論　文

# ピアアセスメントの低次評価者母数を持つ項目反応理論

宇都　雅輝[†]　　植野　真臣[††]

Item Response Theory with Assessors' Lower Order Parameters of Peer Assessment

Masaki UTO[†] and Maomi UENO[††]

**あらまし**　近年，構成主義における学習評価法としてピアアセスメントが注目されている．ピアアセスメントでは，評価の信頼性が評価者の特性に依存する問題が指摘されている．この問題を解決するアプローチの一つとして，評価者の特性を表すパラメータを付加した項目反応理論が提案されてきた．しかし，ピアアセスメントでは，評価者数が学習者数と同程度まで増加するため，パラメータ数に対してデータ数が少なくなり，既存モデルでは高精度なパラメータ推定が期待できない．そこで，本論では，通常の項目反応理論について，できる限り評価者パラメータ数が少なくなるように評価者パラメータを付加した，ピアアセスメントのための新たな項目反応理論を提案する．提案手法の特徴は次の通りである，(1) 既存モデルより高精度なパラメータ推定が可能である．(2) 評価者特性として評価の一貫性と厳しさの影響を反映した学習者の能力推定が可能である．(3) 学習者の正確な能力推定が期待できる．さらに，本論では，シミュレーション実験および被験者実験により提案手法の有効性を示す．

**キーワード**　ピアアセスメント，項目反応理論，信頼性，評価者特性，パラメータ推定

## 1. はじめに

近年，構成主義における学習評価法として，学習者同士による学習成果物の相互評価法，ピアアセスメント [1] が注目されている [2]．ピアアセスメントの利点として，以下のような点が報告されている．

（1） 学習者に評価者の役割を与えることで，モチベーションを向上できる [3].

（2） 学習者が自己の学習に責任感を持ち，自律的に学習を行うことができる [2] [3].

（3） 評価活動が学習の一環に含まれるため，失敗から学ぶ機会が得られる [3].

（4） 評価スキルやディスカッションスキルなど，より社会的で実践的なスキルを育成できる [3] [4].

（5） 他者を評価することで，他者の成果から学ぶとともに，学習者の内省を促進できる [2] [3] [4].

（6） 表面的でない深い学習が促進される [3] [4].

（7） 教師不在でも迅速に大量かつ多様なフィードバックを与えることができ，効果的な学習が促進される [4]．また，経歴が似た学習者からのコメントは理解しやすい [2].

（8） 成人学生の場合，教師一人で採点を行うよりも多人数で評価を行ったほうが信頼性が高くなる. [2]

このような利点を持つことから，ピアアセスメントを支援するシステムがこれまでに多数開発されてきた [2], [5]〜[14].

一方，ピアアセスメントでは，評価の信頼性が評価者の特性に依存する問題が指摘されている [2] [7] [11] [15]. 具体的には，以下のような評価者の特性が信頼性の低下を引き起こすことが報告されている [16].

（1） 評価者間で評価の甘さ/厳しさが存在すること.

（2） 評価者間，および，評価者内で評価基準が一貫している保証がないこと.

同様の課題は，複数の評価者による論述式テストの採点などでも指摘されており [17]- [18]，この問題を解決するために，評価者特性を表すパラメータを付加した項目反応理論が提案されてきた [16] [19] [20]．例えば，Patz et.al [19] は，多値型項目反応モデルの一つである一般化部分採点モデル [21] に対して，課題に対

[†] 長岡技術科学大学，新潟県
Nagaoka University of Technology, 1603-1 Kamitomioka, Nagaoka-shi, Niigata 940-2188, Japan
[††] 電気通信大学大学院 情報システム学研究科，東京都
Graduate School of Information Systems, The University of Electro-Communications, 1-5-1 Chofugaoka, Chofu-shi, Tokyo 182-8585, Japan

する評価者の評価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．また，宇佐美 [16] 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の評価の一貫性パラメータを加えた一般化部分採点モデルの拡張モデルを提案している．ここでは，評価者の特性のみでなく，文字の美醜効果や性別など，受験者側のバイアス要因を推定結果に反映させる手法も提案している．これらのモデルは，Linacre [20] が提案した多層ラッシュモデルの拡張と解釈できる．

一方，ピアアセスメントにおける項目反応理論としては，植野ら [2] があるに留まる．ここでは，段階反応モデル [22] に対し，各評点に対する評価の厳しさを表す評価者パラメータを付加したモデルを提案している．ここでは，評価の厳しさパラメータを用いて，評価の一貫性を表す指標を求める手法も提案している．

しかし，ピアアセスメントでは，評価者数が学習者数と同程度まで増加するため，これらのモデルでは，パラメータ数に対するデータ数が少なくなり，高精度なパラメータ推定が期待できない．

この問題を解決するために，本論では，通常の項目反応理論について，できる限り評価者パラメータ数が少なくなるように，評価者の「評価の一貫性」と「評価の厳しさ」を表すパラメータを付加した，ピアアセスメントのための新たな項目反応理論を提案する．提案手法の特徴は以下の通りである．

（1） 既存手法よりも高精度なパラメータ推定が可能である．

（2） 評価者特性として評価の一貫性と厳しさの影響を反映した学習者の能力推定が可能である．

（3） 学習者の正確な能力推定が期待できる．

さらに，本論では，シミュレーション実験および被験者実験により提案手法の有効性を示す．

## 2. ピアアセスメント

筆者らの一人が開発してきた LMS「Samurai」では，ピアアセスメント機能を持つ掲示板システムを搭載している．学習者は，図 1 のように自身の学習成果物や意見などを投稿できる．さらに，他の学習者は，投稿された成果物に対して評価やコメントを付与できる．図 1 は，e ラーニング授業で提示された課題に対して，学習者がレポートを投稿した画面である．図 1 の下半分に表示される掲示板システムには，このレポートに対する他の学習者からの意見が書き込まれている．また，画面左上に提示されている 5 つのボタンは，ピアアセスメントのための評価ボタンである．評価ボタンは，-2（非常に悪い），-1（やや悪い），0（普通），1（やや良い），2（非常に良い）が用意されている．レポートを提出した学習者は，これらの評価や意見を踏まえて成果物を修正する．

このピアアセスメントシステムから得られる評価データ $\boldsymbol{U}$ は，学習者 $j$（$j = 1, \cdots, J$）の課題 $i$（$i = 1, \cdots, I$）に対する評価者 $r$（$r = 1, \cdots, R$）の評価カテゴリ $k$（$k = 1, ..., K$）で構成される．このとき，評価ボタンによる得点-2，-1，0，1，2 を順序尺度 1，2，3，4，5 に線形変換する．すなわち，

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

と定義できる．

このデータは,「学習者」×「課題」×「評価者」の三相データとなることがわかる．本論では，このようなデータに対して項目反応理論を適用することを想定する．

## 3. 項目反応理論 *(Item Response Theory)*

項目反応理論は，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．項目反応理論の利点として，以下のような点が挙げられる．

（1） 推定精度の低い異質項目の影響を小さくして能力推定を行うことができる．

（2） 異なる項目への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．

項目反応理論はこれまで正誤判定問題や多肢選択式問題などの客観的テストへの利用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形成的評価への応用も進められている [23] [24].

本論で扱うようなリッカート型データのための多値型項目反応モデルとしては，段階反応モデル：Graded Response Model [22](以下，GRM) と一般化部分採点モデル：Generalized Partial Credit Model [21](以下，GPCM) が知られている．以降では，これらの項目反応モデルについて詳述する．


図 1　ピアアセスメントシステム
Fig. 1　Peer Assessment System

### 3.1　段階反応モデル (GRM)

GRM は，学習者 $j$ 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．

$$P_{ijk} = P^*_{ijk-1} - P^*_{ijk} \tag{2}$$

$$\begin{cases} P^*_{ijk} = \frac{1}{1+\exp\left(-\alpha_i(\theta_j - b_{ik})\right)} & k = 1, \cdots, K-1 \\ P^*_{ij0} = 1 \\ P^*_{ijK} = 0 \end{cases} \tag{3}$$

ここで，$\alpha_i$ は項目 $i$ の識別力パラメータ，$b_{ik}$ は項目 $i$ のカテゴリ $k$ の難易度パラメータ，$\theta_j$ は学習者 $j$ の能力パラメータを表す．ただし，$b_{i1} < b_{i2} < \cdots < b_{ik} < \cdots < b_{iK-1}$ と制約される．

例として，$K = 5$，$\alpha_i = 1.0$，$b_{i1} = -3.0$，$b_{i2} = -1.5$，$b_{i3} = 0.0$，$b_{i4} = 3.0$ とした際の，式 (2) で表される項目反応曲線を図 2 に示す．図 2 では横軸に学習者の能力 $\theta_j$，縦軸にその学習者がカテゴリ $k$ と反応する確率を示している．図より，能力が低いほど低いカテゴリへの反応確率が高くなり，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．

### 3.2　一般化部分採点モデル (GPCM)

GPCM は，部分採点モデル Partial Credit Model [25](以下，PCM) を一般化し，項目間で異なる識別力パラメータを導入した項目反応モデルであり，反応関数は次式で与えられる．

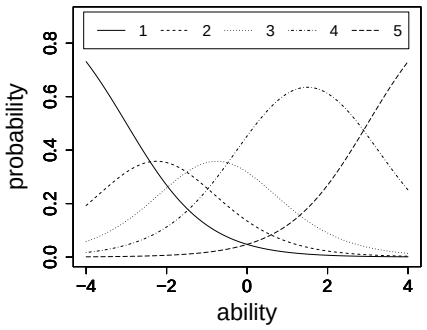



図 2　GRM の項目反応曲線
Fig. 2　Item characteristic curves of the graded response model

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

ここで，$\beta_{ik}$ はステップパラメータと呼ばれ，項目 $i$ においてカテゴリ $k-1$ からカテゴリ $k$ に遷移する難しさを表す．ただし，$\beta_{i1} = 0.0$ とする．

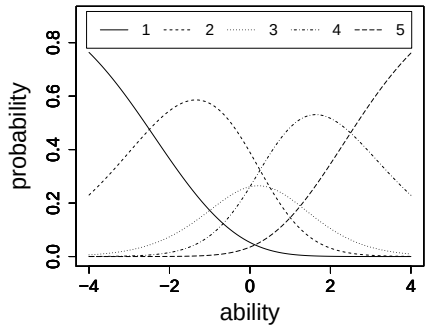



図 3　GPCM の項目反応曲線
Fig. 3　Item characteristic curves of the generalized partial credit model

GPCM は，ステップパラメータ $\beta_{ik}$ を $\beta_i - d_{ik}$ と分解し，次のように反応関数を与えることもある．

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

このとき，$\beta_i$，$d_{ik}$ はそれぞれ位置パラメータ，閾値パラメータと呼ばれる．

GPCM において，$\alpha_i = 1.0$ とすると PCM と一致する．さらに，PCM において，$d_{ik}$ が項目間で等しいと仮定する，すなわち，$\cdots = d_{i-1,k} = d_{ik} = d_{i+1,k} = \cdots$ としたモデルは評定尺度モデル [26] として知られている．

GPCM の例として，カテゴリ数 $K = 5$，$\alpha_i = 1.0$，$\beta_{i2} = -2.5$，$\beta_{i3} = 0.5$，$\beta_{i4} = 0.0$，$\beta_{i5} = 2.5$ とした際の，式 (4) で表される項目反応曲線を図 3 に示す．図 3 では，横軸に学習者の能力 $\theta_j$，縦軸にその学習者がカテゴリ $k$ と反応する確率を示している．

GPCM では，カテゴリ $k-1$ とカテゴリ $k$ の反応曲線が，ステップパラメータ $\beta_{ik} = \theta_j$ の位置で交差する．この現象はステップパラメータがカテゴリ順に並んでいない場合にも生起する．例えば，図 3 の例では，「$\beta_{i3} < \beta_{i4}$」となっている．これは，カテゴリ 3 に移ったときには，カテゴリ 4 に移ることが容易であることを意味し，結果として，カテゴリ 3 への反応確率が $\theta$ 全域で小さくなり，反応曲線は図 3 のように沈んだ形となる．一方，GRM では，項目の難易度パラメータが $b_{i1} < b_{i2} < \cdots < b_{ik} < \cdots < b_{iK-1}$ と制約されるため，反応曲線の頂点が必ずカテゴリ順に並ぶ．

以上から，GPCM は，GRM よりも柔軟な表現が可能である反面，ステップパラメータ $\beta_{ik}$ の取りうる値の範囲が広くなるためパラメータの推定精度では GRM に劣ると考えられる．

本論では，これらの項目反応モデルを拡張して，評価者特性を反映できるピアアセスメントのための項目反応モデルの構築を目指す．

## 4. 評価者特性を加えた IRT モデル

2. で述べたように，ピアアセスメントで蓄積されるデータ $\boldsymbol{U}$ は，「学習者」×「課題」×「評価者」の 3 相データとなる．このようなデータに対して GRM や GPCM などの通常の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ために，GRM や GPCM に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．

### 4.1 評価者パラメータを加えた GPCM

Patz et al. [19] は，GPCM に評価者パラメータを加えた拡張モデルを提案している．このモデルでは，課題 $i$ に対する学習者 $j$ の成果物に，評価者 $r$ が評価カテゴリ $k$ を与える確率 $P_{ijrk}$ を次式で定義する．

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

ここで，$a_i$ は課題 $i$ の識別力，$\beta_{ik}$ は課題 $i$ におけるカテゴリ $k-1$ からカテゴリ $k$ への遷移の難しさを表すステップパラメータ (ただし $\beta_{i1} = 0$)，$\rho_{ir}$ は課題 $i$ における評価者 $r$ の評価の厳しさを表す．以降，このモデルを Patz1999 と呼ぶ．

宇佐美 [16] 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．

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

ここで，$\alpha_r$ は評価者 $r$ の評価の一貫性，$\beta_i$ は課題 $i$ の位置パラメータ，$\beta_r$ は評価者 $r$ の位置パラメータ，$d_{ik}$ は課題 $i$ におけるカテゴリ $k$ の閾値パラメータ（ただし $d_{i1} = 0$），$d_r$ は評価者 $r$ の閾値パラメータを表す．ただし，パラメータの識別性のために，$\Pi_r\alpha_r = 1$，$\sum_r \beta_r = 0$，$\Pi_r d_r = 1$ を仮定する．以降では，このモデルを Usami2010 と呼ぶ．

以上のモデルは，対数オッズ比 $ln(P_{ijrk}/P_{ijrk-1})$

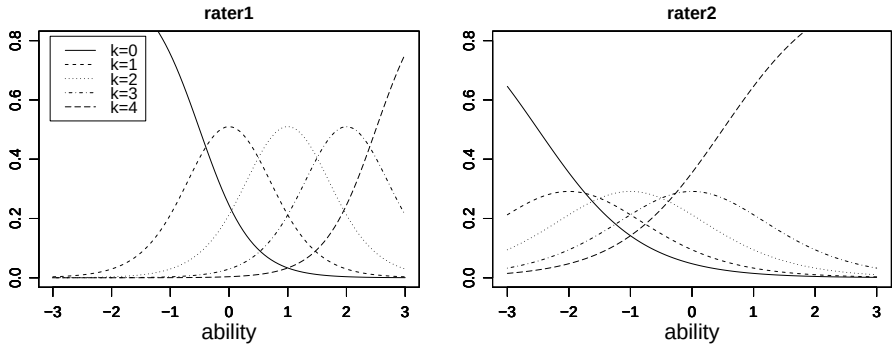



図 4　異なる評価者による提案モデルの反応曲線
Fig. 4　Item characteristic curves of two different raters

を，$\theta_j - b_i - \beta_r - d_k$ のように学習者 $j$，課題 $i$，評価者 $r$，評点 $k$ による特性の線形和として定義した多層ラッシュモデル [20] の拡張とみなせる．

### 4.2　評価者パラメータを加えた GRM

植野ら [2] は，ピアアセスメントのための項目反応理論として，GRM を拡張した以下のモデルを提案している．

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

$$\begin{cases} P^*_{ijrk} = \frac{1}{1+\exp(-\alpha_i(\theta_j - b_i - \varepsilon_{r,k}))} & k = 1, \cdots, K-1 \\ P^*_{ijr0} = 1 \\ P^*_{ijrK} = 0 \end{cases} \tag{6}$$

ここで，$b_i$ は課題 $i$ の難易度，$\varepsilon_{r,k}$ は評価者 $r$ による評点 $k$ への厳しさを表す．ただし，$\varepsilon_{r,1} < \varepsilon_{r,2} < \cdots < \varepsilon_{r,K-1}$ とする．以降では，このモデルを Ueno2008 と呼ぶ．Ueno2008 では，評価者の得点 $k$ への評価確率が，学習者の能力に対して広く等分に分散しているとき，識別力の高い評価が行われていると捉え，評価の一貫性の指標を次式で定義している．

$$R_r = \frac{1}{K} exp(P(\hat{\varepsilon_{r,k}}) log P(\hat{\varepsilon_{r,k}})) \tag{7}$$

ここで，$P(\hat{\varepsilon_{r,k}}) = \frac{1}{1+\exp(\hat{\varepsilon_{r,k-1}})} - \frac{1}{1+\exp(\hat{\varepsilon_{r,k}})}$ とする．

しかし，これらのモデルでは，評価者数が学習者数と同程度まで増加するピアアセスメントでは，パラメータ数に対するデータ数が少なくなり，高精度なパラメータ推定が期待できない．

本論では，この問題を解決するために，通常の項目反応理論に対し，できる限り評価者パラメータ数が少なくなるように，評価者の「評価の一貫性」と「評価の厳しさ」を表すパラメータを付加した項目反応モデルを提案する．

## 5.　提案モデル

ここでは，ピアアセスメントにおける項目反応モデルとして，GRM を拡張した以下のモデルを提案する．

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

$$\begin{cases} P^*_{ijrk} = \frac{1}{1+\exp(-\alpha_i\alpha_r(\theta_j - b_{ik} - \varepsilon_r))} & k = 1, \cdots, K-1 \\ P^*_{ijr0} = 1 \\ P^*_{ijrK} = 0 \end{cases} \tag{8}$$

ここで，$b_{ik}$ は課題 $i$ において評点 $k$ を得る難易度，$\varepsilon_r$ は評価者 $r$ の評価の厳しさを表す．ただし $b_{i1} < b_{i2} < \cdots < b_{iK-1}$ とする．また，パラメータの識別性のために，$\Pi_r \alpha_r = 1$ を仮定する．

提案モデルでは，評価者の厳しさパラメータとして $\varepsilon_r$ を導入した．このパラメータは，評価者一人に一つだけ対応することが特徴である．Ueno2008 や Patz1999 のような多次元の評価者パラメータを採用すると，評価者数の増加に伴い，パラメータ数が急速に増加し，パラメータの推定精度が低下する．それに対し，提案モデルでは，評価者数に対するパラメータ数の増加が比較的緩慢となり，パラメータ推定精度の向上が期待できる．

また，提案モデルでは，Usami2010 以外の既存モデルでは導入されていない評価者の一貫性パラメータを導入した．ピアアセスメントでは，評価者間/評価者内で評価基準が一貫している保証がなく [27]，評価の一貫性の欠如は，評価の信頼性低下を引き起こす [16]．

表 1　各モデルのパラメータ数
Table 1　The number of parameters in each models

| | パラメータ数 |
|---|---|
| Proposed | $IK + 2R + J$ |
| Patz1999 | $I(K + R) + J$ |
| Usami2010 | $I(K + 1) + 3R + J$ |
| Ueno2008 | $2I + R(K - 1) + J$ |

したがって，信頼性向上のためには，評価の一貫性を考慮した学習者の能力推定が必要である．Usami2010 で導入されたパラメータ $\alpha_r$ は，評価の一貫性に対応する評価者特性を適切に表現できることが報告されており [16]，本研究でも，信頼性の高い評価を実現するために，$\alpha_r$ を評価の一貫性パラメータとして採用した．

ここで，提案モデルのパラメータの解釈を示すために，評価者特性の異なる 2 人の評価者の反応曲線を図 4 に示す．ここでは，課題パラメータを $\alpha_i = 1.5, b_{i1} = -1.5, b_{i2} = -0.5, b_{i3} = 0.5, b_{i4} = 1.5$ とし，評価者パラメータを，Rater1(左図) は $\alpha_r = 1.5, \varepsilon_r = 1.0$, Rater2(右図) は $\alpha_r = 0.8, \varepsilon_r = -1.0$ とした．図 4 では，横軸に学習者の能力 $\theta_j$，縦軸に各評価者がそれぞれの評価カテゴリを付与する確率を表す．

図 4 から，評価の一貫性 $a_r$ が大きい Rater1 は，Rater2 に比べて，学習者の能力の違いによって各評価カテゴリへの反応確率が大きく変化していることがわかる．これは，Rater1 の方が，能力の違いを精度よく識別できることを意味する．また，Rater1 は Rater2 に比べて，評価の厳しさパラメータ $\varepsilon_r$ が大きく，反応曲線が全体として右に移動している．これは，Rater1 から高い評点を得るためには，Rater2 から同じ評点を得るよりも高い能力が必要であることを表す．提案モデルを利用することで，課題の特性についても，評価者特性と同様に分析することができる．

## 6. モデルパラメータ

ここでは，提案モデルと既存モデルのパラメータ数を比較する．

表 1 に，提案モデル (Proposed と表記)，Patz1999, Usami2010，Ueno2008 のパラメータ数を示した．表 1 より，評価カテゴリ数 $K > 2$，課題数 $I > 1$ のときに提案モデルの評価者パラメータ数が最小となることがわかる．また，総パラメータ数は，カテゴリ数 $K = 5$ とすると，$2R > 3I$ かつ $I > 2$ の条件において，提案モデルが最小となることがわかる．以上の条件は，課題数に対し評価者数が多いピアアセスメントでは，一般に満たされると考えられる．

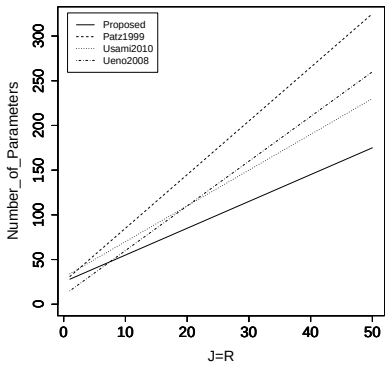



図 5　評価者数 = 学習者数とパラメータ数の関係
Fig. 5　The relationships between the number of raters(=learners) and parameters

さらに，図 5 に，カテゴリ数 $K = 5$，課題数 $I = 5$ としたときの，評価者数 $R =$ 学習者数 $J$ の変化に対する各モデルのパラメータ数の変化を示した．$R = J$ は厳しい制約であるが, コミュニティ内で最大数のピアレビューアーが評価する場合を想定し，パラメータ推定が最も難しい状態を考える．図 5 は，横軸が $R = J$, 縦軸がパラメータ数を表す．図 5 からも，評価者数 = 学習者数が多いときには，提案モデルのパラメータ数が最小となることが確認できる．ただし，評価者数 = 学習者数が少ない時は，Ueno2008 のパラメータ数が最小となっている．提案モデルと Ueno2008 のパラメータ数の大小関係が入れ替わる交点は，課題数 $I$ に依存して決まり，課題数 $I$ が少なくなると交点は 0 に近づく．

以上から，評価者数 = 学習者数が増加する，または，課題数が少なくなるほど，提案モデルのパラメータ数が既存モデルと比較して少なくなり，パラメータの推定精度において，提案モデルが優れることが期待できる．

なお，提案モデルとパラメータ数が等しくなるモデルとして，以下のような GPCM の拡張モデルも考えられる．

$$P_{ijrk} = \frac{exp \sum_{m=1}^{k} [\alpha_i \alpha_r (\theta_j - \beta_{im} - \rho_r)]}{\sum_{l=1}^{K} exp \sum_{m=1}^{l} [\alpha_i \alpha_r (\theta_j - \beta_{im} - \rho_r)]} \tag{9}$$

ここで，$\rho_r$ は評価者 $r$ の評価の厳しさを表す．

しかし，3. で述べたように，GRM は位置パラメータ $b_{ik}$ に制約があり，GPCM に比べて高精度なパラメータ推定が期待できるため，本論では GRM の拡張モデルを採用した．

## 7. パラメータ推定

本章では，提案モデルのパラメータ推定手法について述べる．

項目反応理論におけるパラメータ推定には，ニュートンラフソン法や EM アルゴリズムによる，周辺最尤推定法や最大事後確率推定法が一般に利用される [28] [29]．一方，近年では，計算機性能の向上とともにマルコフ連鎖モンテカルロ (MCMC) が利用されるようになってきた．

MCMC は，より高精度なパラメータ推定を可能とするベイズ推定のための手法である．ベイズ推定は，データに関する情報と未知パラメータに関する事前情報とを統合し，未知パラメータの事後分布を導く．未知パラメータは，ある確率分布 $g(\cdot)$ に従って発生すると仮定する．

ここで，各パラメータの集合をそれぞれ $\boldsymbol{\theta} = \{\theta_1, \cdots, \theta_J\}$，$\boldsymbol{\alpha_i} = \{\alpha_{i=1}, \cdots, \alpha_{i=I}\}$，$\boldsymbol{b} = \{b_{11}, \cdots, b_{IK-1}\}$，$\boldsymbol{\alpha_r} = \{\alpha_{r=1}, \cdots, \alpha_{r=R}\}$，$\boldsymbol{\varepsilon} = \{\varepsilon_1, \cdots, \varepsilon_R\}$ と表す．さらに，各パラメータの事前分布のハイパーパラメータを，$\tau_\theta, \tau_{\alpha_i}, \tau_b, \tau_{\alpha_r}, \tau_\varepsilon$ と表し，事前分布を $g(\theta_j|\tau_\theta)$，$g(a_i|\tau_{\alpha_i})$，$g(b_{ik}|\tau_b)$，$g(\alpha_r|\tau_{\alpha_r})$，$g(\varepsilon_r|\tau_\varepsilon)$ とする．このとき，反応データ $\boldsymbol{U}$ を所与として，未知パラメータの事後分布は以下のように導かれる．

$$
\begin{aligned}
&g(\boldsymbol{\theta}, \tau_\theta, \boldsymbol{\alpha_i}, \tau_{\alpha_i}, \boldsymbol{b}, \tau_b, \boldsymbol{\alpha_r}, \tau_{\alpha_r}, \boldsymbol{\varepsilon}, \tau_\varepsilon|\boldsymbol{U}) \\
&\propto L(\boldsymbol{U}|\boldsymbol{\theta}, \boldsymbol{\alpha_i}, \boldsymbol{b}, \boldsymbol{\alpha_r}, \boldsymbol{\varepsilon}) g(\boldsymbol{\theta}|\tau_\theta) g(\tau_\theta) g(\boldsymbol{\alpha_i}|\tau_{\alpha_i}) g(\tau_{\alpha_i}) \\
&\quad g(\boldsymbol{b}|\tau_b) g(\tau_b) g(\boldsymbol{\alpha_r}|\tau_{\alpha_r}) g(\tau_{\alpha_r}) g(\boldsymbol{\varepsilon}|\tau_\varepsilon) g(\tau_\varepsilon) \quad (10)
\end{aligned}
$$

ここで，

$$
\begin{aligned}
&L(\boldsymbol{U}|\boldsymbol{\theta}, \boldsymbol{\alpha_i}, \boldsymbol{b}, \boldsymbol{\alpha_r}, \boldsymbol{\varepsilon}) \\
&\quad = \Pi_{j=1}^{J} \Pi_{i=1}^{I} \Pi_{r=1}^{R} \Pi_{k=1}^{K} (P_{ijrk})^{z_{ijrk}} \quad (11)
\end{aligned}
$$

$$
z_{ijrk} = \begin{cases} 1 : x_{ijr} = k \text{ のとき} \\ 0 : \text{上記以外} \end{cases} \quad (12)
$$

MCMC では，式 (10) の事後分布をシミュレーションにより求める．ここでは，MCMC の一種である，Gibbs サンプリングと Metropolis Hastings を組み合わせた手法を利用する．

ここで，課題パラメータと評価者パラメータをそれぞれまとめて $\boldsymbol{\lambda} = (\boldsymbol{\alpha_i}, \boldsymbol{b})$，$\boldsymbol{\Pi} = (\boldsymbol{\alpha_r}, \boldsymbol{\varepsilon})$ と表す．課題パラメータ，評価者パラメータ，被験者パラメータはそれぞれ独立と仮定できるため，対象となる事後分布は $g(\boldsymbol{\lambda}, \boldsymbol{\Pi}, \boldsymbol{\theta}) = g(\boldsymbol{\lambda}) g(\boldsymbol{\Pi}) g(\boldsymbol{\theta})$ となる．アルゴリズムの大枠は，$\boldsymbol{\lambda}^t$ と $\boldsymbol{\Pi}^t$ を所与として $\boldsymbol{\theta}^{t+1}$ をサンプリングし，$\boldsymbol{\lambda}^t$ と $\boldsymbol{\theta}^{t+1}$ を所与として $\boldsymbol{\Pi}^{t+1}$ をサンプリング，そして，$\boldsymbol{\Pi}^{t+1}$ と $\boldsymbol{\theta}^{t+1}$ を所与として $\boldsymbol{\lambda}^{t+1}$ をサンプリングすることを繰り返す．ここで，例えば，時点 $t$ における項目パラメータ $\boldsymbol{\lambda}^t$ と評価者パラメータ $\boldsymbol{\Pi}^t$ を所与とすると，学習者の能力パラメータベクトルの $j$ 番目の要素 $\theta_j^t$ は，条件付き分布 $\pi(\theta_j|\boldsymbol{\theta}^t_{-j}, \boldsymbol{\lambda}^t, , \boldsymbol{\Pi}^t, U)$ からサンプリングする．ただし，$\boldsymbol{\theta}^t_{-j} = \boldsymbol{\theta}^t \backslash \theta_j^t$ を表す．ここでは，この条件付き分布からのサンプルを得るために，Metropolis Hastings を用いる．

Metropolis Hastings では，$\theta_j^{t+1}$ を得るために，まず，提案分布 $h(\theta_j^*|\theta_j^t)$ から候補 $\theta_j^*$ をサンプリングする．提案分布には，正規分布 $N(\theta_j^t, \sigma_\theta^2)$，すなわち，

$$
h(\theta_j^*|\theta_j^t) = \frac{1}{\sigma_\theta \sqrt{(2 * \pi)}} exp\left[ -\frac{(\theta_j^* - \theta_j^t)^2}{2\sigma_\theta^2} \right] \quad (13)
$$

を用いる．

提案分布から，サンプリングされた $\theta_j^*$ は，次の採択確率で採択/棄却を決定する．

$$
a(\theta_j^*|\theta_j^t) = min\left( \frac{L(\boldsymbol{U_j}|\theta_j^*, \boldsymbol{\theta}^t_{-j}, \boldsymbol{\lambda}^t, \boldsymbol{\Pi}^t) g(\theta_j^*)}{L(\boldsymbol{U_j}|\theta_j^t, \boldsymbol{\theta}^t_{-j}, \boldsymbol{\lambda}^t, \boldsymbol{\Pi}^t) g(\theta_j^t)}, 1 \right) \quad (14)
$$

ここで，

$$
\begin{aligned}
&L(\boldsymbol{U_j}|\theta_j^*, \boldsymbol{\theta}^t_{-j}, \boldsymbol{\lambda}^t, \boldsymbol{\Pi}^t) \\
&\quad = \Pi_{i=1}^{I} \Pi_{r=1}^{R} \Pi_{k=1}^{K} p(x_{ijr}|\theta_j^*, \boldsymbol{\theta}^t_{-j}, \boldsymbol{\lambda}^t, \boldsymbol{\Pi}^t)^{z_{ijrk}} \quad (15)
\end{aligned}
$$

式 (14) で採択されなければ $\theta_j^{t+1} = \theta_j^t$ とする．これを，$j = 1, \cdots, J$ について行い，$\boldsymbol{\theta^{t+1}}$ が得られる．

以上のように Gibbs サンプリングと Metropolis Hastings を，課題パラメータ $\boldsymbol{\lambda}^{t+1}$ と評価者パラメータ $\boldsymbol{\Pi}^{t+1}$ についても同様にサンプリングする．MCMC では，以上のアルゴリズムを十分に繰り返し，得られた複数のサンプルの平均値を EAP 推定値とする．なお，分布が収束したと推測されるまでのバーンイン期

表 2　シミュレーションによるパラメータ推定の精度評価結果
Table 2　Parameter estimation accuracies of each item response models

| | | $J=R=5$ | $J=R=10$ | $J=R=20$ | $J=R=50$ |
|---|---|---|---|---|---|
| Proposed | $\bar{\alpha_i}$ | 0.384 (0.134) | 0.243 (0.125) | 0.113 (0.058) | 0.067 (0.054) |
| | $\bar{b_{ik}}$ | 0.220 (0.094) | 0.167 (0.085) | 0.119 (0.063) | 0.086 (0.063) |
| | $\bar{\alpha_r}$ | 0.176 (0.067) | 0.146 (0.036) | 0.092 (0.018) | 0.098 (0.022) |
| | $\bar{\varepsilon_r}$ | 0.215 (0.094) | 0.208 (0.072) | 0.171 (0.063) | 0.103 (0.018) |
| | $\bar{\theta_j}$ | 0.293 (0.072) | 0.243 (0.125) | 0.174 (0.107) | 0.128 (0.063) |
| | 全パラメータ | 0.241(0.093) | 0.186(0.086) | 0.127(0.062) | 0.092(0.053) |
| Patz1999 | $\bar{\alpha_i}$ | 0.602 (0.228) | 0.359 (0.174) | 0.182 (0.107) | 0.080 (0.058) |
| | $\bar{\beta_{ik}}$ | 0.453 (0.197) | 0.337 (0.161) | 0.232 (0.080) | 0.121 (0.063) |
| | $\bar{\rho_{ir}}$ | 0.280 (0.139) | 0.242 (0.112) | 0.200 (0.094) | 0.172 (0.067) |
| | $\bar{\theta_j}$ | 0.218 (0.103) | 0.231 (0.085) | 0.245 (0.112) | 0.125 (0.089) |
| | 全パラメータ | 0.374(0.168) | 0.276(0.128) | 0.207(0.093) | 0.165(0.067) |
| Usami2010 | $\bar{\alpha_i}$ | 0.673 (0.125) | 0.429 (0.228) | 0.358 (0.117) | 0.172 (0.134) |
| | $\bar{\beta_i}$ | 0.484 (0.273) | 0.302 (0.224) | 0.294 (0.061) | 0.164 (0.067) |
| | $\bar{d_{ik}}$ | 0.261 (0.130) | 0.244 (0.134) | 0.213 (0.040) | 0.172 (0.036) |
| | $\bar{\alpha_r}$ | 0.122 (0.027) | 0.108 (0.027) | 0.099 (0.006) | 0.092 (0.018) |
| | $\bar{\beta_r}$ | 0.287 (0.201) | 0.222 (0.183) | 0.179 (0.057) | 0.120 (0.058) |
| | $\bar{d_r}$ | 0.153 (0.080) | 0.157 (0.036) | 0.158 (0.017) | 0.186 (0.067) |
| | $\bar{\theta_j}$ | 0.390 (0.206) | 0.370 (0.170) | 0.270 (0.039) | 0.167 (0.067) |
| | 全パラメータ | 0.310(0.136) | 0.255(0.130) | 0.224(0.045) | 0.174(0.054) |
| Ueno008 | $\bar{\alpha_i}$ | 0.388 (0.082) | 0.251 (0.048) | 0.075 (0.020) | 0.041 (0.006) |
| | $\bar{b_i}$ | 0.273 (0.093) | 0.222 (0.118) | 0.148 (0.085) | 0.127 (0.027) |
| | $\bar{\varepsilon_{rk}}$ | 0.194 (0.061) | 0.195 (0.061) | 0.133 (0.031) | 0.141 (0.013) |
| | $\bar{\theta_j}$ | 0.290 (0.102) | 0.332 (0.147) | 0.184 (0.060) | 0.148 (0.019) |
| | 全パラメータ | 0.240(0.073) | 0.223(0.077) | 0.134(0.036) | 0.128(0.013) |
| GPCM 拡張モデル | $\bar{\alpha_i}$ | 0.550 (0.170) | 0.372 (0.188) | 0.280 (0.125) | 0.071 (0.063) |
| | $\bar{\beta_{ik}}$ | 0.417 (0.228) | 0.259 (0.134) | 0.225 (0.101) | 0.110 (0.098) |
| | $\bar{\alpha_r}$ | 0.231 (0.080) | 0.144 (0.045) | 0.115 (0.022) | 0.104 (0.018) |
| | $\bar{\rho_r}$ | 0.216 (0.094) | 0.174 (0.072) | 0.130 (0.040) | 0.094 (0.058) |
| | $\bar{\theta_j}$ | 0.257 (0.089) | 0.252 (0.107) | 0.250 (0.094) | 0.139 (0.139) |
| | 全パラメータ | 0.371(0.175) | 0.248(0.120) | 0.211(0.087) | 0.106(0.085) |

間は，パラメータの初期値の影響が残るため推定に利用しない．

なお，各パラメータの事前分布は，$\theta_j \sim N(0,\sigma_\theta^2)$，$\alpha_i \sim LN(1,\sigma_{\alpha_i}^2)$，$b_{ik} \sim N(0,\sigma_b^2)$，$\alpha_r \sim LN(1,\sigma_{\alpha_r}^2)$，$\varepsilon_r \sim N(1,\sigma_\varepsilon^2)$ とする．ここで，$N(\mu,\sigma^2)$ は平均 $\mu$，標準偏差 $\sigma$ の正規分布を，$LN(\mu',\sigma'^2)$ は平均 $\mu'$，標準偏差 $\sigma'$ の対数正規分布を表す．提案分布の分散は，これらの事前分布の分散に比べて非常に小さい値を用いる．

## 8. シミュレーション実験

本章では，提案モデルにおけるパラメータ推定の精度を評価するために，シミュレーション実験を行う．ここでは，提案モデルと Patz1999，Usami2010，Ueno2008，式 (9) で定義した GPCM の拡張モデルについて，以下の実験を行った．

（1） パラメータの真値をれぞれ以下の分布に従ってランダムに設定した．

- $\alpha_i, \alpha_r \sim LN(1.0, 0.4)$
- $b_{i1}, \varepsilon_{i1} \sim N(-1.5, 0.4)$
- $b_{i2}, \varepsilon_{i2} \sim N(-0.5, 0.4)$
- $b_{i3}, \varepsilon_{i3} \sim N(0.5, 0.4)$
- $b_{i4}, \varepsilon_{i4} \sim N(1.5, 0.4)$,
- $\beta_{ik}, \rho_{ir}, \beta_i, \beta_{ir}, d_{ik}, b_i, \varepsilon_r, \rho_r \sim N(0.0, 0.4)$
- $d_r \sim N(1.0, 0.2)$,
- $\theta_j \sim N(0.0, 1.0)$.

ただし，$b_{i1} < \cdots < b_{i4}$,$\varepsilon_{i1} < \cdots < \varepsilon_{i4}$，$\Pi_r \alpha_r = 1$，$\sum_r \beta_r = 0$，$\Pi_r d_r = 1$ とする．

（2） ランダムに生成したパラメータを用いて，課題数 $I=5$，評価カテゴリ数 $K=5$ とし，学習者数 $J=$ 評価者数 $R$ を $5, 10, 20, 50$ と変化させて，各モデルからデータを発生させた．

（3） 生成したデータを用いて，MCMC によるパラメータ推定を行った．ベイズ推定に用いる事前分布としては，真値の生成に用いた分布と同じ分布を用いた．また，MCMC の提案分布には，標準偏差 0.01 の正規分布を用いた．MCMC のバーンイン期間は 30000 時点とし，自己相関を考慮して，30000 時点から 50000 時点までのサンプルを 1000 間隔で収集し，収集した

サンプルの平均を EAP 推定値とした．

（4） MCMC で推定した各パラメータの推定値と，予め設定した真値との平均平方二乗誤差を算出した．

（5） 上記の手順を 10 回繰り返し，平均平方二乗誤差（以降，RMSE）の平均と標準偏差を算出した．

以上の実験結果を表 2 に示す．表 2 において，「全パラメータ」の行は，全パラメータに対する RMSE の平均値と標準偏差を示している．表 2 より，$J = R = 5$ の Ueno2008 を除くすべての条件で，全パラメータの RMSE の平均値が，提案モデルで最小となっていることがわかる．$J = R = 5$ では，パラメータ数が最小となる Ueno2008 が，最小の RMSE となっている．ただし，このとき，提案モデルと Ueno2008 において，全パラメータに対する RMSE の平均値の差異は 0.01 未満と微小であり，概ね同等の推定精度であるといえる．

また，提案モデルと GPCM 拡張モデルの実験結果から，パラメータ数は同じであるにも関わらず，すべての場合で，提案モデルの方が RMSE が小さいことがわかる．このことから，5. で述べた仮説「GRM は位置パラメータに制約があるため，GPCM の拡張モデルよりも高精度なパラメータ推定が期待できる」が妥当であることを確認できた．

以上より，提案モデルが，既存モデルよりも高精度なパラメータ推定を実現できることを示せた．

## 9. 実データ適用

本章では，実データを用いて，提案モデルの妥当性を確認する．ここでは，次の被験者実験を行った．

（1） 著者らの一人が開講している統計学の e-learning 講義において，2009 年から 2011 年までに共通する 5 つのレポート課題をすべて提出していた学習者から，ランダムに 20 名を選出した．学習者の内訳は，2009 年度 (8 人)，2010 年度 (8 人)，2011 年度 (4 人) であった．

（2） ここでも，パラメータ推定の最も難しい状態を想定し，評価者数がレポート数に一致するように，収集した 5 課題× 20 名分のすべてのレポートを，過去にこの講義を受講した理系大学院生 20 名に評価させた．評価の際には，筆者らが用意したルーブリックを提示し，5 段階の評価カテゴリを用いて採点させた．

（3） 収集した評価データを利用して，提案モデル，Patz1999，Usami2010，Ueno2008 のパラメータを MCMC で推定し，モデル選択基準の一つとして知られる BIC [30] を算出した．BIC は，値が最大となるモデルを最適モデルとみなす．

表 3　実データによるモデル比較
Table 3　Model comparison using the real data

| | BIC (括弧内：対数尤度) | | |
|---|---|---|---|
| | J=R=5 | J=R=10 | J=R=20 |
| Proposed | -224.11 (-115.47) | -412.38 (-225.94) | -1531.39 (-1189.35) |
| Patz1999 | -266.36 (-121.51) | -517.80 (-238.15) | -1736.76 (-1166.70) |
| Usami2010 | -239.48 (-118.78) | -443.30 (-225.79) | -1593.89 (-1175.84) |
| Ueno2008 | -222.68 (-125.80) | -472.75 (-239.70) | -1734.20 (-1183.13) |

（4） 手順 (3) と同様の推定を，学習者数 $J$ = 評価者数 $R = 5, 10$ となるようにデータを減らした場合についても実施した．$J = R = 5$ のデータは $U_{i,0,0}$〜$U_{i,5,5}$，$J = R = 10$ のデータは $U_{i,0,0}$〜$U_{i,10,10}$ とした．

以上の実験結果を，表 3 に示す．表 3 の括弧内は対数尤度を表す．表 3 から，$J = R = 5$ のときは，パラメータ数が最小となる Ueno2008 が最適モデルと推定されているが，その他では，提案モデルが最適なモデルとして推定されていることがわかる．6. で述べたように，課題数 $I$ が評価者 = 学習者数に対して少なくなれば，$J = R$ が小さくても提案モデルのパラメータ数が最小となり，提案モデルが最適モデルと推定されると予測できる．一般に，ピアアセスメントでは，課題数に比べて評価者 = 学習者数が多いため，一般的な条件下では提案モデルが最適なモデルであると判断できる．

## 10. 評価者・課題特性の分析例

本章では，提案モデルを用いた評価者特性と課題特性の分析例を示す．

まず，表 4 に提案モデルを用いて推定した各パラメータの推定値を示す．また，例として，課題 1 と 5 に対する評価者 3,4,8 の反応曲線を図 6 に示す．

図 6 から，*Rater*3 は，評価の一貫性が高く，全体的に評価が厳しい傾向があることがわかる．平均未満の学習者には最低点を付けやすい評価者といえる．*Rater*4 は，評価の一貫性，評価の厳しさともに平均的であり，全体の得点を概ね均等に付ける評価者であるといえる．*Rater*8 は，評価の厳しさは平均的だが，評価の一貫性がやや低い．平均的な能力の学習者に対して，カテゴリ 1 とカテゴリ 5 への反応確率が同程度で

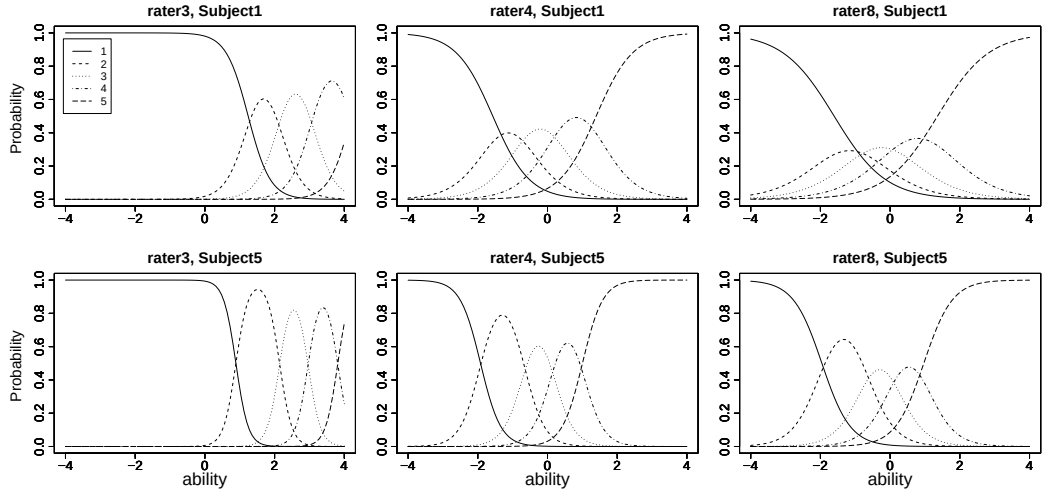



図 6　実データから推定された反応曲線の例
Fig. 6　The examples of the item response curve estimated using the real data

表 4　評価者・課題パラメータの推定例
Table 4　The rater and subject parameters estimated using the real data

| | $\hat{\alpha_i}$ | $\hat{b_{i0}}$ | $\hat{b_{i1}}$ | $\hat{b_{i2}}$ | $\hat{b_{i3}}$ |
|---|---|---|---|---|---|
| 課題 1 | 1.786 | -1.336 | -0.448 | 0.499 | 1.630 |
| 課題 2 | 2.514 | -1.513 | -0.345 | 0.585 | 1.300 |
| 課題 3 | 1.965 | -1.793 | -0.675 | 0.559 | 1.737 |
| 課題 4 | 2.751 | -1.277 | -0.323 | 0.597 | 1.656 |
| 課題 5 | 3.210 | -1.688 | -0.435 | 0.384 | 1.237 |

| | $\hat{\alpha_r}$ | $\hat{\varepsilon_r}$ | | $\hat{\alpha_r}$ | $\hat{\varepsilon_r}$ |
|---|---|---|---|---|---|
| 評価者 1 | 0.812 | 0.231 | 評価者 11 | 1.084 | -0.032 |
| 評価者 2 | 0.925 | 0.196 | 評価者 12 | 0.853 | 0.264 |
| 評価者 3 | 1.765 | 2.578 | 評価者 13 | 1.134 | -0.228 |
| 評価者 4 | 1.066 | -0.234 | 評価者 14 | 1.253 | -0.127 |
| 評価者 5 | 0.871 | -0.122 | 評価者 15 | 1.313 | -0.564 |
| 評価者 6 | 1.474 | 0.011 | 評価者 16 | 0.977 | -0.671 |
| 評価者 7 | 0.810 | -0.180 | 評価者 17 | 0.832 | -0.666 |
| 評価者 8 | 0.760 | -0.279 | 評価者 18 | 1.168 | -0.575 |
| 評価者 9 | 0.974 | 0.160 | 評価者 19 | 0.833 | -0.798 |
| 評価者 10 | 0.958 | -0.047 | 評価者 20 | 0.703 | 0.748 |

| | $\hat{\theta}$ | | $\hat{\theta}$ |
|---|---|---|---|
| 学習者 1 | 0.351 | 学習者 11 | -0.026 |
| 学習者 2 | -0.080 | 学習者 12 | -0.133 |
| 学習者 3 | 0.777 | 学習者 13 | -0.309 |
| 学習者 4 | -0.035 | 学習者 14 | -0.281 |
| 学習者 5 | 0.152 | 学習者 15 | -0.019 |
| 学習者 6 | 0.335 | 学習者 16 | -0.353 |
| 学習者 7 | -0.353 | 学習者 17 | 0.165 |
| 学習者 8 | 0.436 | 学習者 18 | -0.313 |
| 学習者 9 | -0.025 | 学習者 19 | -0.215 |
| 学習者 10 | 0.032 | 学習者 20 | -0.268 |

ある点など評価のばらつきが大きいことがわかる．一方，課題の特性として，$Subject1$ に比べて $Subject5$ は，識別力は高いが，やや評点 2 が付けられやすい傾向の課題であることがわかる．

提案モデルを利用することで，以上のような，評価者による「評価の一貫性」と「評価の厳しさ」の影響と，課題による「項目の識別力」と「困難度」の影響が，学習者の能力推定に反映される．

## 11. 能力推定の信頼性評価

本章では，提案手法を利用することで，ピアアセスメントにおける評価の信頼性が向上するかを評価する．

評価の信頼性とは，ある課題群（または評価者群）から得られた評点がどの程度安定しているかを意味する [31]．すなわち，同一の学習者群に対して，ある課題群（または，評価者群）A を用いて得られた評点が，異なる課題群（評価者群）B から得られた評点と高い相関を示せば，信頼性の高い評価であると判断できる．そこで，ここでは，以下の手順でピアアセスメントにおける評価の信頼性を測定した．

（1） 9. で収集した実データを用いて，提案モデル，Patz1999，Usami2010，Ueno2008 の課題パラメータと評価者パラメータをそれぞれ推定した．

（2） ここで，課題 1〜課題 5 から任意の 2 課題を選択して作成した課題の組を課題群と呼び，全ての課題群パターン（${}_5C_2 = 10$ パターン）を課題群集合と呼ぶ．また，評価者 1〜評価者 20 から任意の 10 名を選択して得られる評価者の集合を評価者群と呼び，全ての評価者群パターン（${}_{20}C_{10} = 184756$ パターン）

表 5　異なる評価者群・課題群の評点データから推定した能力パラメータ間の相関比較
Table 5　Comparing of the correlations between the learner's abilities estimated using the data of different rater and subject groups

| | Proposed | Patz1999 | Usami2010 | Ueno2008 | 平均得点 |
|---|---|---|---|---|---|
| | Mean=.756<br>SD=.103<br>n = 4950 | Mean=.735<br>SD=.115<br>n = 4950 | Mean=.749<br>SD=.106<br>n = 4950 | Mean=0.728<br>SD=.118<br>n = 4950 | Mean=.600<br>SD=.131<br>n = 4950 |
| Patz1999 | t=8.609, p< .01 | - | - | - | - |
| Usami2010 | t=2.765, p< .05 | t=5.844, p < .01 | - | - | - |
| Ueno2008 | t=11.353, p< .01 | t=2.743, p < .05 | t=8.588, p< .01 | - | - |
| 平均得点 | t=67.043, p< .01 | t=58.434, p< .01 | t=64.278, p< .01 | t=55.690, p< .01 | - |

Mean：平均値，SD：標準偏差，n：サンプル数，t：検定統計量

から選択した 10 個の評価者群を評価者群集合と呼ぶ．ここでは，課題群集合と評価者群集合から一群ずつ選択して課題群×評価者群の組を作成し，それに対応する評点データを実データから作成した．同様に，全ての課題群×評価者群の組（100 パターン）に対応する評点データを作成した．

（3） 手順 1 で推定した課題・評価者パラメータと，手順 2 で作成した各課題群×評価者群に対応する評点データを用いて，提案モデル，Patz1999，Usami2010，Ueno2008 により，学習者の能力 $\hat{\boldsymbol{\theta}}$ を推定した．

（4） 各課題群×評価者群から推定された $\hat{\boldsymbol{\theta}}$ について，他の全ての課題群×評価者群から推定された $\hat{\boldsymbol{\theta}}$ との相関を算出した．比較のために，評価カテゴリの平均点（平均得点と呼ぶ）を評価値とした場合についても，同様に相関を求めた．相関にはピアソンの累積相関係数を用いた．

（5） 全課題群×評価者群間の相関係数の平均値について，提案モデル，Patz1999，Usami2010，Ueno2008，平均得点の結果を比較するために，Tukey 法による多重比較を行った．

本実験では，手順 (1) で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ていれば，課題群や評価者群を変更しても，それらの影響を考慮した学習者の能力パラメータ推定がなされるため，群ごとの能力推定値に高い相関が得られると考えられる．

表 5 に実験結果を示す．表 5 から，提案モデル，Patz1999，Usami2010，Ueno2008 を利用した場合と，平均得点を用いた場合を比較すると，項目反応モデルを利用したときに優位に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が評価の信頼性向上に有効であったことが確認できる．

さらに，表 5 から，他の全ての手法と比較して，提案モデルが有意に高い相関を示したことがわかる．このことから，ピアアセスメントにおいて，提案モデルが最も信頼性の高い能力推定を実現できることが示された．

ここで，提案モデルで導入した評価者パラメータの妥当性を評価するために，次の二つのモデルを加えて本章と同様の実験を行った．

（1） 提案モデルから，評価者の評価の一貫性パラメータを削除したモデル（Proposed without $a_r$ と呼ぶ）

（2） Patz1999 において，課題 $i$ に対する評価者 $r$ の厳しさを表す $\rho_{ir}$ を，全課題に対して一定の厳しさを表す $\rho_r$ に変更したモデル（Patz1999 with $\rho_r$ と呼ぶ）

表 6 に，各モデルの基準モデル，評価者/課題パラメータ，本実験におけるパラメータ数，本実験により算出された相関係数の平均と標準編差，さらに，相関係数の平均値に対するモデル間の多重比較結果を示した．

表 6 の Proposed と Proposed without $\alpha_r$ の比較から，評価者の評価の一貫性 $\alpha_r$ を付与したとき優位に高い相関を示したことがわかる．このことから，評価の一貫性パラメータ $\alpha_r$ の利用が信頼性向上に有効であったことが確認できる．

また，表 6 から，Ueno2008 の信頼性が最も低かったことが分かる．Ueno2008 の特徴は，評価者パラメータに評価カテゴリとの交互作用を仮定している点である．これに対して，Ueno2008 以外のモデルでは，課題パラメータに評価カテゴリとの交互作用を付与している．実際のデータでは，評価者特性と評価カテゴリに顕著な交互作用がなかったため，それを仮定した Ueno2008 の信頼性が低下したと解釈できる．

さらに，表 6 の Patz1999 と Patz1999 with $\rho_r$ の比較から，評価者の厳しさが課題ごとに異なると仮定しても，信頼性は向上しなかったことがわかる．これは，課題と評価者の交互作用を仮定すると，データへ

表 6　評価者パラメータの有効性評価結果
Table 6　Evaluation results of effectiveness of the rater parameters

| | Proposed | Proposed without $a_r$ | Ueno2008 | Patz1999 | Patz1999 with $\rho_r$ | Usami2010 |
|---|---|---|---|---|---|---|
| | Mean=.756 SD=.103 n=4950 | Mean=.748 SD=.108 n=4950 | Mean=.728 SD=.118 n=4950 | Mean=.735 SD=.115 n=4950 | Mean=.739 SD=.106 n=4950 | Mean=.749 SD=.106 n=4950 |
| Proposed without $a_r$ | p<.01 | - | - | - | - | - |
| Ueno2008 | p<.01 | p<.01 | - | - | - | - |
| Patz1999 | p<.01 | p<.01 | p<.05 | - | - | - |
| Patz1999 with $\rho_r$ | p<.01 | - | p<.01 | - | - | - |
| Usami2010 | p<.05 | - | p<.01 | p<.01 | p<.01 | - |
| 基準モデル | GRM | GRM | GRM | GPCM | GPCM | GPCM |
| パラメータ数 | 85 | 65 | 110 | 145 | 65 | 100 |
| 評価者の評価の一貫性パラメータ | $\alpha_r$ | - | - | - | - | $\alpha_r$ |
| 評価者の評価の厳しさパラメータ | $\varepsilon_r$ | $\varepsilon_r$ | $\varepsilon_{rk}$ | $\rho_{ir}$ | $\rho_r$ | $\beta_r + d_r$ |
| 課題の識別力パラメータ | $\alpha_i$ | $\alpha_i$ | $\alpha_i$ | $\alpha_i$ | $\alpha_i$ | $\alpha_i$ |
| 課題の難易度パラメータ | $b_{ik}$ | $b_{ik}$ | $b_i$ | $\beta_{ik} + \rho_{ir}$ | $\beta_{ik}$ | $\beta_i + d_{ik}$ |

Mean：平均値，SD：標準偏差，n：サンプル数

の当てはまりが向上する反面，評価者パラメータ数の増加によりパラメータの推定精度が低下するためと解釈できる．

以上の結果から，評価者数が学習者数と同程度まで増加する場合，提案モデルで導入した評価の厳しさパラメータ $\varepsilon_r$ を用いることで，他のパラメータを用いた場合と同等以上の信頼性が得られることがわかった．

以上から，提案した評価者パラメータのピアアセスメント信頼性への有効性が示されたと考えられる．

## 12. むすび

本論では，ピアアセスメントにおいて，評価の信頼性が評価者の特性に依存する問題を解決するために，新たな項目反応理論を提案した．ここでは，評価者パラメータを付加した既存の項目反応モデルをピアアセスメントに適用すると，パラメータ数に対するデータ数が少ないために，高精度なパラメータ推定が期待できないことを指摘した．この問題を解決するために，本論では，通常の項目反応理論について，できる限り評価者パラメータ数が少なくなるように，評価者の「評価の一貫性」と「評価の厳しさ」を表すパラメータを付加した項目反応モデルを提案した．

さらに，シミュレーション実験と被験者実験により，提案モデルの以下の利点を確認した．(1) 既存手法より高精度なパラメータ推定が可能である．(2) 評価の一貫性と厳しさの影響を反映した学習者の能力推定が可能である．(3) 学習者の正確な能力推定が期待できる．

本研究では，学習者群=評価者群となるピアアセスメントに対して，項目反応モデルを適用することを想定した．しかし，実際には，学習者群をいくつかの小グループに分割して，ピアアセスメントを実施することも多い．このような場合，グループごとに推定された各パラメータが，同一尺度上で比較できないという問題が生じる．この問題を解決するアプローチとして，等価法の利用が考えられる．今後はこの問題について研究していきたい．

また，ピアアセスメントの信頼性向上に関する手法として，評価者・課題・学習者ごとの評点の分散をもとに信頼性を定量的に測定し，評価結果の分析や評価デザインの検討に利用する一般化可能性理論 [32] や，複数の評価者の評価結果を一旦集約・分析し，その結果を踏まえて各評価者が再度評価を行うデルファイ法 [33] なども知られている．今後は，これらの手法を本研究に取り入れ，信頼性の詳細な分析や評価活動のデザインなどについても検討したい．

## 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.

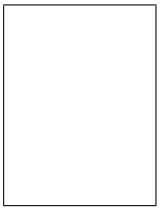
[2] 植野真臣，ソンムァンポクポン，岡本敏雄，永岡慶三，“ピアアセスメントにおける評価者特性を考慮した項目反応理論，” 電子情報通信学会論文誌. D, 情報・システム，vol.91, no.2，pp.377–388，2008.

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

[4] J. Hamer, K.T. Ma, and H.H. Kwong, “A method of automatic grade calibration in peer assessment,” Seventh Australasian Computing Education Conference, vol.42, pp.67–72, 2005.

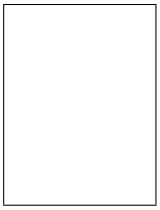
[5] P. Davies, “Review in computerized peer-assessment. will it have an effect on student marking consistency?,” 11th CAA International Computer Assisted Conference, pp.143–151, 2007.
[6] S.S.J. Lin, E.Z.F. Liu, and S.M. Yuan, “Web-based peer assessment:feedback for students with various thinking-styles,” Journal of Computer Assisted Learning, vol.17, no.4, pp.420–432, 2001.
[7] E.F. Gehringer, “Strategies and mechanisms for electronic peer review,” Proceedings of the 30th Annual Frontiers in Education, vol.1, pp.2–7, IEEE Computer Society, 2000.
[8] S. Trahasch, “Towards a flexible peer assessment system,” Information Technology Based Higher Education and Training, pp.516–520, Proceedings of the FIfth International Conference, 2004.
[9] S. Trahasch, “From peer assessment towards collaborative learning,” ASEE/IEEE Frontiers in Education Conference, vol.2, pp.16–20, 2004.
[10] A. Bhalerao and A. Ward, “Towards electronically assisted peer assessment: A case study,” Association for Learning Technology Journal, vol.9, pp.26–37, 2001.
[11] 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.
[12] J. Sitthiworachart and M. Joy, “Effective peer assessment for learning computer programming,” Proceedings of the 9th annual SIGCSE conference on Innovation and technology in computer science education, pp.122–126, 2004.
[13] 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.
[14] M. Ueno, An Item Response Theory for Peer Assessment, Statistics, textbooks and monographs, Mary B.R, 2010.
[15] 藤原康宏，大西　仁，加藤　浩，“公平な相互評価のための評価支援システムの開発と評価：学習成果物を相互評価する場合に評価者の選択で生じる「お互い様効果」,” 日本教育工学会論文誌，vol.31，no.2，pp.125–134，2007.
[16] 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.
[17] R.J. Patz, B.W. Junker, and M.S. Johnson, “The hierarchical rater model for rated test items and its application to large-scale educational assessment data,” Journal of Educational and Behavioral Statistics, vol.27, no.4, pp.341–366, 1999.
[18] N.M. Dato and D. Gruijter, “Two simple models for rater effects,” Applied Psychological Mesurement, vol.8, no.2, pp.213–218, 1984.
[19] 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.
[20] J.M. Linacre, Many-faceted Rasch Measurement, MESA Press, 1989.
[21] E. Muraki, “A generalized partial credit model: Application of an em algorithm,” Applied Psychological Measurement, vol.16, no.2, pp.159–176, 1992.
[22] F. Samejima, “Estimation of latent ability using a response pattern of graded scores,” pp.1–100, no.17, Psychometrika Monography, 1969.
[23] M.Matteucci and L.Stracqualursi, “Student assessment via graded response model,” STATISTICA, pp.435–447, 2006.
[24] T.D. Lawrence, “A model of rater behavior in essay grading based on signal detection theory,” Journal of Educational Mesurement, vol.42, no.1, pp.53–76, 2005.
[25] G. Masters, “A rasch model for partial credit scoring,” Psychometrika, vol.47, no.2, pp.149–174, 1982.
[26] D. Andrich, “A rating formulation for ordered response categories,” Psychometrika, vol.43, no.4, pp.561–573, 1978.
[27] K. Cho, C.D. Schunn, and R. Wilson, “Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives,” Journal of Educational Psychology, vol.98, no.4, pp.891–901, 2006.
[28] F.B. Baker and S.H. Kim, Item Response Theory: Parameter Estimation Techniques, Statistics, textbooks and monographs, Marcel Dekker, 2004.
[29] W.V.D. Linden and R.K. Hambleton, Handbook of modern item response theory, Springer Verlag, 1996.
[30] G. Schwarz, “Estimating the dimension of a model,” Annals of Statistics, vol.6, pp.461–464, 1978.
[31] 宇佐美慧，“論述式テストの運用における測定論的問題とその対処,” 日本テスト学会誌，vol.9，no.1，pp.145–164，2013.
[32] L.J. Cronbach, R. Nageswari, and G.C. Gleser, “Theory of generalizability: A liberation of reliability theory,” The British Journal of Statistical Psychology, vol.16, pp.137–163, 1963.
[33] A.H. Linstone and M. Turoff, “The delphi method: Techniques and applications,” Reading, MA: Addison Wesley, 1975.

（平成 xx 年 xx 月 xx 日受付）



**宇都　雅輝**　（正員）

2009 年電気通信大学人間コミュニケーション学科卒．2011 年同大学大学院博士前期課程修了． 2013 年同大学院博士後期課程修了．博士 (工学)．現在，長岡技術科学大学　特任助教.



**植野　真臣**　（正員）

1993 年神戸大学大学院教育学研究科修了，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），現在，電気通信大学大学院　情報システム学研究科教授.

**Abstract** As an assessment method based on constructivist theory, peer assessment is generating some interest. A certain issue remaining in peer assessment is that some rater biases affect the reliability of ratings. To solve the issue, some item response models that incorporate rater parameters have been proposed. However, accurate estimation of these models is difficult because the number of data, which can be used for estimation, decreases with increasing the number of raters. This paper proposes a new item response model for peer assessment that incorporates rater parameters to avoid increasing the number of parameters.

**Key words** Peer Assessment, Item Response Theory, Reliability, Parameter Estimation