2020年度　修士論文

パフォーマンス評価のための評価者パラメータを持つDeep-IRTモデル

1931067

塩野谷　周平

情報理工学研究科

情報・ネットワーク工学専攻

2021年1月25日

目 次

1　はじめに

近年，大学入試や入社試験，学習評価などの様々な評価場面において，論理的思考力や問題解決力といった受検者の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 [1, 2, 3, 4, 5, 6]．パフォーマンス評価は課題に対する受検者のパフォーマンスを評価者が直接採点する評価方法であり，これまでにも大学入試における論述式テストや外国語のリスニング試験，入社試験における面接やグループディスカッション，学習場面におけるプログラミング課題やレポート課題など，様々な形式で広く活用されてきた．

しかし，パフォーマンス評価では，受検者に与えられる評価点が評価者の特性に強く依存し，能力測定の精度低下を引き起こす問題が指摘されている [5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]．この問題を解決する手法として，テスト理論の一つである項目反応理論（Item Response Theory: IRT）[16] に，評価者特性を表すパラメータを付与したモデルが近年多数提案されている (e.g., [7, 11, 12, 13, 14, 17, 18, 19, 20, 21]) ．これらの項目反応モデルは評価者特性を考慮して能力推定ができるため，受検者の素点の合計や平均といった単純な評価方法よりも高精度な能力測定が可能となる [6, 7, 11, 12]．

また，現実の評価の場面では，異なる受検者に実施された異なる課題へのパフォーマンスを比較するニーズがしばしば生じる [22, 23]．このような状況で IRT モデルを適用する場合，それぞれのテスト結果から推定されるモデルパラメータを同一尺度上に位置付ける「等化」が必要となる．

一般に，等化を行うためには受検者・課題・評価者の 3 相のうち，二つ以上の相について共通部分を含むようテストを設計する必要がある [23, 24]．ただし，共通受検者を含むテスト設計の場合，受検者の回答負担の増加や学習効果の影響を考慮し，共通課題と共通評価者を用いて等化を行うことが望ましいとされる [23, 24, 25]．

IRT を用いて高精度なパフォーマンステストの等化を行うためには，受検者と評価者の同一母集団からの独立ランダムサンプリングを仮定しなければならない [26]．しかし，受検者や評価者が単一の母集団からランダムサンプリングされることは稀で，各学校やグループ単位でテストデータがサンプルされることが多い．一般に，独立ランダムサンプリングを仮定しない母数確率モデルは複雑であり，実用化は困難である．また，同一母集団からの独立ランダムサンプリングが仮定できない状況下では，高精度な等化に多数の共通課題・共通評価者が必要となる [27, 28]．共通課題の増加は課題

内容の露出によるテストの信頼性低下を招き，共通評価者の増加は評価者の採点負担の増加を引き起こすことが知られている (e.g., [29, 30, 31, 32, 33, 34, 35, 36, 37, 38])．

受検者の同一母集団からの独立ランダムサンプリングを仮定しないモデルとして，木下・植野 [39, 40] は深層学習を用い，受検者の項目への正答確率をモデル化した Deep-IRT モデル (Item Deep Response Theory :IDRT) を提案している．IDRT は受検者ネットワークと項目ネットワークの 2 つの独立したニューラルネットワークを持ち，出力される能力パラメータと項目パラメータを用いることで，高精度なパフォーマンス予測が可能となる．しかし，従来の IDRT は「受検者」×「項目」の 2 相データを想定しており，本論のパフォーマンス評価のような「受検者」×「課題」×「評価者」の 3 相データに対しては，適用することはできない．

そこで本研究では，従来の IDRT に評価者ネットワークを追加し，評価者特性パラメータを推定するモデルを提案する．本手法は受検者・課題・評価者の 3 相のパフォーマンステストの等化において，以下の利点が期待できる．

（1）受検者と評価者の独立ランダムサンプリングが成り立たない場合でも，IRT モデルに比べ能力推定精度が向上する．

（2）受検者と評価者の母集団が単一でない場合にも IRT モデルに比べ能力推定精度が向上する．

本論ではシミュレーション・実データ実験により，提案モデルの有効性を示した．

ただし，Deep-IRT などの深層学習を用いた予測モデルは，学習過程における学習者の能力値を把握することで課題への反応を予測する Knowledge Tracing の分野において，広く用いられているが [41, 42, 43, 44, 45, 46, 47, 48, 49, 50]，これらは時系列学習における反応予測を目的としており，パラメータの解釈性がなくテスト理論として用いることができないため，本論の目的とは異なる．

2　パフォーマンス評価データ

本章では，パフォーマンス評価によって得られるデータ $\boldsymbol{U}$ を，課題 $i \in \{1, \cdots, I\}$ における受検者 $j \in \{1, \cdots, J\}$ のパフォーマンスに評価者 $r \in \{1, \cdots, R\}$ が与える評価カテゴリー $k \in \{1, \cdots, K\}$ の集合として，次のように定義する．

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

ここで，u_{ijr} は課題 i における受検者 j の成果物に対する評価者 r の評価カテゴリーを表し，$u_{ijr} = -1$ は欠測データを表す．

本論では上記のデータ行列 $\boldsymbol{U}$ を扱う．

3　項目反応理論

項目反応理論（Item Response Theory: IRT）は，近年コンピュータテスティングの普及に伴い，様々な分野で用いられるテスト理論の一つである [16]．IRT は受検者の能力と項目の特性（困難度や識別力）を推定し，受検者の項目での正答確率を求める確率モデルである．IRT には以下の利点があげられる．

（1）受検者グループに対して不変の項目パラメータをもち，項目データベースの構築などに有効である．

（2）異なる項目への受検者の反応を同一尺度上で評価できる．

このような利点から，IRT は適応型テストや等質テスト自動構成のようなテスト理論の基礎として，TOEFL [51] や IT パスポート試験 [52]，医療系共用試験 [53] など様々な評価場面で用いられてきた．

これまで，IRT は正誤判定問題や多肢選択式問題のように正誤が一意に決まる客観式テストに利用されていたが，近年では論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応する研究も進められている [14, 27, 28]．本研究のパフォーマンス評価データに適応できる多値型項目反応モデルには，段階反応モデル（Graded Response Model: GRM）[54] や一般化部分採点モデル（Generalized Partial Credit Model: GPCM）[55] が知られている．これらの多値型項目反応モデルは「受検者」×「課題」の 2 相データに用いられていたが,「受検者」×「課題」×「評価者」の 3 相のパフォーマンス評価データに適応させるために，評価者特性を表すパラメータを付与した IRT モデルが多数提案されてきた (e.g., [7, 11, 12, 13, 14, 17, 18, 19, 20, 21]) ．

本章では，本研究で用いる評価者特性パラメータを付与したIRTモデルを紹介する．

3.1 多相ラッシュモデル

評価者パラメータを付与したIRTモデルとして最も一般的なモデルは，Linacre[17]が提案した多相ラッシュモデル（Many-facet Rasch Model: MFRM）である．MFRMには幾つかのバリエーションが存在するが (e.g.,[9, 13])，最も代表的なモデルでは，$u_{ijr} = k$ が得られる確率 P_{ijrk} を次式で求める．

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

ここで，θ_j は受検者 j の能力，β_i は課題 i の困難度，β_r は評価者 r の厳しさ，d_k は評価カテゴリー $k-1$ から k に遷移する困難度を表す．パラメータの識別のために $\beta_{r=1} = 0, d_1 = 0, \sum_{k=2}^{K} d_k = 0$ を仮定する．

多相ラッシュモデルではすべての課題について受検者の能力を識別する力（識別力）が一定と仮定するが，この制約を緩めたモデルとして，課題間の識別力の差異を表現できるGPCMやGRMに対して評価者パラメータを付与したモデルが提案されてきた．

3.2 評価者パラメータを付与したGPCM

Patz and Junker[18] は，GPCM[55] に評価者特性を表すパラメータを付与したモデルを提案している．

$$P_{ijrk} = \frac{\exp \sum_{m=1}^{k} [\alpha_i(\theta_j - \beta_{im} - \rho_{ir})]}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} [\alpha_i(\theta_j - \beta_{im} - \rho_{ir})]} \tag{2}$$

ここで，α_i は課題 i における識別力，β_{ik} は，課題 i においてカテゴリー $k-1$ から k に遷移する困難度，ρ_{ir} は課題 i における評価者 r の厳しさを表す．パラメータの識別のために，$\beta_{i1} = 0, \rho_{i1} = 0; \forall i$ を仮定する．

また，宇佐美 [14] は，評価者間の評価が一貫している保証がないことを指摘し，評価者の一貫性を表現したパラメータをもつGPCMを提案している．

$$P_{ijrk} = \frac{\exp \sum_{m=1}^{k} [\alpha_i \alpha_r(\theta_j - (\beta_i + \beta_r) - d_{im} d_r)]}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} [\alpha_i \alpha_r(\theta_j - (\beta_i + \beta_r) - d_{im} d_r)]} \tag{3}$$

ここで，α_r は評価者 r の一貫性を表すパラメータ，d_{ik} は課題 i におけるカテゴリー k の閾値パラメータ，d_r は評価者 r の閾値パラメータを表す．パラメータの識別のために，$\prod_r \alpha_r = 1, \sum_r \beta_r = 0, \prod_r d_r = 1, d_{i1} = 0$ を仮定する．

さらに，Uto and Ueno[11, 12] は，採点基準が他の評価者と極端に異なる異質評価者の特性を考慮した GPCM を提案している．

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

ここで，d_{rk} は評価カテゴリー k に対する評価者 r の厳しさを表す．パラメータの識別のために，$\alpha_{r=1} = 1, \beta_{r=1} = 0, d_{r1} = 0, \sum_{k=2}^{K} d_{rk} = 0$ を仮定する．このモデルでは評価者特性として，新たに尺度範囲の制限（特定の評価カテゴリーを過剰あるいは避けて使用する傾向）を表現でき，異質評価者が含まれる場合でも高精度な能力推定が実現できる．

3.3 評価者パラメータを付与した GRM

Ueno and Okamoto[20] は，GRM[54] に評価者特性を表すパラメータを付与したモデルを提案している．

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

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

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

また，Uto and Ueno[7, 21] は，宇佐美 [14] 同様，能力推定精度に評価の一貫性が依存する問題を指摘し，評価者の一貫性パラメータを付与した GRM を提案している．

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

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

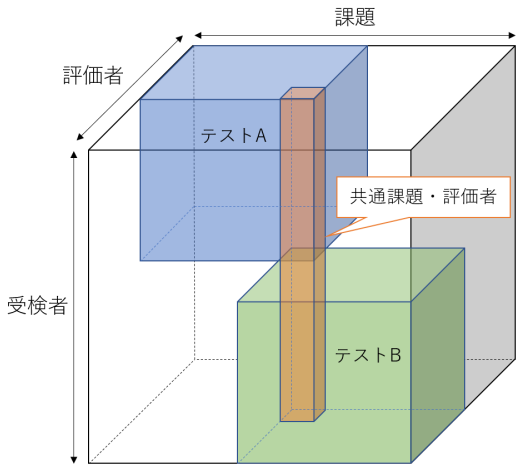

図 1: 共通課題・共通評価者を用いたパフォーマンステストの等化の概要図

ここで，b_{ik} は課題 i において k より大きいカテゴリーを得る困難度を表し，ε_r は評価者 r の厳しさを表す．b_{ik} は順序制約 $b_{i1} < b_{i2} < \cdots < b_{iK-1}$ を仮定し，パラメータの識別のために，$\alpha_{r=1} = 1, \varepsilon_1 = 0$ を仮定する．

3.4　パフォーマンステストの等化

本章で紹介した IRT モデルを用いることにより，パフォーマンス評価データに対し評価者の特性を考慮した能力推定が実現でき，受検者の素点の合計や平均などの単純な評価方法よりも高精度な能力推定が可能となる [6, 7, 11, 12]．一方で，現実の評価場面では，異なる受検者に実施された異なるパフォーマンステストの結果を比較するニーズがしばしば生じる．このような状況で IRT モデルを適用する場合，それぞれのテスト結果から推定されるパラメータを同一尺度上に位置付ける「等化」が必要となる．

パフォーマンステストの等化は，課題と評価者の一部が共通するよう各テストを設計する方法が一般的である [23, 24]．図 1 に共通課題・共通評価者を用いたパフォーマンステストの等化の概要例を示す．図のようにパフォーマンス評価データは三相データであるため，三次元配列で表現する．色付きの領域には受検者の反応データが存在し，それ以外の領域は欠測データを表す．図 1 の例では二つのパフォーマンステスト（テスト A，テスト B と呼ぶ）に対し共通課題と共通評価者を配置し，得られたデータからパラメータを推定する．

IRT モデルを用いて高精度なパフォーマンステストの等化を行うためには，受検者と評価者の同一母集団から独立ランダムサンプリングを仮定しなければならない．しかし，現実には，受検者や評価者は多母集団であり，独立ランダムサンプリングでないことが多く，能力値推定精度を低下させる可能性がある．受検者と評価者の同一母集団からの独立ランダムサンプリングが仮定できない場合，高精度な等化には多数の共通課題・共通評価者が必要となる [27, 28]．共通課題の増加は課題内容の露出によるテストの信頼性低下を招き [29, 30, 31, 32, 33, 34, 35]，共通評価者の増加は評価者の採点負担の増加を引き起こすことが示されている [36, 37, 38]．

4　Deep-IRT モデル

機械学習の分野では，深層学習モデルを用いることで，受検者の母集団と独立性を仮定せずにパフォーマンス予測を行う [39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49]．木下・植野 [39, 40] では，受検者の能力パラメータを出力する受検者ネットワークと項目の難易度パラメータを出力する項目ネットワークを持ち，この二つの独立したネットワークから出力されるパラメータを組み合わせることで，受検者の項目への反応をモデル化した Deep-IRT モデル (Item Deep response Theory :IDRT) を提案している．実用の場面では，受検者が一つの母集団からランダムサンプリングされることは稀で，各学校やグループ単位でテストデータがサンプリングされることが多いが，IDRT を用いることで，受検者の単一母集団からの独立ランダムサンプリングが仮定できない場合や，受検者が多母集団からサンプリングされている場合でも，IRT よりも高精度な能力推定が可能となる．また，IDRT は受検者の能力値や課題の困難度など解釈可能なパラメータを持つため，テスト理論として用いることもできる．

しかし，従来の Deep-IRT モデルは,「受検者」×「項目」の 2 相データへの適応を想定しており，本論で扱うパフォーマンス評価データの「受検者」×「課題」×「評価者」のような 3 相データには直接適応できない．そこで本研究では，パフォーマンス評価データに適応できる Deep-IRT モデルを提案する．

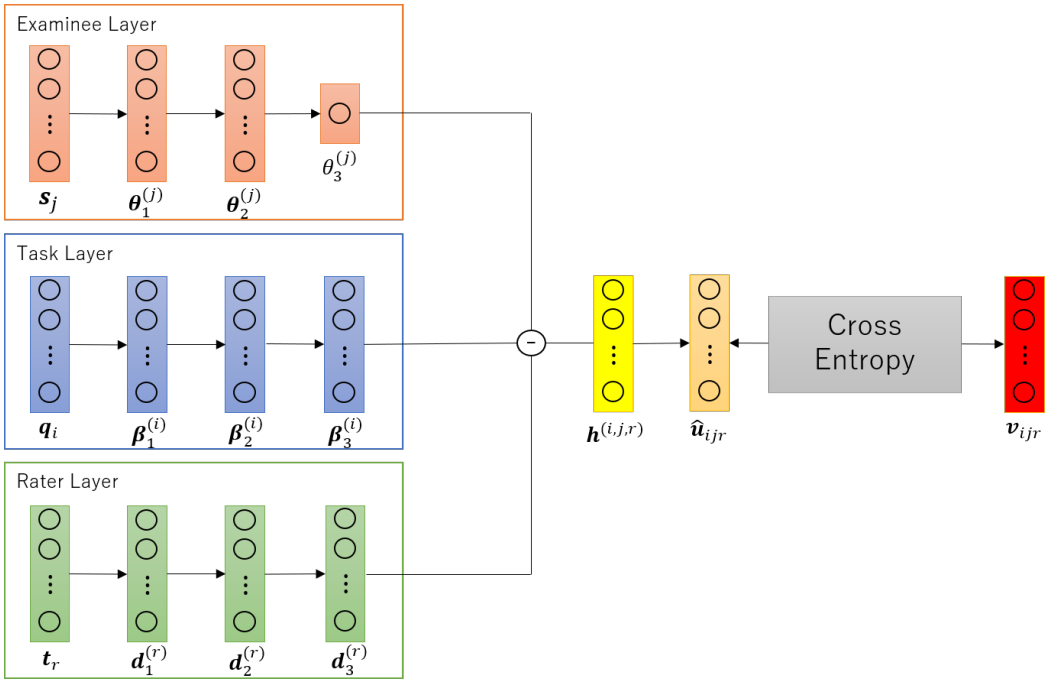

図 2: 提案モデルの概要図

5　評価者パラメータを持つ Deep-IRT モデル

本研究では，パフォーマンス評価データにおいて，受検者と評価者の独立ランダムサンプリングが成り立たない場合でも高精度な能力推定が可能なモデルを提案する．具体的には，木下・植野 [39, 40] の Deep-IRT モデルに評価者ネットワークを追加し，評価者特性を表すパラメータが可能な Deep-IRT モデルを提案する．

5.1　提案モデル概要

提案モデルの概要図を図 2 に示す．提案モデルは受検者ネットワーク（Examinee Network）と課題ネットワーク（Task Network）と評価者ネットワーク（Rater Network）の三つの独立したニューラルネットワークの出力を組み合わせることで，受検者の課題への反応確率をモデル化する．

受検者ネットワークでは j 番目の受検者を表現する one-hot vector $\boldsymbol{s}_j \in \mathbb{R}^J$ を入力とする．s_j は j 番目の要素のみが 1，他の要素が 0 であり，以下のように 4 層のニュー

ラルネットワークを構成する．

$$\boldsymbol{\theta}_1^{(j)} = \tanh\left(\boldsymbol{W}^{(\theta_1)}\boldsymbol{s}_j + \boldsymbol{\tau}^{(\theta_1)}\right) \tag{7}$$

$$\boldsymbol{\theta}_2^{(j)} = \tanh\left(\boldsymbol{W}^{(\theta_2)}\boldsymbol{\theta}_1^{(j)} + \boldsymbol{\tau}^{(\theta_2)}\right) \tag{8}$$

$$\theta_3^{(j)} = \boldsymbol{W}^{(\theta_3)}\boldsymbol{\theta}_2^{(j)} + \tau^{(\theta_3)} \tag{9}$$

ここでは活性化関数として，以下のハイパボリックタンジェント関数を用いる．

$$\tanh(x) = \frac{\exp(x) - \exp(-x)}{\exp(x) + \exp(-x)} \tag{10}$$

$\boldsymbol{W}^{(\theta_1)}, \boldsymbol{W}^{(\theta_2)}$ は以下の重みパラメータ行列である．

$$\boldsymbol{W}^{(\theta_1)} = \begin{pmatrix} w_{11}^{(\theta_1)} & w_{12}^{(\theta_1)} & \cdots & w_{1J}^{(\theta_1)} \\ w_{21}^{(\theta_1)} & w_{22}^{(\theta_1)} & \cdots & w_{2J}^{(\theta_1)} \\ \vdots & \vdots & \ddots & \vdots \\ w_{|\boldsymbol{\theta}_1|1}^{(\theta_1)} & w_{|\boldsymbol{\theta}_1|2}^{(\theta_1)} & \cdots & w_{|\boldsymbol{\theta}_1|J}^{(\theta_1)} \end{pmatrix}$$

$$\boldsymbol{W}^{(\theta_2)} = \begin{pmatrix} w_{11}^{(\theta_2)} & w_{12}^{(\theta_2)} & \cdots & w_{1|\boldsymbol{\theta}_1|}^{(\theta_2)} \\ w_{21}^{(\theta_2)} & w_{22}^{(\theta_2)} & \cdots & w_{2|\boldsymbol{\theta}_1|}^{(\theta_2)} \\ \vdots & \vdots & \ddots & \vdots \\ w_{|\boldsymbol{\theta}_2|1}^{(\theta_2)} & w_{|\boldsymbol{\theta}_2|2}^{(\theta_2)} & \cdots & w_{|\boldsymbol{\theta}_2||\boldsymbol{\theta}_1|}^{(\theta_2)} \end{pmatrix}$$

$\boldsymbol{W}^{(\theta_3)}$ は以下の重みパラメータベクトルである．

$$\boldsymbol{W}^{(\theta_3)} = \left(w_1^{(\theta_3)}, w_2^{(\theta_3)}, \ldots, w_{|\boldsymbol{\theta}_2|}^{(\theta_3)}\right)$$

また，$\boldsymbol{\tau}^{(\theta_1)} = \left(\tau_1^{(\theta_1)}, \tau_2^{(\theta_1)}, \ldots, \tau_{|\boldsymbol{\theta}_1|}^{(\theta_1)}\right)$ および，$\boldsymbol{\tau}^{(\theta_2)} = \left(\tau_1^{(\theta_2)}, \tau_2^{(\theta_2)}, \ldots, \tau_{|\boldsymbol{\theta}_2|}^{(\theta_2)}\right)$ はバイアスパラメータベクトル，$\tau^{(\theta_3)}$ はバイアスパラメータである．本論では受検者ネットワークの出力 $\theta_3^{(j)}$ を受検者 j の能力パラメータとみなす．

図 3 に受検者ネットワークのグラフィカル表現を示す．図 3 で明らかなように，提案モデルは能力パラメータに共通の母集団を仮定しておらず，得られた反応データの予測を最大にするように重みパラメータを更新する．例えば，反応データ u_{ijr} が与えられたとき，すべての重みパラメータが更新され，$\theta_3^{(j)}$ だけでなく他の受検者の θ_3 も更新されるため，受検者パラメータ間の独立性が存在しないことがわかる．

同様に，課題ネットワークでは i 番目の課題を表現する one-hot vector $\boldsymbol{s}_i \in \mathbb{R}^I$ を入力とする．s_i は i 番目の要素のみが 1，他の要素が 0 であり，以下のように 4 層の

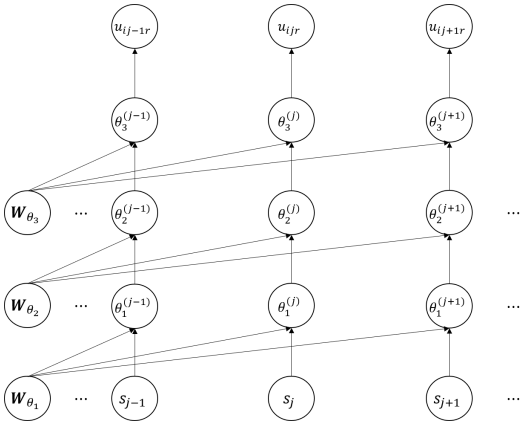

図 3: 受検者ネットワークのグラフィカル表現

ニューラルネットワークを構成する.

$$\boldsymbol{\beta}_1^{(i)} = \tanh\left(\boldsymbol{W}^{(\beta_1)}\boldsymbol{s}_i + \boldsymbol{\tau}^{(\beta_1)}\right) \tag{11}$$

$$\boldsymbol{\beta}_2^{(i)} = \tanh\left(\boldsymbol{W}^{(\beta_2)}\boldsymbol{\beta}_1^{(i)} + \boldsymbol{\tau}^{(\beta_2)}\right) \tag{12}$$

$$\boldsymbol{\beta}_3^{(i)} = \boldsymbol{W}^{(\beta_3)}\boldsymbol{\beta}_2^{(i)} + \boldsymbol{\tau}^{(\beta_3)} \tag{13}$$

$\boldsymbol{W}^{(\beta_1)}, \boldsymbol{W}^{(\beta_2)}, \boldsymbol{W}^{(\beta_3)}$ は以下の重みパラメータ行列である.

$$\boldsymbol{W}^{(\beta_1)} = \begin{pmatrix} w_{11}^{(\beta_1)} & w_{12}^{(\beta_1)} & \dots & w_{1I}^{(\beta_1)} \\ w_{21}^{(\beta_1)} & w_{22}^{(\beta_1)} & \dots & w_{2I}^{(\beta_1)} \\ \vdots & \vdots & \ddots & \vdots \\ w_{|\boldsymbol{\beta}_1|1}^{(\beta_1)} & w_{|\boldsymbol{\beta}_1|2}^{(\beta_1)} & \dots & w_{|\boldsymbol{\beta}_1|I}^{(\beta_1)} \end{pmatrix}$$

$$\boldsymbol{W}^{(\beta_2)} = \begin{pmatrix} w_{11}^{(\beta_2)} & w_{12}^{(\beta_2)} & \dots & w_{1|\boldsymbol{\beta}_1|}^{(\beta_2)} \\ w_{21}^{(\beta_2)} & w_{22}^{(\beta_2)} & \dots & w_{2|\boldsymbol{\beta}_1|}^{(\beta_2)} \\ \vdots & \vdots & \ddots & \vdots \\ w_{|\boldsymbol{\beta}_2|1}^{(\beta_2)} & w_{|\boldsymbol{\beta}_2|2}^{(\beta_2)} & \dots & w_{|\boldsymbol{\beta}_2||\boldsymbol{\beta}_1|}^{(\beta_2)} \end{pmatrix}$$

$$\boldsymbol{W}^{(\beta_3)} = \begin{pmatrix} w_{11}^{(\beta_3)} & w_{12}^{(\beta_3)} & \dots & w_{1|\boldsymbol{\beta}_2|}^{(\beta_3)} \\ w_{21}^{(\beta_3)} & w_{22}^{(\beta_3)} & \dots & w_{2|\boldsymbol{\beta}_2|}^{(\beta_3)} \\ \vdots & \vdots & \ddots & \vdots \\ w_{K1}^{(\beta_3)} & w_{K2}^{(\beta_3)} & \dots & w_{K|\boldsymbol{\beta}_2|}^{(\beta_3)} \end{pmatrix}$$

また，$\boldsymbol{\tau}^{(\beta_1)} = \left(\tau_1^{(\beta_1)}, \tau_2^{(\beta_1)}, \ldots, \tau_{|\boldsymbol{\beta}_1|}^{(\beta_1)}\right)$，$\boldsymbol{\tau}^{(\beta_2)} = \left(\tau_1^{(\beta_2)}, \tau_2^{(\beta_2)}, \ldots, \tau_{|\boldsymbol{\beta}_2|}^{(\beta_2)}\right)$，$\boldsymbol{\tau}^{(\beta_3)} = \left(\tau_1^{(\beta_3)}, \tau_2^{(\beta_3)}, \ldots, \tau_K^{(\beta_3)}\right)$ はバイアスパラメータベクトルである．出力 $\beta_3^{(i)}$ を課題 i において，評価カテゴリー k を得る困難度を表すパラメータとみなす．困難度パラメータを推定する際に課題間の独立性を仮定していないことが特徴である．

同様に，評価者ネットワークでは r 番目の課題を表現する one-hot vector $\boldsymbol{s}_r \in \mathbb{R}^R$ を入力とする．s_r は r 番目の要素のみが 1，他の要素が 0 であり，以下のように 4 層のニューラルネットワークを構成する．

$$\boldsymbol{d}_1^{(r)} = \tanh\left(\boldsymbol{W}^{(d_1)}\boldsymbol{s}_r + \boldsymbol{\tau}^{(d_1)}\right) \tag{14}$$

$$\boldsymbol{d}_2^{(r)} = \tanh\left(\boldsymbol{W}^{(d_2)}\boldsymbol{d}_1^{(r)} + \boldsymbol{\tau}^{(d_2)}\right) \tag{15}$$

$$\boldsymbol{d}_3^{(r)} = \boldsymbol{W}^{(d_3)}\boldsymbol{d}_2^{(r)} + \boldsymbol{\tau}^{(d_3)} \tag{16}$$

$\boldsymbol{W}^{(d_1)}, \boldsymbol{W}^{(d_2)}, \boldsymbol{W}^{(d_3)}$ は以下の重みパラメータ行列である．

$$\boldsymbol{W}^{(d_1)} = \begin{pmatrix} w_{11}^{(d_1)} & w_{12}^{(d_1)} & \ldots & w_{1R}^{(d_1)} \\ w_{21}^{(d_1)} & w_{22}^{(d_1)} & \ldots & w_{2R}^{(d_1)} \\ \vdots & \vdots & \ddots & \vdots \\ w_{|\boldsymbol{d}_1|1}^{(d_1)} & w_{|\boldsymbol{d}_1|2}^{(d_1)} & \ldots & w_{|\boldsymbol{d}_1|R}^{(d_1)} \end{pmatrix}$$

$$\boldsymbol{W}^{(d_2)} = \begin{pmatrix} w_{11}^{(d_2)} & w_{12}^{(d_2)} & \ldots & w_{1|\boldsymbol{d}_1|}^{(d_2)} \\ w_{21}^{(d_2)} & w_{22}^{(d_2)} & \ldots & w_{2|\boldsymbol{d}_1|}^{(d_2)} \\ \vdots & \vdots & \ddots & \vdots \\ w_{|\boldsymbol{d}_2|1}^{(d_2)} & w_{|\boldsymbol{d}_2|2}^{(d_2)} & \ldots & w_{|\boldsymbol{d}_2||\boldsymbol{d}_1|}^{(d_2)} \end{pmatrix}$$

$$\boldsymbol{W}^{(d_3)} = \begin{pmatrix} w_{11}^{(d_3)} & w_{12}^{(d_3)} & \ldots & w_{1|\boldsymbol{d}_2|}^{(d_3)} \\ w_{21}^{(d_3)} & w_{22}^{(d_3)} & \ldots & w_{2|\boldsymbol{d}_2|}^{(d_3)} \\ \vdots & \vdots & \ddots & \vdots \\ w_{K1}^{(d_3)} & w_{K2}^{(d_3)} & \ldots & w_{K|\boldsymbol{d}_2|}^{(d_3)} \end{pmatrix}$$

また，$\boldsymbol{\tau}^{(d_1)} = \left(\tau_1^{(d_1)}, \tau_2^{(d_1)}, \ldots, \tau_{|\boldsymbol{d}_1|}^{(d_1)}\right)$, $\boldsymbol{\tau}^{(d_2)} = \left(\tau_1^{(d_2)}, \tau_2^{(d_2)}, \ldots, \tau_{|\boldsymbol{d}_2|}^{(d_2)}\right)$, $\boldsymbol{\tau}^{(d_3)} = \left(\tau_1^{(d_3)}, \tau_2^{(d_3)}, \ldots, \tau_K^{(d_3)}\right)$ はバイアスパラメータベクトルである．出力 $d_3^{(r)}$ を評価カテゴリー k における評価者 r の厳しさを表すパラメータとみなす．厳しさパラメータを推定する際に評価者の独立性を仮定していないことが特徴である．

次に，受検者の能力パラメータ，課題の困難度パラメータと評価者の厳しさパラメータを用いて受検者の課題への反応をモデル化する．具体的には，以下のように隠れ層 $\boldsymbol{h}^{(i,j,r)} = (h_1^{(i,j,r)}, h_2^{(i,j,r)}, \cdots, h_K^{(i,j,r)})$ を求め，課題 i において受検者 j が評価者 r によって評価カテゴリー k を得る確率 $\hat{\boldsymbol{u}}_{ijr} = [\hat{u}_{ijr1}, \hat{u}_{ijr2}, \cdots, \hat{u}_{ijrK}]$ を算出し，モデルの出力とする．

$$h_k^{(i,j,r)} = \sum_{l=1}^{k} (\theta_3^{(j)} - \beta_{3l}^{(i)} - d_{3l}^{(r)}) \tag{17}$$

$$\begin{aligned} \hat{u}_{ijrk} &= \mathrm{softmax}(\boldsymbol{h}^{(i,j,r)}, k) \\ &= \frac{\exp\left(h_k^{(i,j,r)}\right)}{\sum_{k'} \exp\left(h_{k'}^{(i,j,r)}\right)} \end{aligned} \tag{18}$$

ここでは IRT と同様の解釈ができるようパラメータ構成を模倣した深層学習モデルを提案している．しかし IRT とは異なり，受検者の母集団と独立性を仮定せずに課題への反応予測を最大にするようにモデルが構成されている．これにより，異なるテストの受検者の能力推定値も利用しながら，最も予測精度が高くなるように能力を推定できる．

5.2 パラメータ学習

一般に，深層学習では微分可能な損失関数を定義し，誤差逆伝播法によりパラメータを学習する，提案モデルでは，損失関数として，以下のような分類誤差を表すクロスエントロピー l を用いる．

$$l = -\sum_{k=1}^{K} v_{ijrk} \log \hat{u}_{ijrk} \tag{19}$$

ここで $\boldsymbol{v}_{ijr} \in \mathbb{R}^K$ は，反応データが $u_{ijr} = k$ であったとき，k 番目の要素のみ 1，他を 0 とした one-hot vector を表す．

提案モデルは，パフォーマンス評価データをもとに，adaptive moment estimation (Adam) [56] と呼ばれる最適化アルゴリズムに従い，損失関数が小さくなるようにすべてのパラメータを同時に更新する．

5.3 Adam

本研究で用いたAdamは，各パラメータの学習率を自動で調節する勾配法の一種である．学習率を固定する場合や他の最適化アルゴリズム [57, 58, 59, 60] よりも損失関数が小さくなる傾向にあり，近年，深層学習モデルによく用いられている．

Adamでは，学習途中の勾配が大きなパラメータはよく学習されているとみなし学習率を低くする．t 回目の学習において，各パラメータの勾配 $\boldsymbol{g}_t$ が与えられたとき，それまでの勾配の重み付き平均 $\boldsymbol{m}_t$ と勾配の二乗の重み付き平均 $\boldsymbol{var}_t$ は以下のように算出される．

$$\boldsymbol{m}_t = \gamma_1 \boldsymbol{m}_{t-1} + (1-\gamma_1)\boldsymbol{g}_t \tag{20}$$

$$\boldsymbol{var}_t = \gamma_2 \boldsymbol{var}_{t-1} + (1-\gamma_2)\boldsymbol{g}_t^2 \tag{21}$$

ここで，γ_1, γ_2 はチューニングパラメータであり，任意の値を設定する．

これらの推定バイアスを補正した $\boldsymbol{m}_t^* = \boldsymbol{m}_t/(1-\gamma_1^t), \boldsymbol{var}_t^* = \boldsymbol{var}_t/(1-\gamma_2^t)$ を用いて，t 回目の学習におけるすべてのパラメータベクトル $\boldsymbol{x}_t$ は以下のように更新する．

$$\boldsymbol{x}_t \leftarrow \boldsymbol{x}_{t-1} - \frac{\mu}{\sqrt{\boldsymbol{var}_t^*} + \epsilon}\boldsymbol{m}_t^* \tag{22}$$

μ は学習率の初期値，ϵ は発散を防ぐための微小な定数である．

6　シミュレーション実験

等化に関する研究や多母集団を仮定した研究では，実データの収集に膨大なコストと時間を要するため，できるだけ現実に近い条件に設定して，シミュレーションにより評価を行うことが一般的である [27, 61, 62, 63, 64, 65, 66]．本章ではシミュレーションデータに提案モデルと評価者特性を表すパラメータを付与した IRT を適用し，受検者と評価者の独立ランダムサンプリングが成り立たない場合での提案モデルの有効性を示す．具体的には，受検者と評価者の割り当て方法，受検者数，評価者数，共通課題数，共通評価者数を変化させた際の能力推定精度を比較し，受検者と評価者の独立ランダムサンプリングが成り立たないシミュレーションデータに対して，少ない共通課題・共通評価者数でも提案モデルが高精度な能力推定を行えることを示す．

本実験のシミュレーションデータは，**4.4** の図 1 と同様の二つのパフォーマンステスト（テスト A とテスト B）で構成されるデータを用いた．各テストは I 個の課題，各受検者グループは J 人，各評価者グループは R 人よって採点される状況を想定する．シミュレーションデータの生成は，最も一般的な評価者パラメータを持つ IRT モデルの MFRM より行った．

本実験では，受検者と評価者の同一母集団からのランダムサンプリングを仮定した「ランダム割り当て」と，ランダムサンプリングが仮定できない「システム割り当て」と，各テストの受検者・評価者が異なる母集団からサンプリングした「多母集団割り当て」の三つの方法で受検者と評価者を各テストに割り当て，パフォーマンス評価データを生成した．

ランダム割り当て

1）以下の分布からパラメータを発生させる．

$$\theta_j,\, \beta_i,\, \beta_r,\, d_k \sim N(0.0, 1.0) \tag{23}$$

ここで，$N(\mu, \sigma)$ は平均 μ，標準偏差 σ の正規分布を表す．

2）発生させた能力・困難度・厳しさパラメータを持つ受検者・課題・評価者を各テストにランダムに割り当てる．

システム割り当て

1）式 (23) に従い，パラメータを発生させる．

2）発生させた能力パラメータを持つ受検者をパラメータの昇順に並び替え二分割し，下位をテスト A，上位をテスト B に割り当てる．同様に評価者に関しても厳しさパラ

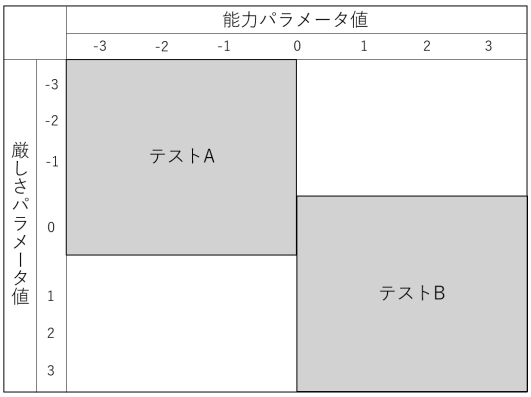

図 4: システム割り当て

メータの昇順に並び替え，各テストに割り当てる．課題に関しては，発生させた困難度パラメータを持つ課題を各テストにランダムに割り当てる．

図 4 にシステム割り当ての概要図を示す．テスト A には能力パラメータ値が下位の受検者と厳しさパラメータ値が下位の評価者を割り当て，テスト B には能力パラメータ値が上位の受検者と厳しさパラメータ値が上位の評価者を割り当てる．

多母集団割り当て

1）テスト A とテスト B について，能力パラメータと厳しさパラメータをそれぞれ異なる分布から発生させる．具体的に，テスト A では能力パラメータと厳しさパラメータはそれぞれ $\theta_j,\ \beta_r \sim N(-1.0, 0.5)$ から，テスト B ではそれぞれ $\theta_j,\ \beta_r \sim N(1.0, 0.5)$ から発生させる．ただし，共通評価者の厳しさパラメータは $\beta_r \sim N(1.0, 0.5)$ から発生させる．課題の困難度パラメータに関しては各テスト共通で，式 (23) の分布から発生させる．

2）発生させた能力・困難度・厳しさパラメータを持つ受検者・課題・評価者をランダムに割り当てる．

上記のシミュレーションデータを用いて，以下の手順で能力推定精度の比較を行った．

（1）生成したデータを用いて，MFRM のパラメータ推定を行った．推定はマルコフ連鎖モンテカルロ法（Markov chain Monte Carlo: MCMC）[7, 14, 18] を用いた期待事後確率推定法（Expected a posteriori）で行い，すべてのパラメータを同時に推定した．なお推定に用いる事前分布は式 (23) の分布を用いた．

（2）生成したデータを用いて，提案モデルのパラメータ推定を行った．なお，本研究

表 1: 共通するチューニングパラメータの値

パラメータ	値
エポック数	300
μ	0.01
ϵ	10^{-8}
γ_1	0.9
γ_2	0.999

の提案モデルの実装は，深層学習のフレームワークの一つである Chainer[1]を用い，バッチ学習でパラメータを学習した．また，すべてのシミュレーション・実データ実験に共通するパラメータは表 1 の値を用いた．これらのパラメータのうち，$\epsilon, \gamma_1, \gamma_2$ に関しては，先行研究 [56] の値を用いた．また提案モデルに関しては，$\boldsymbol{\theta}_1^{(j)}, \boldsymbol{\theta}_2^{(j)}, \boldsymbol{\beta}_1^{(i)}, \boldsymbol{\beta}_2^{(i)}, \boldsymbol{d}_1^{(r)}, \boldsymbol{d}_2^{(r)}$ のノード数はすべて共通の値を用い，ノード数を 100 から 400 まで 20 ずつ変化させ実験を行った．

（３）能力パラメータの真値と，（１）で求めた推定値との平均平方二乗誤差（Root Mean Square Error: RMSE）を算出した．また，同様に能力パラメータの真値と，（２）で求めた推定値との RMSE も算出した．ただし，（２）で求めた能力推定値 $\theta_3^{(j)}$ は，能力推定値の平均 $\mathrm{mean}(\boldsymbol{\theta}_3)$ と標準偏差 $\mathrm{sd}(\boldsymbol{\theta}_3)$ をもとに以下の平均０，分散１の分布に標準化した値を用いた．

$$\hat{\theta}_3^{(j)} = \frac{\theta_3^{(j)} - \mathrm{mean}(\boldsymbol{\theta}_3)}{\mathrm{sd}(\boldsymbol{\theta}_3)} \tag{24}$$

（４）上記の手順を 10 回繰り返し，RMSE の平均を算出した．

以上の実験をそれぞれランダム割り当て，システム割り当て，多母集団割り当ての３つのデータ割り当て方法に対して行い，受検者数，評価者数，共通課題数，共通評価者数を変化させて行った．なお，課題数は $I = 5$，評価カテゴリー数は $K = 5$ とした．

実験結果を表 2 に示す．RMSE は値が小さいほど推定精度が高いとみなせる．表 2 から，受検者・評価者をランダムに割り当てた場合は，多くの条件で，MFRM の方が精度が高いことがわかる．ランダム割り当てではすべての受検者と評価者が同一母

[1] https://chainer.org/

集団からランダムにサンプリングされており，MFRM からのデータ発生条件と同一であるために高精度な推定ができたと考えられる．

一方，システム割り当てでは，多くの条件で提案モデルの能力推定精度が MFRM を上回っている．提案モデルは受検者と評価者の独立性を仮定せず，他の受検者の能力推定値や評価者の評価特性値との関連性を考慮しながら推定を行うことで，テスト間における能力特性や評価特性の違いを自動的に修正できたと考えられる．また，各テストの受検者と評価者が異なる母集団に属する場合でも，多くの条件で提案モデルの能力推定精度が MFRM を上回った．

また表 2 から，共通評価者が 0 人のときは MFRM よりも提案モデルの能力推定精度は低くなっているが，共通評価者が 0 人の場合は提案モデル，MFRM のどちらを用いた場合でも，RMSE は共通評価者を含む場合に比べ大きくなっており，高精度な等化ができないことがわかる．また共通課題数，共通評価者数が増えた場合，MFRM の方が精度が高くなっている条件もある．共通課題数や共通評価者数を増やした場合，IRT モデルの推定精度は向上ことが知られており [24]，提案モデル，MFRM のどちらを用いても高精度な等化が可能となる．しかし，共通課題数や共通評価者数の増加はテストの信頼性低下や評価者の採点負担の増加を引き起こすため [29, 30, 31, 32, 33, 34, 35, 36, 37, 38]，少ない共通課題数や共通評価者数でも高い推定精度を示す提案モデルが有効であることがわかる．

表 2: シミュレーションによる能力パラメータの推定精度

共通課題数	共通評価者数	受検者数 25																	
		評価者数 5						評価者数 10						評価者数 25					
		ランダム割り当て		システム割り当て		多母集団割り当て		ランダム割り当て		システム割り当て		多母集団割り当て		ランダム割り当て		システム割り当て		多母集団割り当て	
		MFRM	提案モデル	MFRM	提案モデル	MFRM	提案モデル	MFRM	提案モデル	MFRM	提案モデル	MFRM	提案モデル	MFRM	提案モデル	MFRM	提案モデル	MFRM	提案モデル
0	0	**0.309**	0.347	**0.791**	0.928	**0.935**	1.025	**0.305**	0.328	**0.758**	0.951	**0.955**	1.160	**0.194**	0.257	**0.751**	0.869	**0.943**	0.984
	1	**0.309**	0.343	0.710	**0.581**	0.859	**0.488**	**0.249**	0.283	0.688	**0.305**	0.906	**0.342**	**0.218**	0.258	0.658	**0.238**	0.911	**0.259**
	2	0.315	**0.314**	0.684	**0.568**	0.879	**0.573**	**0.229**	0.273	0.617	**0.330**	0.862	**0.332**	**0.188**	0.238	0.661	**0.282**	0.891	**0.365**
	3	**0.301**	0.330	0.718	**0.637**	0.882	**0.608**	**0.245**	0.299	0.697	**0.433**	0.902	**0.437**	**0.206**	0.266	0.701	**0.265**	0.836	**0.209**
1	0	**0.287**	0.329	**0.724**	0.921	**0.877**	1.074	**0.231**	0.268	**0.688**	0.885	**0.805**	1.023	**0.219**	0.257	**0.603**	0.919	**0.738**	1.001
	1	**0.295**	0.308	0.385	**0.334**	0.457	**0.268**	**0.242**	0.267	0.346	**0.258**	0.411	**0.238**	**0.167**	0.187	0.343	**0.227**	0.415	**0.239**
	2	**0.296**	0.325	0.357	**0.339**	0.430	**0.308**	**0.260**	0.276	0.272	**0.228**	0.355	**0.276**	**0.184**	0.226	0.227	**0.219**	0.292	**0.198**
	3	**0.252**	0.265	**0.381**	0.393	0.412	**0.277**	**0.221**	0.258	0.265	**0.235**	0.272	**0.240**	**0.187**	0.218	0.214	**0.208**	0.286	**0.201**
2	0	**0.299**	0.325	**0.715**	0.945	**0.868**	1.106	**0.233**	0.261	**0.658**	0.871	**0.794**	1.019	**0.154**	0.175	**0.576**	0.936	**0.773**	1.135
	1	**0.272**	0.294	0.334	**0.284**	0.396	**0.301**	**0.222**	0.253	0.295	**0.284**	0.403	**0.288**	**0.186**	0.214	0.236	**0.229**	0.412	**0.267**
	2	**0.294**	0.307	0.258	**0.249**	0.315	**0.267**	**0.219**	0.251	0.269	**0.257**	0.296	**0.222**	**0.168**	0.185	0.247	**0.229**	0.327	**0.261**
	3	**0.286**	0.319	**0.355**	0.383	0.329	**0.288**	**0.220**	0.235	0.223	**0.211**	0.288	**0.231**	**0.125**	0.170	**0.232**	0.240	0.235	**0.216**
3	0	**0.304**	0.350	**0.655**	0.952	**0.862**	1.116	**0.239**	0.263	**0.658**	0.892	**0.787**	1.091	**0.218**	0.252	**0.589**	0.909	**0.806**	1.114
	1	**0.294**	0.295	0.332	**0.289**	0.408	**0.285**	**0.266**	0.269	0.301	**0.260**	0.419	**0.312**	**0.179**	0.206	0.290	**0.245**	0.364	**0.311**
	2	**0.292**	0.303	0.327	**0.299**	0.347	**0.277**	**0.199**	0.214	0.213	**0.208**	0.317	**0.276**	**0.177**	0.195	**0.205**	0.228	0.305	**0.239**
	3	**0.292**	0.329	0.301	**0.283**	0.315	**0.301**	**0.214**	0.232	0.235	**0.221**	0.248	**0.219**	**0.175**	0.213	**0.173**	0.187	0.227	**0.183**
共通課題数	共通評価者数	受検者数 50																	
		評価者数 5						評価者数 10						評価者数 25					
		ランダム割り当て		システム割り当て		多母集団割り当て		ランダム割り当て		システム割り当て		多母集団割り当て		ランダム割り当て		システム割り当て		多母集団割り当て	
		MFRM	提案モデル	MFRM	提案モデル	MFRM	提案モデル	MFRM	提案モデル	MFRM	提案モデル	MFRM	提案モデル	MFRM	提案モデル	MFRM	提案モデル	MFRM	提案モデル
0	0	**0.303**	0.314	**0.798**	0.996	**0.998**	0.919	**0.193**	0.225	**0.795**	0.949	**0.970**	1.193	**0.164**	0.208	**0.767**	0.854	**0.987**	1.062
	1	**0.286**	0.299	0.740	**0.482**	0.934	**0.557**	**0.234**	0.258	0.713	**0.465**	0.917	**0.402**	**0.199**	0.259	0.750	**0.289**	0.933	**0.311**
	2	**0.282**	0.312	0.769	**0.516**	0.943	**0.604**	0.223	**0.220**	0.741	**0.547**	0.918	**0.468**	**0.168**	0.213	0.728	**0.282**	0.949	**0.363**
	3	**0.287**	0.290	0.766	**0.587**	0.882	**0.916**	**0.228**	0.277	0.730	**0.388**	0.979	**0.572**	**0.160**	0.241	0.726	**0.374**	0.967	**0.361**
1	0	**0.274**	0.311	**0.751**	0.962	**0.925**	1.070	**0.218**	0.263	**0.712**	0.989	**0.875**	1.098	**0.172**	0.203	**0.609**	0.898	**0.838**	1.128
	1	**0.279**	0.300	0.381	**0.315**	0.459	**0.256**	**0.223**	0.247	0.347	**0.269**	0.451	**0.321**	**0.149**	0.163	0.245	**0.211**	0.430	**0.390**
	2	**0.285**	0.302	0.307	**0.270**	0.434	**0.312**	**0.220**	0.222	0.251	**0.215**	0.289	**0.208**	**0.136**	0.194	0.200	**0.178**	0.306	**0.298**
	3	0.298	**0.294**	0.339	**0.281**	0.355	**0.280**	**0.198**	0.210	0.222	**0.193**	0.273	**0.229**	**0.147**	0.175	**0.167**	0.158	**0.214**	0.234
2	0	**0.265**	0.290	**0.734**	0.916	**0.937**	1.156	**0.219**	0.232	**0.666**	0.907	**0.881**	1.073	**0.162**	0.190	**0.638**	0.927	**0.786**	1.044
	1	**0.267**	0.283	0.311	**0.254**	0.435	**0.284**	**0.213**	0.211	0.253	**0.210**	0.448	**0.311**	**0.127**	0.168	0.242	**0.234**	0.318	**0.309**
	2	**0.245**	0.253	0.254	**0.238**	0.332	**0.283**	**0.203**	0.208	0.214	**0.204**	0.271	**0.232**	**0.137**	0.162	**0.211**	0.230	0.247	**0.229**
	3	**0.258**	0.278	0.283	**0.268**	0.305	**0.263**	**0.206**	0.218	0.232	**0.213**	0.254	**0.232**	**0.134**	0.154	**0.194**	0.246	**0.203**	0.178
3	0	**0.282**	0.301	**0.724**	0.908	**0.937**	1.124	**0.216**	0.245	**0.704**	0.902	**0.878**	1.061	**0.158**	0.185	**0.658**	0.943	**0.805**	1.083
	1	**0.284**	0.294	0.316	**0.277**	0.379	**0.286**	**0.201**	0.217	0.303	**0.262**	0.332	**0.230**	**0.151**	0.172	0.231	**0.196**	0.321	**0.237**
	2	**0.268**	0.279	0.277	**0.255**	0.329	**0.282**	**0.212**	0.219	0.228	**0.209**	0.257	**0.197**	**0.144**	0.162	0.195	**0.182**	**0.286**	0.308
	3	**0.261**	0.271	**0.284**	0.293	0.278	**0.260**	**0.210**	0.219	0.194	**0.203**	0.227	**0.193**	**0.166**	0.184	**0.167**	0.185	**0.248**	0.298
共通課題数	共通評価者数	受検者数 100																	
		評価者数 5						評価者数 10						評価者数 25					
		ランダム割り当て		システム割り当て		多母集団割り当て		ランダム割り当て		システム割り当て		多母集団割り当て		ランダム割り当て		システム割り当て		多母集団割り当て	
		MFRM	提案モデル	MFRM	提案モデル	MFRM	提案モデル	MFRM	提案モデル	MFRM	提案モデル	MFRM	提案モデル	MFRM	提案モデル	MFRM	提案モデル	MFRM	提案モデル
0	0	**0.282**	0.314	**0.812**	1.059	**1.006**	1.097	**0.206**	0.227	**0.791**	1.052	**0.996**	1.095	**0.161**	0.197	**0.800**	0.990	**0.972**	1.041
	1	**0.256**	0.284	0.792	**0.548**	0.995	**0.721**	**0.208**	0.247	0.807	**0.594**	0.953	**0.542**	**0.143**	0.176	0.772	**0.356**	0.970	**0.316**
	2	**0.257**	0.276	0.774	**0.594**	0.987	**0.523**	**0.190**	0.211	0.791	**0.464**	0.941	**0.505**	**0.145**	0.248	0.756	**0.243**	0.967	**0.331**
	3	**0.262**	0.286	0.806	**0.677**	0.982	**0.632**	**0.235**	0.259	0.766	**0.589**	0.988	**0.467**	**0.156**	0.222	0.785	**0.333**	0.952	**0.356**
1	0	**0.267**	0.289	**0.815**	0.991	**0.969**	1.107	**0.201**	0.224	**0.740**	0.986	**0.930**	1.102	**0.147**	0.203	**0.665**	0.976	**0.878**	1.088
	1	**0.260**	0.268	0.329	**0.247**	0.473	**0.270**	**0.180**	0.194	0.346	**0.292**	0.376	**0.277**	**0.133**	0.171	0.246	**0.245**	0.311	**0.224**
	2	**0.261**	0.274	0.329	**0.247**	0.473	**0.270**	**0.180**	0.194	0.346	**0.292**	0.376	**0.277**	**0.126**	0.149	0.234	**0.218**	0.228	**0.200**
	3	**0.257**	0.258	0.327	**0.290**	0.413	**0.292**	**0.197**	0.205	0.278	**0.259**	0.251	**0.193**	**0.127**	0.145	**0.152**	0.199	**0.216**	0.237
2	0	**0.264**	0.298	**0.775**	1.006	**0.948**	1.103	**0.211**	0.235	**0.730**	0.952	**0.922**	1.092	**0.145**	0.188	**0.680**	0.982	**0.850**	1.103
	1	**0.254**	0.256	0.266	**0.243**	0.443	**0.342**	**0.203**	0.219	0.254	**0.215**	0.376	**0.277**	**0.127**	0.168	0.242	**0.234**	0.318	**0.309**
	2	0.258	0.258	0.270	**0.242**	0.364	**0.270**	**0.196**	0.205	0.213	**0.196**	0.271	**0.236**	**0.130**	0.173	**0.187**	0.239	0.209	**0.191**
	3	**0.254**	0.263	0.274	**0.225**	0.330	**0.289**	**0.186**	0.202	0.218	**0.192**	0.222	**0.209**	**0.117**	0.149	**0.150**	0.175	**0.187**	0.198
3	0	**0.268**	0.295	**0.795**	1.013	**0.974**	1.100	**0.216**	0.240	**0.731**	1.012	**0.915**	1.130	**0.140**	0.183	**0.639**	0.951	**0.840**	1.090
	1	**0.255**	0.266	0.308	**0.251**	0.434	**0.328**	**0.194**	0.209	0.256	**0.237**	0.343	**0.283**	**0.141**	0.168	0.224	**0.161**	0.308	**0.286**
	2	**0.241**	0.247	0.244	**0.234**	0.287	**0.252**	**0.182**	0.188	0.198	**0.186**	0.244	**0.226**	**0.129**	0.152	0.162	**0.159**	0.214	**0.191**
	3	**0.253**	0.259	0.261	**0.249**	0.284	**0.271**	**0.187**	0.203	0.203	**0.194**	**0.253**	0.260	**0.134**	0.148	**0.151**	0.170	**0.211**	0.283

7　実データ実験

本章では実データを用い，提案モデルが等化処理を含むパフォーマンス評価データに対しても有効であることを示す．具体的には，受検者の能力推定精度の比較を IRT モデルと提案モデルとで行う．以降では簡単のために，式 (2)〜(6) のモデルをそれぞれ「Patz1999」,「Usami2010」,「Uto2020」,「Ueno2008」,「Uto2016」表記する．

7.1　実データ概要

本節では，実データの概要を説明する．

レポートデータ

レポートデータは，大学生が提出した e ラーニングでのレポート課題を，コースチューターが採点したパフォーマンス評価データ行列である [67]．受検者数は 30，課題数は 5，評価者数は 5，評価カテゴリー数は 5 であり，欠測値の割合は 9.7%である．

ピアアセスメントデータ

ピアアセスメントデータは，ライティング課題を大学生が互いに評価したパフォーマンス評価データ行列である [67]．学習者数は 34，課題数は 4，評価者数は 30，評価カテゴリー数は 5 であり，欠測値の割合は 0%である．

7.2　能力推定値の信頼性

本節では，実データにおける能力推定値の信頼性を提案モデルと IRT モデルで比較する．実験は以下の手順で行った．

（1）提案モデル，MFRM，Patz1999，Usami2010，Uto2020，Ueno2008，Uto2016 を用いて，実データから能力パラメータを推定した．なお，IRT モデルに関しては MCMC を用いた EAP 推定法で推定を行った．なお，事前分布は表 3 の値を用いた．

（2）図 1 のように，レポートデータの場合は各テスト受検者数 15，課題数 2，評価者数 2 のテストに，ピアアセスメントデータの場合は学習者数 17，課題数 2，評価者数 15 のテストにそれぞれ分割する．このとき，課題と評価者は完全データからランダムに選択する．

（3）提案モデル，IRT モデルを用い，（1）で求めた課題・評価者パラメータを所与として能力パラメータを算出する．

表 3: IRT モデルのパラメータ分布

$\log \alpha_i \sim N(0.1, 0.4), \log \alpha_r \sim N(0.0, 0.5)$
$\beta_i, \beta_r, \beta_{ik}, \epsilon_r, \rho_{ir}, d_{ik}, d_{rk}, d_k, b_i, d_r, \theta_j \sim N(0.0, 1.0)$
$b_{ik}, \epsilon_{rk} \sim MN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
$\boldsymbol{\mu} = \{-2.0, -0.75, 0.75, 2.0\}$
$\boldsymbol{\Sigma} = \begin{pmatrix} 0.16 & 0.10 & 0.04 & 0.04 \\ 0.10 & 0.16 & 0.10 & 0.04 \\ 0.04 & 0.10 & 0.16 & 0.10 \\ 0.04 & 0.04 & 0.10 & 0.16 \end{pmatrix}$

（4）（1）で求めた能力パラメータと（3）で求めた能力パラメータの相関係数を算出する．相関係数には，ピアソンの積率相関係数を用いる．

（5）（2）～（4）の手順を 10 回繰り返し，相関係数の平均を求める．

以上の実験をテスト間の共通課題数，共通評価者数を変化させて行った．なお提案モデルは **7.** と同様にネットワークのノード数を変化させて実験を行った．また，能力推定値を比較する指標には相関係数の他に RMSE も広く用いられているが，本実験で RMSE を用いると，能力値分布に標準正規分布を仮定している IRT モデルでは推定値を中央に縮約して得られるので RMSE が減じられ正しい比較ができない．そこで信頼性評価で用いられる相関係数を用いて評価する．

実験結果を表 4 に示す．一般にテスト理論では，本実験の相関係数の値が大きいほど，能力推定値の信頼性が高いとみなせる．表 4 から多くの条件で提案モデルが従来の IRT モデルより高い信頼性を示したことがわかる．

さらに，各モデルの能力推定値分布の非正規性を示すため，実験手順（3）で求めた能力推定値の歪度と尖度の平均を表 5 に示す．どちらの指標も 0 に近ければ正規分布に近い分布であることを示す．表 5 から，提案モデルはどちらのデータセットに対しても，能力推定値分布は正規分布から大きく乖離していることがわかる．このことから，提案モデルは能力分布が正規分布から乖離するほど，信頼性の高い能力推定が行えることがわかる．

表 4: 実データにおける能力推定精度

共通	共通評	レポートデータ						
課題数	価者数	提案モデル	MFRM	Patz1999	Usami2010	Uto2020	Ueno2008	Uto2016
0	0	**0.766**	0.708	0.712	0.743	0.747	0.738	0.740
	1	**0.819**	0.742	0.718	0.796	0.799	0.801	0.785
	2	**0.813**	0.739	0.749	0.748	0.755	0.765	0.751
1	0	**0.831**	0.758	0.770	0.791	0.800	0.808	0.789
	1	**0.787**	0.755	0.737	0.744	0.765	0.756	0.739
	2	**0.796**	0.746	0.754	0.744	0.748	0.753	0.726
2	0	**0.816**	0.745	0.771	0.776	0.781	0.775	0.755
	1	**0.846**	0.768	0.775	0.803	0.806	0.807	0.789
	2	**0.825**	0.782	0.814	0.780	0.787	0.782	0.758
共通	共通評	ピアアセスメントデータ						
課題数	価者数	提案モデル	MFRM	Patz1999	Usami2010	Uto2020	Ueno2008	Uto2016
0	0	**0.881**	0.861	0.789	0.879	0.872	0.863	0.867
	1	**0.874**	0.854	0.780	0.873	0.862	0.861	0.863
	2	**0.877**	0.857	0.748	0.873	0.864	0.857	0.857
	3	**0.877**	0.849	0.773	0.873	0.863	0.862	0.860
1	0	0.836	0.821	0.772	0.849	**0.850**	0.835	0.844
	1	0.870	0.870	0.814	**0.872**	**0.872**	0.863	0.870
	2	**0.885**	0.847	0.820	0.868	0.858	0.856	0.850
	3	**0.869**	0.845	0.809	0.866	0.861	0.854	0.843
2	0	0.849	0.827	0.790	**0.856**	0.845	0.842	0.837
	1	**0.877**	0.856	0.839	0.870	0.860	0.851	0.844
	2	0.873	0.856	0.833	**0.883**	0.872	0.870	0.869
	3	**0.863**	0.844	0.822	0.856	0.848	0.842	0.841

表 5: 能力推定値の非正規性

	レポートデータ		ピアアセスメントデータ	
	歪度	尖度	歪度	尖度
提案モデル	1.953	7.166	-1.317	2.571
MFRM	0.620	0.649	-0.941	1.456
Patz1999	0.921	1.948	-0.768	1.032
Usami2010	0.628	0.676	-1.118	2.017
Uto2020	0.889	1.990	-1.027	1.699
Ueno2008	0.685	0.920	-1.021	1.607
Uto2016	0.657	0.955	-1.102	2.062

8 むすび

本研究では，受検者の母集団と独立ランダムサンプリングを仮定しないパフォーマンス予測モデルの Deep-IRT モデルに，評価者パラメータを含んだモデルを提案した．

提案モデルは受検者，課題，評価者の 3 つの独立したニューラルネットワークを入力とし，3 つのネットワークから出力されるパラメータを組み合わせて，課題への正答確率をモデル化する．受検者ネットワークの出力を能力パラメータとみなした．

シミュレーション・実データ実験により以下の利点があることがわかった．

１）受検者と評価者の独立ランダムサンプリングが仮定できない場合でも，能力を高精度に推定できる．

２）受検者と評価者の母集団が単一でない場合でも，能力を高精度に推定できる．

これらにより，提案モデルは等化の際必要な共通課題や共通評価者が少ない場合で特に有効であり，実データに対しても IRT モデルよりも信頼性の高い能力推定が行えることが明らかとなった．

提案モデルは受検者・課題・評価者の 3 相データへの適応を目的としたが，今後はルーブリック評価における評価観点を含めた 4 相データへの適応や，提案モデルを用いたピアアセスメントにおけるグループ構成の最適化手法の開発，e ラーニングを用いたピアアセスメント [68, 69, 70, 71, 71, 72] における提案モデルの適応などを行っていきたい．

謝辞

本論文を作成するにあたり，指導教員の植野真臣教授から，丁寧かつ熱心なご指導を賜りました．ここに感謝の意を表します．また，研究についてご助言をいただきました宇都雅輝准教授，川野秀一准教授，研究に関する議論や論文執筆についてご指摘いただきました先輩方，研究室の皆様に感謝いたします．

参考文献

[1] R. Schendel and A. Tolmie,“Beyond translation: adapting a performance-task-based assessment of critical thinking ability for use in Rwanda,” Assessment & Evaluation in Higher Education, vol.42, no.5, pp.673–689, 2017.

[2] Y. Abosalem, “Assessment techniques and students’ higher-order thinking skills,” Int. J. Secondary Education, vol.4, no.1, pp.1–11, 2016.

[3] Y. Rosen and M. Tager, “Making student thinking visible through a concept map in computer-based assessment of critical thinking,” J. Educational Computing Research, vol.50, no.2, pp.249–270, 2014.

[4] O.L. Liu, L. Frankel, and K.C. Roohr, “Assessing critical thinking in higher education: Current state and directions for next-generation assessment,” ETS Research Report Series, vol.2014, no.1, pp.1–23, 2014.

[5] H.J. Bernardin, S. Thomason, M.R. Buckley, and J.S. Kane,“Rater rating-level bias and accuracy in performance appraisals: The impact of rater personality, performance management competence, and rater accountability,” Human Resource Management, vol.55, no.2, pp.321–340, 2016.

[6] 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.

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

[8] N.L.A. Kassim, “Judging behaviour and rater errors: An application of the many-facet Rasch model,”GEMA Online Journal of Language Studies, vol.11, no.3, pp.179–197, 2011.

[9] C.M. Myford and E.W. Wolfe, “Detecting and measuring rater effects using many-facet Rasch measurement: Part I,” J. Appl. Meas., vol.4, pp.386–422, 2003.

[10] T. Eckes, “Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis,” Language Assessment Quarterly, vol.2, no.3, pp.197–221, 2005.

[11] 宇都雅輝，植野真臣，“ピアアセスメントにおける異質評価者に頑健な項目反応理論，” 信学論（D），vol.J101-D, no.1, pp.211–224, 2018.

[12] M. Uto and M. Ueno, ”A generalized many-facet Rasch model and its Bayesian estimation using Hamiltonian Monte Carlo,” Behaviormetrika, vol.47, issue.2, pp.469–496, 2020.

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

[14] 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.

[15] 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.

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

[17] J.M. Linacre, Many-faceted Rasch Measurement, MESA Press, 1989.

[18] R.J. Patz and B.W. Junker, “Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses,” J. Educational and Behavioral Statistics, vol.24, no.4, pp.342–366, 1999.

[19] R.J. Patz, B.W. Junker, M.S. Johnson, and L.T. Mariano, “The hierarchical rater model for rated test items and its application to large-scale educational assessment data,” J. Educational and Behavioral Statistics, vol.27, no.4, pp.341–366, 2002.

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

[21] 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, 2015.

[22] E. Muraki, C.M. Hombo, and Y.W. Lee,“Equating and linking of performance assessments,”Applied Psychological Measurement, vol.24, pp.325–337, 2000.

[23] G. Engelhard,“Constructing rater and task banks for performance assessments,”J. Outcome Measurement, vol.1, no.1, pp.19–33, 1997.

[24] J.M. Linacre,“A user’s guide to FACETS Rasch-model computer programs,”2014.

[25] 泉 毅，山野井真児，山田剛史，金森保智，対馬英樹,“共通項目数が等化の精度に及ぼす影響:大規模学力テストデータを用いた探索的研究，”教育実践学論集，vol.13, pp.49–57, 2012.

[26] W.J. van der Linden and M.D. Barrett,“Linking item response model parameters,”Psychometrika, vol.81, no.3, pp.650–673, Sept. 2016.

[27] 宇都雅輝,“評価者特性パラメータを付与した項目反応モデルに基づくパフォーマンステストの等化精度，”信学論（D），vol.J101-D, no.6, pp.895–905, 2018.

[28] M. Uto,“Accuracy of performance-test linking based on a many-facet Rasch model,”Behavior Research Methods, Springer, 2020.

[29] T. Ishii, P. Songmuang, and M. Ueno,“Maximum clique algorithm for uniform test forms assembly,”Artificial Intelligence in Education - 16th International Conference, AIED, pp.451–462, 2013.

[30] 石井隆稔，ソンムァン・ポクポン，植野真臣,“最大クリーク問題を用いた複数等質テスト自動構成，”信学論（D），vol.J97-D, no.2, pp270–280, 2014.

[31] T. Ishii, P. Songmuang, and M. Ueno, “Maximum clique algorithm and its approximation for uniform test form assembly,” IEEE Trans. Learning Technologies, vol.7, no.1, pp.83–95, 2014.

[32] 石井隆稔，植野真臣，“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，vol.11, no.1, pp.131–149, 2015.

[33] T. Ishii and M. Ueno,“Clique algorithm to minimize item exposure for uniform test forms assembly,” Artificial Intelligence in Education - 17th International Conference, AIED, pp.638–641, 2015.

[34] 石井隆稔，赤倉貴子，植野真臣，“複数等質テスト構成における整数計画問題を用いた最大クリーク探索の近似法，” 信学論（D），vol.J100-D, no.1, pp.47–59, 2017.

[35] T. Ishii and M. Ueno,“Algorithm for uniform test assembly using a maximum clique problem and integer programming,” Artificial Intelligence in Education – 18th International Conference, AIED, pp.102–112, 2017.

[36] W.D. Way, “Protecting the integrity of computerized testing item pools,” Educational Measurement: Issues and Practice, vol.17, no.4, pp.17–27, 1998.

[37] M. Uto, T. Nguyen and M. Ueno, “Group Optimization to Maximize Peer Assessment Accuracy Using Item Response Theory,” Artificial Intelligence in Education - 18th International Conference, AIED, pp.393–405, 2017.

[38] M. Uto, T. Nguyen and M. Ueno, “Group optimization to maximize peer assessment accuracy using item response theory and integer programming,” IEEE Transactions on Learning Technologies, vol.13, issue.1, pp.91–106, 2020.

[39] 木下涼，植野真臣，“深層学習によるテスト理論：Item Deep Response Theory，” 信学論（D），vol.J103, no.4, pp.314–329, 2020.

[40] 植野真臣，木下涼，“ポスト項目反応理論：深層学習によるテスト理論，”Precision Medicine, vol.3, no.5, 5 月号，pp.56–62, 2020.

[41] C. Piech, J. Bassen, J. Huang, S. Ganguli, M. Sahami, L.J. Guibas, and J. Sohl-Dickstein, "Deep knowledge tracing," Advances in Neural Information Processing Systems 28, eds. by C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama, and R. Garnett, pp.505–513, Curran Associates, Inc., 2015.

[42] J. Zhang, X. Shi, I. King, and D.-Y. Yeung, "Dynamic Key-Value Memory Networks for Knowledge Tracing," Proc. 26th International Conference on World Wide Web, pp.765–774, WWW '17, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2017.

[43] C.V. Le, Z.A. Pardos, S.D. Meyer, and R. Thorp, "Communication at scale in a MOOC using predictive engagement analytics," Artificial Intelligence in Education - 19th International Conference, AIED, pp.239-252, 2018.

[44] Y. Jiang, N. Bosch, R.S. Baker, L. Paquette, J. Ocumpaugh, J.M.A.L. Andres, A.L. Moore, and G. Biswas, "Expert feature-engineering vs. deep neural networks: Which is better for sensor free affect detection?," Artificial Intelligence in Education, AIED, pp.198–211, 2018.

[45] S. Ruseti, M. Dascalu, A.M. Johnson, R. Balyan, K.J. Kopp, D.S. McNamara, S.A. Crossley, and S. Trausan-Matu, "Predicting question quality using recurrent neural networks," Artificial Intelligence in Education, AIED, pp.491–502, 2018.

[46] T.I. Dhamecha, S. Marvaniya, S. Saha, R. Sindhgatta, and B. Sengupta, "Balancing human efforts and performance of student response analyzer in dialog-based tutors," Artificial Intelligence in Education, AIED, pp.70–85, 2018.

[47] X. Yang, Y. Huang, F. Zhuang, L. Zhang, and S. Yu, "Automatic chinese short answer grading with deep autoencoder," Artificial Intelligence in Education, AIED, pp.399–404, 2018.

[48] Y. Su, Q. Liu, Q. Liu, Z. Huang, Y. Yin, E. Chen, C. H. Q. Ding, S. Wei, and G. Hu, "Exercise-enhanced sequential modeling for student performance prediction," in AAAI, pp.2435–2443, 2018.

[49] C.-K. Yeung, “Deep-IRT: Make Deep Learning Based Knowledge Tracing Explainable Using Item Response Theory,” in Proceedings of the 12th International Conference on Educational Data Mining, EDM, 2019.

[50] 堤瑛美子, 木下涼, 植野真臣, “Knowledge Tracing のための Sliding Window 隠れマルコフ IRT,” 信学論（D）, vol.J103, no.12, pp.894–905, 2020.

[51] Educational Testing Service, “The TOEFL Test,” https://www.ets.org/toefl/

[52] 独立行政法人情報処理推進機構, “IT パスポート試験,” https://www3.jitec.ipa.go.jp/JitesCbt/

[53] 公益社団法人医療系大学間共用試験実施評価機構, “臨床実習開始前の「共用試験」第 13 版（平成 27 年度）,” http://www.cato.umin.jp/e-book/13/index.html

[54] F. Samejima, “Estimation of latent ability using a response pattern of graded scores,” Psychometrika Monography, vol.17, pp.1–100, 1969.

[55] E. Muraki, “A generalized partial credit model,” in Handbook of Modern Item Response Theory, ed. W.J. van derLinden and R.K. Hambleton, pp.153–164, Springer, 1997.

[56] D.P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv:1412.6980, 2014.

[57] J. Duchi, E. Hazan, and Y. Singer, “Adaptive subgradient methods for online learning and stochastic optimization,” J. Mach. Learn. Res., vol.12, pp.2121–2159, July 2011.

[58] M.D. Zeiler, “Adadelta: An adaptive learning rate method,” CoRR, vol.abs/1212.5701,2012.

[59] A. Graves, “Generating sequences with recurrent neural networks,” CoRR, vol.abs/1308.0850, 2013.

[60] T. Schaul, S. Zhang, and Y. LeCun, “No more pesky learning rates,” Proceedings of the 30th International Conference on Machine Learning, eds. by

S. Dasgupta and D. McAllester, vol.28, pp.343-351, Proceedings of Machine Learning Research, PMLR, Atlanta, Georgia, USA, 17–19 Jun 2013.

[61] 藤森 進,“同時尺度調整法による垂直的等化の検討,” 人間科学研究, vol.20, pp.34–47, 1998.

[62] 藤森 進,“共通項目の部分得点モデル化によるテストの等化,” 人間科学研究, vol.27, pp.77–81, 2005.

[63] S. Arai and S. Mayekawa,“A comparison of equating methods and linking designs for developing an item pool under item response theory,”Behaviormetrika, vol.38, pp.1–16, 2011.

[64] 光永悠彦, 前川眞一,“項目反応理論に基づくテストにおける項目バンク構築時の等化方法の比較,” 日本テスト学会誌, vol.8, no.1, pp.31–48, 2012.

[65] S. Kilmen and N. Demirtasli, “Comparison of test equating methods based on item response theory according to the sample size and ability distribution,” Social and Behavioral Sciences, vol.46, no.Supplement C, pp.130–134, 2012.

[66] I. Uysal and S. Kilmen, “Comparison of item response theory test equating methods for mixed format tests,” International Online Journal of Educational Sciences, vol.8, no.2, pp.1–11, 2016.

[67] M. Uto and M. Ueno, “Empirical comparison of item response theory models with rater's parameters,” Heliyon, Elsevier, vol.4, no.5, pp.1–32, 2018.

[68] M. Ueno and K. Nagaoka, “Learning log database and data mining system for e-Learning-On-Line statistical outlier detection of irregular learning processes,” Proceedings of the International Conference on Advanced Learning Technologies 2002, pp.436–438, 2002.

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

[70] M. Ueno, “Animated agent to maintain learner's attention in e-learning,” E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, pp.194–201, 2004.

[71] M. Ueno, “Online outlier detection system for learning time data in E-Learning and It's evaluation,” Proc. of Computers and Advanced Technology in Education (CATE2004), pp.248–253, 2004.

[72] M. Ueno, “Animated pedagogical agent based on decision tree for e-Learning,” Fifth IEEE International Conference on Advanced Learning Technologies (ICALT'05), pp.188–192, 2005.