な項目反応理論について解説する．

3. 項目反応理論

項目反応理論（Item Response Theory: 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ひとつである（Lord, 1980）．IRTの特徴として，以下のような点が挙げられる（Ueno & Okamoto, 2008）．

1) 推定精度の低い異質項目の影響を小さくして能力推定を行うことができる．
2) 異なる項目への受験者の反応を同一尺度上で評価できる．
3) 欠測データから容易にパラメータを推定できる．

IRTは，適応型テスト（豊田, 2013）や等質テスト自動構成（石井・植野, 2015）といった現在のテスト理論の基礎をなす理論であり，情報処理技術者試験のひとつであるITパスポート試験（独立行政法人情報処理推進機構, 2013）や医療系大学間共用試験実施評価機構による臨床実習開始前の共用試験（公益社団法人医療系大学間共用試験実施評価機構, 2014）などの国内の大規模試験を含め，様々な評価場面において広く活用されている．

これまで，IRTは，正誤判定問題や多肢選択式問題などの2値の反応データを扱う客観テストに利用されることが一般的であったが，近年では，多値型項目反応モデルを利用した小論文などの評価への応用も進められている（例えば，DeCarlo, 2005; Matteucci & Stracqualursi, 2006）．

以降では，IRTにおける基礎的なモデルとして，2値型データを扱うモデルを解説し，その拡張モデルである多値モデルを紹介する．

3.1. 2値型項目反応モデル

最も基礎的な項目反応モデルとして，受験者の能力とテスト項目の困難度との関係をロジスティックモデルで定義したラッシュモデル（Rasch, 1980）が知られている．ラッシュモデルでは，受験者jが項目iに正答する確率を次式で表す．

$$P_{ij} = \frac{\exp(\theta_j - b_i)}{1 + \exp(\theta_j - b_i)} \quad (1)$$

ここで，b_iは項目iの困難度を，θ_jは受験者jの能力を表す．困難度パラメータb_iは，正答確率が0.5となる能力値θを表す．これらの項目パラメータと能力パラメータは，テストへの反応データから推定することができる．

ラッシュモデルは，少数のパラメータで記述されており，安定的にパラメータを推定できる点が特徴である．このため，ラッシュモデルは，受験者数が少なく反応データ数が十分に得られない場合に利用されることが多い（豊田，2013）．一方で，ラッシュモデルは，モデルが単純であるために，複雑な項目特性を表現できず，データへの適合が悪いことも多い（村木, 2011）．

そのため，実際のテスト場面では，ラッシュモデルに項目の識別力パラメータを加えた2母数ロジスティックモデルが利用されることが一般的である．2母数ロジスティックモデルでは，項目反応確率を次式で表す．

$$P_{ij} = \frac{\exp(\alpha_i(\theta_j - b_i))}{1 + \exp(\alpha_i(\theta_j - b_i))} \quad (2)$$

ここで，α_iは項目iの識別力を表す．識別力α_iは，能力値$\theta = b_i$付近の能力をどの程度の精度で識別できるかを表す．

ラッシュモデルや2母数ロジスティックモデルは，項目に対する受験者の反応が，正誤のような2値データで表される場合に利用できる．しかし，パフォーマンス評価では，多段階の評価カテゴリが対応付けられた評価基準に基づいて採点を行うことが多く，評価データは多値の順序尺度データとなることが一般的である．このような多値データに適用できるモデルとして，2値型項目反応モデルを拡張した多値型項目反応モデルが提案されてきた．

3.2. 部分採点モデル（PCM）

Masters（1982）によって提案された多値型項目反応モデルとして，部分採点モデル（Partial Credit Model: PCM）が知られている．PCMでは，受験者jが項目iに対してカテゴリ$k \in \{1 \cdots K\}$と反応する確率P_{ijk}を次式で与える．

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

ここで，β_{ik}は項目iにおいてカテゴリ$k-1$からカテゴリ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．PCMでは，モデルの識別性のために，$\beta_{i1} = 0 : \forall i$を所与とする．

PCMは，カテゴリkへの反応確率P_{ijk}とカテゴリ$k-1$への反応確率P_{ijk-1}のロジット比$\log(P_{ijk}/P_{ijk-1})$を受験者の能力とカテゴリkにおける項目困難度の線形和$\theta_j - \beta_{ik}$で定義しており，項目への正答確率と誤答確率のロジット比を$\theta_j - b_i$と定義するラッシュモデルの多値への一般化と解釈できる．

3.2. 一般化部分採点モデル（GPCM）

Muraki（1997）は，PCMにおける項目識別力一定の制約を緩和したモデルとして，一般化部分採点モデル（Generalized Partial Credit Model: GPCM）を提案している．GPCMでは，受験者jが項目iに対してカテゴリkと反応する確率P_{ijk}を次式で与える．

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

PCMと同様に，モデルの識別性のために，$\beta_{i1} = 0 : \forall i$

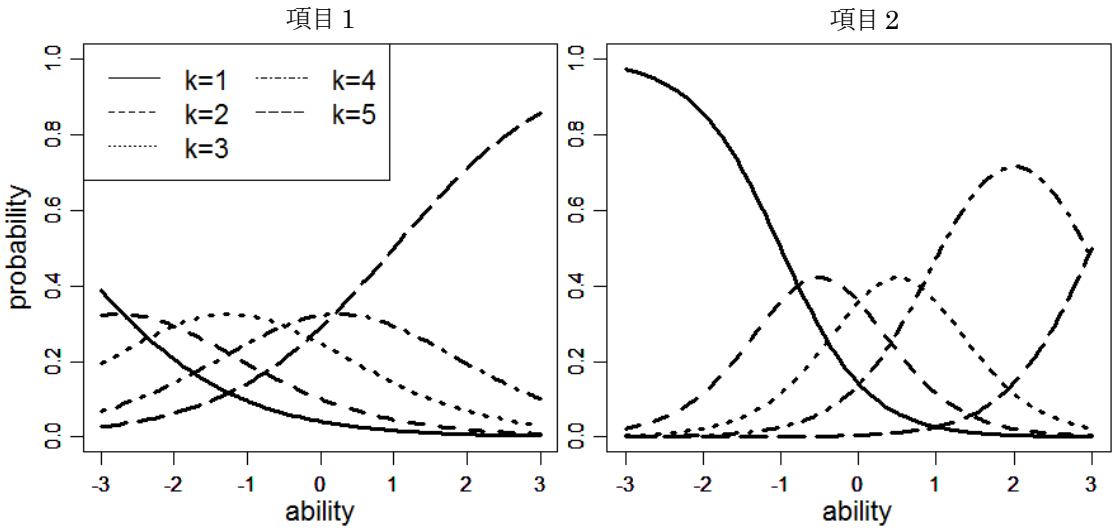

図 1. 段階反応モデルの項目特性曲線

を所与とする.

GPCMは，Andrich（1978）による評定尺度モデルと同様に，ステップパラメータ β_{ik} を $\beta_i + d_k$，あるいは $\beta_i + d_{ik}$ と分解することができ，以下のようなモデルで表されることもある.

$$P_{ijk} = \frac{\exp\sum_{m=1}^{k}[\alpha_i(\theta_j - \beta_i - d_{im})]}{\sum_{l=1}^{K}\exp\sum_{m=1}^{l}[\alpha_i(\theta_j - \beta_i - d_{im})]} \quad (5)$$

$$P_{ijk} = \frac{\exp\sum_{m=1}^{k}[\alpha_i(\theta_j - \beta_i - d_m)]}{\sum_{l=1}^{K}\exp\sum_{m=1}^{l}[\alpha_i(\theta_j - \beta_i - d_m)]} \quad (6)$$

ここで，β_i は項目 i の位置パラメータ，d_{ik} は項目 i のカテゴリ k に対する閾値パラメータ，d_k はカテゴリ k のカテゴリパラメータと呼ばれる．モデルの識別性のために，$d_{i1} = 0, \sum_{k=2}^{K} d_{ik} = 0 : \forall i, d_1 = 0, \sum_{k=2}^{K} d_k = 0$ を所与とする.

3.3. 段階反応モデル（GRM）

段階反応モデル（Graded Response Model：GRM）は，Samejima（1969）が考案した多値型項目反応モデルである．GRMでは，受験者 j が項目 i に対してカテゴリ k と反応する確率 P_{ijk} を次式で与える.

$$P_{ijk} = P^*_{ijk-1} - P^*_{ijk} \quad (7)$$

$$P^*_{ijk} = \frac{\exp\left(\alpha_i(\theta_j - b_{ik})\right)}{1 + \exp\left(\alpha_i(\theta_j - b_{ik})\right)} : \; k = 1, \cdots, K-1 \quad (8)$$

ここで，b_{ik} は項目 i において k より大きいカテゴリに反応する困難度を表す．また，$P^*_{ij0} = 1, P^*_{ijK} = 0$ とする．GRMでは，困難度パラメータ b_{ik} に順序制約 $b_{i1} < b_{i2} < \cdots < b_{iK-1}$ を課す.

ここで，多値型項目反応モデルのパラメータの解釈を説明するために，異なる項目パラメータを持つ二つの項目について，GRMの反応曲線を図1に示した．図1では，カテゴリ数 $K = 5$ とし，項目パラメータは項目1（左図）では $\alpha_i = 0.9, \boldsymbol{b_{ik}} = \{-3.5, -2.0, -0.5, 1.0\}$ を，項目2（右図）では $\alpha_i = 1.8, \boldsymbol{b_{ik}} = \{-1.0,\ 0.0, 1.0, 3.0\}$ を与えた．図1では，横軸が受験者の能力 θ，縦軸が能力 θ の受験者が各カテゴリ k に反応する確率を示す.

図1より，能力が低いほど低いカテゴリへの反応確率が高く，能力が高いほど高いカテゴリへの反応確率が高くなることがわかる．また，困難度パラメータが全体的に大きい項目2では，項目1に比べて，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，項目2では，能力の低い受験者が高いカテゴリに反応しにくいことを意味する．また，識別力が大きい項目2では，能力 θ の変化に伴う各カテゴリへの反応確率の変化が，項目1に比べて大きいことが確認できる．これは，項目2の方が，能力の微小な違いを精度よく識別できることを意味する.

4. 課題と評価者の特性を考慮した項目反応モデル

これまで紹介してきた項目反応モデルは，受験者とテスト項目の2相データへの適用を想定している．一方で，本論で想定するパフォーマンス評価データ $\boldsymbol{X}$ は，パフォーマンス課題 $i \in \{1 \cdots I\}$ に対する受験者 $j \in \{1 \cdots J\}$ のパフォーマンスに，評価者 $r \in \{1 \cdots R\}$ が与える評点 $k \in \{1 \cdots K\}$ の集合であり，以下のような3相データとして定義される.

$$\boldsymbol{X} = \{x_{ijr} \,|\, x_{ijr} \in \{-1,1 \cdots K\}\} \quad (i = 1 \cdots I, j = 1 \cdots J, r = 1 \cdots R) \tag{9}$$

ここで，$x_{ijr} = -1$は欠測データを表す．

このような3相データに対して，これまで紹介した一般的な項目反応モデルを直接適用することはできない．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項目反応モデルが提案されてきた（例えば，DeCarlo, Kim, & Johnson, 2011; Linacre, 1989; Lu & Wang, 2006; Patz & Junker, 1999; Patz et al., 2002; Ueno & Okamoto, 2008; 宇佐美，2010; Uto & Ueno, 2015）．これらの項目反応モデルは，従来の項目反応モデルにおいて，項目パラメータを課題の特性を表すパラメータとみなし，評価者特性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．

4.1. 多相ラッシュモデル (Many-Facet Rasch Model)

多相データのための項目反応モデルとして，最も広く知られているモデルは，Linacre（1989）が提案した多相ラッシュモデルである．多相ラッシュモデルは，ラッシュモデルに課題と受験者以外の要因を表すパラメータを付与したモデルである．例えば，評価者rの評価の厳しさを表すパラメータβ_rを付与した多相ラッシュモデルでは，評価者rが課題iにおける受験者jのパフォーマンスにポジティブな判定を与える確率を次式で与える．

$$P_{ijr} = \frac{\exp(\theta_j - b_i - \beta_r)}{1 + \exp(\theta_j - b_i - \beta_r)} \tag{10}$$

ここで，b_iは課題iの困難度を表す．

多相ラッシュモデルは2値データを扱うモデルであるが，ラッシュモデルと同様に，PCMを用いた多値への拡張モデルも提案されている．多相ラッシュモデルのPCMによる拡張モデルは，受験者・課題・評価者・評点間にどのような交互作用を仮定するかによって，複数のモデル化が考えられる（Myford & Wolfe, 2003; 野口・大隅, 2014）．ここでは，最も単純な多値型多相ラッシュモデルであるCommon stepモデルを紹介する．

Common stepモデルは，評価者rが課題iにおける受験者jのパフォーマンスに評点kを与える確率を次式で定義する．

$$P_{ijrk} = \frac{\exp \sum_{m=1}^{k} (\theta_j - b_i - \beta_r - d_m)}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} (\theta_j - b_i - \beta_r - d_m)} \tag{11}$$

ここで，d_kは評点$k-1$から評点kに遷移する困難度を表す．パラメータの識別性のために$\beta_1 = 0, d_1 = 0$を仮定する

Common stepモデルに代表される多相ラッシュモデルは，以降で紹介する項目反応モデルと比べて，比較的少ないパラメータ数で記述され，パラメータ推定値が安定的に得られやすいという特徴がある．さらに，FACETSやConQuestといった実用的な推定用ソフトウェアが普及していることから，多相ラッシュモデルはパフォーマンス評価データの分析に広く利用されてきた（Kassim, 2011; Myford & Wolfe, 2004; 野口・大隅, 2014; Thomas, 2005）．

多相ラッシュモデルでは，全ての課題で識別力が一定であること，また，全ての評価者について評価の一貫性の特性が一定であることが仮定される．しかし，一般に，パフォーマンス評価では，これらの仮定は成り立たないことが指摘されている (宇佐美, 2013a; Uto & Ueno, 2015)．そこで，この制約を緩めたモデルとして，多相ラッシュモデルをGPCMにより拡張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（1999）は，課題iにおける評価者rの評価の厳しさを表すパラメータρ_{ir}を付与したGPCMの拡張モデルを提案している．このモデルでは，評価者rが課題iにおける受験者jのパフォーマンスに評点kを与える確率を次式で定義する．

$$P_{ijrk} = \frac{\exp \sum_{m=1}^{k} [\alpha_i (\theta_j - \rho_{ir} - \beta_{im})]}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} [\alpha_i (\theta_j - \rho_{ir} - \beta_{im})]} \tag{12}$$

ここで，α_iは課題iの識別力を表し，β_{ik}は課題iにおいて評点$k-1$から評点kに遷移する困難度を表す．モデルの識別性のために，$\beta_{i1} = 0, \rho_{i1} = 0 : \forall i$を仮定する．

宇佐美（2010）は，評価者内/評価者間で評価が一貫している保証がないことを指摘し，これに対応する評価者パラメータを加えた以下のGPCM拡張モデルを提案している．

$$P_{ijrk} = \frac{\exp \sum_{m=1}^{k} [\alpha_i \alpha_r (\theta_j - (\beta_i + \beta_r) - d_{im} d_r)]}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} [\alpha_i \alpha_r (\theta_j - (\beta_i + \beta_r) - d_{im} d_r)]} \tag{13}$$

ここで，α_rは評価者rの評価の一貫性，β_iは課題iの位置パラメータ，d_{ik}は課題iにおける評点kに対する閾値パラメータ，d_rは評価者rによる評点のばらつきの大きさを表す．モデルの識別性のために，

$\prod_r \alpha_r = 1, \sum_r \beta_r = 0, \prod_r d_r = 1, d_{i1} = 0, \sum_{k=2}^{K} d_{ik} = 0 : \forall i$ を仮定する．このモデルでは，α_r により評価者の一貫性を考慮できるだけでなく，d_r により評価者の中心化傾向も考慮できる点が特徴といえる．これらのパラメータの解釈は5.1節で詳述する．

4.3. 評価者パラメータを付与したGRM

Ueno and Okamoto（2008）は，GRMに評価者パラメータを付与した以下のモデルを提案している．

$$P_{ijrk} = P^*_{ijrk-1} - P^*_{ijrk} \tag{14}$$

$$P^*_{ijrk} = \frac{\exp\left(\alpha_i\left(\theta_j - b_i - \epsilon_{rk}\right)\right)}{1 + \exp\left(\alpha_i\left(\theta_j - b_i - \epsilon_{rk}\right)\right)} \tag{15}$$

$$: k = 1, \cdots, K-1$$

ここで，ϵ_{rk}は評価者 r の評点 k に対する厳しさを表す．また，$P^*_{ijr0} = 1, P^*_{ijrK} = 0$ とする．評価の厳しさパラメータϵ_{rk}には順序制約$\epsilon_{r1} < \epsilon_{r2} < \cdots < \epsilon_{rK-1}$を課す．モデルの識別性のために，$\epsilon_{11} = -2.0$を仮定する．このモデルでは，$\epsilon_{rk}$により評価者の中心化傾向と尺度範囲の制限の特性を考慮できる．この特徴については5.1節で詳述する．

Uto and Ueno（2015）は，GRMに評価者の厳しさと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．

$$P_{ijrk} = P^*_{ijrk-1} - P^*_{ijrk} \tag{16}$$

$$P^*_{ijrk} = \frac{\exp\left(\alpha_i\alpha_r\left(\theta_j - b_{ik} - \epsilon_r\right)\right)}{1 + \exp\left(\alpha_i\alpha_r\left(\theta_j - b_{ik} - \epsilon_r\right)\right)} \tag{17}$$

$$: k = 1, \cdots, K-1$$

ここで，ϵ_rは，評価者 r の評価の厳しさを表し，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．また，$P^*_{ijr0} = 1, P^*_{ijrK} = 0$ とする．このモデルでは，課題の困難度パラメータ b_{ik} に順序制約 $b_{i1} < b_{i2} < \cdots < b_{iK-1}$を課す．モデルの識別性のために，$\alpha_{r=1} = 1, \epsilon_{r=1} = 0$を仮定する．

4.4. 階層評価者モデル（Hierarchical Rater Model）

以上で紹介したモデルは，従来の項目反応モデルに評価者パラメータを直接付与したモデルとして定義された．一方で，これとは異なるアプローチのモデル化として，階層評価者モデル（Hierarchical Rater Model: HRM）が提案されてきた（例えば，DeCarlo, Kim, & Johnson, 2011; Lu & Wang, 2006; Patz et al., 2002）．

このアプローチの特徴は，課題 i に対する受験者 j のパフォーマンスに対し，理想的な評点ξ_{ij}を仮定する点にある．HRMでは，課題 i に対する受験者 j のパフォーマンスに評価者 r が与える観測評点 x_{ijr}は，理想評点 ξ_{ij} に評価者特性が加味されて得られると仮定し，観測評点 x_{ijr} の生成過程を階層モデルとして定式化する．

例えば，Patz et al.（2002）は以下のHRMを提案している．

1) 課題 i における受験者 j のパフォーマンスに理想評点 $\xi_{ij} = k$を与える確率 P'_{ijk}を以下のPCMで定義する．

$$P'_{ijk} = \frac{\exp\sum_{m=1}^{k}[\theta_j - \beta_i - d_{im}]}{\sum_{l=1}^{K}\exp\sum_{m=1}^{l}[\theta_j - \beta_i - d_{im}]} \tag{18}$$

モデルの識別性のために，$d_{i1} = 0, \sum_{k=2}^{K} d_{ik} = 0 : \forall i$ を所与とする．

2) 理想評点 ξ_{ij}を所与として，課題 i に対する受験者 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．

$$P_{ijrk} \propto \exp\{\frac{-k + \xi_{ij} + \sigma_r}{2\psi_r^2}\} \tag{19}$$

ここで，σ_rは評価者 r の評価の厳しさ，$1/\psi_r^2$は評価者 r の評価の一貫性を表す．

HRMの特徴は，項目反応モデルを組み込んだ階層モデルとして観測評点の生成過程を表現した点にある．同様の階層モデル化を用いたアプローチとして，項目反応モデルを利用しない手法がデータサイエンス分野において提案されている（例えば，Goldin, 2012; Piech, Huang, Chen, Do, Ng, & Koller , 2013）．

例えば，Piech et al.（2013）は，以下のモデルを提案している．

$$\begin{aligned}
&\tau_{ir} \sim G(\alpha_0, \beta_0) \ \forall i, r \\
&b_{ir} \sim N\left(0, \frac{1}{\eta_0}\right) \ \forall i, r \\
&\xi_{ij} \sim N\left(\mu_0, \frac{1}{\gamma_0}\right) \ \forall i, j \\
&x_{ijr} \sim N\left(\xi_{ij} - b_{ir}, \frac{1}{\tau_{ir}}\right) \ \forall i, j, r
\end{aligned} \tag{20}$$

ここで，t_{ir}は課題 i における評価者 r の評価の一貫性，b_{ir}は課題 i における評価者 r の評価の厳しさ，$1/\eta_0$は評価者の厳しさパラメータb_{ir}の分布の分散，μ_0と$1/\gamma_0$は理想評点 ξ_{ij} の分布の平均と分散を表す．また，$G(\alpha_0, \beta_0)$は α_0, β_0 をパラメータとするガンマ分布を表す．

しかし，Piech et al.（2013）やGoldin（2012）のモ

表1. モデルの比較

モデル	考慮する特性		要因間の交互作用	パラメータ数
	課題	評価者		
MFRM	困難度	厳しさ	なし	$I+K+R+J-2$
Patz1999	識別力 困難度	厳しさ	評価者・課題間 課題・評点間	$I(K+R-1)+J$
Usami2010	識別力 困難度	厳しさ 一貫性 中心化傾向	課題・評点間	$IK+3(R-1)+J$
Ueno2008	識別力 困難度	厳しさ 中心化傾向 尺度範囲の制限	評価者・評点間	$2I+R(K-1)+J$
Uto2015	識別力 困難度	厳しさ 一貫性	課題・評点間	$IK+2(R-1)+J$
HRM2002	識別力 困難度	厳しさ 一貫性	課題・評点間 課題・受験者間	$I(K-1+J)+2R+J$

デルは，IRTを用いたHRMと異なり，受験者の能力を表すパラメータを持たない．本論では，課題と評価者の特性を考慮して受験者の真の能力を測定する問題を扱っており，受験者の能力を扱うことができないこれらのモデルは，本論の目的には適していない．

5. モデルの比較

本節では，これまでに紹介してきた評価者と課題パラメータを付与した項目反応モデルについて，それぞれの特徴を解説する．以降では，式（11）の多相ラッシュモデルをMFRM，式（12）のPatz and Junker（1999）のモデルをPatz1999，式(13)の宇佐美(2010)のモデルをUsami2010，式（14）（15）のUeno and Okamoto (2008) のモデルをUeno2008，式(16)(17)のUto and Ueno (2015) のモデルをUto2015，式 (18)（19）のPatz et al.（2002）のモデルをHRM2002と呼ぶ．

5.1. 評価者・課題特性パラメータの比較

表1に，各項目反応モデルの特徴を整理した．ここでは，各モデルで考慮される課題・評価者特性と，それらのパラメータによりモデル化される要因間の相互作用についてまとめた．

表1より，MFRMを除くすべてのモデルにおいて，課題の識別力と困難度パラメータが付与されていることがわかる．

評価者特性パラメータとしては，評価の厳しさパラメータがすべてのモデルに共通して採用されており，評価の一貫性パラメータはUsami2010とUto2015のみで採用されている．さらに，Usami2010とUeno2008では中心化傾向を，Ueno2008では尺度範囲の制限を考慮できる点が特徴的である．

評価の厳しさと一貫性パラメータは，3.3節で示したGRMの項目困難度と識別力と同様の解釈が可能である．以降では，Usami2010とUeno2008における中心化傾向と尺度範囲の制限の解釈について説明する．

Usami2010のモデルでは，パラメータ d_r によって中心化傾向を表現できる．図2では，Usami2010のモデルについて，d_r 以外のパラメータを固定したとき，$d_r = 0.3$ とした場合の項目特性曲線を左図に，$d_r = 1.3$ とした場合の項目特性曲線を右図に示した．図2では，横軸が受験者の能力 θ を，縦軸が能力 θ の受験者に対し評価者が評点 k を与える確率を示す．図2から，d_r が大きい場合，各評点への反応曲線が $\theta = 0$ から一様に離れ，結果として，$\theta = 0$ において最も反応確率が高い評点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とが確認できる．これにより，評価尺度の平均値付近の評点に評価が集中する中心化傾向を表現できていることがわかる．また，d_r が小さい場合には，各評点に対する反応曲線が $\theta = 0$ に一様に近づいていることがわかる．これは，中心化傾向とは反対に，極端な評点に評価が集中する極端化傾向と呼ばれる特性を表すと解釈できる．

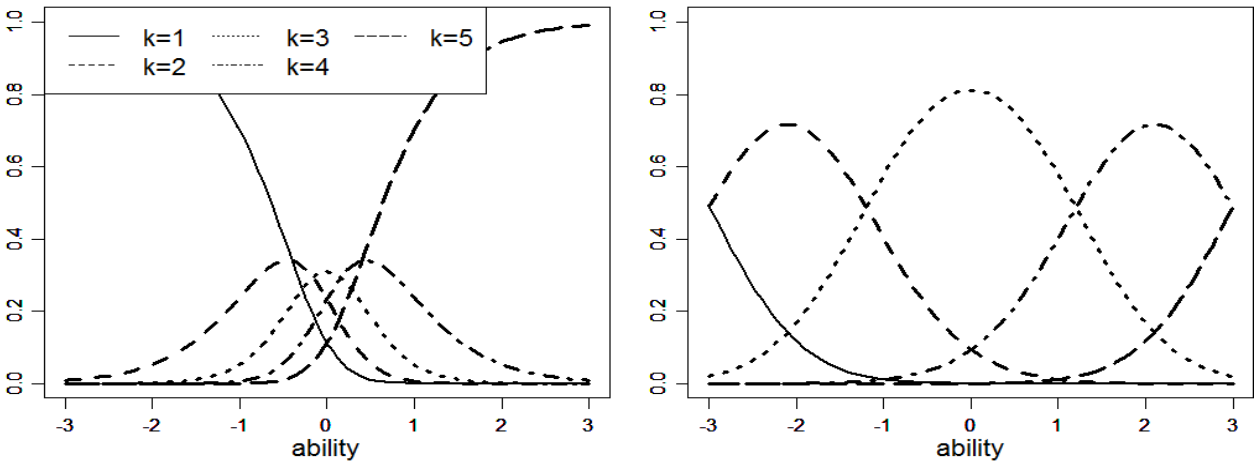

図 2. Usami2010 の項目特性曲線

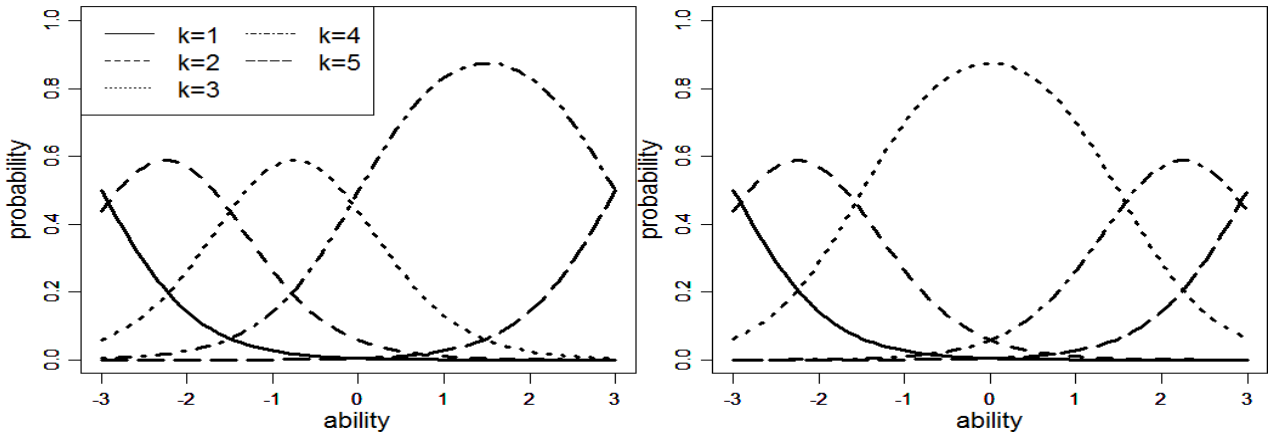

図 3. Ueno2008 の項目特性曲線

他方, Ueno2008では評価者と評点間の交互作用パラメータ ϵ_{rk}により, 中心化傾向と尺度範囲の制限を表現できる. 図3では, Ueno2008のモデルについて, ϵ_{rk} 以外のパラメータを固定したとき, $\epsilon_{rk} = \{-3.0, -1.5, 0.0, 3.0\}$とした場合の項目特性曲線を左図に, $\epsilon_{rk} = \{-3.0, -1.5, 1.5, 3.0\}$とした場合の項目特性曲線を右図に示した. 図3では, 横軸が受験者の能力 θを, 縦軸が能力 θ の受験者に対し評価者が評点 k を与える確率を示す. Ueno2008のモデルでは, $\epsilon_{rk=2} < 0 < \epsilon_{rk=3}$の条件で, $\epsilon_{rk=2}$と$\epsilon_{rk=3}$の差が大きくなるようにパラメータを設定することで, 図3右のように評価カテゴリ3への反応確率が高くなる中心化傾向を表現できる. また, $\epsilon_{rk=3}$と$\epsilon_{rk=4}$の差を大きくすることで, 図3左のように評点4を与えやすいといった尺度範囲の制限の特性についても表現できる.

他方, 各モデルが採用している要因間の交互作用に着目すると, Patz1999における評価者・課題間の交互作用パラメータ ρ_{ir}と, Ueno2008における評価者・評点間の交互作用パラメータ ϵ_{rk} が特徴的といえる.

Patz1999が採用している評価者・課題間の交互作用パラメータ ρ_{ir}は, 課題ごとに評価者の厳しさが異なることを表している. 課題間で, 評価に必要な背景知識や考え方が異なるような場合には, 課題ごとに評価者の評価基準が変化しうると考えらえる. このような課題を扱う場合には, Patz1999のモデルが適切であるといえる.

Ueno2008では, 他のモデルが, 各評点の与えられ方が課題に依存すると仮定するのに対し, 評点の与えられ方は評価者に依存すると仮定している. 一般に, 評価活動は, 課題に対応した評価基準に準拠して行われる. 評価基準には評点ごとに具体的な達成水準が示されていることから, 各評点 k の生起しやすさは, 課題に対応する評価基準に強く依存すると想定される. しかし, 評価者の主観的な判断を要する評価基準を用い

る場合や評価基準が存在しない場合，評価者が評価基準から逸脱した評価を行っていると想定される場合には，各評点の得られやすさは，評価者に強く依存すると考えられる．このような場合，上述した中心化傾向や尺度範囲の制限といった特性が生起しやすいと予想できるため，これらの特性を考慮できるUeno2008はデータへの適合がよくなると考えられる．

以上のように，要因間に複雑な交互作用を仮定したモデルや多様な特性パラメータを付与したモデルは，現実の評価者や課題の特性をより正確にモデル化できるため，データへの適合が良く，受験者の真の能力を精度よく測定できると考えられる．一方で，このようなモデルでは，受験者や課題，評価者数の増加に伴うパラメータ数の増加量が大きくなる．一般に，パラメータ数が増加し，パラメータ数に対するデータ数が減少すると，パラメータの推定精度が低下するため（Bishop, 2006），モデルの複雑化による表現力の向上とパラメータ推定精度はトレードオフの関係にあるといえる．すなわち，モデルを適用する場合には，扱うデータの特徴をモデルが適切に表現しているかと，パラメータ数に対して十分なデータ数が得られているかを考慮してモデルを選択することが重要となる．そこで，次節では，モデルのパラメータ数とパラメータ推定精度について議論する．

5.2. パラメータ数の比較

ここでは，各モデルのパラメータ数を比較するために，MFRM，Patz1999，Usami2010，Ueno2008，Uto2015，HRM2002におけるパラメータ数を表1に示した．

表1より，全ての条件において，MFRMが最もパラメータ数が少ないモデルであることがわかる．ただし，4.1節で述べたように，このモデルでは，全ての課題について識別力が一定，かつ全ての評価者について評価の一貫性の特性が一定であることを仮定している．しかし，これらの仮定は，一般的なパフォーマンス評価において成立しないため，MFRMでは，評価者と課題のバイアスを十分に取り除いた受験者の能力推定は困難であるといえる．

MFRM以外のモデルを比較すると，課題数に対して評価者数が多い場合，具体的には，評点数$K = 5$としたとき，課題数Iが2以上であり，評価者数Rと課題数Iが$2(R+1) > 3I$を満たす場合，Uto2015のパラメータ数が最小となる．他方，課題数が評価者数に対して多くなる場合，具体的には，$2(R+1) < 3I$の条件では，Ueno2008のパラメータ数が最小となる．表1に挙げたように，Ueno2008以外のモデルは，各評点kの得られやすさが課題に依存すると仮定し，課題・評点間の交互作用パラメータ（具体的には，d_{ik}やb_{ik}）を採用している．これらのパラメータは，課題数の増加に伴い評点数Kに比例して増加するが，Ueno2008では，課題と評点間の交互作用は仮定しないため，課題数の増加に伴うパラメータ数の増加が緩慢になる．

また，HRM2002では，各受験者に対する課題ごとの理想評点ξ_{ij}もデータから推定する必要があり，他のモデルと比べてパラメータ数が極端に多くなるという特徴がある．パフォーマンス評価では，評価者数や課題数に比べて受験者数が多くなることが一般的であるが，これらのモデルでは，受験者数の増加に伴い，推定すべき理想評点パラメータの数が急速に増加するため，実際の評価場面において高精度なパラメータ推定は期待できない．

5.3. パラメータ推定精度の比較

ここでは，MFRM，Patz1999，Usami2010，Ueno2008，Uto2015，HRM2002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し，パラメータ推定精度がモデルの複雑さに依存することを示す．

IRTのパラメータ推定は，一般に，周辺最尤推定（Marginal Maximum Likelihood Estimation: MMLE）やベイズ推定により実行される（Baker & Kim, 2004; 豊田, 2005）．ベイズ推定は，事後分布の最大値をパラメータの点推定値とする最大事後確率推定（Maximum a Posteriori: MAP）と，事後分布の期待値を点推定値とするExpected a Posteriori（EAP）推定に細分化できる．ベイズ推定は，周辺最尤推定に比べて頑健な推定を実現できることが知られている（Fox, 2010; 宇都・植野, 2015; Uto & Ueno, 2015）．そこで，本論では，多次元項目反応モデルなどの複雑な項目反応モデルにおいて，項目パラメータの推定に広く利用されてきた（Fox, 2010）マルコフ連鎖モンテカルロ法（Markov Chain Monte Carlo: MCMC; Brooks, Gelman, Jones, & Meng, 2011）によるEAP推定をパラメータ推定法として採用する．IRTにおけるMCMCアルゴリズムとしては，Patz and Junker (1999)が提案したMetropolis Hastings within Gibbs Samplingが広く利用される．評価者と課題のパラメータを付与した項目反応モデルにおけるMCMCアルゴ

リズムの詳細については，宇佐美（2010），宇都・植野（2015），Uto and Ueno（2015）を参照されたい．

ここでは各モデルのパラメータ推定精度を評価するために，以下のシミュレーション実験を行った．

1. 以下の5つの条件について，各モデルのパラメータ真値をランダムに生成し，それらのパラメータを所与として，各モデルから評価データを発生させた．

   A)　$I = 5, R = 10, J = 100$

   B)　$I = 5, R = 5, J = 100$

   C)　$I = 5, R = 5, J = 50$

   D)　$I = 5, R = 3, J = 50$

   E)　$I = 3, R = 3, J = 50$

これらの条件は，パフォーマンス評価では，受験者数に対し，課題数や評価者数が少ないことが一般的であることを踏まえて設定した．ただし，条件Aは，一つのパフォーマンスに対し10名の評価者が評価を行うと仮定しており，実践場面におい

表 2. パラメータ推定精度

		$J=100$ $R=10$ $I=5$		$J=100$ $R=5$ $I=5$		$J=50$ $R=5$ $I=5$		$J=50$ $R=3$ $I=5$		$J=50$ $R=3$ $I=3$	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
MFRM	b_i	0.066	0.055	0.084	0.059	0.132	0.088	0.116	0.084	0.143	0.087
	β_r	0.058	0.037	0.059	0.051	0.102	0.085	0.115	0.077	0.119	0.088
	d_k	0.046	0.036	0.061	0.053	0.109	0.081	0.118	0.100	0.106	0.075
	all	0.057	0.043	0.069	0.056	0.116	0.086	0.116	0.089	0.121	0.084
Patz1999	α_i	0.064	0.053	0.092	0.061	0.123	0.091	0.112	0.091	0.158	0.127
	β_{ik}	0.103	0.091	0.129	0.106	0.162	0.137	0.241	0.201	0.259	0.200
	ρ_{ir}	0.110	0.091	0.106	0.087	0.157	0.132	0.174	0.143	0.169	0.166
	all	0.105	0.090	0.115	0.095	0.155	0.131	0.203	0.180	0.219	0.188
Usami2010	α_i	0.113	0.097	0.102	0.074	0.116	0.093	0.171	0.113	0.207	0.199
	d_{ik}	0.129	0.100	0.092	0.065	0.158	0.151	0.173	0.141	0.187	0.301
	α_r	0.109	0.122	0.134	0.113	0.169	0.156	0.240	0.224	0.271	0.247
	β_i	0.110	0.073	0.077	0.051	0.161	0.110	0.144	0.111	0.130	0.113
	β_r	0.067	0.071	0.065	0.050	0.134	0.152	0.115	0.126	0.165	0.140
	d_r	0.116	0.082	0.092	0.077	0.142	0.183	0.104	0.083	0.207	0.173
	all	0.106	0.100	0.109	0.095	0.155	0.149	0.201	0.191	0.227	0.233
Ueno2008	α_i	0.034	0.027	0.050	0.033	0.068	0.043	0.110	0.070	0.075	0.042
	b_i	0.156	0.100	0.147	0.103	0.179	0.103	0.162	0.120	0.237	0.158
	ϵ_{rk}	0.111	0.087	0.130	0.106	0.180	0.140	0.188	0.149	0.206	0.173
	all	0.108	0.089	0.119	0.102	0.161	0.130	0.163	0.132	0.188	0.164
Uto2015	α_i	0.059	0.036	0.078	0.060	0.070	0.062	0.088	0.071	0.118	0.079
	b_{ik}	0.125	0.095	0.126	0.103	0.159	0.118	0.179	0.149	0.192	0.145
	α_r	0.054	0.047	0.062	0.032	0.099	0.063	0.082	0.060	0.130	0.078
	ϵ_r	0.096	0.072	0.074	0.059	0.085	0.095	0.138	0.091	0.164	0.107
	all	0.096	0.083	0.105	0.091	0.130	0.110	0.154	0.136	0.171	0.131
HRM2002	β_i	0.348	0.233	0.424	0.387	0.300	0.236	0.449	0.291	0.304	0.300
	d_{ik}	0.428	0.299	0.640	0.496	0.485	0.349	0.704	0.539	0.791	0.590
	σ_r	0.026	0.021	0.217	0.300	0.065	0.054	0.347	0.245	0.245	0.185
	ψ_r	0.070	0.047	0.141	0.116	0.134	0.125	0.291	0.193	0.457	0.287
	ξ_{ij}	0.029	0.169	0.298	0.473	0.160	0.367	0.425	0.546	0.497	0.609
	all	0.252	0.283	0.477	0.467	0.349	0.331	0.588	0.487	0.596	0.532

ては限定的な条件といえる．ここでは，課題数に対して評価数が増加する場合の性能を評価するためにこの条件を含めた．すべての条件において評点数$K=5$とした．

2. 生成したデータを用いて，MCMCにより，各モデルのパラメータ推定値を求めた．ここで，事前分布には，真値の生成に用いた分布と同一の分布を用いた．MCMCのバーンイン期間は10000とし，自己相関を考慮して10000時点から20000時点までのサンプルを100間隔で収集し，有効サンプルとした．推定には，Javaプログラムを用いた．
3. 手順2で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（以降，RMSE）を算出した．
4. 上記の手順を10回繰り返し，RMSEの平均と標準偏差を算出した．紙幅の都合から，ここでは，個々のパラメータについてではなく，パラメータ群ごと（例えば，$\alpha_i=\{\alpha_{i=1},\cdots,\alpha_{i=I}\}$）にRMSEの平均値と標準偏差を算出した．なお，RMSEは，値が小さいほど真値と推定値との誤差が小さく，推定精度が良いことを意味する．

表2に評価者と課題パラメータに関する結果を，表3に能力パラメータに関する結果を示した．表2，表3ではMeanの列がRMSEの平均を，SDの列が標準偏差を表す．また，表2において，allの行は，対象のモデルの全評価者・課題パラメータに対するRMSEの平均と標準偏差を表す．

表2から，パラメータの推定精度は受験者数が増加するほど改善することがわかる．また，パラメータ数が少ないモデルほど，パラメータの推定精度がよい傾向も確認できる．具体的には，すべての条件において，最もパラメータ数が少ない，MFRMの推定精度が最もよく，最もパラメータ数が多いHRM2002の精度が悪い．HRM2002では，受験者数が増加すると，パラメータξ_{ij}の数が急速に増加するため，受験者数が増加してもパラメータ推定の精度が大きくは改善しなかったと解釈できる．なお，HRM2002では，評価者数の増加に伴うパラメータ数の増加が比較的緩慢であることから，受験者数に対して評価者数が増加する条件Aにおいては，推定精度が改善している．MFRMとHRM2002以外のモデルについては，パラメータ数には若干の差異があるものの，受験者数がパラメータ数に対して比較的多いため，どのモデルも比較的高い推定精度を示しており，パラメータ数と推定精度の間に明らかな相関は認められなかった．

また，表3より，能力パラメータの推定精度は，課題数や評価者数が増加することで向上することが確認できる．モデル間で精度を比較すると，極端に課題・評価者パラメータの推定精度が悪いHRM2002では，能力パラメータの推定精度が著しく低下していることがわかる．それ以外のモデルについては，概ね同等の推定精度を示していることがわかる．

6. 実データを用いた信頼性の比較

ここでは，本論で紹介してきた項目反応モデルを実際のパフォーマンス評価データに適用し，各項目反応モデルを評価の信頼性の観点で比較する．

ここでは，eラーニング講義において，4つの課題に対して提出された40名分のレポートを5名の大学院生が5段階で採点したデータを用いた．

6.1. 記述統計量

実データの記述統計量として，表 4 に評価者・課題ごとの評点の平均値と分散，合計点との相関，各評点

表 3. 能力パラメータの推定精度

	$J=100$ $R=10$ $I=5$		$J=100$ $R=5$ $I=5$		$J=50$ $R=5$ $I=5$		$J=50$ $R=3$ $I=5$		$J=50$ $R=3$ $I=3$	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
MFRM	0.145	0.115	0.205	0.162	0.222	0.178	0.267	0.198	0.309	0.242
Patz1999	0.139	0.114	0.168	0.126	0.160	0.130	0.232	0.175	0.314	0.251
Usami2010	0.166	0.130	0.179	0.149	0.211	0.168	0.240	0.182	0.305	0.254
Ueno2008	0.164	0.117	0.211	0.161	0.214	0.151	0.273	0.206	0.348	0.280
Uto2015	0.134	0.101	0.177	0.125	0.193	0.147	0.226	0.164	0.331	0.265
HRM2002	0.371	0.299	0.379	0.290	0.385	0.295	0.398	0.315	0.479	0.359

表 4 各評価者・課題における評点の平均，分散，合計点との相関，および各評点の出現頻度

	平均	分散	合計点との相関	各評点の出現頻度				
				評点 1	評点 2	評点 3	評点 4	評点 5
評価者 1	1.738	0.855	0.859	16	55	56	21	12
評価者 2	1.875	0.811	0.828	9	59	48	31	13
評価者 3	2.237	0.649	0.768	2	28	72	46	12
評価者 4	2.069	0.750	0.894	8	36	71	27	18
評価者 5	1.956	0.544	0.795	5	29	102	16	8
課題 1	2.095	0.903	0.833	15	41	80	38	26
課題 2	1.890	0.863	0.544	4	66	88	32	10
課題 3	1.850	0.765	0.481	13	62	80	32	13
課題 4	2.065	0.838	0.480	8	38	101	39	14

の出現頻度を，表 5 に各課題における各評価者の評点の平均値を示した．ここで，評価者ごとの評点の平均値は「評価者の厳しさ」を，評点の分散は「中心化傾向」を，合計点との相関は「評価の一貫性」を示す指標と解釈できる（Saal, Downey, & Lahey, 1980）．評点の分散については，値が小さいほど中心化傾向が強いと解釈できる．また，各評点の出現頻度からは，尺度範囲の制限の傾向を読み取ることができる．同様に，課題ごとの評点の平均値は「課題の困難度」を，合計点との相関は「識別力」を表すと解釈できる．

表 4 より，本研究で用いた実データでは，各課題の評点と合計点との相関が，課題間で大きくばらついていることがわかる．このことから，課題ごとに識別力が大きく異なっていると推測できる．すなわち，このデータに対しては，課題の識別力をモデル化しているモデルの当てはまりがよいと予測できる．

さらに，評価者特性としては，評点の分散が評価者間で比較的ばらついていることが確認できる．このことから，中心化傾向が評価者間で異なっていると考えられる．さらに，各評価者における評点の出現頻度に着目すると，極端ではないものの尺度範囲の制限の傾向が読み取れる．具体的には，評価者 1 や評価者 2 は，評点 2 と 3 を付与しやすい傾向が確認でき，評価者 3 は評点 4 を若干与えやすい傾向が読み取れる．以上から，中心化傾向や尺度範囲の制限を考慮できる Usami2010 や Ueno2008 の利用が有効であると考えられる．

さらに，表 5 から，各評価者の評点の与え方が課題ごとに異なる傾向も確認できる．例えば，評価者 1 と 5 は，課題 3 に対して他の課題よりも低い評点を与える傾向があるが，他の評価者は課題 3 に対して他の課題と同程度かそれ以上の評価を与える傾向がある．このことから，課題と評価者の交互作用を仮定した Patz1999 も，データへの当てはまりが良く，信頼性の改善に有効であると解釈できる．

表 5. 各評価者の各項目に対する評点の平均値

	課題 1	課題 2	課題 3	課題 4
評価者 1	1.800	1.775	1.350	2.025
評価者 2	1.925	1.700	1.900	1.975
評価者 3	2.400	2.050	2.175	2.325
評価者 4	2.100	2.000	2.075	2.100
評価者 5	2.250	1.925	1.750	1.900

次節では，本節の考察を踏まえて，各モデルの信頼性について比較を行う．

6.2. 項目反応モデルによる能力評価の信頼性

ここでは，項目反応モデルによる能力評価の信頼性を示すために，実データを利用して，MFRM, Patz1999, Usami2010，Ueno2008，Uto2015，HRM2002について，以下の実験を行った．

1. 実データを用いて，各モデルの課題・評価者パラメータと受験者の能力パラメータを MCMC により推定した．パラメータの事前分布には，$\log \alpha_i \sim N(0.1,0.4)$，$\log \alpha_r \sim N(0.0,0.5)$，$\log \psi_r \sim N(0.0,0.4)$，$\beta_i$，$\beta_r$，$\beta_{ik}$，$b_i$，$\epsilon_r$，$\sigma_r$，$\rho_{ir}$，$d_{ik}$，$d_r$，$d_k$，$\theta_j \sim N(0.0,1.0)$ を用いた．ここで，$N(\mu,\sigma)$は，平均μ，標準偏差σ の正規分布を表す．ここでは，事前分布の分散を大きめの値に設定している．これはパラメータに対する事前知識が少ないことを反映している（宇佐美, 2010）．他

方, Ueno2008 と Uto2015 の評点に関するパラメータ ϵ_{rk}, b_{ik} の事前分布には，$K-1$ 次元正規分布を仮定する（Uto & Ueno, 2015）．この分布のパラメータである平均値ベクトルと共分散行列は, 評点パラメータ ϵ_{rk}, b_{ik} の順序制約（例えば，$b_{i1} < \cdots < b_{iK-1}$）を考慮して決定する必要がある．しかし，これらの適切な選択は容易ではないため，本実験では，階層ベイズモデルを用いたベイズ推定法（Fox, 2010; Uto & Ueno, 2015）により ϵ_{rk} と b_{ik} の事前分布のパラメータをデータから推定した．

2. 以下の条件でデータの一部を欠測させた．
   (条件1) 各受験者が提出した 4 つのレポートから，ランダムに 2 つのレポートを選択し，それらに対する評価データを欠測させた．
   (条件2) 各レポートに対する 5 人分の評価データの内，ランダムに選択した 3 人分の評価データを欠測させた．
3. 手順 2 の各条件で生成したデータと，手順 1 で推定した課題・評価者パラメータを所与として，各モデルにより受験者の能力パラメータを推定した．比較のために，評点の平均値も求めた．
4. 手順 3 で推定された能力パラメータと，手順 1 で推定された能力パラメータとの相関を求めた．相関にはピアソンの積率相関係数を用いた．平均点についても同様に相関を求めた．
5. 手順 2〜4 を，手順 2 で欠測させるデータをランダムに変更しながら10回繰り返した．
6. 得られた相関係数の平均と標準偏差を条件ごとに求めた．

この実験では，手順 1 において課題パラメータと評価者パラメータが高精度に推定され，モデルが評価者と課題の特性を適切に表現していれば，評価者や課題が変化しても安定して受験者の能力値が得られるため, 高い相関を示すと予想できる．なお，この実験では，手順 1 で完全データから推定した能力値を真値，手順 3 で欠測データから推定した能力値を観測値とみなし，それらの間の相関係数を求めており，この値は，能力の真値と観測値との相関で定義される信頼性インデックス（村木, 2011）に準じる指標と解釈できる．

実験の結果を表 6 に示す．表 6 では Mean の列が相関係数の平均を，SD の列が標準偏差を表す．表 6 から，HRM2002 を除き，項目反応モデルを利用して推定した能力値は，素点の平均点と比べて高い相関を示していることがわかる．このことから，項目反応モデルの利用は，評価の信頼性向上に有効であることが確認できる．HRM2002 は，パラメータ数が極端に多いことから，パラメータの推定精度が著しく低下するため，信頼性が低下したと解釈できる．

表6. 信頼性の評価結果

	条件1		条件2	
	Mean	SD	Mean	SD
MFRM	0.880	0.015	0.722	0.094
Patz1999	0.956	0.005	0.757	0.120
Usami2010	0.944	0.010	0.759	0.110
Ueno2008	0.930	0.015	0.736	0.068
Uto2015	0.936	0.013	0.747	0.129
HRM2002	0.807	0.039	0.655	0.070
平均点	0.878	0.016	0.718	0.095

また, 表6より, 本実験では, Usami2010 と Patz1999 が高い信頼性を示したことがわかる. 6.1 節で示したように，Usami2010 では，評価者の中心化傾向を考慮していること，Patz1999 では，評価者と課題の交互作用を考慮している点が，信頼性の改善に有効であったと解釈できる．

6.1 節の議論から, Ueno2008 も高い信頼性を示すと予想されたが，本実験では，Patz1999，Usami2010, Uto2015 より低い信頼性を示した. この原因としては, Ueno2008 が考慮していない，評価者の一貫性と課題・評点間の交互作用の特性の影響が考えられる. 表 4 より，極端ではないものの，評価者間で評価の一貫性にばらつきが確認でき，課題間での各評点の得られやすさにも特性差が認められる．本研究で扱った実データでは，これらの特性が信頼性に与える影響が大きかったため，Ueno2008 の信頼性が低下したと解釈できる．

また，実験手順 1 で述べたように，本実験では，Uto2015 と Ueno2008 のパラメータ b_{ik}，ϵ_{rk} について，事前分布のパラメータもデータから推定した．一般に，事前分布は分析者の主観を反映して決定されるが，不適切な事前分布を採用した場合，評価の信頼性が低下することが知られている（Uto & Ueno, 2015）．例えば，Ueno2008 のパラメータ ϵ_{rk} の事前分布として，平均値ベクトルを $\{-2.0, -0.75, 0.75, 2.0\}$，共分散行列を $\begin{bmatrix} 0.25, 0.16, 0.16, 0.16 \\ 0.16, 0.25, 0.16, 0.16 \\ 0.16, 0.16, 0.25, 0.16 \\ 0.16, 0.16, 0.16, 0.25 \end{bmatrix}$ とする $K-1$ 次元正規

分布を与え，本節と同様の実験を行ったところ，条件 2 において，平均点を下回る信頼性を示した．適切な事前分布の決定が難しい場合には，1) 複数の異なる事前分布を用いて推定を行い，性能が良い事前分布を採用する，2) 階層ベイズモデルを用いたベイズ推定法（Fox, 2010; Uto & Ueno, 2015）によりデータから事前分布のパラメータを推定する，などのアプローチにより，適切な事前分布を求めることが望ましい．

7. 総合考察

7.1. 各モデルの特徴と適用場面の考察

これまでに述べてきたように，項目反応モデルを用いてパフォーマンス評価の信頼性を改善するためには，1) 現実の評価場面において想定される評価者と課題の特性が適切にモデル化されていること，2) 得られたデータからモデルパラメータを高精度に推定できること，が重要となる．5.1 節で述べたように，この 2 点はトレードオフの関係にあると解釈できることから，状況に応じて適切なモデルを選択することが重要といえる．そこで，ここでは，これまでの議論を踏まえて各モデルの総合的な考察を行い，それぞれのモデルに適した評価場面について述べる．

多相ラッシュモデルの一種である MFRM の特徴は，すべてのモデルの中で最も少数のパラメータで記述される点である．このモデルでは，評価の信頼性低下を引き起こす評価者・課題特性として知られる，評価者の一貫性や中心化傾向，課題の識別力などをモデル化していないため，一般に実データへの当てはまりは悪いと考えられる．しかし，モデルが単純であるために，少数データからも安定したパラメータ推定が期待でき，十分な評価データ数が得られない場合には有効なモデルと解釈できる．

GPCMに評価者パラメータを加えたモデルの一つであるPatz1999の特徴は，課題ごとに評価者の厳しさが異なるという特性をモデル化している点にある．評価に必要な背景知識や考え方が課題ごとに異なると想定される場合には，このモデルの利用が適切であると考えられる．ただし，課題数と評価者数が共に増加する場合には，パラメータ数が極端に増加するため，このモデルでは，パラメータの推定精度が低下すると考えられる．

GRM 拡張モデルの一つである Ueno2008 は，各評点 k の得られやすさが，課題ではなく評価者に依存すると仮定しており，これにより評価者の中心化傾向と尺度範囲の制限の特性をモデル化できる点が特徴である．5.1 節で議論したように，一般には各評点の得られやすさは課題に対応する評価基準に依存すると考えられる．しかし，評価基準が曖昧または存在しない場合や，評価者が評価基準から逸脱した評価を行っていると想定される場合には，各評点の得られ方が評価者に強く依存すると考えられる．このような場合には，Ueno2008 が有効であると考えらえる．

Uto2015 の特徴としては，評価者数が増加する場合，具体的には，評点数 $K = 5$ としたとき，課題数 I が 2 以上であり，評価者数 R と課題数 I が $2(R+1) > 3I$ の条件を満たす場合，パラメータ数が MFRM に次いで少なくなる点が挙げられる．さらに，MFRM とは異なり，他の多くのモデルで共通して採用している特性パラメータを残しつつ，パラメータ数を軽減しているため，評価者数が増加する大規模評価環境のように，パラメータ数の増加によりパラメータ推定精度が低下する可能性がある場合に，有効なモデルと解釈できる．

Usami2010 の特徴としては，課題の識別力と困難度に加え，評価者の一貫性，厳しさ，中心化傾向を同時に考慮して受験者の能力を推定できる点が挙げられる．さらに，このモデルは，比較的少ないパラメータ数で記述されている点も特徴である．Usami2010 は，様々な評価場面に適用可能な汎用性が高いモデルと解釈できる．

HRM2002 の特徴は，各課題に対する受験者のパフォーマンスに理想的な評点が存在すると仮定する点にある．しかし，これらのモデルでは，受験者数の増加に伴いパラメータ数が急速に増加し，パラメータの推定精度が著しく低下することが問題となる．現実のパフォーマンス評価では，課題や評価者数に対して受験者数が多くなることが一般的であるため，この性質は好ましくない．一方，これらのモデルでは，評価者数 R の増加に伴うパラメータ数の増加が比較的緩慢であるため，受験者数や課題数に対して評価者数が多い場合には利用可能といえる．しかし，実践場面においてこのような条件は一般的でないため，適用可能な場面は限定的である．

7.2. 課題数と評価者数が信頼性に与える影響の考察

1 節において，評価の信頼性を改善するためには，

適切な評価者数と課題数の選択が重要であることを述べた．ここでは，評価者数と課題数がIRTの信頼性に与える影響について，一般化可能性理論（Generalizability Theory; Cronbach, Nageswari, & Gleser, 1963; Brennan, 2000）に基づく既存研究の結果と対比させて議論する．

平井（2006, 2007）や宇佐美（2011, 2013b, 2015）は，一般化可能性理論を用いて，記述式テストの評価データを分析し，評点の分散に占める誤差成分の分散の割合として，評価者の成分よりも課題の成分が大きい傾向があることを報告している．これは，パフォーマンス評価の評点が，評価者に比べて課題の特性に強く依存する傾向があることを意味する．このことから，評価の信頼性を改善するためには，評価者数よりも課題数を優先して増加させ，課題特性の影響を軽減することが効率的であると指摘されている（宇佐美，2011, 2015）．

しかし，IRTにおける信頼性の改善には，課題数の増加が必ずしも効率的であるとは解釈できない．以降では，これについて議論するために，まずIRTの信頼性係数を導入する．

6.2節では，評価の信頼性を，信頼性インデックス（村木, 2011）に準ずる方法で評価した．一方で，IRTでは，情報量関数を用いて評価の信頼性を推定する手法も提案されており（豊田，1989），パフォーマンス評価の信頼性分析にも既に活用されている（Nguyen et al., 2015; 渡部・平井，1993）．

IRTにおける信頼性係数ρ_Iは，古典的テスト理論の拡張として次式で定義される（豊田，1989）．

$$\rho_I = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_e^2} \tag{21}$$

ここで，σ_θ^2は受験者の能力分布$g(\theta)$の分散，σ_e^2は次式で定義される能力推定の誤差分散である．

$$\sigma_e^2 = \int g(\theta)/I(\theta)\, d\theta \tag{22}$$

ここで，$I(\theta)$は項目反応モデルの情報量関数を表す．

IRTでは，能力分布$g(\theta)$に標準正規分布を仮定することが一般的である．したがって，$\sigma_\theta^2 = 1$となり，信頼性係数は以下で表される．

$$\rho_I = \frac{1}{1+\sigma_e^2} = \frac{1}{1+\int g(\theta)/I(\theta)\, d\theta} \tag{23}$$

式（23）から，IRTでは，情報量関数$I(\theta)$がθ全域に対して大きくなるほど，また，$I(\theta)$が受験者の能力分布$g(\theta)$に近い形状をとるほど，測定誤差σ_e^2が小さくなり，信頼性が高く推定されることがわかる．

本論で紹介したIRTにおける情報量関数は，課題iと評価者rが能力θ_jの受験者に対して与える情報量$I_{ir}(\theta_j)$の，全評価者・課題に関する総和として以下で定義される．

$$I(\theta_j) = \sum_i \sum_r I_{ir}(\theta_j) \tag{24}$$

$$I_{ir}(\theta_j) = -\mathrm{E}\left[\frac{\partial^2 \log P_{ijrk}}{\partial \theta_j^2}\right] = \sum_k I_{irk}(\theta_j) P_{ijrk} \tag{25}$$

$$I_{irk}(\theta_j) = \frac{\partial^2 \log P_{ijrk}}{\partial \theta_j^2} \tag{26}$$

式（24）から，情報量は課題数と評価者数の増加に伴って増加する性質を持つことがわかる．前述のように，IRTでは，情報量が大きくなるほど信頼性が高く推定されるため，課題数と評価者数が増加すると信頼性が改善されることがわかる．これは，一般化可能性理論の信頼性と一致する性質である．

一方で，一般化可能性理論では，誤差分散が大きい要因，すなわち課題を増加させると効率よく信頼性を改善できると述べた．しかし，IRTの信頼性においては，これは必ずしも成立しない．IRTにおける信頼性を効率よく改善するためには，受験者の能力分布$g(\theta)$における平均値$\theta = 0$周辺の能力を精度よく識別できる評価者や課題を増やすことが重要であり，課題と評価者に優先順位はないと考えられる．

例で説明する．図4に，異なる特性を持つ二つの課題について，Uto2015の項目特性曲線P_{ijrk}と情報量関数$I_{ir}(\theta)$，測定誤差関数$g(\theta)/I_{ir}(\theta)$を示した．課題パラメータには，左図では$\alpha_i = 1.3$, $\boldsymbol{b_{ik}} = \{-3.0, -2.0, -1.0, 0.0\}$，右図では$\alpha_i = 1.8$, $\boldsymbol{b_{ik}} = \{-1.5, -0.7, 0.7, 1.5\}$を所与とした．図4左は，識別力がやや低く，全体として難易度が低い異質な課題の例である．この課題では，情報量が全体として低く推定され，誤差関数も全体として高い値を示していることがわかる．他方，図4右の課題は，困難度が全体に均等に配置されており，識別力も高いことから好ましい特性の課題と解釈できる．この課題では，情報量が全体として高く，測定誤差は小さく推定されている．それぞれの課題について，信頼性係数を算出すると，左図の課題では0.68，右図の課題では0.84となる．このように，IRTでは，好ましい特性の課題や評価者は信頼性向上への寄与が大きく，異質な課題や評価者は寄与が小さくなる．

以上から，IRTにおける信頼性の改善には，必ずしも課題数の増加が効率的とは言えず，追加される評価

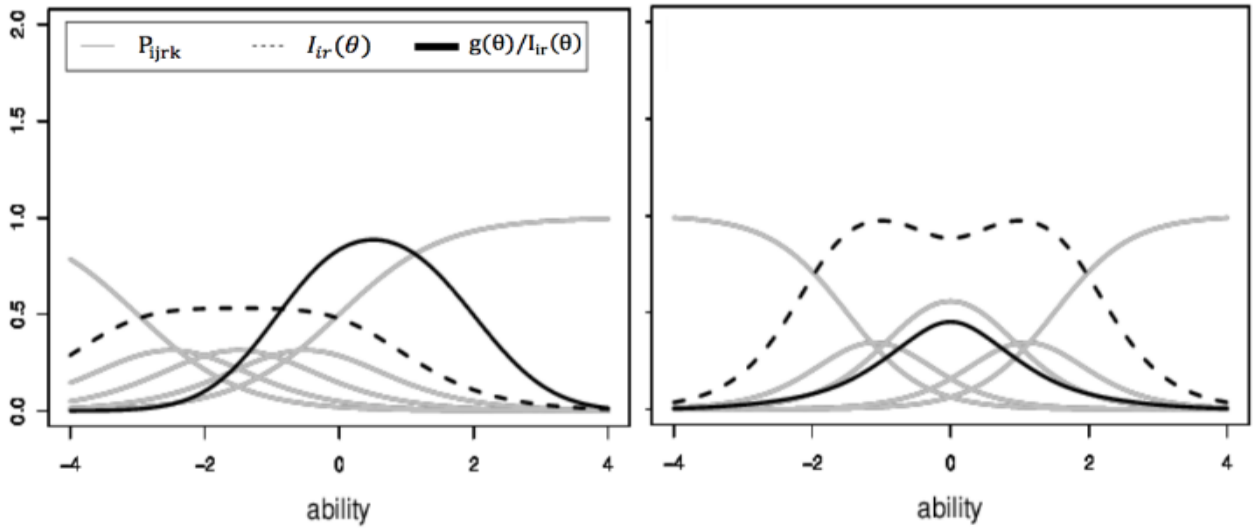

図 4. IRT における情報量関数と信頼性

者あるいは課題の質が強く影響を与えることがわかる．なお，評価者や課題の特性が信頼性係数 ρ_I に与える影響は，モデルが採用している評価者・課題パラメータに依存して異なりうる．しかし，能力分布の平均値付近を精度よく識別しない異質な課題や評価者が信頼性の改善に大きく寄与しないという性質は，式（23）の信頼性係数の定義から明らかであり，モデルに依存しない性質であると解釈できる．

8. まとめと今後の課題

本論では，パフォーマンス評価の信頼性を改善する手法として，評価者と課題の特性パラメータを付与した項目反応モデルを紹介し，各モデルの特徴を比較した．

本論では，まず，パフォーマンス評価の信頼性低下を引き起こすバイアス要因として，課題と評価者の特性について概説した．次に，課題と評価者の特性を考慮して受験者の能力を推定できる項目反応モデルについて，既存モデルを概説した．さらに，各項目反応モデルで採用しているパラメータとそれらの特徴について解説し，パラメータ数とパラメータ推定精度について，モデル間の比較を行った．さらに，実際のパフォーマンス評価への適用を通して，各モデルにおける評価の信頼性を比較した．最後に，総合考察として，本論で議論したモデルの特徴を整理し，各モデルに適した評価場面について議論を行った．また，IRTにおける信頼性改善において，課題数と評価者数のどちらを増加させることが効率的かについて議論し，これらは一意には定まらないことを述べた．ただし，妥当性の観点も考慮した場合，評価者よりも課題を増加させることが重要である可能性がある．妥当性の問題もパフォーマンス評価の重要課題の一つであり（宇佐美，2015），今後は妥当性も考慮した項目反応モデルの比較・分析を検討したい．

1節で述べたように，評価の信頼性を改善するためには，評価者数や課題数，課題内容や評価基準を十分に吟味することが不可欠である．一方で，評価環境を十分に吟味した上で，課題や評価者のバイアスが残ると想定される場合には，本論で紹介した項目反応モデルの活用が有効であると考えられる．今後，評価環境のデザインと項目反応モデルの活用の両面から，パフォーマンス評価の信頼性に関する様々な知見が積み重ねられることを期待したい．

また，本論で紹介した項目反応モデルは，テストの文脈のみでなく，教育場面における評価データの分析や，オンラインショッピングにおけるレイティングの分析，クラウドソーシングの品質評価など様々な場面に適用可能である．教育分野においては既にいくつかの応用例が提案なされているが（Nguyen et al., 2015; 宇都, 2015; Uto & Ueno, 2015; 山本・宇都・植野, 2015），今後は様々な領域において活発な応用が行われることを期待する．

謝辞

本研究はJSPS科研費15K16256の助成を受けたものです．

参考文献

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561-573.

Baker, F. B. & Kim, S. H. (2004). *Item response theory: parameter estimation techniques*. Boca Raton, FL: CRC Press.

Bishop, C. M. (2006). *Pattern recognition and machine learning (Information science and statistics)*. New York: Springer-Verlag.

Brennan, R. L. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement*, *24*, 339-353.

Brooks, S., Gelman, A., Jones, G. & Meng, X. (2011). *Handbook of Markov chain Monte Carlo*. Boca Raton, FL: CRC Press.

Cronbach, L. J., Nageswari, R. & Gleser, G. (1963). Theory of generalizability: A liberation of reliability theory. *The British Journal of Statistical Psychology, 16*, 137-163.

DeCarlo, L. T. (2005). A model of rater behavior in essay grading based on signal detection theory. *Journal of Educational Measurement, 42*, 53-76.

DeCarlo, L. T., Kim, Y. K. & Johnson, M. S. (2011). A hierarchical rater model for constructed responses, with a signal detection rater model. *Journal of Educational Measurement, 48*, 333-356.

De Gruijter, D. N. M. (1984). Two simple models for rater effects. *Applied Psychological Measurement,* 8, 213-218.

独立行政法人情報処理推進機構 (2013). IT パスポート試験. https://www3.jitec.ipa.go.jp/JitesCbt/html/about/range.html, 閲覧日 2015 年 11 月 29 日.

Fox, J. P. (2010). *Bayesian item response modeling: Theory and applications*. New York: Springer-Verlag.

Goldin, I. M. (2012). *Accounting for peer reviewer bias with Bayesian models*. Paper presented at the Proceedings of the Workshop on Intelligent Support for Learning Groups at the 11th International Conference on Intelligent Tutoring Systems, Chania, Greece.

平井洋子 (2006). パフォーマンス・アセスメントによる高次思考能力の測定. 平成 14-16 年度科学研究費成果報告書.

平井洋子 (2007). 主観的評定における評定基準, 評定者数, 課題数の効果について: 一般化可能性理論による定量的研究. 人文学報, *380*, 25-64.

石井隆稔・植野真臣 (2015). e テスティングにおける複数等質テスト自動構成手法の展望. 日本テスト学会誌, *11*, 131-149.

Kassim, N. L. A. (2011). Judging behaviour and rater errors: an application of the many-facet Rasch model. *GEMA Online Journal of Language Studies, 11(3)*, 179-197.

Kim, S. (2012). A note on the reliability coefficients for item response model-based ability estimates. *Psychometrika, 77*, 153-162.

公益社団法人医療系大学間共用試験実施評価機構 (2014). 臨床実習開始前の「共用試験」第 12 版 (平成 26 年度). http://www.cato.umin.jp/e-book/12/index.html, 閲覧日 2016 年 4 月 4 日.

高大接続システム改革会議 (2015). 高大接続システム改革会議「中間まとめ」. http://www.mext.go.jp/b_menu/shingi/chousa/shougai/033/toushin/__icsFiles/afieldfile/2015/09/15/1362096_01_2_1.pdf, 閲覧日 2016 年 3 月 31 日.

Linacre, J. M. (1989). *Many-faceted Rasch measurement*. Chicago, IL: MESA Press.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Mahwah, NJ: Lawrence Erlbaum Associates.

Lu, Y. & Wang, X. (2006). A hierarchical Bayesian framework for item response theory models with applications in ideal point estimation. *The society for political methodology*, 1-19.

Lurie, S. J., Nofziger, A. C., Meldrum, S., Mooney, C. & Epstein, R. M. (2006). Effects of rater selection on peer assessment among medical students. *The International Journal of Medical Education, 40*, 1088-1097.

Masters, G. (1982). A rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

松下佳代 (2012). パフォーマンス評価による学習の質の評価 ― 学習評価の構図の分析に基づいて. 京都大学高等教育研究, *18*, 75-114.

Matteucci, M. & Stracqualursi, L. (2006). Student

assessment via graded response model. *Statistica*, *66*, 435-447.

文部科学省中央教育審議会 (2014). 新しい時代にふさわしい高大接続の実現に向けた高等学校教育、大学教育、大学入学者選抜の一体的改革について (答申). http://www.mext.go.jp/b_menu/shingi/chukyo/chukyo0/toushin/__icsFiles/afieldfile/2015/01/14/1354191.pdf, 閲覧日 2016 年 4 月 4 日.

Muraki, E. (1997). A generalized partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 153-164). New York: Springer-Verlag.

村木英治 (2011). 項目反応理論 (シリーズ<行動計量の科学>). 朝倉書店.

Muraki, E., Hombo, C. & Lee, Y. (2000). Equating and linking of performance assessments. *Applied Psychological Measurement*, *24*, 325–337.

Myford, C. M. & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*, 386–422.

Myford, C. M. & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement, 5*, 189–227.

Nguyen, T., Uto, M., Abe, Y. & Ueno, M. (2015). *Reliable peer assessment for team project based learning using item response theory.* Paper presented at the Proceedings of the 23rd International Conference on Computers in Education, Hangzhou, China.

日本テスト学会 (2007). テスト・スタンダード-日本のテストの将来に向けて. 金子書房.

野口裕之・大隅敦子 (2014). テスティングの基礎理論：基礎理論から最先端理論まで. 研究社.

Patz, R. J. & Junker, B. W. (1999). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, *24*, 342-366.

Patz, R. J., Junker, B. W., Johnson, M. S. & Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics, 27*, 341-384.

Piech, C., Huang, J., Chen, Z., Do, C., Ng, A. & Koller, D. (2013). *Tuned models of peer assessment in MOOCs*. Paper presented at the Proceedings of the Sixth International conference on Educational Data Mining, Tennessee, USA.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: The University of Chicago Press.

Saal, F. E., Downey, R. G. & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, *88*, 413-428.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, *17*, 1-100.

Thomas, E. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly*, *2*, 197-221.

豊田秀樹 (1989). 項目反応モデルにおける信頼性係数の推定法. 教育心理学研究, *37,* 283-285.

豊田秀樹 (2005). 項目反応理論 [理論編]. 朝倉書店.

豊田秀樹 (2013). 項目反応理論 [中級編]. 朝倉書店.

Ueno, M. & Okamoto, T. (2008). *Item response theory for peer assessment*. Paper presented at the Proceedings of the Eighth IEEE International Conference on Advanced Learning Technologies, Santander, Cantabria, Spain.

植野真臣・荘島宏二郎 (2010). 学習評価の新潮流. 朝倉書店.

宇佐美慧 (2010). 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, *58*, 163-175.

宇佐美慧 (2011). 小論文評価データの統計解析-制限字数を考慮した測定論的課題の検討. 行動計量学, *38*, 33-50.

宇佐美慧 (2013a). 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, *9*, 145-164.

宇佐美慧 (2013b). 論述式テストを通した評価と選抜

の信頼性に関わる諸要因の影響力についての定量的比較検討. 日本教育工学会論文誌, *36*, 451-464.

宇佐美慧 (2015). 論述式テストの測定論的問題再考-主要な論点の整理とその現実的解決のために. 日本テスト学会第 13 回大会発表論文抄録集, 56-59.

宇都雅輝 (2015). ピアアセスメントのための項目反応理論を用いた評価者選択. 教育システム情報学会第 40 回全国大会講演論文集, 136-137.

宇都雅輝・植野真臣 (2015). 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. 電子情報通信学会論文誌 *D*, *J98-D*, 3-16.

Uto, M. & Ueno, M. (2015). Item response theory for peer assessment. *IEEE Transactions on Learning Technologies*, PrePrint.

山本美紀・宇都雅輝・植野真臣 (2015). 項目反応理論によるルーブリックの自己評価力への影響分析-評価者特性と目標志向性, 学習観, 動機づけ, 学習方略に着目して. 日本教育工学会第 31 回全国大会講演論文集, 459-460.

Wang, Z. & Yao, L. (2013). The effects of rater severity and rater distribution on examinees' ability estimation for constructed response items. *ETS Research Report Series*, *2*, 1-22.

渡部洋・平井洋子 (1993). 段階反応モデルによる小論文データの解析. 東京大学教育学部紀要, *33*, 143-150.