THE IEICE TRANSACTIONS ON INFORMATION AND SYSTEMS (JAPANESE EDITION)

電子情報通信学会
論文誌 D 情報・システム

DOI:10.14923/transinfj.2019JDP7068
早期公開日:2019/12/27

情報・システムソサイエティ
一般社団法人 電子情報通信学会
THE INFORMATION AND SYSTEMS SOCIETY
THE INSTITUTE OF ELECTRONICS, INFORMATION AND COMMUNICATION ENGINEERS

論　文

深層学習によるテスト理論：Item Deep Response Theory

木下　涼[†a)]　植野　真臣[†b)]

A Test Theory based on Deep Learning: Item Deep Response Theory

Ryo KINOSHITA[†a)] and Maomi UENO[†b)]

あらまし　項目反応理論では異なるテストを受検した受検者を同一尺度上で評価することができる．しかし，そのためには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，共通項目を含む複数のテストデータをもとにリンケージと呼ばれる処理が必要である．リンケージはテスト実施に膨大な作業を伴うだけでなく，前述の仮定により理論的に最適値を得る保証はない．この問題を解決するため，本研究では深層学習を用い，受検者の母集団と独立性を仮定しないテスト理論として Item Deep Response Theory を提案する．提案モデルはリンケージを必要とせず，受検者が単一母集団からのランダムサンプリングでない場合にも精度の低下が抑えられ，複数の母集団からサンプリングされている場合に頑健である．ここではシミュレーション・実データ実験より，提案モデルは未知の項目への反応予測精度が高く，特に受検者が同一の母集団からランダムにサンプリングできない場合に項目反応理論に対して大きく能力推定精度と反応予測精度を向上させることを示す．

キーワード　教育工学，テスト理論，深層学習，項目反応理論，リンケージ

1. まえがき

テスト理論 (test theory) は，大別して以下の二つがある．

1) 記述統計学の一つである古典的テスト理論 [1], [2].
2) 能力パラメータを組み込んだ数理モデルである項目反応理論 (Item Response Theory; IRT) [3]〜[5].

特に，後者は以下のような利点をもち，近年着目されている [6].

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，項目データベース構築などに有効である．
2) 異なるテストを受検した受検者同士を同一尺度上で評価することができる．

これらの利点から，IRT はハイステークスな大規模テストや e テスティングの基盤技術として用いられている [7]〜[10]．しかし，IRT を用いて大規模な項目データベースを構築したり，異なるテストの受検者を同一尺度上で評価するためには，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，リンケージ (linkage) と呼ばれる処理が必要である．リンケージは共通項目を含む複数のテストを用意するといった膨大な作業を伴うだけでなく，理論的に最適値を得る保証がなく [11]，推定パラメータのバイアスや標準誤差が大きくなる可能性も高い [5], [12]〜[15].

さらに，現実に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，各学校やクラス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分布モデルは複雑であり実用化は困難である．一方，機械学習分野では母数確率モデルを用いず深層学習を用いて，大規模で複雑なモデルを比較的容易に構築できるようになってきた．例えば，深層学習により学習者の反応履歴と各項目の対応するスキルから学習者の知識状態を推定する Deep Knowledge Tracing(DKT) が提案され，関連研究も盛んに行われている [16]〜[21]. これらのモデルは深層学習を用いることで，受検者の母集団や独立性を仮定することなく，未知の項目への反応予測を高性能に行えることが報告されている．未知の項目への反応を高精度に予測することで，受検者に適応した項目を出題することができる．しかし，これらのモデルは解釈可能な学習者パラメータ，項目パ

† 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: kinoshita@ai.is.uec.ac.jp
b) E-mail: ueno@ai.is.uec.ac.jp

ラメータを持たず，テスト理論として扱うことができない．

本論では IRT に代わる新しいテスト理論として深層学習を用い，受検者の母集団と独立ランダムサンプリングを仮定せず，受検者の能力パラメータと項目の難易度パラメータにより受検者の項目への正答確率をモデル化した Item Deep Response Theory(IDRT) を提案する．本手法は以下の利点が期待される．

1)IRT のリンケージ手法では，通常テスト間に共通する項目を利用する．本手法ではテスト間に共通する項目がない場合にも能力推定精度の低下が抑えられる．

2)IRT で仮定される受検者の独立ランダムサンプリングが成り立たない場合にも能力推定精度の低下が抑えられる．

3)IRT で仮定される受検者の母集団が単一でない場合にも能力推定精度の低下が抑えられる．

これらは，IRT によるリンケージの仮定が成り立たない状況でも，提案モデルが推定精度の低下を抑えて能力を推定することが可能であることを意味する．

本論ではシミュレーション・実データより，提案モデルの有効性を示した．また，近年，代表的なパフォーマンス予測モデルである Bayesian Knowledge Tracing や DKT と比較して IRT の未知の項目への反応予測精度が高いことが報告されている [22]．しかし，提案モデルは IRT よりさらに予測精度が高いことを示した．

2. 項目反応理論 (IRT)

本章では，最もよく用いられるテスト理論である IRT を紹介する．

2.1 テストデータ

今，受検者 $i \in \{1, ..., I\}$ が項目 $j \in \{1, ..., J\}$ に正答したとき $u_{ij} = 1$，誤答したとき $u_{ij} = 0$ とする．ただし，$u_{ij} = -1$ のときは欠測値を表す．また，受検者 i の全ての項目に対する項目反応パターン $\boldsymbol{u}_i$ を以下のように書く．

$$\boldsymbol{u}_i = (u_{i1}, u_{i2}, .., u_{iJ}) \tag{1}$$

さらに，全ての受検者の項目反応パターンを用いてテストデータ行列 $\boldsymbol{U}$ を以下のように表す．

$$\boldsymbol{U} = \begin{bmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \\ \vdots \\ \boldsymbol{u}_I \end{bmatrix} = \begin{bmatrix} u_{11} & u_{12} & \dots & u_{1J} \\ u_{21} & u_{22} & \dots & u_{2J} \\ \vdots & \vdots & & \vdots \\ u_{I1} & u_{I2} & \dots & u_{IJ} \end{bmatrix}$$

本論では，このテストデータ行列を扱う．

2.2 モデル

IRT は，様々な評価場面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一つである [3]〜[5]．

IRT のモデルとして，以下の 2 パラメータロジスティックモデル（2-Parameter Logistic Model; 2PLM）が最も広く利用されてきた．

$$\begin{aligned} P_j(\theta_i) &= P(u_{ij} = 1 \mid \theta_i) \\ &= \frac{1}{1 + exp(-1.7a_j(\theta_i - b_j))} \end{aligned} \tag{2}$$

ここで，$P_j(\theta_i)$ は能力パラメータ $\theta_i \in (-\infty, \infty)$ の受検者 i が項目 j に正答する確率を示している．また，$a_j \in [0, \infty)$ は項目 j の識別力パラメータ，$b_j \in (-\infty, \infty)$ は項目 j の困難度パラメータと呼ばれる項目パラメータである．

受検者の能力パラメータ θ の事後分布 $g(\theta|\boldsymbol{u})$ は θ の事前分布 $f(\theta)$ と尤度関数 $L(\theta|\boldsymbol{u})$ からベイズの定理を用いて以下のように導かれる．

$$g(\theta|\boldsymbol{u}) = \frac{L(\theta|\boldsymbol{u})f(\theta)}{h(\boldsymbol{u})} \tag{3}$$

ここで，$h(\boldsymbol{u})$ は項目反応パターン $\boldsymbol{u}$ の周辺分布であり，次式で求められる．

$$h(\boldsymbol{u}) = \int_{-\infty}^{\infty} L(\theta|\boldsymbol{u})f(\theta)d\theta \tag{4}$$

能力パラメータ推定では，理論的に最も予測精度が高い，期待事後確率推定法（Expected a Priori; EAP）を用いて $\hat{\theta}$ を推定する．

$$\hat{\theta} = \int_{-\infty}^{\infty} \theta g(\theta|\boldsymbol{u})d\theta \tag{5}$$

事後分布を解析的に計算するのは困難なため，マルコフ連鎖モンテカルロ法（Markov Chain Monte Carlo methods; MCMC）などを用いて推定値を数値的に求めることが一般的である．

ここで，事前分布 $f(\theta)$ は，あらかじめ定めた θ の母集団分布を示し，同一集団からの受検者の独立サンプリングを仮定していることとなる．そのため，異なる項目からなる複数のテストデータからパラメータを推定する場合には，項目パラメータの推定値を同一基準に補正するリンケージ処理を行うことが一般的である．

2.3 リンケージ

複数のテストデータのリンケージを行う際には，テ

ストを実施する前に次のいずれかのリンケージ計画を立てる必要がある [23].
1) 複数のテストを受検する受検者によってリンケージする共通受検者計画.
2) 複数のテストに共通するテスト項目によってリンケージする共通項目計画.
3) 係留テストと呼ばれる共通項目群を用意し，係留テストと各尺度に共通受検者を用意してリンケージする係留テスト計画.

テスト実施後，得られたテストデータ行列から共通する受検者・項目をもとに，各パラメータを同一尺度上に変換する．その際に用いられる手法として，1) 共通する受検者・項目をもとに，特定のテストの尺度に他のテストの尺度を変換する等化係数推定法 [24]〜[27]，2) 全てのテストデータに対して一度にパラメータの推定を行う同時尺度推定法 [28]，3) 既知のリンケージ後の項目パラメータを所与とした上で，未知のパラメータの推定を行う固定項目パラメータ法などが知られている.

リンケージの実施には綿密な等化計画と莫大なコストが必要になることが多い [13]．さらに，同時確率分布を完全に表現できる IRT のリンケージは存在しないことが知られており，リンケージ精度は保証されない [11]．特に，現実には受検者の能力分布は多母集団かつ独立サンプリングができないことが多く，このような場合には大きく能力推定値の精度が損なわれてしまう.

3. Item Deep Response Theory

本論では上述のリンケージの問題を踏まえ，受検者間の独立性を仮定せず，他の受検者の反応データを利用することでリンケージを用いなくとも高精度に項目の難易度パラメータ，受検者の能力パラメータが推定できるモデルを提案する．具体的には，深層学習を用い，受検者の母集団および独立サンプリングを仮定しない Item Deep Response Theory(IDRT) を提案する.

3.1 Deep Response Model

IDRT の予測モデルである Deep Response Model (DRM) の概要図を図 1 に示す．提案モデルでは，受検者ネットワーク（Examinee Network) と項目ネットワーク (Item Network) の二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.

受検者ネットワークでは i 番目の受検者を表現するため，i 番目の要素のみが 1，他の要素が 0 の one-hot vector $\boldsymbol{s}_i \in \mathbb{R}^I$ を入力として，次のように 4 層のニューラルネットワークを構成する.

$$\boldsymbol{\theta}_1^{(i)} = tanh\left(\boldsymbol{W}^{(\theta_1)}\boldsymbol{s}_i + \boldsymbol{\tau}^{(\theta_1)}\right) \tag{6}$$

$$\boldsymbol{\theta}_2^{(i)} = tanh\left(\boldsymbol{W}^{(\theta_2)}\boldsymbol{\theta}_1^{(i)} + \boldsymbol{\tau}^{(\theta_2)}\right) \tag{7}$$

$$\theta_3^{(i)} = \boldsymbol{W}^{(\theta_3)}\boldsymbol{\theta}_2^{(i)} + \tau^{(\theta_3)} \tag{8}$$

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ている.

$$tanh(x) = \frac{exp(x) - exp(-x)}{exp(x) + exp(-x)} \tag{9}$$

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

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

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

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

$$\boldsymbol{W}^{(\theta_3)} = \begin{pmatrix} w_1^{(\theta_3)}, & w_2^{(\theta_3)}, & \cdots & w_{|\boldsymbol{\theta}_2|}^{(\theta_3)} \end{pmatrix}$$

また，$\boldsymbol{\tau}^{(\theta_1)} = \left(\tau_1^{(\theta_1)}, \tau_2^{(\theta_1)} ... \tau_{|\boldsymbol{\theta}_1|}^{(\theta_1)}\right)$ および，$\boldsymbol{\tau}^{(\theta_2)} = \left(\tau_1^{(\theta_2)}, \tau_2^{(\theta_2)} ... \tau_{|\boldsymbol{\theta}_2|}^{(\theta_2)}\right)$ 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 $\theta_3^{(i)}$ を受検者 i の能力パラメータとみなす.

ここで，提案モデルにおいて，受検者ネットワークの入力から重みパラメータを通じて θ_3 が推定され，θ_3 から各項目への反応が発生する過程のグラフィカルモデルを図 2 に示す．図 2 から明らかなように，提案モデルでは IRT と異なり，能力パラメータに共通の母集団を仮定していない．また，提案モデルでは，得られた反応データの予測を最大にするよう重みパラメータ $\boldsymbol{W}_{\theta_1}, \boldsymbol{W}_{\theta_2}, \boldsymbol{W}_{\theta_3}, \boldsymbol{W}_y$ を更新する．例えば，反応データ u_{ij} が与えられた時，全ての重みパラメータが更新

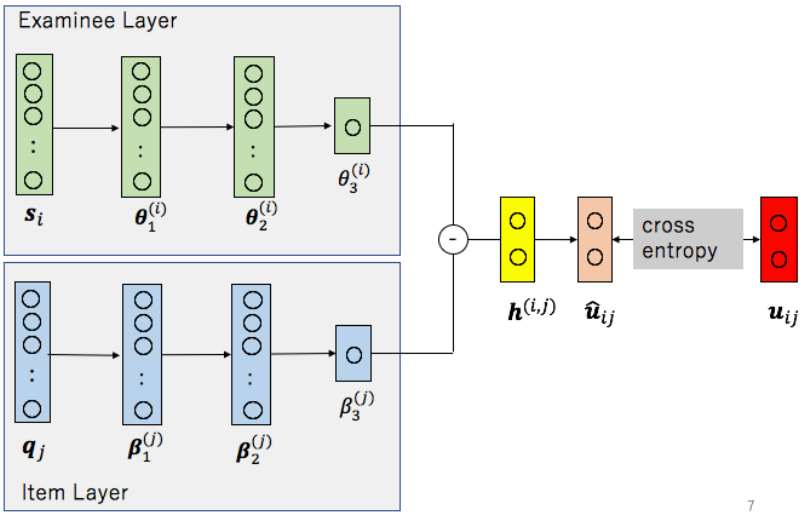


図 1　提案モデルの概要図
Fig. 1　proposed model

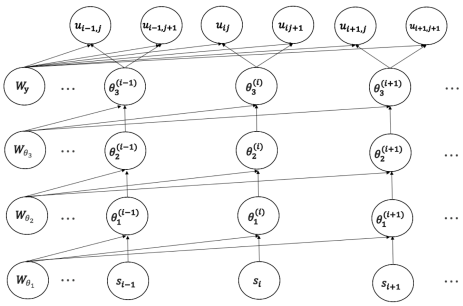


図 2　受検者ネットワークのグラフィカルモデル
Fig. 2　Graphical Model of Examinee Network

され，重みパラメータを介して $\theta_3^{(i)}$ 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，項目ネットワークでは j 番目の項目を表現するため，j 番目の要素のみが 1，他の要素は 0 の one-hot vector　$\boldsymbol{q}_j \in \mathbb{R}^J$ を入力として，次のように 4 層のニューラルネットワークを構成する．

$$\boldsymbol{\beta}_1^{(j)} = tanh\left(\boldsymbol{W}^{(\beta_1)}\boldsymbol{q}_j + \boldsymbol{\tau}^{(\beta_1)}\right) \tag{10}$$

$$\boldsymbol{\beta}_2^{(j)} = tanh\left(\boldsymbol{W}^{(\beta_2)}\boldsymbol{\beta}_1^{(j)} + \boldsymbol{\tau}^{(\beta_2)}\right) \tag{11}$$

$$\beta_3^{(j)} = \boldsymbol{W}^{(\beta_3)}\boldsymbol{\beta}_2^{(j)} + \tau^{(\beta_3)} \tag{12}$$

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

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

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

また，$\boldsymbol{W}^{(\beta_3)}$ は以下の重みパラメータベクトルである．

$$\boldsymbol{W}^{(\beta_3)} = \begin{pmatrix} w_1^{(\beta_3)}, & w_2^{(\beta_3)}, & \dots & w_{|\boldsymbol{\beta}_2|}^{(\beta_3)} \end{pmatrix}$$

さらに，$\boldsymbol{\tau}^{(\beta_1)} = \left(\tau_1^{(\beta_1)}, \tau_2^{(\beta_1)} \dots \tau_{|\boldsymbol{\beta}_1|}^{(\beta_1)}\right)$，$\boldsymbol{\tau}^{(\beta_2)} = \left(\tau_1^{(\beta_2)}, \tau_2^{(\beta_2)} \dots \tau_{|\boldsymbol{\beta}_2|}^{(\beta_2)}\right)$ はバイアスパラメータベクトル，$\tau^{(\beta_3)}$ はバイアスパラメータである．

本論では，項目ネットワークの出力 $\beta_3^{(j)}$ を項目 j の難易度パラメータとみなす．難易度パラメータを推定する際に項目間の独立性を仮定していないことが特徴である．

次に，IRT のパラメータ解釈に倣い，受検者の能力パラメータと項目の難易度パラメータの差を用いて受検者の項目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 $\boldsymbol{h}^{(i,j)} = (h_0^{(i,j)}, h_1^{(i,j)})$ を求め，

式 (14) にしたがって受検者 i の項目 j への反応確率 $\hat{\boldsymbol{u}}_{ij} = [1-\hat{u}_{ij}, \hat{u}_{ij}]$ を算出し，モデルの出力とする．

$$\boldsymbol{h}^{(i,j)} = (\boldsymbol{W}^{(y)})^T(\theta_3^{(i)} - \beta_3^{(j)}) + \boldsymbol{\tau}^{(y)} \tag{13}$$

$$\begin{aligned}\hat{u}_{i,j} &= softmax(\boldsymbol{h}^{(i,j)}) \\ &= \frac{exp(h_1^{(i,j)})}{exp(h_0^{(i,j)}) + exp(h_1^{(i,j)})}\end{aligned} \tag{14}$$

$\boldsymbol{W}^{(y)} = (w_1^{(y)}, w_2^{(y)})$，$\boldsymbol{\tau}^{(y)} = (\tau_1^{(y)}, \tau_2^{(y)})$ は，それぞれ重みパラメータベクトル，バイアスパラメータベクトルである．

ここでは，IRT と同様の解釈ができるようにその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．ただし，このモデルでは受検者の母集団・独立性を仮定せず，受検者の項目への反応予測を最大化するようにモデルが構成されている．そのため，異なるテストの受検者の能力推定値も利用し，最も予測精度が高くなるように能力を推定する．

また，未知の受検者 $I+1$ の反応データが得られた場合には，以下の手順で重みパラメータ

$$\boldsymbol{W}_{I+1}^{(\theta_1)} = \begin{pmatrix} w_{1I+1}^{(\theta_1)} \\ w_{2I+1}^{(\theta_1)} \\ \vdots \\ w_{|\boldsymbol{\theta}_1|I+1}^{(\theta_1)} \end{pmatrix}$$

のみを新たに推定すればよい．具体的には以下のように推定する．

（1） 式 (6) の $\boldsymbol{W}^{(\theta_1)}$ を以下のように置き換える．

$$\boldsymbol{W}^{(\theta_1)*} = \left[\boldsymbol{W}^{(\theta_1)}; \boldsymbol{W}_{I+1}^{(\theta_1)}\right]$$

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

ここで，; は行列の結合を示す．

（2） 学習済みの $\boldsymbol{W}^{(\theta_1)}, \boldsymbol{W}^{(\theta_2)}, \boldsymbol{W}^{(\theta_3)}$ と項目ネットワークを所与として受検者 $I+1$ の項目反応パターン $\boldsymbol{u}_{I+1}$ から $\boldsymbol{W}_{I+1}^{(\theta_1)}$ を学習する．

（3） 式 (6) から式 (8) に従って $\theta_3^{(I+1)}$ を計算する．

3.2 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，受検者 i が項目 j に正答する予測確率 $\hat{\boldsymbol{u}}_{ij}$ と反応データ $\boldsymbol{u}_{ij} = [1-u_{ij}, u_{ij}]$ から分類誤差を表すクロスエントロピー ℓ を算出する．

$$\ell(u_{ij}, \hat{u}_{ij}) = -u_{ij}\log\hat{u}_{ij} - (1-u_{ij})\log(1-\hat{u}_{ij}) \tag{15}$$

一般の機械学習手法と同様に，深層学習もデータの偏りに大きな影響を受けることが知られている．しかし，現実のテストデータでは受検者や項目の反応に偏りがあることが多い．この問題を解決するため，出現頻度が少ないデータの重みを大きくする cost-sensitive learning が一般に用いられている [29]．本研究でも ℓ を基に以下の損失関数 $Loss_{class}$ を定義する．

$$\begin{aligned} Loss_{class} = &\sum_i \sum_j \ell(u_{ij}, \hat{u}_{ij}) \\ &+ \gamma_1 \sum_{i \in L-examinee} \sum_{j \in (u_{ij}=1)} \ell(u_{ij}, \hat{u}_{ij}) \\ &+ \gamma_2 \sum_{i \in H-examinee} \sum_{j \in (u_{ij}=0)} \ell(u_{ij}, \hat{u}_{ij}) \\ &+ \gamma_3 \sum_{j \in L-item} \sum_{i \in (u_{ij}=1)} \ell(u_{ij}, \hat{u}_{ij}) \\ &+ \gamma_4 \sum_{j \in H-item} \sum_{i \in (u_{ij}=0)} \ell(u_{ij}, \hat{u}_{ij}) \end{aligned} \tag{16}$$

ここで，$L-examinee$ は正答率が $\alpha_{L-examinee}$ 以下の受検者集合，$H-examinee$ は正答率が $\alpha_{H-examinee}$ 以上の受検者集合，$L-item$ は正答率が α_{L-item} 以下の項目集合，$H-item$ は正答率が α_{H-item} 以上の項目集合である．また，$\gamma_1, \gamma_2, \gamma_3, \gamma_4$ と $\alpha_{L-examinee}, \alpha_{H-examinee}, \alpha_{L-item}, \alpha_{H-item}$ はチューニングパラメータであり，任意の値に設定する．

提案モデルは，テストデータ行列をもとに，adaptive moment estimation(Adam) [30] と呼ばれる最適化アルゴリズムに従って，損失関数が小さくなるように全てのパラメータを同時に更新する．

3.3 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．深層学習で最もよく用いられているアルゴリズムであり，学習率を固定する場合や他の最適化アルゴリズム [31]〜[34] よりも損失関数が低くなる傾向が報告されている [30]．

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．具体的に，t 回目の学習における各パラメータの勾配 $\boldsymbol{g}_t$ が得られた時，それまでの勾配の重み付き平均 $\boldsymbol{m}_t$ と勾配の 2 乗の重み付き平均 $\boldsymbol{var}_t$ を以下のように算出する．

$$\boldsymbol{m}_t = \alpha_1 \boldsymbol{m}_{t-1} + (1-\alpha_1)\boldsymbol{g}_t \tag{17}$$

$$\boldsymbol{var}_t = \alpha_2 \boldsymbol{var}_{t-1} + (1-\alpha_2)\boldsymbol{g}_t^2 \tag{18}$$

ここで，α_1, α_2 はチューニングパラメータであり，任意の値に設定する．

これらの推定バイアスを補正した $\boldsymbol{m}_t^* = \boldsymbol{m}_t/(1-\alpha_1^t)$，$\boldsymbol{var}_t^* = \boldsymbol{var}_t/(1-\alpha_2^t)$ を用いて，t 回目の学習では全てのパラメータベクトル $\boldsymbol{x}_t$ を以下のように更新する．

$$\boldsymbol{x}_t \leftarrow \boldsymbol{x}_{t-1} - \frac{\mu}{\sqrt{\boldsymbol{var}_t^*} + \epsilon}\boldsymbol{m}_t^* \tag{19}$$

μ は学習率の初期値であり，ϵ は発散を防ぐための微小な定数である．

4. シミュレーション実験

リンケージを伴う研究や多母集団を仮定した研究では実データを収集するのに多大なコストを要するため，現実に近い条件のもと，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 [35]〜[37]．したがって，本章ではシミュレーションデータに提案モデルと IRT を適用し，リンケージが必要な状況での能力推定精度と未知の項目への反応予測精度を比較する．

4.1 実験条件

本研究では提案モデルの実装にニューラルネットワークのフレームワークの一つである Chainer(注1) を用い，バッチ学習でパラメータを更新した．

全ての実験に共通するパラメータの値を表 1 に示す．これらのパラメータのうち，ϵ ，α_1 ，α_2 には先行研究 [30] で提案されている値を用いた．

IRT のモデルには 2PLM を採用し，パラメータ推定は MCMC 法を用いた EAP 推定で行なった．ここでリンケージには同時尺度推定法を用い，すべてのパラメータを同時に推定した．また，各パラメータの事前分布には次の分布を用いた．

$$\theta \sim N(0,1), \quad a \sim LN(0,1), \quad b \sim N(1, 0.4) \tag{20}$$

(注1)：https://chainer.org/

表 1　共通するチューニングパラメータの値
Table 1　The values of tuning parameters

パラメータ	値	パラメータ	値
$\boldsymbol{\theta}_1^{(i)}$ のノード数	50	γ_1	0.1
$\boldsymbol{\theta}_2^{(i)}$ のノード数	50	γ_2	0.1
$\boldsymbol{\beta}_1^{(j)}$ のノード数	50	γ_3	0.1
$\boldsymbol{\beta}_2^{(j)}$ のノード数	50	γ_4	0.1
エポック数	300	$\alpha_{L-examinee}$	0.2
μ	0.01	$\alpha_{H-examinee}$	0.8
ϵ	10^{-8}	α_{L-item}	0.2
α_1	0.9	α_{H-item}	0.8
α_2	0.999		

$N(\mu,\sigma), LN(\mu,\sigma)$ は平均 μ，標準偏差 σ の正規分布と対数正規分布を表す．

4.1.1 能力推定精度

本論では能力推定精度として，能力推定値と真の能力値との平均平方二乗誤差 (Root Mean Square Error; RMSE)，Pearson の積率相関係数と Kendall の順位相関係数を用いる．ただし，RMSE を求める際には能力推定値 $\theta_3^{(i)}$ を全ての能力推定値の平均 $mean(\boldsymbol{\theta}_3)$ と標準偏差 $sd(\boldsymbol{\theta}_3)$ をもとに以下の平均 0，分散 1 の分布に標準化したのち算出する．

$$\hat{\theta}_3^{(i)} = \frac{\theta_3^{(i)} - mean(\boldsymbol{\theta}_3)}{sd(\boldsymbol{\theta}_3)} \tag{21}$$

4.1.2 未知の項目への反応予測精度

また，人工知能分野におけるアダプティブラーニング研究では学習者の未知の反応予測が重要である．近年，IRT の未知の項目への反応予測精度が最も高いことが報告されている [22]．そこで，本論では未知の項目への反応予測精度も IRT と比較する．具体的には，未知の項目への反応予測精度として 10 分割交差検証を用いてテストデータ行列を訓練データと評価データに分割し，訓練データ内の反応から推定した能力値と難易度を利用して評価データの全ての反応予測を行い，次式の F 値を算出する．

$$F = \frac{2\mathrm{Recall}\cdot\mathrm{Precision}}{\mathrm{Recall}+\mathrm{Precision}} \tag{22}$$

ここで，Precision は反応予測の正解率を示し，Recall は実際の反応のうち，正しく予測された反応の割合を示す．F 値は正答・誤答ごとに予測精度の評価が可能であり，予測精度の評価に最もよく用いられている指標の一つである．本論でもこの F 値を予測精度として用いる．

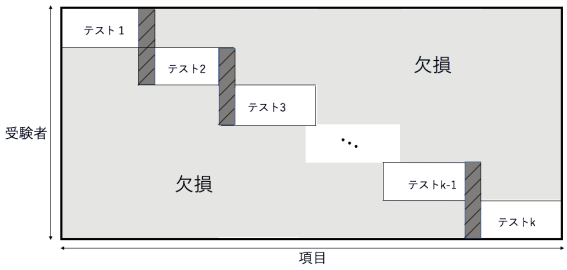


図 3　シミュレーションデータのテストデータ行列
Fig. 3　An example of Item Response Pattern Matrix

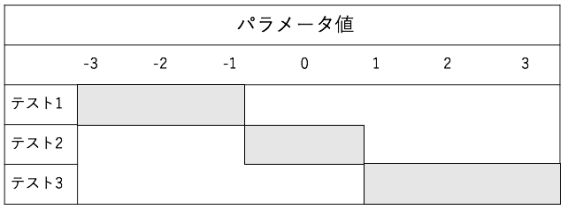


図 4　システム割り当て
Fig. 4　System Generation

4.2　受検者のランダムサンプリングが仮定できない場合の能力推定精度

本節では受検者の割り当て方法，テスト項目数，テスト間の共通項目数，テストの受検人数を変化させた際の能力推定精度を比較し，受検者のランダムサンプリングが仮定できない場合の提案モデルの優位性を明らかにする．

本実験では J 個の項目で構成された 10 個のテストを，それぞれ I 人からなる受検者グループが反応した状況を想定し，k 番目のテストが $k-1$ 番目，$k+1$ 番目のテストとのみ共通項目を持つシミュレーションデータを生成する．ここで，k 個のテストからなるシミュレーションデータの例を図 3 に示す．各テスト間の斜線部が共通項目である．

シミュレーションデータの生成は式 (23) に従ってパラメータを発生させ，式 (2) の 2PLM により行う．

$$\theta \sim N(0,1), \quad a \sim LN(0,1), \quad b \sim N(1,0.4) \quad (23)$$

なお，これらのパラメータ分布は式 (20) に示した 2PLM の事前分布と一致しているため，2PLM による推定精度が最も高くなる条件であることに注意されたい．

本実験では，受検者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の二つの方法で受検者を各テストに割り当て，テストデータを生成する．ここで，項目パラメータは共通項目を含む全ての項目について，式 (23) の母集団から発生させた．

ランダム割り当て

1) 式 (23) に従って，受検者数だけ能力値を発生させる．
2) 発生させた能力パラメータを持つ受検者を各テストに無作為に割り当てる．

システム割り当て

1) 式 (23) に従って，受検者数だけ能力値を発生させる．
2) 発生させた能力パラメータを持つ受検者を能力値の昇順に並び替え，テスト数だけ等分割し受検者グループを作成する．
3) k 番目のテストに，k 番目の受検者グループを割り当てる．

図 4 にテスト数が 3 の場合のシステム割り当ての概要図を示す．各テストに対して，色のついている範囲の能力値を持つ受検者を割り当てる．

ここで，表 2 に各テストを構成する受検者の割り当て方法・項目数・共通項目数・受検人数を変化させた場合について，4.1.1 の RMSE・積率相関係数・順位相関係数を示す．ただし，表中の受検者数，項目数は 10 個の各テストの受検者数，項目数を示し，総受検者数，総項目数はすべてのテストの受検者数，項目数を示す．

表 2 から，ランダムに受検者を割り当てた場合は多くの条件で，IRT の精度が高いことがわかる．この条件では，全てのテストについて受検者が同一母集団からランダムにサンプリングされており，受検者数が十分に大きくなると各テストごとの能力推定値が真の能力分布をよく近似できるので，リンケージなしでも IRT で高い精度が得られたと考えられる．ただし，ランダム割り当てであっても，テスト項目数・受検人数が少ない場合は提案モデルと IRT の精度差が小さくなっていくことがわかる．つまり，IRT の理想的な状況から乖離するほど両者の差が小さくなることが示された．

一方，システム的に受検者を割り当てた場合は，全ての条件で提案モデルの精度が IRT を上回っている．すなわち，システム割り当てではテスト間の共通項目を元にリンケージを行える状況でも，提案モデルが IRT

表 2　各条件を変化させた場合の能力推定精度
Table 2　Parameter estimation accuracies when the way to generate student parameters, the number of items and students per test and common items were changed

ランダム割り当て							システム割り当て						
項目数	共通項目数（総項目数）	受検者数（総受検者数）	手法	RMSE	Pearson	Kendall	項目数	共通項目数（総項目数）	受検者数（総受検者数）	手法	RMSE	Pearson	Kendall
10	5 (55)	50 (500)	DRM	0.469	0.890	0.748	10	5 (55)	50 (500)	DRM	**0.665**	**0.778**	**0.568**
			IRT	**0.420**	**0.912**	**0.781**				IRT	1.111	0.381	0.237
		100 (1000)	DRM	0.447	0.900	0.766			100 (1000)	DRM	**0.622**	**0.807**	**0.629**
			IRT	**0.438**	**0.904**	**0.770**				IRT	0.779	0.696	0.466
		500 (5000)	DRM	0.434	0.907	0.769			500 (5000)	DRM	**0.611**	**0.812**	**0.639**
			IRT	**0.432**	0.907	**0.776**				IRT	0.792	0.702	0.499
		1000 (10000)	DRM	0.424	0.908	**0.771**			1000 (10000)	DRM	**0.621**	**0.822**	**0.651**
			IRT	**0.411**	**0.911**	0.733				IRT	0.712	0.702	0.501
	0 (100)	50 (500)	DRM	0.458	0.896	0.747		0 (100)	50 (500)	DRM	**0.997**	**0.502**	**0.267**
			IRT	**0.456**	0.896	**0.751**				IRT	1.170	0.314	0.184
		100 (1000)	DRM	0.455	0.832	0.765			100 (1000)	DRM	**0.721**	**0.740**	**0.561**
			IRT	**0.440**	**0.903**	**0.767**				IRT	1.176	0.308	0.197
		500 (5000)	DRM	0.433	0.852	0.785			500 (5000)	DRM	**0.701**	**0.761**	**0.591**
			IRT	**0.423**	**0.861**	**0.789**				IRT	1.016	0.498	0.277
		1000 (10000)	DRM	0.412	0.910	**0.799**			1000 (10000)	DRM	**0.698**	**0.782**	**0.591**
			IRT	**0.403**	**0.914**	0.794				IRT	0.808	0.673	0.457
30	5 (255)	50 (500)	DRM	0.328	0.921	0.855	30	5 (255)	50 (500)	DRM	**0.561**	**0.835**	**0.696**
			IRT	**0.301**	**0.941**	**0.865**				IRT	0.613	0.786	0.622
		100 (1000)	DRM	0.319	0.949	0.865			100 (1000)	DRM	**0.501**	**0.875**	**0.716**
			IRT	**0.292**	**0.957**	**0.870**				IRT	0.573	0.836	0.672
		500 (5000)	DRM	0.339	0.942	0.834			500 (5000)	DRM	**0.499**	**0.878**	**0.722**
			IRT	**0.290**	**0.958**	**0.873**				IRT	0.553	0.846	0.679
		1000 (10000)	DRM	0.329	0.947	0.844			1000 (10000)	DRM	**0.495**	**0.892**	**0.731**
			IRT	**0.298**	**0.968**	**0.879**				IRT	0.534	0.851	0.691
	0 (300)	50 (500)	DRM	0.328	0.946	**0.860**		0 (300)	50 (500)	DRM	**0.661**	**0.781**	**0.586**
			IRT	**0.308**	**0.952**	0.858				IRT	0.786	0.691	0.489
		100 (1000)	DRM	0.339	0.943	0.851			100 (1000)	DRM	**0.579**	**0.832**	**0.664**
			IRT	**0.314**	**0.951**	**0.858**				IRT	0.762	0.709	0.506
		500 (5000)	DRM	0.321	0.941	0.853			500 (5000)	DRM	**0.561**	**0.852**	**0.684**
			IRT	**0.299**	**0.945**	**0.873**				IRT	0.732	0.705	0.512
		1000 (10000)	DRM	0.302	0.938	0.853			1000 (10000)	DRM	**0.539**	**0.850**	**0.644**
			IRT	**0.281**	**0.948**	**0.881**				IRT	0.712	0.709	0.506
50	5 (455)	50 (500)	DRM	0.317	0.950	0.882	50	5 (455)	50 (500)	DRM	**0.376**	**0.929**	**0.802**
			IRT	**0.251**	**0.969**	**0.895**				IRT	0.426	0.909	0.760
		100 (1000)	DRM	0.312	0.964	0.891			100 (1000)	DRM	**0.393**	**0.923**	**0.811**
			IRT	**0.243**	**0.970**	**0.896**				IRT	0.805	0.750	0.543
		500 (5000)	DRM	0.288	0.959	0.894			500 (5000)	DRM	**0.372**	**0.930**	**0.810**
			IRT	**0.232**	**0.973**	**0.901**				IRT	1.044	0.454	0.282
		1000 (10000)	DRM	0.278	0.961	0.894			1000 (10000)	DRM	**0.392**	**0.914**	**0.798**
			IRT	**0.234**	**0.973**	**0.901**				IRT	0.923	0.512	0.342
	0 (500)	50 (500)	DRM	0.360	0.935	0.856		0 (500)	50 (500)	DRM	**0.635**	**0.798**	**0.599**
			IRT	**0.274**	**0.962**	**0.876**				IRT	0.782	0.694	0.489
		100 (1000)	DRM	0.261	0.966	0.884			100 (1000)	DRM	**0.408**	**0.916**	**0.785**
			IRT	**0.251**	**0.968**	**0.892**				IRT	0.612	0.812	0.532
		500 (5000)	DRM	0.341	0.942	0.887			500 (5000)	DRM	**0.421**	**0.891**	**0.765**
			IRT	**0.241**	**0.971**	**0.899**				IRT	0.598	0.822	0.495
		1000 (10000)	DRM	0.266	0.968	0.889			1000 (10000)	DRM	**0.411**	**0.901**	**0.785**
			IRT	**0.241**	**0.972**	**0.901**				IRT	0.602	0.829	0.498

よりも高精度であることがわかった．今回の実験では一つの母集団からランダムサンプリングされた受検者を能力順に分割して各テストに割り当てており，共通項目がある場合，同時尺度推定法は欠測データによる漸近一致性を持つ能力値推定として正当化できる．しかし，実際には受検者が受けていない多くの項目への反応を推定しながら能力推定することになり誤差が大きくなることがわかる．さらに，共通項目がなく，同時尺度推定法が理論的に適用できない場合でも，提案モデルでは共通項目がある場合に比べ，精度の低下が抑えられていることもわかる．これらは，提案モデルが受検者の独立性を仮定しておらず，異なるテストを受検した他の受検者の反応と能力推定値を考慮しながら受検者の項目への反応予測精度が高くなるように能力推定し，各テストの受検者能力特性の違いを自動的に補正できたことによると考えられる．

4.3　多母集団に対する能力推定値の頑健性

前述の通り，IRT では単一の受検者母集団からの独

表 3　多母集団に対する頑健性
Table 3　The robustness for multi-population data

項目数	受検者数	μ_1	μ_2	σ^2	IRT	DRM
50	1000	-0.3	0.3	0.7	**0.186****	0.216
		-0.5	0.5	0.5	**0.184****	0.232
		-0.7	0.7	0.3	0.210	**0.206**
		-0.9	0.9	0.1	0.207	**0.195***
	10000	-0.3	0.3	0.7	**0.180****	0.215
		-0.5	0.5	0.5	**0.201***	0.214
		-0.7	0.7	0.3	0.217	**0.211***
		-0.9	0.9	0.1	0.207	**0.185****

**p<.01　*p<.05

立ランダムサンプリングを仮定している．この単一母集団の仮定が IRT が同時確率分布を完全には表現できない原因であり，完全なリンケージを阻害する理由にもなっている [11]．一方，提案モデルは受検者に母集団を仮定しておらず，各テストの受検者が異なる母集団からなる場合にも能力推定の精度が高いことをシミュレーション実験により示す．

本実験では，二つの受検者グループに異なる母集団を仮定する．具体的には受検者グループごとに式 (24) の異なる正規分布から能力値を受検者に割り当てる．

$$N_1(\mu_1, \sigma^2), \quad N_2(\mu_2, \sigma^2) \qquad (24)$$

さらに，各受検者グループが共通項目を含む異なるテストを受検する状況を仮定し，テストデータ行列を 2PLM により生成する．ここでは共通項目を 5 項目とした．

生成した各テストデータ行列について能力を推定し，4.1.2 の RMSE を算出した結果を表 3 に示す．なお，σ^2 は分布全体の標準偏差が 1 に近づくように設定した．

ここで，提案モデルと項目反応理論のパラメータ推定精度に差があることを確かめるため，ウィルコクソンの符号順位検定を行ない，各条件の平均に有意差があるか確かめた．

表より，母集団の平均の差が小さく単一の母集団に近い時は IRT の RMSE が有意に小さいが，母集団の平均の差が大きくなると提案モデルの RMSE が有意に小さくなることがわかった．したがって，提案モデルは大きく異なる複数の母集団から受検者がサンプリングされる場合に頑健に能力推定できることがわかる．

次に，提案モデルが多峯性を表現できるモデルであることを示すため，図 5 に $N_1(-0.7, 0.3), N_2(0.7, 0.3)$ から受検者をサンプリングし生成したテストデータに

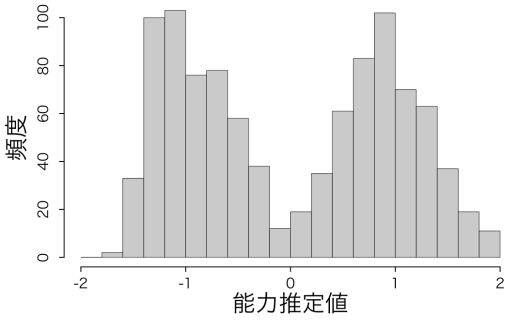


図 5　提案モデルの多母集団に対する能力推定値のヒストグラム
Fig. 5　A histogram of estimated student parameters using proposed method toward multi-population data

表 4　多母集団に対する未知の項目への反応予測精度 (F 値)
Table 4　The robustness for multi-population data

項目数	受検者数	μ_1	μ_2	σ^2	IRT	DRM
50	1000	-0.3	0.3	0.7	0.800	**0.817***
		-0.5	0.5	0.5	0.767	**0.783***
		-0.7	0.7	0.3	0.789	**0.800****
		-0.9	0.9	0.1	0.833	**0.867****
	10000	-0.3	0.3	0.7	0.815	**0.822***
		-0.5	0.5	0.5	0.767	**0.786***
		-0.7	0.7	0.3	0.782	**0.798****
		-0.9	0.9	0.1	0.801	**0.830****

**p<.01　*p<.05

ついて，提案モデルの能力推定値の分布を示す．図から提案モデルは明らかに二峯分布で能力を推定しており，母集団の多峯性を十分に表現しているため，能力推定精度が高くなることがわかる．

4.4　多母集団に対する未知の項目への予測精度

4.3 より，提案モデルは母集団の多峯性を十分に表現できることが示された．このことから，多母集団の受検者に対しても未知の項目への反応予測精度が高いことが期待される．そこで，本節では 4.3 のデータについて 4.1.2 同様に 10 分割交差検証を用いてテストデータ行列を訓練データと評価データに分割し，訓練データ内の反応から推定した能力値と難易度を利用して評価データの全ての反応予測を行い，予測精度を示す F 値を算出する．これを各条件について 10 回ずつ繰り返した F 値の平均を表 4 に示す．ここで，提案モデルと IRT の反応予測精度に差があることを確かめるため，ウィルコクソンの符号順位検定を行ない，各条件の平均に有意差があるか確かめた．

表 4 より，提案モデルは全ての条件で IRT よりも有意に F 値が高く，高精度な反応予測ができることがわかる．したがって，提案モデルは能力推定精度が高い条件に限らず，高精度な反応予測が可能なモデルであることが示唆された．

5. 実データ実験

本章では実データを用い，提案モデルが現実で得られるデータにも有効であることを示す．具体的には，能力推定の信頼性と未知の項目への反応予測精度を IRT と比較する．なお，これらの実験はすべて 4.1 と同様の条件を用いた．

5.1 実データ概要

本節では，実データの概要を説明する．ただし，データの概要は表 5 の 1 列目から 4 列目にもまとめて示してある．

情報データ

情報データは，植野 [38]〜[40] が開発した e ラーニングシステム"Samurai"を用いて行った情報に関する二種類の期末テストデータ行列であり，欠測値がなく単一の受検者グループからなる．受検者数はそれぞれ 169,266 であり，項目数はどちらも 50 である．

批判的思考データ

批判的思考データは，批判的思考を主題として大学生を対象に行われたテストから得られたテストデータ行列である [41]．受検者数は 1221，項目数は 179 であり，欠測値の割合が 87.8%と高い．また，共通受検者によるリンケージが行われている．

プログラミングデータ

プログラミングデータは，プログラミング問題を主題とし，大学生を対象に web システムを用いて収集された二種類のテストデータ行列である [42], [43]．受講者はそれぞれ 93,74，項目数はそれぞれ 13,19 であり，欠測値の割合はそれぞれ 0%, 6.8%と低い．いずれのデータ収集の際も，正答できない受検者にはヒントを提示しているが，本研究ではヒントを用いずに正答した反応のみを正答とみなした．

模試データ

模試データは，駿台模試に対する高校生のテストデータ行列であり，数学 1A と物理の二科目からなる．受検者はそれぞれ 12348,9172 であり，項目数は 48,24 である．模試の特性上，項目間に強い依存性があるがすべてを独立の項目として扱っている．欠測値の割合はそれぞれ 16.4%,12.0%である．

ASSISTMENT データ

ASSISTMENT データは一般に公開されているオンライン教育サービス ASSISTMENT(注2)の数学の問題に対する解答データである．ここでは解答数が 2 以上の受検者 3941，解答数が 30 以上の項目 2921 を用いる．欠測値の割合は 84.4%と高い．

CDM データ

項目分析や Knowledge Tracing 研究のために一般に公開されているデータとして，R の CDM パッケージ [44] に含まれる ECPE データと TIMSS データがある．ECPE データは言語に関するテストの解答データである．受検者は 2922，項目数は 28 であり欠測値はない．また，TIMSS データは数学に関するテストの解答データである．受検者は 757，項目数は 24 であり，欠測値はない．

統計データ

統計データは前述の"Samurai"上で行われた大学生を対象とした講義の確認テストデータであり，受検者は 26，項目数は 25 である．また，欠測値の割合は 33.8%である．

情報倫理データ

情報倫理データも前述の"Samurai"上で行われた大学生を対象とした講義の確認テストデータであり，受検者は 31，項目数は 90 である．また，欠測値の割合は 46.3%である．

技術者倫理データ

技術者倫理データも前述の"Samurai"上で行われた大学生を対象とした講義の確認テストデータであり，受検者は 85，項目数は 69 である．また，欠測値の割合は 26.4%である．

Classi データ

Classi データはタブレット上で Classi(注3)の開発するシステムを用い，高校生が授業中に解答した確認テストからなる．ここでは物理・化学・生物の 3 科目のテストデータ行列を用いる．受検者はそれぞれ 239,1139,192 であり，項目数は 119,364,114 である．それぞれのデータの欠測値の割合は 82.4%,96.4%,93.5%であり，欠測値の多いデータである．

5.2 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評価する．実データでは真値との RMSE を直接求め

(注2)：https://sites.google.com/site/assistmentsdata/home/assistment-2009-2010-data
(注3)：https://classi.jp/

ることができないため，以下のように信頼性を求める．1) テストデータ行列の項目をランダムに二等分し，テストデータ行列を二つの項目グループへの反応データに二等分する．2) 分割した各反応データから受検者の能力を推定する．3) 能力推定値間の RMSE と相関係数を求める．

これらの手順を提案モデルと IRT を用い，それぞれ 10 回繰り返して算出した RMSE と相関係数の平均を各テストデータの受検者数，項目数，欠測割合とともに表 5 に示す．算出した RMSE が低い，または相関係数が高い場合に信頼性が高いとみなせる．ここで，DRM w/o cost は式 (16) の有効性を評価するために損失関数として式 (16) の代わりに式 (15) のみを用いた提案モデルである．さらに，提案モデルの能力推定値分布と正規分布との乖離度（非正規性）を示すため，歪度と尖度の平均も表 5 に示す．どちらの指標も 0 に近い分布ほど正規分布に近い分布であると判断できる．

本実験では各指標に関して全てのデータセットの平均を算出し，ウィルコクソンの符号順位検定を用いて有意差があるか確かめた．

表 5 より，多くのデータセットにおいて提案モデルが IRT より高い信頼性を示し，各指標の平均も提案モデルが高いことがわかる．特に異常値に影響を受けにくい順位相関係数では各データセットの平均に有意差が見られた．また，能力推定値の信頼性については提案モデルと DRM w/o の間に有意差はみられず，少数ラベルの考慮は能力推定値の信頼性に大きな影響を与えないことが示された．

ここで，二つ以上の指標において IRT の信頼性が提案モデルと同等以上であるデータセットは情報 1，模試_数学，模試_物理，ASSISTMENTS，TIMSS データであった．これらのデータセットの能力推定値分布の歪度はいずれも 0 に近く，歪度の観点から正規分布に近い分布であることが示された．例として，図 6 に模試_数学データについて各手法の能力推定値のヒストグラムを示す．図 6 に代表されるように，これらのデータでは提案モデルを用いても正規分布に近い形状で能力を推定していることが確認できる．

一方，歪度がこれらのデータよりも 0 に近いにも関わらず，提案モデルの精度が高いデータセットとして Classi_物理, Classi_化学があげられる．これらのデータセットについて各手法の能力推定値のヒストグラムを図 7, 図 8 に示す．図 7, 図 8 と表 5 から，これらのデータセットは提案モデルの能力推定値分布の尖度は極端に小さく，正規分布とは大きく異なる能力分布を推定していることがわかる．

提案モデルの能力推定値分布が正規分布と乖離してしまう場合の原因として，1) 真の能力分布が正規分布に従っていない場合，2) 真の能力分布は正規分布であるがランダムサンプリングされていない，またはデータ数が小さいなどの理由で正しく推定できてない場合が考えられる．このような状況では，提案モデルが IRT よりも信頼性の高い能力推定値を得ることができることが示された．

5.3 未知の項目への反応予測精度

本節では，提案モデルが IRT と比較して，実データに対しても高精度に未知の項目への反応予測が可能であることを示す．未知の項目への反応予測精度として，4.1.2 同様に 10 分割交差検証を用いて各テストデータ行列を訓練データと評価データに分割し，訓練データ内の反応から推定した能力値と難易度を利用して評価データの全ての反応予測を行い予測精度を示す F 値を算出した結果を表 5 に示す．

ここでも，全てのデータセットの予測精度の平均について，ウィルコクソンの符号順位検定を用いて有意差があるか確かめた．その結果，IRT・DRM w/o cost に対して，提案モデルの F 値が 5%有意に高いことが明らかとなった．

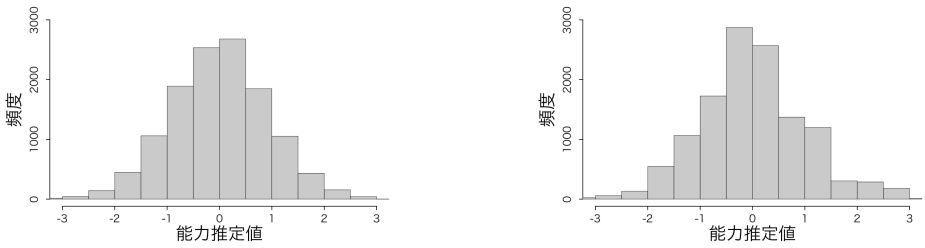
したがって，提案モデルは IRT と比較して，多くの実データにおいて未知の項目への反応予測も高精度であることが示された．また，式 (16) の損失関数を用いたモデルが式 (15) の損失関数を用いたモデルよりも有意に予測精度が高いことから，少数ラベルの考慮が反応予測に重要であることがわかる．

提案モデルと比較し，IRT の精度が高いデータセットとして批判的思考，模試_数学，模試_物理，ASSISTMENTS データがあげられる．表 5 からこれらのデータセットはいずれも受検者が多いデータセットであることがわかる．ただし，例外として ECPE データは受検者数が多いが提案モデルの予測精度が高い．ここで図 9 に，ECPE データについて各手法の能力推定値のヒストグラムを示す．表 5, 図 9 より，提案モデルが推定した ECPE データの能力推定値分布は歪度が比較的高く，尖度が極端に小さく，正規分布とは大きく乖離した分布であることがわかる．したがって，受検者数が多く，その能力分布が正規分布に従っているデータセットについては IRT の予測精度が提

表 5　実データ実験結果
Table 5　The results of actual data

データセット				手法	信頼性（5.2 節）			予測精度（5.3 節）	DRM 能力推定値の非正規性	
実データ	受検者数	項目数	欠測割合		RMSE	Pearson	Kendall	F 値	歪度	尖度
情報 1	169	50	0%	2PLM	**0.466**	**0.891**	0.685	0.734	0.111	0.752
				DRM w/o	0.496	0.881	0.685	**0.737**		
				DRM	0.514	0.867	**0.687**	**0.737**		
情報 2	266	50	0%	2PLM	0.562	0.841	**0.668**	0.699	0.436	-0.427
				DRM w/o	**0.555**	**0.846**	0.663	**0.700**		
				DRM	**0.555**	0.845	0.662	**0.700**		
批判的思考	1221	179	87.8%	2PLM	1.064	0.464	0.318	**0.695**	-0.387	0.848
				DRM w/o	**1.023**	**0.474**	0.325	**0.695**		
				DRM	1.025	**0.474**	**0.327**	0.689		
プログラミング 1	94	13	0%	2PLM	0.890	0.599	0.403	0.719	-0.522	0.896
				DRM w/o	**0.850**	**0.630**	**0.427**	0.717		
				DRM	0.864	0.622	0.417	**0.729**		
プログラミング 2	74	19	6.8%	2PLM	0.752	0.713	0.468	0.676	0.446	-0.711
				DRM w/o	0.730	0.721	0.470	0.682		
				DRM	**0.720**	**0.737**	**0.475**	**0.685**		
模試_数学	12348	48	16.4%	2PLM	**0.589**	**0,748**	**0.533**	**0.783**	0.221	0.406
				DRM w/o	0.721	0.731	0.515	0.780		
				DRM	0.744	0.723	0.521	0.780		
模試_物理	9172	24	12.0%	2PLM	**0.884**	**0.609**	**0.424**	0.721	-0.174	-0.756
				DRM w/o	0.899	0.591	0.416	0.710		
				DRM	0.911	0.585	0.411	0.710		
ASSISTMENTS	3941	2921	84.4%	2PLM	**0.827**	**0.658**	0.441	**0.685**	0.241	0.309
				DRM w/o	0.828	0.650	0.445	**0.685**		
				DRM	0.849	0.639	**0.478**	0.679		
ECPE	2922	28	0%	2PLM	0.875	0.615	0.435	0.719	0.532	-1.261
				DRM w/o	0.877	0.615	0.436	0.719		
				DRM	**0.874**	**0.618**	**0.440**	**0.729**		
TIMSS	757	24	0%	2PLM	0.753	**0.716**	**0.525**	0.711	-0.203	-0.657
				DRM w/o	**0.752**	0.713	0.521	0.710		
				DRM	0.753	**0.716**	0.523	**0.712**		
統計	26	25	33.8%	2PLM	0.619	0.801	0.398	0.852	0.740	-0.573
				DRM w/o	0.568	0.822	0.494	0.862		
				DRM	**0.545**	**0.846**	**0.582**	**0.893**		
情報倫理	31	90	46.3%	2PLM	0.394	0.920	0.643	0.746	-0.937	-0.042
				DRM w/o	0.390	0.920	0.675	0.782		
				DRM	**0.382**	**0.925**	**0.712**	**0.803**		
技術者倫理	85	69	26.4%	2PLM	0.544	0.850	**0.403**	0.634	0.790	-0.282
				DRM w/o	0.522	0.854	0.305	0.593		
				DRM	**0.517**	**0.865**	0.313	**0.685**		
Classi_物理	239	119	92.4%	2PLM	1.053	0.444	0.299	0.720	-0.189	-0.947
				DRM w/o	0.958	0.557	0.425	0.719		
				DRM	**0.943**	**0.554**	**0.403**	**0.721**		
Classi_化学	1139	364	96.4%	2PLM	1.077	0.420	0.297	0.710	0.122	-1.407
				DRM w/o	0.950	0.552	0.413	**0.711**		
				DRM	**0.923**	**0.574**	**0.439**	**0.711**		
Classi_生物	192	114	93.5%	2PLM	1.020	0.475	0.326	0.722	0.358	-1.308
				DRM w/o	0.829	0.502	0.410	**0.725**		
				DRM	**0.748**	**0.717**	**0.531**	**0.725**		
平均				2PLM	0.764	0.680	0.451	0.719		
				DRM w/o	0.747	0.692	0.470	0.719		
				DRM	**0.742**	**0.707**	**0.495***	**0.728***		

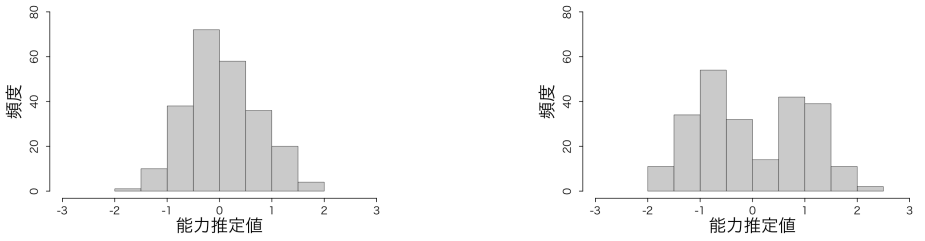
*p<.05



(a) 2PLM による能力推定値　　(b) DRM による能力推定値

図 6　模試_数学データにおける能力推定値のヒストグラム

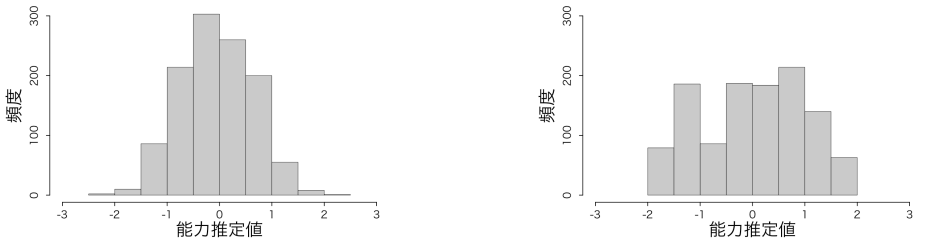
Fig. 6　Histograms of estimated abilities toward Test_Math data



(a) 2PLM による能力推定値　　(b) DRM による能力推定値

図 7　Classi_物理データにおける能力推定値のヒストグラム

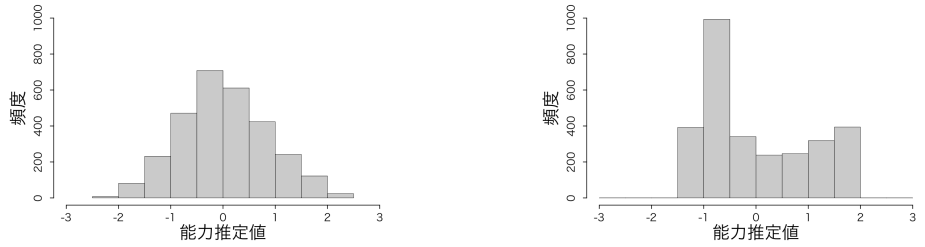
Fig. 7　Histograms of estimated abilities toward Classi_Physics data



(a) 2PLM による能力推定値　　(b) DRM による能力推定値

図 8　Classi_化学データにおける能力推定値のヒストグラム

Fig. 8　Histograms of estimated abilities toward Classi_Chemistory data



(a) 2PLM による能力推定値　　(b) DRM による能力推定値

図 9　ECPE データにおける能力推定値のヒストグラム

Fig. 9　Histograms of estimated abilities toward Critical Thinking data

案モデルを上回るが，それ以外の場合には提案モデルが IRT の精度を上回ることが示唆された．

6. むすび

本研究では，受検者に母集団とランダムサンプリングを仮定しない深層学習を用いたテスト理論である Item Deep Response Theory を提案した．

具体的に，提案モデルは受検者と項目を独立したニューラルネットワークの入力とし，それぞれの出力を組み合わせて項目への正答確率を予測する深層学習モデルである．本研究では受検者を入力としたニューラルネットワークの出力を受検者の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，提案モデルには以下の利点があることが明らかとなった．

1) テスト間に共通する項目がなくとも能力を高精度に推定することができる．
2) 受検者にランダムサンプリングが仮定できない場合にも能力を高精度に推定することができる．
3) 受検者の母集団が単一でない場合にも能力を高精度に推定することができる．
4) 過去の反応履歴から未知の項目への反応予測をする際に提案モデルが最も高精度な予測が可能である．

これらより，提案モデルはリンケージを十分に行うことができないデータや単一の母集団を仮定できないデータに特に有効であり，実データへの当てはまりが良いことが明らかとなった．

提案モデルは IRT と比較して反応予測精度が有意に高いモデルであるため，アダプティブテスティングやアダプティブラーニング [42], [45], [46] への応用が期待される．また，提案モデルは多段階反応データや時系列反応データ，多次元データへの適応が容易であるため，さらなるモデルの拡張を行う．

今回，アイテムバンクから各テストについて必要数だけ項目をランダムサンプリングしてテスト構成を行なった．しかし，情報量を用いたテスト構成法などが実際には用いられ，その構成法も多く存在する [8], [47]〜[50]．テストの構成法による影響分析については今後の課題とする．

謝辞　本研究は JSPS 科研費 19H05663, 19K21751 の助成を受けたものです．

文　献

[1] J. P. Guilford, Psychometric Method, vol.XIV, 01 1936.
[2] H. Gulliksen, “A course in the theory of mental tests,” Psychometrika, vol.8, no.4, pp.223–245, Dec. 1943.
[3] F.M. Lord and M.R. Novick, Statistical Theories of Mental Test Scores, Addison-Wesley, 1968.
[4] F.B. Baker and S.H. Kim, Item Response Theory: Parameter Estimation Techniques, Second Edition, Statistics: A Series of Textbooks and Monographs, Taylor & Francis, 2004.
[5] W.J. van der Linden., Handbook of Item Response Theory, Volume Two: Statistical Tools, Chapman and Hall/CRC Statistics in the Social and Behavioral Sciences, Chapman and Hall/CRC, 2016.
[6] 植野真臣，知識社会における e ラーニング，培風館，2007.
[7] 植野真臣，永岡慶三，e テスティング，培風館，2009.
[8] T. Ishii, P. Songmuang, and M. Ueno, “Maximum clique algorithm and its approximation for uniform test form assembly,” IEEE Transactions on Learning Technologies, vol.7, no.1, pp.83–95, Jan. 2014.
[9] 仁田善雄，齋藤宣彦，後藤英司，高木　康，石田達樹，江藤一洋，“医療系大学間共用試験における e テスティング，” 日本テスト学会第 12 回大会発表論文抄録集，第 33 巻，pp.58–59，2014.
[10] 谷澤明紀，本多康弘，“情報処理技術者試験における e テスティング，” 日本テスト学会第 12 回大会発表論文抄録集，第 33 巻，pp.54–57，2014.
[11] W.J. van derLinden and M.D. Barrett, “Linking item response model parameters,” Psychometrika, vol.81, no.3, pp.650–673, Sept. 2016.
[12] F.M. Lord, Applications of item response theory to practical testing problems, L. Erlbaum Associates Hillsdale, N.J, 1980.
[13] 豊田秀樹，岩間徳兼，中村彩子，齋藤康寛，“項目反応理論を用いたテスト運用への切り替えコスト軽減の試み—多数の潜在特性尺度の同時等化法を利用して—，” 日本オペレーションズ・リサーチ学会和文論文誌，vol.58，pp.122–147，2015.
[14] W.J. van der Linden, Handbook of Item Response Theory, Volume Three: Applications, Chapman and Hall/CRC Statistics in the Social and Behavioral Sciences, Chapman and Hall/CRC, 2016.
[15] S.-H. Joo, P. Lee, and S. Stark, “Evaluating anchor-item designs for concurrent calibration with the ggum,” Applied Psychological Measurement, vol.41, no.2, pp.83–96, 2017.
[16] 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.
[17] 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

2018, London, UK, June 27-30, 2018, Proceedings, Part I, pp.239–252, 2018.

[18] 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, pp.198–211, Springer International Publishing, Cham, 2018.

[19] 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, pp.491–502, Springer International Publishing, Cham, 2018.

[20] 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, pp.70–85, Springer International Publishing, Cham, 2018.

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

[22] K.H. Wilson, Y. Karklin, B. Han, and C. Ekanadham, "Back to the basics: Bayesian extensions of irt outperform neural networks for proficiency estimation," vol.1, pp.539–544, 06 2016.

[23] M. J. Kolen and R. L. Brennan, Test Equating, Scaling, and Linking: Methods and Practices (2nd ed.), 01 2004.

[24] G.L. Marco, "Item characteristic curve solutions to three intractable testing problems," Journal of Educational Measurement, vol.14, no.2, pp.139–160, 1977.

[25] B.H. Loyd and H.D. Hoover, "Vertical equating using the rasch model," Journal of Educational Measurement, vol.17, no.3, pp.179–193, 1980.

[26] T. Haebara, "Equating logistic ability scales by a weighted least squares method," Japanese Psychological Research, vol.22, no.3, pp.144–149, 1980.

[27] M.L. Stocking and F.M. Lord, "Developing a common metric in item response theory," Applied Psychological Measurement, vol.7, no.2, pp.201–210, 1983.

[28] R.D. Bock and M.F. Zimowski, "Multiple group irt," pp.433–448, Springer New York, New York, NY, 1997.

[29] W. Shen, X. Wang, X. Bai, and Z. Zhang, "Deepcontour: A deep convolutional feature learned by positive-sharing loss for contour detection," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.3982–3991, June 2015.

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

[31] 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.

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

[33] A. Graves, "Generating sequences with recurrent neural networks.," CoRR, vol.abs/1308.0850, 2013.

[34] 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.

[35] S. Kilmen and N. Demirtasli, "Comparison of test equating methods based on item response theory according to the sample size and ability distribution," Procedia - Social and Behavioral Sciences, vol.46, pp.130–134, 2012. 4th WORLD CONFERENCE ON EDUCATIONAL SCIENCES (WCES-2012) 02-05 February 2012 Barcelona, Spain.

[36] I. Uysal and S. Kilmen, "Comparison of item response theory test equating methods for mixed format tests," International Online Journal of Educational Sciences, vol.2016, 06 2016.

[37] 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，" 電子情報通信学会論文誌. D, 情報・システム = The IEICE transactions on information and systems (Japanese edition)，vol.101，no.6，pp.895–908，June 2018.

[38] M. Ueno, "Animated agent to maintain learner's attention in e-learning," Proceedings of E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2004，eds. by J. Nall and R. Robson，pp.194–201，Association for the Advancement of Computing in Education (AACE)，Washington, DC, USA，2004.

[39] M. Ueno, "Data mining and text mining technologies for collaborative learning in an ilms "samurai".," ICALT '04 Proceedings of the IEEE International Conference on Advanced Learning Technologies, pp.1052–1053, 01 2004.

[40] M. Ueno, "Intelligent lms with an agent that learns from log data," Proceedings of E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2005, ed. by G. Richards, pp.3169–3176, Association for the Advancement of Computing in Education (AACE), Vancouver, Canada, Oct. 2005.

[41] 若山 昇，宮澤芳光，梶谷真司，宇都雅輝，植野真臣，"クリティカルシンキングの設問における識別力・困難度，" 教育テスト研究センター年報，vol.3，pp.28–30，2018.

[42] M. Ueno and Y. Miyazawa, "Irt-based adaptive hints to scaffold learning in programming," IEEE Transactions on Learning Technologies, vol.11, no.4, pp.415–

428, Oct. 2018.
[43] 堤瑛美子，宇都雅輝，植野真臣，“ダイナミックアセスメントのための隠れマルコフ irt モデル，” 電子情報通信学会論文誌. D, 情報・システム = The IEICE transactions on information and systems (Japanese edition)，vol.102，no.2，pp.79–92，2019.
[44] A.C. George, A. Robitzsch, T. Kiefer, J. Groß, and A. Ünlü, “The r package cdm for cognitive diagnosis models,” Journal of Statistical Software, Articles, vol.74, no.2, pp.1–24, 2016.
[45] 植野真臣，松尾淳哉，“項目反応理論を用いて適応型ヒントを提示する足場かけシステム，” 電子情報通信学会論文誌. D, 情報・システム = The IEICE transactions on information and systems (Japanese edition)，vol.98，pp.17–29，2015.
[46] M. Ueno and Y. Miyasawa, “Probability based scaffolding system with fading,” Artificial Intelligence in Education, eds. by C. Conati, N. Heffernan, A. Mitrovic, and M.F. Verdejo, pp.492–503, Springer International Publishing, Cham, 2015.
[47] T. Ishii, P. Songmuang, and M. Ueno, “Maximum clique algorithm for uniform test forms assembly,” Artificial Intelligence in Education, eds. by H.C. Lane, K. Yacef, J. Mostow, and P. Pavlik, pp.451–462, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.
[48] T. Ishii and M. Ueno, “Clique algorithm to minimize item exposure for uniform test forms assembly,” Artificial Intelligence in Education, eds. by C. Conati, N. Heffernan, A. Mitrovic, and M.F. Verdejo, pp.638–641, Springer International Publishing, Cham, 2015.
[49] T. Ishii and M. Ueno, “Algorithm for uniform test assembly using a maximum clique problem and integer programming,” Artificial Intelligence in Education, ElisabethAndr¥’e，R. Baker，X. Hu，M.M.T. Rodrigo，B. duBoulay（編），pp.102–112, Springer International Publishing, Cham, 2017.
[50] 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，” 電子情報通信学会論文誌. D, 情報・システム = The IEICE transactions on information and systems (Japanese edition)，vol.100，no.1，pp.47–59，2017.

（平成 xx 年 xx 月 xx 日受付）

木下　涼

2018 年電気通信大学大学院情報理工学研究科博士前期課程修了，同年電気通信大学大学院情報理工学研究科博士後期課程課程入学，在学中.

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.

Abstract Item Response Theory (IRT) can evaluate examinees who responded different items on the same scale. It needs to assume independently random sampling of examinees and to linkage using common items in different tests. However, it costs enormous time to linkage and theoretically, we cannot achieve the optimal scores. To solve this problem, we propose a test theory based on deep learning, Item Deep Response Theory(IDRT), which does not assume a population or independence of examinees. Without linkage, IDRT can estimate the parameters accurately under the condition where a population of examinees are multi-population. From the results of experiments using simulation and real data, we show IDRT can predict unknown item responses more correctly than IRT, especially in the case the examinees are not sampled randomly from the same population.

Key words educational technology, test theory, linkage, deep learning, item response theory, e-testing