論　文

ピアアセスメントにおける項目反応理論を用いたグループ構成最適化

グエン ドク ティエン[†a)]　宇都　雅輝[††]　植野　真臣[††]

Group Optimization Using Item Response Theory for Peer Assessment

Thien Duc NGUYEN[†a)], Masaki UTO[††], and Maomi UENO[††]

あらまし　近年，社会構成主義に基づく学習評価法としてピアアセスメントが注目されている．一般に，MOOCs のように学習者数が多い場合のピアアセスメントは，評価の負担を軽減するために学習者を複数のグループに分割してグループ内のメンバ同士で行うことが多い．しかし，この場合，学習者の能力測定精度がグループ構成の仕方に依存する問題が残る．この問題を解決するために，本研究では，項目反応理論を用いて，学習者の能力測定精度を最大化するようにグループを構成する手法を提案する．しかし，実験の結果，ランダムにグループを構成した場合と比べ，提案手法が必ずしも高い能力測定精度を示すとは限らないことが明らかとなった．そこで，本研究では，グループ内の学習者同士でのみ評価を行うという制約を緩和し，各学習者に対して少数のグループ外評価者を割り当てる外部評価者選択手法を提案する．シミュレーションと被験者実験から，提案手法を用いて数名の外部評価者を追加することで，グループ内の学習者のみによる評価に比べ，能力測定精度が改善されることが確認された．

キーワード　ピアアセスメント，項目反応理論，グループ構成，評価者選択，能力測定精度，最適化問題

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が注目されている [1]〜[8]．ピアアセスメントの利点としては次のような点が挙げられる [1], [4], [6]．1）評価者の役割を与えることにより学習モチベーションが向上する．2）他者からのコメントやフィードバックにより学習者の内省が促進される．3）他者を評価することにより，他者の成果物からの学びが促される．4）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．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，教師一人で採点を行うよりも，複数名の学習者で評価を行ったほうが信頼性が高くなる．

これらの利点を有することから，ピアアセスメントはこれまでに様々な実践場面で活用されてきた（例えば, [6], [9]〜[13]）．

実践場面におけるピアアセスメントの主な利用法として，学習者同士でコメントを与え合わせることによる学習支援ツールとしての利用が挙げられる．このようなピアアセスメントは，特に協調学習や課題解決型学習（problem-based/project-based learning）などのグループ学習に組み込まれることが多い [14]〜[18]．他方，近年では，Massive Open Online Courses（MOOCs）に代表される大規模 e ラーニング環境の普及に伴い，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急速に高まっている [9], [14]．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，学習者を複数のグループに分割してグループ内のメンバ同士で評価を行わせることで，教師や学習者の負担を大きく増加させることなく評価を実施できる [9], [10], [14]．また，社会構成主義の考え方に基づくと，学習者の能力とは，属するコミュニティのメンバが判断するものとみなせるため [1], [19]，ピアアセスメントを用いた能力測定は妥当であるといえる．以

† 電気通信大学大学院情報システム学研究科，調布市
Graduate School of Information Systems, The University of Electro-Communications, 1–5–1 Chofugaoka, Chofu-shi, 182–8585 Japan
†† 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: thien@ai.is.uec.ac.jp
DOI:10.14923/transinfj.2017JDP7040

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を研究対象とする．

ピアアセスメントに基づく能力測定の課題として，その測定精度が評価者の甘さ・厳しさなどの特性に強く依存する問題が指摘されてきた [1], [2], [4], [6], [17]．この問題を解決する手法の一つとして，評価者特性を表すパラメータを付加した項目反応モデルが提案されてきた [1], [4], [6], [20], [21]．これらの項目反応モデルでは，評価者特性を考慮して学習者の能力を推定できるため，素点の平均や合計などの単純な得点化法より高精度な能力測定を実現できる [2], [6]．

一方で，上述のように，学習者数が多い場合のピアアセスメントは，評価の負担を軽減するために学習者を複数のグループに分割してグループ内のメンバ同士で行わせることが多い．しかし，この場合，項目反応モデルによる能力測定精度がグループ構成の仕方に依存する問題が残る．例えば，評価の一貫性が低い評価者で構成されたグループでは，そのグループ内の学習者に対して高精度な能力測定は期待できない [22]．そこで，本研究では，このようなピアアセスメントの精度を改善するためにグループ構成の最適化を行う．

ピアアセスメントの精度に着目したグループ構成法の研究としては，Nguyen ら [3] があるに留まる．この研究では，全ての学習者に対してできる限り同等な能力測定精度を与えることを目標とし，複数回のピアアセスメントを通して各学習者ができる限り多様な評価者に評価されるようにグループを構成する手法を提案している．この手法は，学習者間での能力測定精度の差異を小さくできるが，能力測定精度を最大化する保証はない．

この問題を解決するために，本論文では，評価者特性パラメータを付与した項目反応モデルを用いて，各学習者に対する能力測定精度を最大化するようにグループを構成する手法を開発する．具体的には，項目反応理論の能力測定精度を表すフィッシャー情報量を用い，各学習者に対する情報量の下限を最大化する整数計画問題としてグループ構成問題を定式化する．提案手法を用いることで，互いに精度良く評価できる学習者が同一のグループに配置されると期待できる．しかし，実験の結果，ランダムにグループを構成した場合と比べて，提案手法が必ずしも高い能力測定精度を示すとは限らないことが明らかとなった．これは，グループ内の学習者同士でのみ評価を行うという制約下では，全ての学習者に情報量の高い評価者を割り当てるようなグループ構成は困難であることを示唆する．

そこで，本論文では，この制約を緩和し，グループ外から数名の外部評価者を導入し各学習者に割り当てる外部評価者選択手法を提案する．具体的には，学習者に対する外部評価者のフィッシャー情報量の下限を最大化する整数計画問題として外部評価者選択手法を定式化する．提案手法では，各学習者の能力を精度良く評価できる外部評価者を採用できるため，少数の外部評価者を追加するだけで能力測定精度を改善できると考えられる．実験から，提案手法を用いて数名の外部評価者を追加することで，グループ内の学習者のみによる評価に比べて，学習者の能力測定精度が改善されることが示された．一般に，外部評価は，評価の精度や客観性の改善に有効であることが知られており [23]，本研究の結果もそれを支持したと解釈できる．

グループ構成の最適化手法は，協調学習などのグループ学習の効果を改善する目的において，これまでにも広く研究されてきた（例えば, [24]～[30]）．本研究ではグループで行うピアアセスメントの精度のみに着目するが，学習者同士が精度よく評価できるグループでは，各学習者に対して正当な評価とフィードバックが与えられるため [8], [31]～[34]，学習効果も高くなると考えられる．したがって，ピアアセスメントの精度を最適化するグループ構成は，大局的には学習効果の改善も期待できる．

2. ピアアセスメント

筆者らの一人は，ピアアセスメント機能をもつ掲示板システムを搭載した Learning Management System (LMS) “Samurai” [13] を開発してきた．この掲示板システムでは，各学習者が自身の学習成果物を自由に投稿でき，他の学習者の成果物に対するコメント付けや評価も行うことができる．図 1 は，e ラーニング講義中に出題された課題に対して，学習者がレポートを投稿した画面の例である．図 1 の下半分の画面には，提出されたレポートに対する他の学習者からのコメントへのリンクが一覧表示されている．画面左上に表示されている五つの星形のボタンは，ピアアセスメントに使用する評価ボタンである．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，これらの評価やコメントを踏まえて成果物を修正する．

本研究では，講義の開講期間中に複数の課題が提示

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

され，個々の課題が完了するたびにピアアセスメントを行う場合を考える．このとき，ピアアセスメントは，学習者を複数のグループ g（$g = 1, \cdots, G$）に分割し，グループ内の学習者同士で行うとする．また，グループ構成は課題ごとに変更すると仮定する．

ここで，x_{igjr} を，課題 i（$i = 1, \cdots, N$）において学習者 j（$j = 1, \cdots, J$）と評価者 r（$r = 1, \cdots, J$）が同一のグループ g に属する場合に 1，そうでない場合に 0 をとるダミー変数とすると，課題 i におけるグループ構成 $\boldsymbol{X}_i$ は次のように定義できる．

$$\boldsymbol{X}_i = \{x_{igjr} \mid x_{igjr} \in \{0, 1\}\}, \forall i, g, j, r. \quad (1)$$

また，課題 i における学習者 j の成果物に評価者 r が与える評価カテゴリー $k \in \{1, \cdots, K\}$ を u_{ijr} で表すと，ピアアセスメントにより得られる評価データ $\boldsymbol{U}$ は以下のように定義できる．

$$\boldsymbol{U} = \{u_{ijr} \mid u_{ijr} \in \{\text{-}1, 1, \cdots, K\}\}, \forall i, j, r. \quad (2)$$

ここで，$u_{ijr} = -1$ は欠測データを表す．グループ内の学習者同士でのみ評価を行う場合，グループ構成 $\boldsymbol{X}_i$ を所与として，$\sum_{g=1}^{G} x_{igjr} = 0$ となる j と r に対応する評価データ u_{ijr} は欠測データとなる．また，本研究では，ピアアセスメントシステムの評価ボタン $\{-2, -1, 0, 1, 2\}$ を順序尺度 $\{1, 2, 3, 4, 5\}$ に変換した 5 段階カテゴリーを評価カテゴリーとして用いる．

本研究では，グループ構成 $\boldsymbol{X} = \{\boldsymbol{X}_1, \cdots, \boldsymbol{X}_N\}$ を最適化することにより，評価データ $\boldsymbol{U}$ から学習者の能力を高精度に測定することを目指す．このために，本研究では項目反応理論を用いる．

3.　項目反応理論

項目反応理論は，数理モデルを用いたテスト理論の一つであり [35]，資格試験や人事考課などの様々な分野で実用化が進められている [5], [36]．項目反応理論の利点としては，次のような点が挙げられる [5], [6], [36]．1）推定精度の低い異質項目の影響を小さくして高精度な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推定を行うことができる．

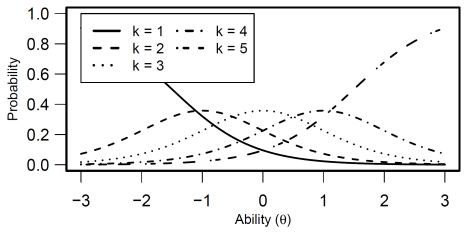

図 2　GRM の項目反応曲線例
Fig. 2　Item characteristic curves of the graded response model for five categories.

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のように正誤が一意に判定できる客観式テストに用いられることが一般的であった．一方で，近年では，記述式試験などのパフォーマンス評価に多値型項目反応モデルを適用する研究も進められている [20], [37]．

本研究で扱うようなリッカート型データに適用できる多値型項目反応モデルとしては，段階反応モデル（Graded Response Model: GRM）[38] や一般化部分採点モデル（Generalized Partial Credit Model: GPCM）[39] が広く利用されてきた．次節では，本研究で用いるピアアセスメントのための項目反応モデルの基礎モデルとなる GRM について述べる．

3.1　段階反応モデル（GRM）

GRM では，学習者 j が課題 i に対してカテゴリー k と反応する確率 P_{ijk} を次式で表す．

$$P_{ijk} = P^*_{ijk-1} - P^*_{ijk} \quad (3)$$

$$P^*_{ijk} = \left(1 + exp\left[-\alpha_i(\theta_j - \beta_{ik})\right]\right)^{-1} \quad (4)$$

ただし，$P^*_{ij0} = 1$，$P^*_{ijK} = 0$ とする．ここで，θ_j は学習者 j の能力パラメータ，α_i は課題 i の識別力パラメータ，β_{ik} は課題 i において評価カテゴリー k を得る難易度パラメータを表す．ただし，$\beta_{i1} < \beta_{i2} < \cdots < \beta_{iK-1}$ と制約する．

例として，$K = 5$，$\alpha_i = 1.5$，$\beta_{i1} = -1.5$，$\beta_{i2} = -0.5$，$\beta_{i3} = 0.5$，$\beta_{i4} = 1.5$ としたときの，GRM の項目反応曲線を図 2 に示す．図 2 では，横

軸が能力 θ_j を表し，縦軸が各カテゴリー k への反応確率 P_{ijk} を表す．図より，能力が低いほど低いカテゴリーへの反応確率が高くなり，能力が高いほど高いカテゴリーへの反応確率が高くなることがわかる．

3.2 ピアアセスメントにおける項目反応理論

本研究で扱う評価データ $\boldsymbol{U}$ は「学習者」×「課題」×「評価者」の三相データとなるが，上述の一般的な項目反応モデルはこのような多相データに直接には適用できない．この問題を解決するアプローチの一つとして，評価者特性を表すパラメータを付与した項目反応モデルが提案されてきた [1], [5], [20], [21], [40]．以降では，ピアアセスメントにおける項目反応モデルとして，最も高精度な能力推定が報告されている Uto and Ueno [1], [2], [4] のモデルについて述べる．

Uto and Ueno [1], [2], [4] のモデルは，GRM に評価者の一貫性と厳しさを表すパラメータを付与したモデルとして定式化される．このモデルでは，課題 i に対する学習者 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} を次式で定義する．

$$P_{ijrk} = P^*_{ijrk-1} - P^*_{ijrk} \tag{5}$$

$$P^*_{ijrk} = \big(1+exp\big[-\alpha_i\alpha_r(\theta_j-\beta_{ik}-\epsilon_r)\big]\big)^{-1} \tag{6}$$

ただし，$P^*_{ijr0} = 1$，$P^*_{ijrK} = 0$ とする．ここで，α_r は評価者 r の評価の一貫性を表し，ϵ_r は評価の厳しさを表す．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．また，GRM と同様に，$\beta_{i1} < \beta_{i2} < \cdots < \beta_{iK-1}$ と制約する．

評価者パラメータの解釈を示すために，評価者特性の異なる 2 人の評価者に関する Uto and Ueno モデルの項目反応曲線を図 3 に示す．ここでは，課題パラメータを $\alpha_i = 1.5$，$\beta_{i1} = -1.5$，$\beta_{i2} = -0.5$，$\beta_{i3} = 0.5$，$\beta_{i4} = 1.5$ とし，評価者パラメータを，Rater1（左図）は，$\alpha_r = 1.5$，$\epsilon_r = 1.0$，Rater2（右図）は，$\alpha_r = 0.8$，$\epsilon_r = -1.0$ とした．図では，横軸が能力 θ_j を表し，縦軸が反応確率 P_{ijrk} を表す．

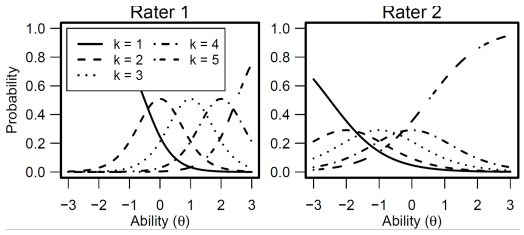

図 3 異なる評価者によるピアアセスメントにおける項目反応モデルの反応曲線例

Fig. 3 Item characteristic curves for two different raters.

図 3 から，一貫性 α_r が高い Rater1 では，能力値 θ の変化に伴う各評点への反応確率の変化が，Rater2 に比べて大きいことがわかる．これは，Rater1 の方が，能力の微小な差異を精度よく識別できることを意味する．また，厳しさパラメータ ϵ_r が大きい Rater1 の反応曲線は，Rater2 の反応曲線と比べて，全体として右に移動していることがわかる．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高い能力が必要であることを意味する．

評価者パラメータを付与した項目反応モデルでは，上記のような評価者特性の影響を考慮して学習者の能力を推定できるため，素点の平均や合計などの単純な得点化法より高精度な能力測定が可能である [1], [2], [4]．また，項目反応モデルをピアアセスメントデータに適用して得られた能力値は，教師一人による評価より精度が高いことも報告されている [6]．学習者の能力測定値は学習者の成績判定 [17], [31] や評価能力の判定 [10]，優秀な他者の推薦 [13] などの様々な目的に活用されるため，能力測定精度の改善は重要な課題といえる．本節で紹介した Uto and Ueno のモデルは，評価者数が増加するピアアセスメントにおいて，類似モデルの中で最も高精度な能力測定が期待できるため [1], [2], [4]，本研究でもこのモデルを用いる．

筆者らは，これまで LMS Samurai を用いたピアアセスメントに対して，学習者をグループに分割してこれらの項目反応モデルを適用する実践を行ってきたが [1], [4], [6]，グループ化の仕方については考慮してこなかった．本研究では，このようなグループで行うピアアセスメントにおいて，項目反応モデルを用いた学習者の能力測定精度を更に改善するために，能力測定精度が最大化されるようなグループを構成することを目的とする．

3.3 フィッシャー情報量

項目反応理論における能力推定の予測誤差は，フィッシャー情報量の逆数に漸近的に一致することが知られている [35]．そのため，項目反応理論では，能力測定精度を表す指標としてフィッシャー情報量が一般に利用される．Uto and Ueno のモデルでは，課題 i において，能力 θ_j をもつ学習者 j に対して評価者 r が与えるフィッシャー情報量 $I_{ir}(\theta_j)$ を以下で定義する．

$$I_{ir}(\theta_j)=\alpha_i^2\alpha_r^2\sum_{k=1}^{K}\frac{\left(P^*_{ijrk-1}Q^*_{ijrk-1}-P^*_{ijrk}Q^*_{ijrk}\right)^2}{P^*_{ijrk-1}-P^*_{ijrk}} \tag{7}$$

ここで，$Q^*_{ijrk}=1-P^*_{ijrk}$ とする．

学習者 j に対するフィッシャー情報量 $I_{ir}(\theta_j)$ が高い評価者 r は，能力 θ_j の学習者 j を精度よく評価できると解釈できる．

また，グループ内の学習者同士でのみ評価を行う場合，グループ構成 $\boldsymbol{X}_i$ を所与として，課題 i における学習者 j へのフィッシャー情報量 $I_i(\theta_j)$ は，同一グループに属する全ての評価者の学習者 j に対する情報量 $I_{ir}(\theta_j)$ の和として，次のように定義できる．

$$I_i(\theta_j)=\sum_{\substack{r=1\\r\neq j}}^{J}\sum_{g=1}^{G}I_{ir}(\theta_j)x_{igjr} \tag{8}$$

すなわち，各学習者に対するフィッシャー情報量 $I_i(\theta_j)$ が最大になるようにグループを構成することで，グループで行うピアアセスメントの能力測定精度を改善できると期待できる．

4. グループ構成手法

ここでは，評価者特性を考慮した項目反応モデルを用いて各学習者に対するフィッシャー情報量ができる限り高くなるようにグループを構成する手法を開発する．具体的には，各学習者に対するフィッシャー情報量の下限を最大化する Max-Min 型の目的関数をもつ整数計画問題としてグループ構成手法を定式化する．

4.1 項目反応理論を用いたグループ構成手法

提案手法では，課題 i におけるグループ構成を以下の整数計画問題として定式化する．

$$\text{Maximize}\quad y_i \tag{9}$$

Subject to

$$\sum_{\substack{r=1\\r\neq j}}^{J}\sum_{g=1}^{G}I_{ir}(\theta_j)x_{igjr}\geq y_i,\ \forall j, \tag{10}$$

$$\sum_{g=1}^{G}x_{igjj}=1,\ \forall j, \tag{11}$$

$$\sum_{g=1}^{G}(1-x_{igjj})\sum_{r=1}^{J}x_{igjr}=0,\ \forall j, \tag{12}$$

$$n_{lb}\leq\sum_{j=1}^{J}x_{igjj}\leq n_{ub},\ \forall g, \tag{13}$$

$$n_{lb}\leq\sum_{g=1}^{G}x_{igjj}\sum_{r=1}^{J}x_{igjr}\leq n_{ub},\ \forall j, \tag{14}$$

$$x_{igjr}=x_{igrj},\ \forall g,j,r, \tag{15}$$

$$x_{igjr}\in\{0,1\},\ \forall g,j,r. \tag{16}$$

ここで，式 (10) は，各学習者に対するフィッシャー情報量の下限 y_i を制約する．また，式 (11) と式 (12) は，各学習者は一つのグループのみに属することを保証する制約である．更に，式 (13) と式 (14) は，各グループに属する学習者数を制約する．ここで，n_{lb} と n_{ub} はそれぞれ，グループを構成する学習者数の下限と上限を表す．本研究では，各グループの構成人数を均等にすることを想定し，$n_{lb}=\lfloor J/G\rfloor$ と $n_{ub}=\lceil J/G\rceil$ とする．$\lfloor\ \rfloor$ と $\lceil\ \rceil$ は床関数と天井関数を表す．目的関数の式 (9) は，上記の制約を満たしつつ，各学習者に対する情報量の下限 y_i を最大化する関数として定義される．

この整理計画問題を解くことにより，学習者に対して情報量の高い評価者を割り当てるようなグループ構成が得られると期待できる．

4.2 シミュレーション実験

提案手法では，互いに精度良く評価できる学習者が同一グループに配置されるため，能力測定精度が改善されると期待できる．このことを確認するために，本節では，提案手法で構成したグループを用いることで，ランダムに構成したグループを用いる場合と比べて能力測定精度が向上するかをシミュレーション実験により評価する．また，本実験では，学習者数 J，課題数 N，グループ数 G が能力測定精度に与える影響を分析するために，これらの条件を変化させて提案手法の性能を評価する．実験手順は以下のとおりである．

（1） 学習者数 $J\in\{15,30\}$，課題数 $N\in\{4,5\}$，評価カテゴリー数 $K=5$ とし，項目反応モデルのパラメータ真値を，表 1 の分布に従いランダムに生成

表 1 実験に用いたパラメータ分布
Table 1 Priors used for experiments.

$\theta_j\sim N(0.0,1.0)$
$log\alpha_r\sim N(0.0,0.5)$, $\epsilon_r\sim N(0.0,0.8)$
$log\alpha_i\sim N(0.1,0.4)$, $\beta_{ik}\sim MN(\boldsymbol{\mu},\boldsymbol{\Sigma})$
$\boldsymbol{\mu}=\{-2.0,-0.75,0.75,2.0\}$
$\boldsymbol{\Sigma}=\begin{vmatrix}0.16&0.10&0.04&0.04\\0.10&0.16&0.10&0.04\\0.04&0.10&0.16&0.10\\0.04&0.04&0.10&0.16\end{vmatrix}$

した．ここで，学習者数と課題数は，著者らの一人が2007年から2013年にLMS Samurai上で開講した2種類のeラーニング講義の実績に基づいて設定した．課題数は各講義における課題数が4と5であったことを元に設定し，学習者数は各講義の開講年度ごとの平均受講者数が12.9（標準偏差4.2）と32.9（標準偏差14.6）であったことを踏まえて設定した．

（2） グループ数$G \in \{3,4,5\}$について，提案グループ構成手法（以降，*MxFiG*と呼ぶ）とランダムグループ構成手法（以降，*RndG*と呼ぶ）を用いてそれぞれグループ構成を行った．グループで行うピアアセスメントでは，各グループの構成人数が3〜14人程度になるようにグループ数を定めることが多く[16],[17],[24],[41]，本実験でもこの範囲に収まるようにグループ数を設定した．また，理想的な条件下における提案手法の性能を評価するために，提案手法で用いるフィッシャー情報量は手順（1）で生成したパラメータ真値を用いて計算した．

（3） 得られたグループと，手順（1）で設定したパラメータ真値を所与とし，評価データをランダムに生成した．

（4） 評価者パラメータと課題パラメータの真値を所与として，生成したデータから学習者の能力パラメータを推定した．推定にはEAP（Expected A Posteriori）推定法[42]を用いた．

（5） 手順（4）で得られた能力パラメータの推定値と真値との2乗平均平方根誤差（以降，RMSEと呼ぶ）を計算した．

（6） 以上の手順を10回繰り返し，RMSEの平均と標準偏差を計算した．

本実験では，整数計画問題のソルバとして*IBM ILOG CPLEX Optimization Studio*[43]を利用した．また，本実験では，5分以内に最適解が得られなかった場合，可能解を採用した．

実験結果を表2の「グループ構成手法」列に示す．表の値はRMSEの平均値と標準偏差（括弧内）を表

表2 シミュレーション実験によるグループ構成手法と外部評価者選択手法の能力測定精度の評価結果

Table 2 RMSEs (SDs) of experiments using simulated data given the true values of parameters.

			グループ構成手法		$n^R=1$			$n^R=2$			$n^R=3$		
J	G	N	RndG	MxFiG	ExRnd	ExFi	ExFiRs	ExRnd	ExFi	ExFiRs	ExRnd	ExFi	ExFiRs
15	3	4	0.338	0.360	0.316	0.272	0.282	0.294	0.245	0.269	0.257	0.227	0.271
			(0.052)	(0.088)	(0.082)	(0.062)	(0.063)	(0.077)	(0.052)	(0.041)	(0.107)	(0.057)	(0.049)
		5	0.293	0.350	0.314	0.263	0.265	0.280	0.232	0.248	0.241	0.220	0.255
			(0.084)	(0.103)	(0.098)	(0.062)	(0.062)	(0.091)	(0.063)	(0.055)	(0.063)	(0.063)	(0.051)
	4	4	0.374	0.432	0.371	0.340	0.356	0.304	0.280	0.314	0.274	0.265	0.323
			(0.137)	(0.118)	(0.092)	(0.096)	(0.094)	(0.076)	(0.058)	(0.049)	(0.078)	(0.058)	(0.073)
		5	0.374	0.386	0.315	0.307	0.325	0.271	0.256	0.290	0.256	0.241	0.292
			(0.114)	(0.118)	(0.057)	(0.093)	(0.100)	(0.056)	(0.069)	(0.072)	(0.090)	(0.069)	(0.073)
	5	4	0.473	0.483	0.376	0.328	0.340	0.355	0.285	0.349	0.311	0.264	-
			(0.123)	(0.075)	(0.100)	(0.053)	(0.051)	(0.105)	(0.062)	(0.080)	(0.056)	(0.063)	-
		5	0.402	0.454	0.353	0.309	0.317	0.334	0.260	0.309	0.281	0.242	-
			(0.097)	(0.103)	(0.081)	(0.066)	(0.064)	(0.078)	(0.080)	(0.095)	(0.060)	(0.081)	-
30	3	4	0.269	0.259	0.240	0.241	0.229	0.236	0.219	0.227	0.223	0.210	0.220
			(0.066)	(0.049)	(0.042)	(0.043)	(0.048)	(0.052)	(0.033)	(0.036)	(0.038)	(0.029)	(0.029)
		5	0.226	0.231	0.218	0.209	0.204	0.203	0.188	0.201	0.209	0.186	0.196
			(0.033)	(0.042)	(0.044)	(0.038)	(0.043)	(0.037)	(0.028)	(0.036)	(0.034)	(0.032)	(0.025)
	4	4	0.321	0.306	0.281	0.259	0.280	0.265	0.244	0.254	0.249	0.218	0.238
			(0.029)	(0.068)	(0.056)	(0.044)	(0.069)	(0.062)	(0.041)	(0.065)	(0.050)	(0.038)	(0.038)
		5	0.252	0.264	0.253	0.235	0.239	0.230	0.216	0.223	0.216	0.197	0.212
			(0.025)	(0.065)	(0.072)	(0.044)	(0.062)	(0.057)	(0.034)	(0.051)	(0.057)	(0.037)	(0.037)
	5	4	0.357	0.347	0.326	0.282	0.294	0.286	0.267	0.276	0.272	0.250	0.253
			(0.068)	(0.070)	(0.072)	(0.055)	(0.080)	(0.043)	(0.050)	(0.063)	(0.046)	(0.047)	(0.047)
		5	0.298	0.307	0.299	0.253	0.254	0.259	0.241	0.247	0.244	0.225	0.222
			(0.043)	(0.045)	(0.045)	(0.051)	(0.060)	(0.048)	(0.039)	(0.051)	(0.038)	(0.037)	(0.037)

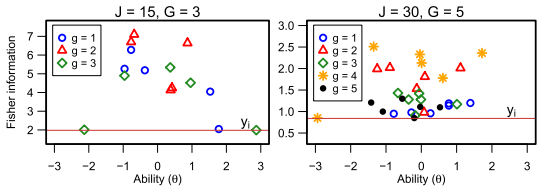

図 4　提案グループ構成手法による学習者の情報量の分布例

Fig. 4　An example of learner's information in groups formed by the proposed grouping method.

す．RMSE が小さいほど能力測定精度が高いことを意味する．

一般に能力測定精度は学習者一人あたりの評価データ数が増加するほど高くなることが知られており [1], [2]，表 2 の結果からも同様の傾向が確認できる．具体的には，グループ数 G が小さくなるほど RMSE が小さくなる傾向が確認できる．これは，グループ数 G が小さいほど各グループの構成人数が増加し，学習者一人あたりの評価者数が増加するためと考えられる．同様に，課題数 N が大きくなるほど RMSE が小さくなる傾向も確認できる．これは課題数の増加によりピアアセスメントの機会が増えるためと解釈できる．

グループ構成手法間で結果を比較すると，提案手法がランダム構成手法に比べて必ずしも高い能力測定精度を示すとは限らなかったことが確認できる．これは，各学習者が単一のグループに属し，グループ内の学習者同士のみで評価を行うという制約下では，必ずしも全ての学習者に対して高い情報量の評価者を割り当てることはできないことを示している．

ここで，提案手法で構成したグループにおける各学習者に対する情報量の分布として，$J = 15, G = 3$ と $J = 30, G = 5$ の場合の例を図 4 に示す．図の横軸は学習者の能力 θ を表し，縦軸は情報量 $I_i(\theta_j)$ を表す．個々の点が個々の学習者を表し，同じ記号で表された学習者は同一のグループに属していることを意味する．図より，情報量が学習者ごとに大きくばらついており，全ての学習者に対して必ずしも高い情報量が与えられていないことが確認できる．

5. 外部評価者選択手法

4.2 の実験から，グループ内の学習者同士のみでピアアセスメントを行う限り，必ずしも能力測定精度を改善できるとは限らないことが明らかとなった．そこで，本研究では，グループ内でのみピアアセスメントを行うという制約を緩和し，グループ外から情報量が高い数名の評価者を割り当てる外部評価者選択手法を開発する．

5.1 項目反応理論を用いた外部評価者選択手法

本研究では，各学習者 j に対するフィッシャー情報量をできる限り高くするように外部評価者を選択する手法を提案する．具体的には，各学習者に対する情報量の下限を最大化する整数計画問題として定式化する．提案手法は，課題 i におけるグループ構成 $\boldsymbol{X}_i$ を所与として，次のように定式化する．

$$\text{Maximize} \quad y_i' \tag{17}$$

Subject to

$$\sum_{r \in \boldsymbol{C}_{ij}}^{J} I_{ir}(\theta_j) w_{ijr} \geq y_i', \ \forall j, \tag{18}$$

$$\sum_{r \in \boldsymbol{C}_{ij}}^{J} w_{ijr} = n^R, \ \forall j, \tag{19}$$

$$\sum_{j=1}^{J} w_{ijr} \leq n^J, \ \forall r, \tag{20}$$

$$w_{ijj} = 0, \ \forall j, \tag{21}$$

$$w_{ijr} \in \{0, 1\}, \ \forall j, r. \tag{22}$$

ここで，w_{ijr} は，課題 i において，学習者 j に評価者 r が割り当てられた場合に 1，そうでない場合に 0 をとる変数を表し，$\boldsymbol{C}_{ij} = \{r \mid r \in \{1, \cdots, J\}, \sum_{g=1}^{G} x_{igjr} = 0\}$ は，課題 i における学習者 j の外部評価者集合を表す．また，n^R は各学習者に割り当てる外部評価者数，n^J は一人の評価者が担当するグループ外学習者数の上限を表す．

式 (18) の制約式は，学習者 j に対して外部評価者集団が与えるフィッシャー情報量の下限 y_i' を制約する．更に，式 (19) により，各学習者に対する外部評価者数が制約される．目的関数の式 (17) は，これらの制約上で各学習者に対する外部評価者が与える情報量の下限の最大化として定義される．

この整数計画問題を解くことにより，グループ外から情報量の高い評価者を割り当てることができ，能力測定精度を改善できると期待できる．

5.2 シミュレーション実験

本研究で提案した外部評価者選択手法では，情報量の高い外部評価者を各学習者に割り当てることができるため，少数の外部評価者を導入するだけで能力測定

精度を改善できると予想できる．このことを確認するために，本節では次の三つのシミュレーション実験を行う．1）**5. 2. 1** では，ランダムに外部評価者を導入した場合と比べて，提案手法が能力測定精度を改善できるかを評価する．ここでは，**4. 2** と同様に，理想的な条件下における提案手法の性能を評価するために，提案手法におけるフィッシャー情報量の計算にはモデルパラメータの真値を用いる．2）項目反応理論における能力測定精度は，一般に評価者数が増加するほど向上する．したがって，上記の実験のみでは，グループ構成手法と提案外部評価者選択手法間に性能差が確認された場合に，その要因が評価者数の増加によるものか，各学習者に適切な評価者が割り当てられたことによるものかを区別できない．そこで，**5. 2. 2** では，提案手法により追加した外部評価者と同数の内部評価者を取り除き，学習者一人あたりの評価者数をグループ構成時と同数にした場合について能力測定精度を求める．この結果がグループ構成時の能力測定精度より高ければ，適切な外部評価者の導入が能力測定精度の改善に有効であることが示せる．本実験でもフィッシャー情報量の計算にはモデルパラメータの真値を用いる．3）上述のとおり以上の二つの実験では，項目反応モデルのパラメータに真値を用いている．しかし，実際にはこれらのパラメータ値は未知であり，データから推定する必要がある．そこで，**5. 2. 3** では，現実の学習場面におけるピアアセスメントに提案手法を適用する方法を述べ，その適用方法を模したシミュレーション実験により提案手法が期待通りに動作するかを評価する．

5. 2. 1　ランダム外部評価者選択手法との比較

本節では，提案手法を用いた場合とランダムに外部評価者を導入した場合の能力測定精度を比較する．このために，ここでは，**4. 2** と同様のシミュレーション実験を行った．ただし，本実験では，手順（2）において提案グループ構成手法でグループを構成した後，全ての学習者に対して $n^R \in \{1,2,3\}$ 名の外部評価者を割り当てた．外部評価者の選択は，提案手法（以降，*ExFi* と呼ぶ）とランダム選択手法（以降，*ExRnd* と呼ぶ）を用いた．なお，提案手法のパラメータである n^J は 12 とした．

実験結果を表 2 の「ExFi」列と「ExRnd」列に示す．表 2 では，「$n^R = 1,2,3$」列がそれぞれ外部評価者数が $1,2,3$ のときの結果を表している．表 2 より，どちらの外部評価者選択手法でも，外部評価者数が増加するほど能力測定精度が改善していることがわかる．

ExFi と *ExRnd* の測定精度を比較すると，$J = 30, G = 3, N = 4, n^R = 1$ を除く全ての場合において，*ExFi* の RMSE が低い値を示したことがわかる．*ExRnd* の方が高い精度を示した $J = 30, G = 3, N = 4, n^R = 1$ では，*ExFi* との差異は 0.001 程度と微小であり，有意な差異ではないと解釈できる．このことから，外部評価者を導入する場合には，フィッシャー情報量の高い外部評価者を選択する提案手法が，能力推定精度の改善において有効であることが確認できる．

5. 2. 2　学習者一人あたりの評価者数を固定した場合の能力測定精度

表 2 からも確認できるように，項目反応理論における能力測定精度は，一般に評価者数が増加するほど向上する．したがって，**5. 2. 1** の実験のみでは，グループ構成手法と比べて提案手法の能力測定精度が改善した要因が，評価者数の増加によるものか，各学習者に適切な評価者が割り当てられたことによるものかを明確には区別できない．そこで，本節では，提案外部評価者選択手法により外部評価者を追加したあと，追加した人数と同数の内部評価者を取り除くことで，学習者一人当たりの評価者数が増加しないように制約した場合について能力測定精度を評価する．具体的には，提案外部評価者選択手法を適用して n^R 人の外部評価者を追加し，各学習者に対する内部評価者を情報量が低い順に n^R 人を取り除く．以降，この手法を *ExFiRs* と呼ぶ．*ExFiRs* による能力測定精度が外部評価者を追加する前より高ければ，適切な外部評価者の追加が能力測定精度の改善に寄与したとみなせる．

そこで，ここでは，*ExFiRs* に対して **5. 2. 1** と同様の実験を行った．実験結果を，表 2 の「*ExFiRs*」列に示した．「$n^R = 1,2,3$」列は外部評価者数を $1,2,3$ 名追加し，同数の内部評価者を取り除いたときの結果を表す．ただし，$J = 15$，$G = 5$ の場合には内部評価者が 2 名となり，3 名の内部評価者を取り除くことができないため，$n^R = 3$ の結果は示していない．

表において，*ExFiRs* と外部評価者の増加前に対応する *MxFiG* の結果を比較すると，学習者一人当たりの評価者数は変化していないにもかかわらず，全ての場合で能力測定精度が改善していることがわかる．このことから，外部から適切な評価者を導入することが，能力測定精度の改善に寄与したことが確認できた．

ただし，本実験結果から，*ExFiRs* では n^R を増加させても測定精度が改善しないことがあることがわか

る．*ExFiRs* では，各学習者に n^R 人の外部評価者を割り当て，グループ内から情報量が低い順に n^R 人の内部評価者を取り除くが，この際に，外部評価者より情報量が高い内部評価者を取り除いたため，情報量が改善されなかったと考えられる．このことを確認するために，n^R が増加したときの *ExFiRs* における情報量の増加量の平均を求めた．結果を表 3 に示す．表 3 の $n^R = 2,3$ の列は，$n^R - 1$ と n^R を比較したときの情報量の増加量を表す．$n^R = 1$ の列は，*MxFiG* と *ExFiRs* の $n^R = 1$ を比較したときの情報量の増加量を表す．表 3 より，$J = 15$，$G = 3,4$ において，n^R を 2 から 3 に増加させた場合に情報量が低下していることがわかる．これは，追加した外部評価者より取り除いた内部評価者の情報量が高かったことを意味する．また，表 2 より，これらの場合には能力測定精度が低下していることが確認でき，情報量の高い内部評価者を取り除いたことによる情報量の低下が能力測定精度の低下を引き起こしたことがわかる．ただし，$J = 15$，$G = 5$，$N = 4$ において n^R を 1 から 2 に増加させた場合では，情報量が増加しているにもかかわらず，能力測定精度は 0.009 とわずかに低下している．しかし，表 3 から，この条件における情報量の増加量は他の条件と比べて小さいことが確認できる．すなわち，追加された外部評価者と取り除かれた内部評価者の情報量がおおむね同程度であったことがわかる．以上から，この条件では，パラメータ推定誤差の影響により能力測定精度がわずかに低下したと解釈できる．

5.2.3 パラメータ推定を含む提案手法の有効性評価

本研究の提案手法を利用するためには，項目反応モデルのパラメータ推定値が必要である．前節までのシミュレーション実験ではパラメータの真値を所与としていたが，実際の評価場面ではこれらの値は一般に未知である．そこで，本節では，現実の評価場面を想定した，パラメータ推定を含む提案手法について述べ，シミュレーション実験によりその妥当性を示す．

表 3　n^R の増加に伴う *ExFiRs* の情報量の増加量
Table 3　Increased amount of information by *ExFiRs*.

J	G	N	$n^R = 1$	$n^R = 2$	$n^R = 3$
15	3	4	4.062	2.003	−0.803
		5	4.969	2.281	−0.866
	4	4	3.940	1.100	−0.814
		5	4.810	1.360	−0.995
	5	4	4.129	0.384	-
		5	5.027	0.496	-
30	3	4	4.293	1.814	1.139
		5	5.095	2.394	1.355
	4	4	4.443	2.006	0.942
		5	5.447	2.609	1.048
	5	4	4.452	2.041	0.729
		5	5.549	2.484	0.959

提案手法では，課題ごとにグループ構成や外部評価者割当を最適化するため，課題パラメータは過去の学習者による評価データなどを用いてあらかじめ推定されている必要がある．ここでは，同一の課題が過去の講義などで出題され，各課題に対する過去の学習者によるピアアセスメントデータが得られている場合を想定する．筆者らが開発した LMS Samurai では，過去の講義で出題された課題に対するピアアセスメントデータが蓄積されており [1], [2], [4], [6]，これらのデータを課題パラメータの推定に利用できる．

一方で，一般に，学習者集団は開講のたびに異なるため，学習者の能力と評価者パラメータを講義開講前に推定しておくことは難しい．そこで，本研究では，講義中の最初の幾つかの課題ではランダムに構成したグループでピアアセスメントを行わせ，以降の課題では，その評価データから推定した学習者の能力と評価者パラメータを所与として提案手法を利用することを想定する．ただし，提案手法は所与とするパラメータ値に基づいて評価者選択を最適化するため，この場合の提案手法の性能は，最初の少数課題に対する評価データから得られるパラメータの推定精度に依存する．一般に，パラメータの推定精度はパラメータ数に対するデータ数が増加するほど向上することから [1]，この運用方法において，パラメータ推定精度を改善するためには以下の二つのアプローチが考えられる．

（1） ランダムグループでピアアセスメントを行わせる課題数（以降では InitN と呼ぶ）を増加させる．

（2） ランダムグループでピアアセスメントを行わせる場合のグループ数（以降では InitG と呼ぶ）を減少させる．

ただし，InitN を増加させると，提案手法を適用できる課題数が減少するため，提案手法による能力測定精度の改善量は減少する．また，InitG を少なくすると，評価者一人当たりの評価対象数が急速に増加し，ピアアセスメントの実施が難しくなる．したがって，実践的には，提案手法が理想通りに動作する範囲内で，できる限り InitN を小さく，InitG を大きく設定することが望まれる．しかし，InitN と InitG が，提

案手法の性能にどのように影響を与えるかは明らかではない．

そこで，本節では，InitN と InitG を変化させたときの提案手法の性能を評価するために，次のシミュレーション実験を行った．

（1） 学習者数 $J = 30$，課題数 $N = 4$，評価カテゴリー数 $K = 5$ とし，項目反応モデルのパラメータの真値を，表 1 の分布に従いランダムに生成した．

（2） 設定したパラメータの真値を所与として評価データをランダムに生成した．

（3） InitN $\in \{1, 2\}$，InitG $\in \{1, 2, 3\}$ の各条件に対して，グループをランダムに構成した．ここで，InitG $= 1$ は，学習者をグループに分割せず全て学習者同士でピアアセスメントを行わせることを意味する．

（4） 構成されたグループを所与として，評価者が割り当てられていない箇所に対応するデータを欠測データに変換した．

（5） 課題パラメータの真値を所与として，データから評価者パラメータと能力パラメータを推定した．推定には Metropolis Hastings within Gibbs Sampling による Markov chain Monte Carlo（MCMC）法 [1], [4] を利用した．各パラメータの事前分布には表 1 の分布を用いた．

（6） 残りの課題に対し，グループ数 $G \in \{3, 4, 5\}$ として，*MxFiG* と *RndG* の手法によりグループ構成を行った．また，*MxFiG* で得られたグループを所与として，*ExRnd*，*ExFi*，*ExFiRs* の三つの手法による外部評価者の割り当てを行った．このとき，*MxFiG*，*ExFi*，*ExFiRs* は，手順（5）で推定した評価者パラメータと能力パラメータを所与として実行した．

（7） 得られたグループ及び外部評価者割当において，評価者が割り当てられていない箇所に対応するデータを欠測データに変換した．

（8） 課題パラメータの真値と評価者パラメータの推定値を所与として，手順（7）のデータから EAP 推定法により能力パラメータを推定した．

（9） 手順（8）で求めた能力パラメータの推定値と真値との RMSE を計算した．

（10） 以上の手順を 10 回繰り返し，RMSE の平均と標準偏差を計算した．

実験結果を表 4 に示す．表 4 より，真のパラメータ値を用いた前節までの実験と同様の傾向が確認できる．具体的には，グループ構成手法間の測定精度の比較では，**4. 2** の実験と同様に，ランダム構成法に比べて，提案グループ構成手法が必ずしも高い能力測定精度を示すとは限らなかったことが確認できる．一方で，提案グループ構成法と外部評価者選択手法を比較すると，外部評価者選択手法が全ての場合で高い能力測定精度を示した．更に，*ExFi* では，**5. 2. 2** の実験と同様に，外部評価者数が増加するほど能力測定精度が改善したことが確認できる．*ExRnd* では，InitN $= 1$, InitG $= 1, G = 4$ と InitN $= 2$, InitG $= 1, G = 4, 5$ の条件において，$n^R = 2$ の精度が $n^R = 1$ より微小に低下している．*ExRnd* では必ずしも情報量が高い外部評価者が選択されるとは限らないため，外部評価者の追加による能力測定精度の改善量が推定誤差を上回ることができなかったことが原因と考えられる．また，*ExFiRs* でも，幾つかの条件で $n^R = 3$ の精度が $n^R = 2$ より低下している．**5. 2. 1** で述べたように，増加された外部評価者より情報量の高い内部評価者が取り除かれたことが要因と考えられる．最後に，*ExFiRs* と *MxFiG* を比較すると，どちらも学習者一人あたりの評価者数は同じであるにもかかわらず，全ての場合で *ExFiRs* が高い測定精度を示したことがわかる．

以上から，本実験で行った InitN $\in \{1, 2\}$ と InitG $\in \{1, 2, 3\}$ の条件において，提案手法は真パラメータを用いた実験と同様の挙動を示しており，理想通りに動作しているとみなせる．

InitG を変化させたときの結果を比較すると，InitG が小さいほど能力測定精度が向上することがわかる．これは，InitG が小さいほど，評価者パラメータと能力パラメータの推定に利用できるデータ数が増加するため，これらのパラメータの推定精度が改善し，結果として提案手法がより適切な評価者を選択できたためと解釈できる．ただし，上述のとおり，InitG を小さくすると評価者一人あたりの評価対象の数が増加するため，実際には学習者による評価の負担を考慮しつつ，できる限り小さい InitG を設定することが望ましい．

また，上述のとおり，提案手法が正常に動作する範囲で InitN をできる限り少なく設定した方が，提案手法を多くの課題で利用できるため，最終的な能力測定精度が向上する．本実験の条件であれば，InitN $= 1$ でも提案手法が正常に動作しているため，InitN $= 1$ の採用が望ましいといえる．

以上より，本節で述べた運用方法により，モデルパラメータが未知の場合でも提案手法を適用でき，能力測定精度も改善されることが示せた．

表 4　シミュレーション実験による評価者と能力のパラメータを推定した場合の能力測定精度
Table 4　RMSEs (SDs) of experiments using simulated data with rater and ability parameters estimation.

			グループ構成手法		$n^R=1$			$n^R=2$			$n^R=3$		
InitN	InitG	G	RndG	MxFiG	ExRnd	ExFi	ExFiRs	ExRnd	ExFi	ExFiRs	ExRnd	ExFi	ExFiRs
1	1	3	0.336	0.330	0.321	0.308	0.312	0.312	0.290	0.297	0.282	0.278	0.293
			(0.050)	(0.065)	(0.070)	(0.053)	(0.055)	(0.062)	(0.053)	(0.052)	(0.048)	(0.048)	(0.049)
		4	0.372	0.346	0.331	0.325	0.329	0.333	0.291	0.304	0.310	0.290	0.313
			(0.064)	(0.051)	(0.046)	(0.051)	(0.059)	(0.041)	(0.046)	(0.050)	(0.050)	(0.039)	(0.047)
		5	0.410	0.401	0.377	0.338	0.343	0.354	0.308	0.328	0.330	0.292	0.318
			(0.073)	(0.076)	(0.064)	(0.051)	(0.049)	(0.060)	(0.054)	(0.056)	(0.044)	(0.052)	(0.047)
	2	3	0.359	0.381	0.367	0.356	0.360	0.339	0.330	0.342	0.334	0.312	0.330
			(0.067)	(0.079)	(0.067)	(0.067)	(0.067)	(0.081)	(0.067)	(0.061)	(0.073)	(0.063)	(0.055)
		4	0.385	0.401	0.368	0.355	0.365	0.360	0.327	0.344	0.343	0.320	0.345
			(0.049)	(0.072)	(0.068)	(0.080)	(0.076)	(0.063)	(0.056)	(0.063)	(0.057)	(0.065)	(0.071)
		5	0.432	0.406	0.372	0.373	0.382	0.369	0.340	0.350	0.345	0.336	0.355
			(0.076)	(0.064)	(0.067)	(0.048)	(0.047)	(0.046)	(0.047)	(0.053)	(0.056)	(0.056)	(0.063)
	3	3	0.356	0.395	0.381	0.357	0.364	0.370	0.335	0.342	0.348	0.323	0.329
			(0.090)	(0.089)	(0.089)	(0.083)	(0.083)	(0.079)	(0.080)	(0.075)	(0.088)	(0.080)	(0.068)
		4	0.390	0.392	0.363	0.356	0.363	0.362	0.322	0.342	0.344	0.307	0.337
			(0.071)	(0.057)	(0.050)	(0.046)	(0.046)	(0.063)	(0.041)	(0.040)	(0.042)	(0.050)	(0.061)
		5	0.460	0.460	0.417	0.411	0.420	0.405	0.384	0.413	0.384	0.363	0.401
			(0.073)	(0.097)	(0.090)	(0.080)	(0.086)	(0.086)	(0.061)	(0.065)	(0.069)	(0.068)	(0.074)
2	1	3	0.377	0.366	0.350	0.324	0.327	0.335	0.314	0.319	0.333	0.304	0.311
			(0.076)	(0.060)	(0.062)	(0.067)	(0.065)	(0.068)	(0.070)	(0.067)	(0.080)	(0.068)	(0.069)
		4	0.441	0.400	0.371	0.364	0.370	0.373	0.333	0.344	0.341	0.332	0.353
			(0.107)	(0.081)	(0.073)	(0.097)	(0.099)	(0.088)	(0.063)	(0.072)	(0.078)	(0.099)	(0.097)
		5	0.477	0.459	0.408	0.388	0.395	0.410	0.364	0.382	0.391	0.353	0.380
			(0.067)	(0.123)	(0.104)	(0.120)	(0.121)	(0.100)	(0.098)	(0.102)	(0.100)	(0.094)	(0.093)
	2	3	0.387	0.396	0.385	0.356	0.364	0.352	0.338	0.353	0.347	0.323	0.337
			(0.104)	(0.059)	(0.060)	(0.058)	(0.063)	(0.064)	(0.052)	(0.059)	(0.065)	(0.057)	(0.060)
		4	0.441	0.445	0.414	0.398	0.405	0.382	0.376	0.382	0.380	0.375	0.393
			(0.107)	(0.077)	(0.071)	(0.085)	(0.089)	(0.067)	(0.087)	(0.102)	(0.075)	(0.081)	(0.094)
		5	0.495	0.490	0.457	0.424	0.426	0.429	0.383	0.396	0.403	0.351	0.373
			(0.102)	(0.130)	(0.129)	(0.121)	(0.121)	(0.101)	(0.069)	(0.072)	(0.109)	(0.088)	(0.099)
	3	3	0.393	0.392	0.376	0.365	0.362	0.379	0.354	0.361	0.362	0.345	0.362
			(0.093)	(0.086)	(0.089)	(0.091)	(0.092)	(0.090)	(0.081)	(0.080)	(0.087)	(0.085)	(0.086)
	3	4	0.449	0.430	0.407	0.399	0.407	0.400	0.370	0.388	0.384	0.362	0.389
			(0.103)	(0.093)	(0.092)	(0.100)	(0.101)	(0.094)	(0.086)	(0.087)	(0.104)	(0.075)	(0.079)
		5	0.460	0.487	0.451	0.427	0.434	0.418	0.419	0.434	0.387	0.389	0.417
			(0.125)	(0.114)	(0.107)	(0.083)	(0.083)	(0.102)	(0.092)	(0.092)	(0.107)	(0.081)	(0.079)

6.　実データ適用

以上のシミュレーション実験では，評価データが項目反応モデルに従って得られると仮定した場合に，提案手法が能力測定精度の改善に有効であることを示した．しかし，実際のピアアセスメントでは，データの生成プロセスが項目反応モデルに従う保証はない．そこで，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．

6.1　実 験 手 順

本研究で行った被験者実験では，まず，34 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 [44] と 2007 年 [45] で出題された課題を日本語に翻訳したものである．これらの課題の解答において専門知識や特別な事前知識は必要としない．各課題に対して提出された各被験者の成果物を，他の全ての被験者に評価させた．評価は NAEP grade 12 [45] 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．

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，**5.2.3** と同様に，評価者パラメータと学習者の能力が未知である場合を想定した実験を行った．ただし，シミュレーション実験と異なり，本実験ではモデルパラメータの真値が与えられないため，RMSE の計算には，能力パラメータの真値の代わりに全データから推定した能力値を用いた．また，課題パラメータは全データを用いて推定された値とした．

また，**5.2.3** のシミュレーション実験から，実データと同程度の課題数と学習者数においては InitN = 1 を採用することが望ましいことを確認した．そこで，ここでは，InitN = 1 のみについて実験を行った．

6.2 実験結果

実験結果を表 5 に示す．表 5 より，シミュレーション実験と同様の傾向が確認できる．具体的には，グループ構成手法間を比較すると，ランダムグループ構成に比べ，提案グループ構成手法が必ずしも高い能力推定精度を達成するとは限らなかったことがわかる．グループ構成手法と外部評価者選択手法の結果を比較すると，外部評価者選択手法が全ての場合で高い能力測定精度を示したことがわかる．更に，*ExFi* では外部評価者が増えるほど能力推定精度が向上することが確認できる．また，**5.2.3** のシミュレーション実験と同様に，*ExRnd* と *ExFiRs* では外部評価者数が増加しても能力測定精度が向上しないことがあることも確認できる．

外部評価者選択手法を比較すると，全ての場合において，*ExFi* が *ExRnd* より高い能力測定精度を示したことがわかる．具体的には，*ExFi* では外部評価者を 1 名追加するだけで，ランダムに 3 名追加した場合と同程度の能力測定精度を達成できており，提案手法が能力測定精度の改善に寄与することが確認できた．また，*ExFiRs* では，学習者あたりの評価者数が増加していないにもかかわらず，*MxFiG* より高精度な能力測定が達成できたことがわかる．以上から，適切な外部評価者の導入が能力測定精度の改善に不可欠であることが示された．

また，シミュレーション実験と同様に，InitG が小さいほど提案手法による能力測定精度が改善したことがわかる．このことから，実践的には，学習者の評価の負担を考慮しつつできる限り InitG を少なく設定することが重要であることが示された．

ここで，提案手法により各学習者に対する情報量が向上したことを確認するために，各学習者に対する情報量の分布を図 5 に示す．ここでは，紙幅の都合上，(1) InitG = 1，$G = 3$，(2) InitG = 1，$G = 5$，(3) InitG = 3，$G = 3$ における *MxFiG* と *ExFiRs* の結果のみを示した．図 5 から，*ExFiRs* と *MxFiG* は学

表 5　評価者パラメータと学習者の能力を推定した場合の実データ適用実験の結果
Table 5　RMSEs (SDs) of experiments using real data with rater and ability parameters estimation.

		グループ構成手法		$n^R = 1$			$n^R = 2$			$n^R = 3$		
InitG	G	RndG	MxFiG	ExRnd	ExFi	ExFiRs	ExRnd	ExFi	ExFiRs	ExRnd	ExFi	ExFiRs
1	3	0.266	0.268	0.254	0.237	0.247	0.258	0.218	0.237	0.230	0.206	0.233
		(0.038)	(0.022)	(0.020)	(0.022)	(0.024)	(0.018)	(0.026)	(0.030)	(0.018)	(0.021)	(0.030)
	4	0.290	0.299	0.289	0.273	0.281	0.276	0.244	0.263	0.267	0.224	0.255
		(0.024)	(0.042)	(0.047)	(0.035)	(0.037)	(0.040)	(0.027)	(0.031)	(0.044)	(0.025)	(0.026)
	5	0.316	0.336	0.308	0.282	0.297	0.289	0.257	0.289	0.274	0.239	0.294
		(0.024)	(0.032)	(0.024)	(0.034)	(0.032)	(0.026)	(0.031)	(0.031)	(0.041)	(0.028)	(0.029)
2	3	0.280	0.273	0.259	0.243	0.245	0.260	0.241	0.246	0.255	0.230	0.241
		(0.025)	(0.035)	(0.029)	(0.037)	(0.039)	(0.040)	(0.039)	(0.040)	(0.033)	(0.030)	(0.036)
	4	0.310	0.326	0.316	0.299	0.309	0.304	0.276	0.296	0.290	0.243	0.275
		(0.029)	(0.056)	(0.050)	(0.045)	(0.046)	(0.054)	(0.039)	(0.037)	(0.038)	(0.043)	(0.040)
	5	0.351	0.349	0.335	0.314	0.327	0.317	0.296	0.321	0.307	0.278	0.306
		(0.049)	(0.045)	(0.035)	(0.053)	(0.057)	(0.040)	(0.047)	(0.049)	(0.032)	(0.034)	(0.037)
3	3	0.304	0.303	0.296	0.278	0.283	0.279	0.262	0.271	0.276	0.252	0.267
		(0.056)	(0.040)	(0.044)	(0.045)	(0.048)	(0.030)	(0.051)	(0.054)	(0.027)	(0.058)	(0.061)
	4	0.349	0.363	0.340	0.325	0.330	0.362	0.308	0.321	0.328	0.292	0.316
		(0.034)	(0.042)	(0.044)	(0.046)	(0.049)	(0.049)	(0.063)	(0.067)	(0.050)	(0.052)	(0.061)
	5	0.375	0.399	0.384	0.359	0.370	0.348	0.328	0.344	0.355	0.309	0.339
		(0.065)	(0.064)	(0.057)	(0.046)	(0.046)	(0.058)	(0.043)	(0.041)	(0.055)	(0.055)	(0.059)

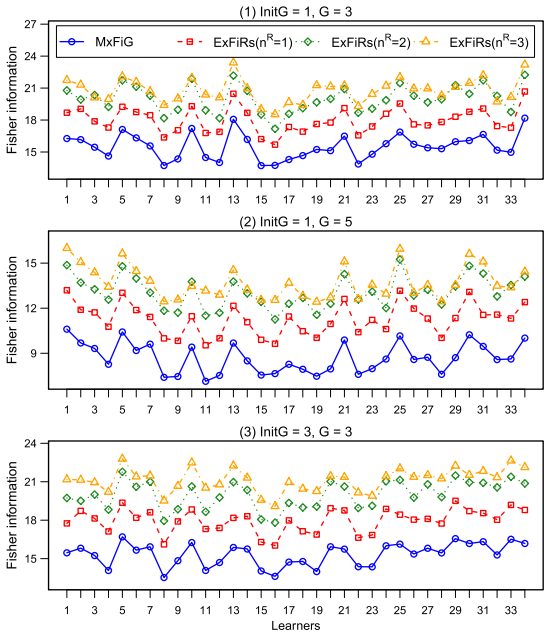

図 5　実データ実験における各学習者に対する情報量の例
Fig. 5　Learners' information of the real data.

習者一人あたりの評価者数は同じであるにもかかわらず，*ExFiRs* は全ての学習者に対して情報量が向上していることがわかる．また，他の条件でも同様の傾向が確認できた．

以上より，提案手法を用いることで，各学習者に対して情報量の高い評価者を割り当てることができ，これにより能力測定精度が改善されることが示された．

7. むすび

本論文では，学習者を複数のグループに分割して行う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手法を提案した．具体的には，評価者特性を考慮した項目反応モデルを用いて，学習者に対するフィッシャー情報量の下限を最大化するようにグループを構成する手法を提案した．しかし，実験から，提案グループ構成手法が必ずしも能力測定精度を向上できるとは限らないことが示され，内部評価の限界が明らかとなった．

この問題を解決するために，グループ内の学習者のみで評価を行わせるという制約を緩和し，グループ外からも少数の評価者を各学習者に割り当てる外部評価者選択手法を提案した．具体的には，グループ外からフィッシャー情報量の高い評価者を選択する整数計画問題として外部評価者選択手法を定式化した．最後に，シミュレーション実験と被験者実験から，提案外部評価者選択手法がピアアセスメントの能力測定精度を改善できることを示した．外部評価は，評価の精度改善に有効であることが指摘されており [23]，本研究の結果はデータからこれに裏付けを与えたといえる．

本研究では，ピアアセスメントの精度最適化のみに着目した．しかし，**1.** で議論したように，互いに精度よく評価できるグループで学習を行った場合，正当な評価と効果的なフィードバックが与えられるため [8], [31]～[34]，高い学習効果が期待できる．今後は，この観点でも提案手法の有効性を評価したい．

謝辞　本研究は JSPS 科研費 15K16256, 17H04726, 15H01772，15K12407 の助成を受けたものです．

文　献

[1] M. Uto and M. Ueno, "Item response theory for peer assessment," IEEE Trans. Learn. Technol., vol.9, no.2, pp.157–170, April 2016.
[2] 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.
[3] T. Nguyen, M. Uto, Y. Abe, and M. Ueno, "Reliable peer assessment for team-project-based learning using item response theory," Proc. 23rd Int. Conf. Comp. Educ., pp.144–153, 2015.
[4] 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.
[5] 植野真臣，荘島宏二郎，学習評価の新潮流，朝倉書店，2010.
[6] 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，" 信学論（D），vol.J91-D, no.2, pp.377–388, Feb. 2008.
[7] K. Topping, "Peer assessment between students in colleges and universities," Rev. Educ. Res., vol.68, no.3, pp.249–276, 1998.
[8] K.J. Topping, E.F. Smith, I. Swanson, and A. Elliot, "Formative peer assessment of academic writing between postgraduate students," Assess. & Eval. High. Educ., vol.25, no.2, pp.149–169, 2000.
[9] H.K. Suen, "Peer assessment for massive open online courses (MOOCs)," The Int. Rev. Res. Open Distributed Learn., vol.15, no.3, pp.312–327, 2014.
[10] C. Piech, J. Huang, Z. Chen, C. Do, A. Ng, and D. Koller, "Tuned models of peer assessment in MOOCs," Proc. 6th Int. Conf. Educ. Dat. Min., pp.153–160, 2013.
[11] K. Cho and C.D. Schunn, "Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system," Comput. & Educ., vol.48, no.3,

pp.409–426, 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," Comput. & Educ., vol.45, no.2, pp.187–202, 2005.
[13] M. Ueno, "Data mining and text mining technologies for collaborative learning in an ILMS "Samurai"," Proc. 4th IEEE Int. Conf. Adv. Learn. Technol., pp.1052–1053, Aug. 2004.
[14] L. Moccozet and C. Tardy, "An assessment for learning framework with peer assessment of group works," Proc. 14th Int. Conf. Inf. Technol. High. Educ. Train., pp.1–5, 2015.
[15] C.-H. Lan, S. Graf, K.R. Lai, and K. Kinshuk, "Enrichment of peer assessment with agent negotiation," IEEE Trans. Learn. Technol., vol.4, no.1, pp.35–46, 2011.
[16] T. Papinczak, L. Young, and M. Groves, "Peer assessment in problem-based learning: A qualitative study," Adv. Heal. Sci. Educ., vol.12, no.2, pp.169–186, 2007.
[17] D.M. Sluijsmans, G. Moerkerke, J.J. van Merriënboer, and F.J. Dochy, "Peer assessment in problem based learning," Stud. Educ. Eval., vol.27, no.2, pp.153–173, 2001.
[18] M. Freeman, "Peer assessment by groups of group work," Assess. & Eval. High. Educ., vol.20, no.3, pp.289–300, 1995.
[19] J. Lave and E. Wenger, Situated Learning. Legitimate Peripheral Participation, R. Pea and J.S. Brown, eds., Cambridge University Press, 1991.
[20] 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，" 教育心理学研究，vol.58, no.2, pp.163–175, 2010.
[21] 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. Educ. Behav. Stat., vol.27, no.4, pp.341–384, 2002.
[22] Z. Wang and L. Yao, "The effects of rater severity and rater distribution on examinees' ability estimation for constructed response items," ETS Res. Rep. Ser., vol.2013, no.2, pp.i–22, 2013.
[23] M. Conley-Tyler, "A fundamental choice: Internal or external evaluation?," Eval. J. Australas, vol.4, pp.3–11, 2005.
[24] Y.S. Lin, Y.C. Chang, and C.P. Chu, "Novel approach to facilitating tradeoff multi-objective grouping optimization," IEEE Trans. Learn. Technol., vol.9, no.2, pp.107–119, April 2016.
[25] I. Srba and M. Bielikova, "Dynamic group formation as an approach to collaborative learning support," IEEE Trans. Learn. Technol., vol.8, no.2, pp.173–186, April 2015.
[26] H. Sadeghi and A.A. Kardan, "A novel justice-based linear model for optimal learner group formation in computer-supported collaborative learning environments," Comput. Hum. Behav., vol.48, pp.436–447, 2015.
[27] J. Moreno, D.A. Ovalle, and R.M. Vicari, "A genetic algorithm approach for group formation in collaborative learning considering multiple student characteristics," Comput. & Educ., vol.58, no.1, pp.560–569, 2012.
[28] R. Hübscher, "Assigning students to groups using general and context-specific criteria," IEEE Trans. Learn. Technol., vol.3, no.3, pp.178–189, 2010.
[29] Y.-T. Lin, Y.-M. Huang, and S.-C. Cheng, "An automatic group composition system for composing collaborative learning groups using enhanced particle swarm optimization," Comput. & Educ., vol.55, no.4, pp.1483–1493, 2010.
[30] N. Khandaker and L.-K. Soh, "Improving group selection and assessment in an asynchronous collaborative writing application," Int. J. Artif. Intell. Educ., vol.20, no.3, pp.231–268, 2010.
[31] P.M. Sadler and E. Good, "The impact of self-and peer-grading on student learning," Educ. Assess., vol.11, no.1, pp.1–31, 2006.
[32] M. van Zundert, D. Sluijsmans, and J. van Merriënboer, "Effective peer assessment processes: Research findings and future directions," Learn. Instr., vol.20, no.4, pp.270–279, 2010.
[33] D.F. Baker, "Peer assessment in small groups: A comparison of methods," J. Manag. Educ., vol.32, no.2, pp.183–209, 2008.
[34] P. Black and D. Wiliam, "Assessment and classroom learning," Assess. Educ. Princ. Policy & Pract., vol.5, no.1, pp.7–74, 1998.
[35] F.M. Lord, Applications of item response theory to practical testing problems, Lawrence Erlbaum Associates, 1980.
[36] 植野真臣，永岡慶三，e テスティング，培風館，2009.
[37] M. Matteucci and L. Stracqualursi, "Student assessment via graded response model," Stat., vol.66, no.4, pp.435–447, 2006.
[38] F. Samejima, "Estimation of latent ability using a response pattern of graded scores," Psychom., vol.34, no.1, pp.1–97, March 1969.
[39] E. Muraki, "A generalized partial credit model: Application of an EM algorithm," Appl. Psychol. Meas., vol.16, no.2, pp.159–176, June 1992.
[40] L.T. DeCarlo, "A model of rater behavior in essay grading based on signal detection theory," J. Educ. Meas., vol.42, no.1, pp.53–76, 2005.
[41] Y. Cho, S. Je, Y.S. Yoon, H.R. Roh, C. Chang, H. Kang, and T. Lim, "The effect of peer-group size on the delivery of feedback in basic life support refresher training: a cluster randomized controlled trial," BMC Med. Educ., vol.16, no.1, p.167, 2016.

[42] F.B. Baker and S.-H. Kim, Item response theory: Parameter estimation techniques, Marcel Dekker, 2004.
[43] IBM Corp., IBM ILOG CPLEX Optimization Studio: CPLEX User's Manual, 12.6 ed., IBM Corp., 2015.
[44] H. Persky, M. Daane, and Y. Jin, "The nation's report card: Writing 2002," Technical report, National Center for Education Statistics, 2003.
[45] D. Salahu-Din, H. Persky, and J. Miller, "The nation's report card: Writing 2007," Technical report, National Center for Education Statistics, 2008.

（平成 29 年 4 月 11 日受付，8 月 23 日再受付，
10 月 24 日早期公開）

グエン ドク ティエン

2004 年ハノイ工科大学情報工学部卒．2014 年電気通信大学大学院情報システム研究科博士前期課程修了，同年同大学院博士後期課程入学，在学中．e テスティング，ベイズ統計，人工知能などの研究に従事．

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）

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），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，2016 年に電気通信大学教授に着任，現在に至る．人工知能，e テスティング，e ラーニング，ベイズ統計などの研究に従事．