修 士 論 文 の 和 文 要 旨

研究科・専攻	大学院　情報理工学研究科　情報・ネットワーク工学専攻　博士前期課程		
氏　　名	石川　文弥	学籍番号	2431016
論 文 題 目	問題項目への正誤反応と所要時間を同時に予測する Multi-task Deep Neural Network		

要　旨

e-testing とは, 異なる問題項目で構成されたテストセットを用いた場合でも, 項目反応理論 (Item Response Theory; IRT) に基づき受検者の潜在能力を同一精度で測定可能なコンピュータテストのことである. 時間制約下における e-testing の公平性を確保するためには, 問題項目の正誤反応の予測に加えて所要時間の予測が重要である. 実際に, 所要時間の予測は, 不正行為の検知や時間制約を組み込んだテスト構成手法など教育評価に関する様々な課題に広く応用されている. 代表的な所要時間予測手法として, 問題項目の正誤反応と所要時間を同時に予測する階層ベイズモデル (Log-normal Response Time IRT; LNIRT) が提案されている. しかし, LNIRT は受検者の潜在変数が正規分布に従うと仮定するため, 実データがこの仮定から逸脱する場合には予測精度が制限される可能性がある.

一方, 近年では深層学習に基づく手法が確率モデルよりも優れた予測性能を示すことが報告されている. 特に, Tsutsumi らの Deep-IRT は, 高精度な正誤反応予測に加え, 受検者の能力を表す受検者ネットワークと項目の困難度を表す項目ネットワークを独立に組み込むことで, IRT と同様に解釈可能なパラメータを推定できる.

そこで, 本論では, 問題項目の正誤反応と所要時間を同時に予測する新たな Deep-IRT を提案する. 提案手法は, マルチタスク学習で高い性能が報告されている Multi-gate Mixture-of-Experts に LNIRT として解釈可能なパラメータを推定する受検者ネットワークおよび項目ネットワークを組み込む. 受検者ネットワークは受検者の能力と回答速度を, 項目ネットワークは項目の困難度と時間困難度をデータから直接推定し, 推定された能力と困難度から正誤反応を, 回答速度と時間困難度から所要時間を予測する. これにより, 提案手法は解釈性を維持しつつ, 正誤反応と所要時間の予測精度を向上させる. 評価実験では, 提案手法と従来手法の正誤反応および所要時間の予測精度を比較し, 提案手法の有効性を示す. 最後に, 提案手法と LNIRT による相関分析を行い, 提案手法のパラメータ推定値の解釈性を評価する.

2025 年度 情報数理工学 (MI) プログラム
修士論文

問題項目への正誤反応と所要時間を同時に予測する Multi-task Deep Neural Network

2026 年 3 月 2 日

電気通信大学 情報数理工学プログラム
学籍番号 2431016

石川 文弥

主任指導教員: 植野 真臣
副指導教員: 宇都 雅輝

目次

1　まえがき

近年，Computer Based Testing（CBT）や Computerized Adaptive Testing（CAT）などを含む e-testing が，教育評価を実施するための枠組みとして広く利用されている [1–7]．e-testing の主要な利点は，異なる問題項目（以降，項目と呼ぶ）で構成されたテストセットを用いた場合でも，項目反応理論 (Item Response Theory; IRT) に基づいて受検者の潜在能力を同一精度で推定できる点にある [1–7]．IRT は，回答データから潜在能力を推定することで，同一尺度上での同一精度による測定を実現する．その結果，e-testing は異なる時点や場所で実施されたテストに対して等質な評価を行うことが可能となる．

しかし，IRT の予測精度には本質的な限界がある．IRT は，一般的に受検者の能力が正規分布からランダムサンプリングされると仮定しているが，実際には能力の分布はこの仮定から逸脱することが多い．この場合，IRT の最適性は理論的に保証されない．

これらの仮定を緩和するために，深層学習に基づき IRT を拡張した Deep-IRT [8–13] が提案されている．深層学習は，統計的な分布の仮定に依存しないため，IRT のような確率モデルより高い予測精度を達成する．しかし，e-testing では解釈性が重要であるにもかかわらず，深層学習はその解釈性が低いという課題がある．この課題に対処するために，Deep-IRT は独立した受検者ネットワークと項目ネットワークを組み込んでいる．受検者ネットワークでは受検者の能力を，項目ネットワークでは項目の困難度を推定する．その結果，Deep-IRT は解釈性を維持しながら実データに対して頑健な能力推定が可能となり，IRT よりも高い予測精度を実現した．

一方で，時間制約下における e-testing の公平性を確保するためには，項目の正誤反応だけでなく所要時間の予測も重要である．実際に，所要時間の予測は，カンニングなどの不正行為の検知 [14–17]，受検者モデリングの改善 [18, 19]，時間制約を組み込んだ CAT 手法への導入 [20, 21] など，教育評価に関する様々な課題に広く応用されている．さらに，所要時間予測はアダプティブ・ラーニングにおいても同様に応用が検討されている (例えば，[22])．

この目的のために，van der Linden [23] は，項目の正誤反応と所要時間を同時に予測可能な階層ベイズモデル (Log-normal Response Time IRT; LNIRT) を提案した．LNIRT は，IRT ベースのアプローチの中で最も高い予測精度を達成している [24–31]．しかし，LNIRT は従来の IRT と同様に，統計的分布の仮定による予測精度の限界を抱えている．

この課題を解決するために，本研究では，LNIRT に匹敵する解釈可能なパラメータを組み込んだ，Multi-gate Mixture-of-Experts (MMoE) [32] に基づく Multi-task Deep-IRT

を提案する．MMoE は，項目の正誤反応や所要時間といったタスク間の相補的な関係を学習するために広く用いられているフレームワークである．具体的には，MMoE は，タスク間で共通の特徴を学習するエキスパートネットワークと，各タスク独自の特徴を抽出するタスク固有のタワーネットワークを用いることで，予測精度を向上させる．しかし，MMoE はその高い予測精度にもかかわらず，解釈性に欠ける．

解釈性の欠如に対処するために，本研究では解釈可能な受検者の潜在特性を学習する受検者ネットワークと項目ネットワークを組み込むことで，MMoE におけるタスク固有のタワーネットワークを拡張した手法を提案する．受検者ネットワークは受検者の能力と回答速度パラメータを推定し，項目ネットワークは項目の困難度と時間困難度パラメータを推定する．推定された受検者の能力と項目の困難度パラメータは項目の正誤反応の予測に使用され，回答速度と時間困難度パラメータは所要時間の予測に使用される．その結果，提案手法は解釈性を維持しながら，項目の正誤反応と所要時間の両方の予測精度を向上させる．

本研究では，提案手法の有効性を実データを用いた比較実験により示した．その結果，提案手法は正誤反応の予測精度を低下させることなく，従来手法と比較して所要時間の予測精度を向上させることを示した．特に，提案手法は，実データにおける受検者の能力と回答速度が正規分布に従わない場合に，予測精度を向上させた．また，提案手法による項目パラメータの推定値は LNIRT に同様の解釈性を持つことを示した．受検者の潜在変数においては，提案手法は観測データから直接分布を学習することで，LNIRT よりも多様な受検者の能力分布を捉えられることが示された．

2　項目反応理論における正誤反応と所要時間の同時予測

本章では，項目反応理論 (Item Response Theory;IRT) に基づく項目の正誤反応と所要時間の同時予測手法について説明する．

2.1　項目反応理論

IRT [33,34] では，受検者 $i \in \{1,\ldots,I\}$ の項目 $j \in \{1,\ldots,J\}$ に対する正誤反応 u_{ij} を以下のように定義する．

$$u_{ij} = \begin{cases} 1 & \text{受検者 } i \text{ が項目 } j \text{ に正答したとき,} \\ 0 & \text{それ以外} \end{cases} \tag{1}$$

広く用いられる IRT として，3 母数ロジスティックモデル (3-Parameter Logistic Model; 3PLM) がある．3PLM では，受検者の潜在能力 $\theta_i \in (-\infty,\infty)$ に対して，項目 j の正答確率は以下のように定義される．

$$p(u_{ij}=1 \mid \theta_i) = c_j + \frac{1-c_j}{1+\exp\{-a_j(\theta_i - b_j)\}}, \tag{2}$$

ここで $a_j \in (0,\infty)$ は識別力，$b_j \in \mathbb{R}$ は困難度，$c_j \in [0,1)$ は当て推量を表す．特に，$c_j = 0$ の場合は 2 母数ロジスティックモデル (2-Parameter Logistic Model; 2PLM) に簡約される．

2.2　所要時間予測のための対数正規分布モデル

Van der Linden [31] は，所要時間を対数正規分布に従う確率変数としてモデル化した手法 (Log-normal Response Time Theory; LNRT) を提案した．LNRT では，受検者 i の項目 j への所要時間 t_{ij} の確率密度関数を以下のように定義する．

$$f(t_{ij};\zeta_i,\phi_j,\lambda_j) = \frac{\phi_j}{t_{ij}\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}\Big[\phi_j\big(\ln t_{ij} - (\lambda_j - \zeta_i)\big)\Big]^2 \right\}, \tag{3}$$

ここで，$\zeta_i \in (-\infty,\infty)$ は受検者 i の回答速度を表す潜在変数であり，$\lambda_j \in (-\infty,\infty)$ および $\phi_j \in (0,\infty)$ は項目 j の時間困難度および時間識別力を表すパラメータである．

2.3 正誤反応と所要時間を同時に予測する階層ベイズモデル

現在，代表的かつ高精度な所要時間予測手法として，階層ベイズモデルに基づく LNIRT (Log-normal Response Time IRT) [23] が提案されている．LNIRT は，受検者レベルの潜在変数と項目レベルのパラメータを通じて，項目の正誤反応と所要時間の関係をモデル化する．具体的には，第一層では，項目の正誤反応および所要時間を予測するモデルが条件付き独立に定義される．各モデルはそれぞれ独自の受検者潜在変数と項目パラメータを持つ．第二層では，これらの潜在変数と項目パラメータは，それぞれ受検者および項目の集団全体で多変量正規分布に従うと仮定される．これにより，LNIRT は正誤反応と所要時間の関係を捉えることができる．

第一層において，LNIRT は受検者と項目の各組み合わせに対して，IRT および LNRT を条件付き独立に定義する．なお，本研究では，IRT として 2PLM を採用する．LNIRT は，受検者の潜在変数 (能力および回答速度) と項目パラメータ (識別力，困難度，時間識別力，時間困難度) がそれぞれ母集団分布からサンプリングされると仮定する．この仮定の下，第二層では以下の同時分布が定義される．

まず，受検者の能力と回答速度の同時分布は二変量正規分布として定義される．

$$
\begin{aligned}
(\theta_i, \zeta_i) &\sim \mathcal{N}_2(\boldsymbol{\mu}_{(\theta_i,\zeta_i)}, \boldsymbol{\Sigma}_{(\theta_i,\zeta_i)}), \\
\boldsymbol{\mu}_{(\theta_i,\zeta_i)} &= (\mu_\theta, \mu_\zeta), \\
\boldsymbol{\Sigma}_{(\theta_i,\zeta_i)} &= \begin{pmatrix} \sigma_\theta^2 & \sigma_{\theta\zeta} \\ \sigma_{\theta\zeta} & \sigma_\zeta^2 \end{pmatrix},
\end{aligned}
\tag{4}
$$

ここで，μ_θ，μ_ζ はそれぞれ受検者全体における能力と回答速度の平均を表し，σ_θ^2，σ_ζ^2 は受検者全体の能力と回答速度の分散を表す．また，共分散項 $\sigma_{\theta\zeta}$，$\sigma_{\zeta\theta}$ は受検者全体における能力と回答速度の共分散を表す．

次に，式 (4) の受検者レベルの分布と同様に，項目パラメータは以下の多変量正規分布に従うと仮定される．

$$
\begin{aligned}
(a_j, b_j, \phi_j, \lambda_j) &\sim \mathcal{N}_4(\boldsymbol{\mu}_{(a_j,b_j,\phi_j,\lambda_j)}, \boldsymbol{\Sigma}_{(a_j,b_j,\phi_j,\lambda_j)}), \\
\boldsymbol{\mu}_{(a_j,b_j,\phi_j,\lambda_j)} &= (\mu_a, \mu_b, \mu_\phi, \mu_\lambda) \\
\boldsymbol{\Sigma}_{(a_j,b_j,\phi_j,\lambda_j)} &= \begin{pmatrix} \sigma_a^2 & \sigma_{ab} & \sigma_{a\phi} & \sigma_{a\lambda} \\ \sigma_{ba} & \sigma_b^2 & \sigma_{b\phi} & \sigma_{b\lambda} \\ \sigma_{\phi a} & \sigma_{\phi b} & \sigma_\phi^2 & \sigma_{\phi\lambda} \\ \sigma_{\lambda a} & \sigma_{\lambda b} & \sigma_{\lambda\phi} & \sigma_\lambda^2 \end{pmatrix},
\end{aligned}
\tag{5}
$$

ここで，μ_a，μ_b，μ_ϕ，μ_λ はそれぞれ項目全体における識別力，困難度，時間識別力，時間困難度の平均を表し，共分散行列 $\boldsymbol{\Sigma}_{(a_j,b_j,\phi_j,\lambda_j)}$ は，これらの項目パラメータの分散お

よび項目パラメータ間の共分散を表す．また，対角要素 (例えば，σ_a^2，σ_b^2) はそれぞれのパラメータの分散に対応し，非対角要素 (例えば，σ_{ab}，$\sigma_{a\phi}$) は対応する項目パラメータ間の共分散を表す．

LNIRT では，第一層および第二層における潜在変数と項目パラメータは，ギブスサンプリング [35] によって推定される．第二層では，受検者パラメータ (θ, ζ) および項目パラメータ (a, b, ϕ, λ) の同時分布の共分散行列に逆ウィシャート事前分布が仮定され，対応する平均ベクトルには条件付き正規事前分布が与えられる．

$$\boldsymbol{\Sigma}_{(\theta_i,\zeta_i)} \sim \text{Inv-Wishart}_{\nu_{(\theta_i,\zeta_i)}}(V^{-1}_{(\theta_i,\zeta_i)}), \tag{6}$$

$$\boldsymbol{\mu}_{(\theta_i,\zeta_i)} \mid \boldsymbol{\Sigma}_{(\theta_i,\zeta_i)} \sim \mathcal{N}(\boldsymbol{\mu}_{0,(\theta_i,\zeta_i)}, \boldsymbol{\Sigma}_{(\theta_i,\zeta_i)}/\kappa_{0,(\theta_i,\zeta_i)}), \tag{7}$$

$$\boldsymbol{\Sigma}_{(a_j,b_j,\phi_j,\lambda_j)} \sim \text{Inv-Wishart}_{\nu_{(a_j,b_j,\phi_j,\lambda_j)}}(V^{-1}_{(a_j,b_j,\phi_j,\lambda_j)}), \tag{8}$$

$$\boldsymbol{\mu}_{(a_j,b_j,\phi_j,\lambda_j)} \mid \boldsymbol{\Sigma}_{(a_j,b_j,\phi_j,\lambda_j)} \sim \mathcal{N}(\boldsymbol{\mu}_{0,(a_j,b_j,\phi_j,\lambda_j)}, \boldsymbol{\Sigma}_{(a_j,b_j,\phi_j,\lambda_j)}/\kappa_{0,(a_j,b_j,\phi_j,\lambda_j)}). \tag{9}$$

ここで，$\nu_{(\theta_i,\zeta_i)}$，$\nu_{(a_j,b_j,\phi_j,\lambda_j)}$ は自由度，$V_{(\theta_i,\zeta_i)}$，$V_{(a_j,b_j,\phi_j,\lambda_j)}$ は逆ウィシャート事前分布のスケール行列を表す．$\boldsymbol{\mu}_{0,(\theta_i,\zeta_i)}$ および $\boldsymbol{\mu}_{0,(a_j,b_j,\phi_j,\lambda_j)}$ は条件付き正規事前分布の平均ベクトルを表し，$\kappa_{0,(\theta_i,\zeta_i)}$ および $\kappa_{0,(a_j,b_j,\phi_j,\lambda_j)}$ はスケーリングパラメータを表す．なお，推定手順の詳細は [23] に示されている．

3　提案手法

LNIRT [23] は正誤反応と所要時間を同時に扱える一方で，予測精度には限界がある．LNIRT は，受検者の能力や回答速度が正規分布に従う仮定しているが，実際にはこの仮定を満たさない場合が多く，予測精度が制限される．一方で，深層学習に基づく手法が確率モデルよりも高い予測精度を示すことが報告されている [8–13]．Tsutsumi らが提案した Deep-IRT [9–13] は，深層学習に基づき IRT を拡張することで，従来の IRT よりも高い予測精度を達成した．さらに，Deep-IRT は IRT として解釈可能なパラメータを推定する受検者ネットワークと項目ネットワークを組み込むことで，e-testing において重要な解釈性を維持している．したがって，本研究では Multi-gate Mixture-of-Experts (MMoE) [32] に基づく Multi-task Deep-IRT を提案する．提案手法は，最先端のマルチタスク深層学習手法である MMoE に LNIRT として解釈可能なパラメータを推定する受検者ネットワークと項目ネットワークを組み込むことで，解釈性を維持しつつ正誤反応と所要時間の予測精度を向上させる．

3.1　Multi-gate Mixture-of-Experts

近年, 関連するタスク間での予測精度を向上させるために, 様々なマルチタスク深層学習手法が提案されている [32,36,37]．これらの手法の中で，Multi-gate Mixture-of-Experts (MMoE) は最も高い予測精度を示すことで知られている．MMoE は，航空宇宙 [38]，産業故障検出 [39]，交通データ解析 [40] など，様々な分野で優れた性能を示している．

MMoE の概要を図 1 に示す．MMoE はエキスパートネットワークと，タスク固有のゲーティングネットワークおよびタワーネットワークの 3 つの主要なネットワークにより構成される．エキスパートネットワークは，d 次元の特徴量で構成される入力ベクトル $\boldsymbol{x} \in \mathbb{R}^d$ から，タスク間で共通する特徴を学習する．次に，各タスク固有のゲーティングネットワークは，エキスパートに異なる重みを割り当てることで，各タスクに対する共通特徴の重要度を学習する．これらの重み付けされた特徴量は，対応するタワーネットワークへの入力として用いられ，予測のためのタスク固有の特徴を学習する．

MMoE は，タスク間で共有される特徴表現を学習するために，E 個のエキスパートネットワーク $(e = 1, 2, \ldots, E)$ で構成される．各エキスパートネットワークは，R 個の隠れ層 $(r = 1, 2, \ldots, R)$ を持つ多層フィードフォワードニューラルネットワークとして実装される．e 番目のエキスパートネットワークにおいて，r 番目の隠れ層の出力 $\boldsymbol{L}_r^e$ は，

線形変換と ReLU 活性化関数を用いて以下のように計算される．

$$\boldsymbol{L}_r^e = \begin{cases} \mathrm{ReLU}(\boldsymbol{W}_r^e \boldsymbol{x} + \boldsymbol{b}_r^e), & r = 1, \\ \mathrm{ReLU}(\boldsymbol{W}_r^e \boldsymbol{L}_{r-1}^e + \boldsymbol{b}_r^e), & r = 2, \ldots, R, \end{cases} \tag{10}$$

ここで，$\boldsymbol{W}_r^e$ と $\boldsymbol{b}_r^e$ はそれぞれ，e 番目のエキスパートネットワークにおける r 番目の隠れ層の重み行列とバイアスベクトルを表す．さらに，各エキスパートから出力される特徴表現は以下の通りとなる．

$$\boldsymbol{f}_e(\boldsymbol{x}) = \boldsymbol{L}_R^e. \tag{11}$$

次に，各タスク $t \in \{1, 2, \ldots, T\}$ に対して，タスク固有のゲーティングネットワーク

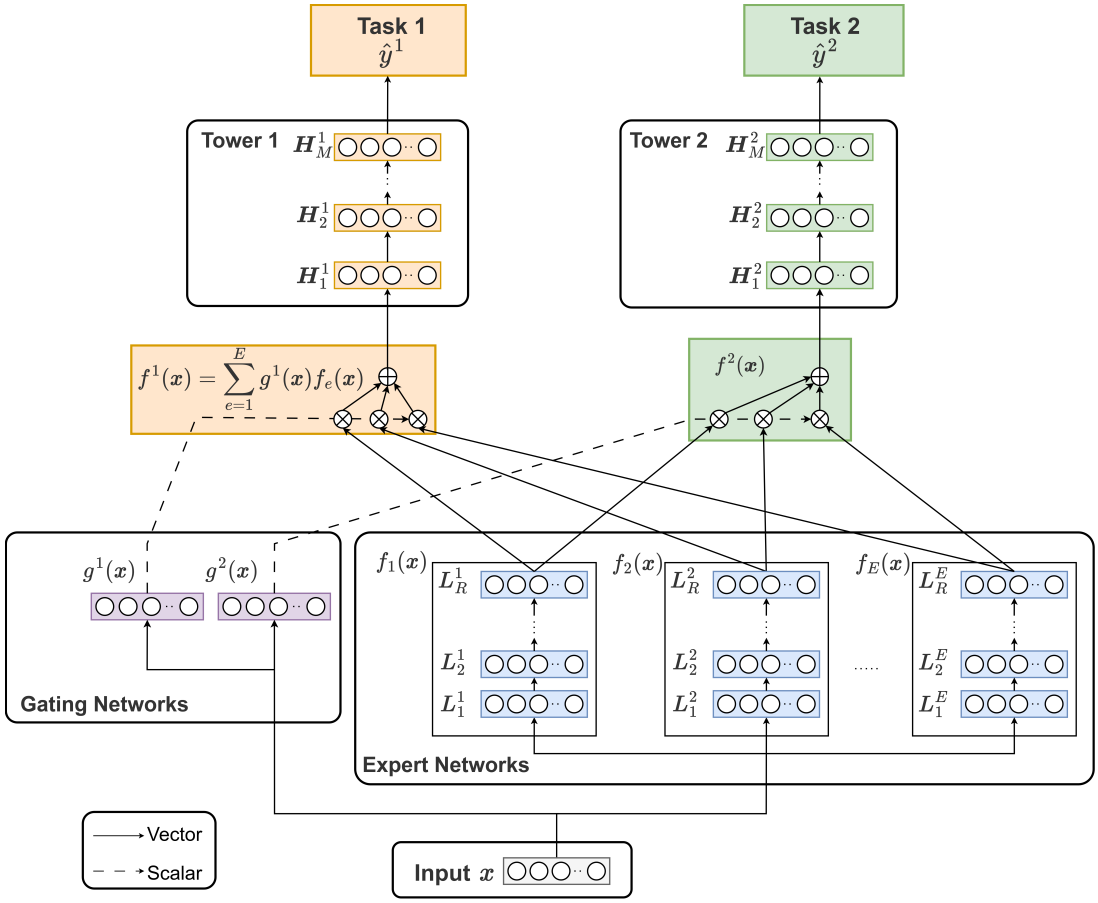


図 1: MMoE の概要図

は，各エキスパートの特徴表現 $\boldsymbol{f}_e(\boldsymbol{x})$ の重要度を学習する．

$$g_e^t(\boldsymbol{x}) = \mathrm{softmax}(\boldsymbol{W}_e^t\ \boldsymbol{x}) \tag{12}$$

$$\boldsymbol{f}^t(\boldsymbol{x}) = \sum_{e=1}^{E} g_e^t(\boldsymbol{x})\, \boldsymbol{f}_e(\boldsymbol{x}), \tag{13}$$

ここで，$\boldsymbol{W}_e^t$ はタスク t 固有の重み行列である．

次に，ゲーティングネットワークから得られたタスク固有の特徴表現 $\boldsymbol{f}^t(\boldsymbol{x})$ は，M 個の隠れ層 $(m = 1, 2, \ldots, M)$ からなるタスク固有のタワーネットワークに入力される．タスク t のタワーネットワークにおいて，m 番目の隠れ層の出力 $\boldsymbol{H}_m^t$ は，線形変換と ReLU 活性化関数を用いて以下のように計算される．

$$\boldsymbol{H}_m^t = \begin{cases} \mathrm{ReLU}(\boldsymbol{W}_m^t\ \boldsymbol{f}^t(\boldsymbol{x}) + \boldsymbol{b}_1^t), & m = 1, \\ \mathrm{ReLU}(\boldsymbol{W}_m^t\ \boldsymbol{H}_{m-1}^t + \boldsymbol{b}_m^t), & m = 2, \ldots, M, \end{cases} \tag{14}$$

ここで，$\boldsymbol{W}_m^t$ と $\boldsymbol{b}_m^t$ はそれぞれ，タスク t の m 番目の隠れ層の重み行列とバイアスベクトルを表す．

最後に，タスク固有の予測 $\hat{y}^t$ は，最終層の出力 $\boldsymbol{H}_M^t$ にタスク依存の出力活性化関数 $\psi^t(\cdot)$ を適用して得られる．

$$\hat{y}^t = \psi^t\left(\boldsymbol{H}_M^t\right). \tag{15}$$

ここで，$\psi^t(\cdot)$ はタスク t の予測目的に対応する出力活性化関数を表す．例えば，二値分類にはシグモイド関数，回帰には恒等関数が用いられる．

MMoE では，各タスク t に対して，タスク固有の予測 $\hat{y}^t$ と真のラベル y^t との乖離度を測定する損失関数 $\mathcal{L}^{(t)}(\hat{y}^t, y^t)$ が定義される．損失関数 $\mathcal{L}^{(t)}(\hat{y}^t, y^t)$ は，分類にはバイナリークロスエントロピー (Binary Cross Entropy; BCE)，回帰には平均二乗誤差 (Mean Squared Error; MSE) など，タスクの予測目的に応じて設定される．MMoE 全体の損失は，全タスクの損失の加重和として以下のように定義される．

$$\mathcal{L}_{\mathrm{total}} = \sum_{t=1}^{T} \alpha_t\, \mathcal{L}^{(t)}(\hat{y}^t, y^t), \quad \text{s.t.} \quad \sum_{t=1}^{T} \alpha_t = 1,\ \alpha_t \geq 0. \tag{16}$$

ここで，α_t はタスク t における全体の損失に対する重要度を制御するハイパーパラメータである．

MMoE は，エキスパートネットワークが関連タスク間で共有される特徴表現を学習し，タスク固有のゲーティングネットワークが各タスクに対する各エキスパートの影響度を

決定する．ゲーティングネットワークによって生成されたタスク固有の特徴表現は，タワーネットワークによって処理され，タスク固有の最終的な予測値を算出する．その結果，MMoE は正誤反応と所要時間などのタスク間の相補的な関係を効果的に捉え，全体的な予測精度を向上させる．

3.2 正誤反応と所要時間を同時に予測する Multi-task Deep-IRT

MMoE は正答率や所要時間といったタスク間の相補的な関係を学習することで予測精度を向上させるが，解釈性が低いという課題がある．そこで，提案手法は，解釈性を高めるために MMoE のタワーネットワークに受検者ネットワークと項目ネットワークを組み込む．受検者ネットワークは受検者の能力および回答速度を推定し，項目ネットワークは項目の困難度および時間困難度を推定する．推定された受検者の能力および項目の困難度は正誤反応の予測に用いられ，回答速度および時間困難度は所要時間の予測に用いられる．これにより，提案手法は解釈可能なパラメータを維持しつつ，正誤反応と所要時間の予測精度を向上させる．

提案手法の概要を図 2 に示す．提案手法は，入力ベクトル $\boldsymbol{x}$ から 4 つのパラメータ $\hat{\theta}_i$, $\hat{\zeta}_i$, $\hat{\beta}_j$, $\hat{\lambda}_j$ を学習する．$\hat{\theta}_i$ と $\hat{\zeta}_i$ は，LNIRT における潜在変数に対応し，それぞれ受検者の能力および回答速度を表す．また，$\hat{\beta}_j$ と $\hat{\lambda}_j$ は項目パラメータに対応し，それぞれ項目の困難度および時間困難度を表す．

提案手法において，入力ベクトルは 2 つのベクトル $\boldsymbol{x} = \{\boldsymbol{s}^i, \boldsymbol{q}^j\}$ により構成される．ベクトル $\boldsymbol{s}^i = \{s_1^i, s_2^i, ..., s_I^i\}$ は受検者 i を表す one-hot ベクトルであり，$i' = i$ (対象受検者) の場合に $s_{i'}^i = 1$, それ以外の場合は $s_{i'}^i = 0$ となる．同様に，$\boldsymbol{q}^j = \{q_1^j, q_2^j, ..., q_J^j\}$ は項目 j を表す one-hot ベクトルである．

入力ベクトル $\boldsymbol{x}$ は，MMoE と同様の構造を持つエキスパートネットワークに入力される．具体的には，提案手法は E 個のエキスパートネットワーク $(e = 1, 2, \ldots, E)$ から構成され，各エキスパートネットワークは，関連タスク間の共有特徴表現を学習するために R 個の隠れ層 $(r = 1, 2, \ldots, R)$ を持つ多層フィードフォワードニューラルネットワークとして実装される．

e 番目のエキスパートネットワークにおける r 番目の隠れ層の出力 $\boldsymbol{L}_r^e$ は，線形変換と ReLU 活性化関数を用いて以下のように計算される．

$$\boldsymbol{L}_r^e = \begin{cases} \mathrm{ReLU}(\boldsymbol{W}_r^e \boldsymbol{x} + \boldsymbol{b}_r^e), & r = 1, \\ \mathrm{ReLU}(\boldsymbol{W}_r^e \boldsymbol{L}_{r-1}^e + \boldsymbol{b}_r^e), & r = 2, \ldots, R, \end{cases} \tag{17}$$

ここで，$\boldsymbol{W}_r^e$ と $\boldsymbol{b}_r^e$ はそれぞれ，e 番目のエキスパートネットワークにおける r 番目の隠

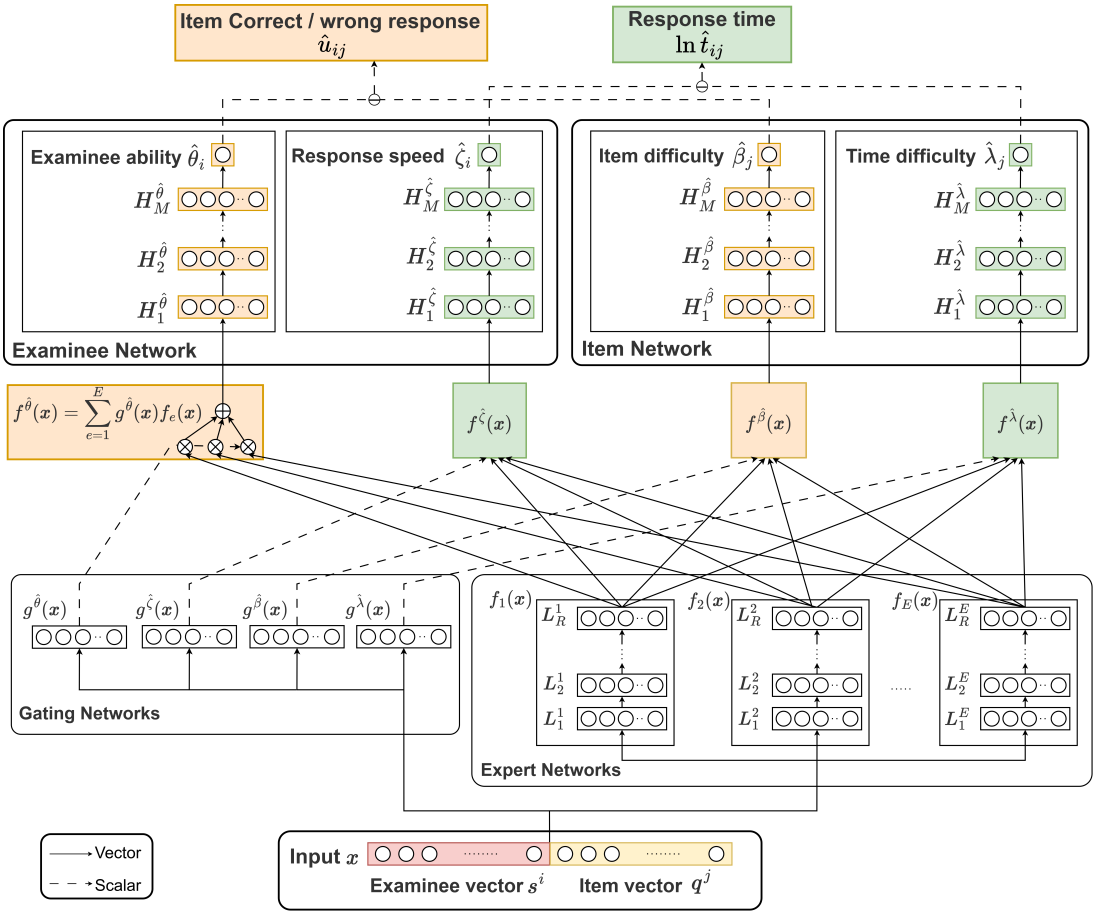


図 2: 提案手法の概要図

れ層の重み行列とバイアスベクトルを表す．さらに，各エキスパートにより出力される特徴表現は以下の通りとなる．

$$\boldsymbol{f}_e(\boldsymbol{x}) = \boldsymbol{L}_R^e. \tag{18}$$

提案手法は，解釈性を高めるために各パラメータに対応する個別のゲーティングネットワークを使用する．具体的には，受検者の能力 $\hat{\theta}$, 回答速度 $\hat{\zeta}$, 項目の困難度 $\hat{\beta}$, および時間困難度 $\hat{\lambda}$ に対応する各エキスパートの重みを出力するための 4 つのゲーティングネットワークが構成される．各ネットワークは，以下のようにソフトマックス活性化関数を用

いた単層フィードフォワードニューラルネットワークとして実装される．

$$g_e^{\hat{\theta}}(\boldsymbol{x}) = \text{softmax}(\boldsymbol{W}_e^{\hat{\theta}}\boldsymbol{x}), \tag{19}$$

$$g_e^{\hat{\zeta}}(\boldsymbol{x}) = \text{softmax}(\boldsymbol{W}_e^{\hat{\zeta}}\boldsymbol{x}), \tag{20}$$

$$g_e^{\hat{\beta}}(\boldsymbol{x}) = \text{softmax}(\boldsymbol{W}_e^{\hat{\beta}}\boldsymbol{x}), \tag{21}$$

$$g_e^{\hat{\lambda}}(\boldsymbol{x}) = \text{softmax}(\boldsymbol{W}_e^{\hat{\lambda}}\boldsymbol{x}), \tag{22}$$

ここで，$\boldsymbol{W}_e^{\hat{\theta}}$, $\boldsymbol{W}_e^{\hat{\zeta}}$, $\boldsymbol{W}_e^{\hat{\beta}}$, $\boldsymbol{W}_e^{\hat{\lambda}}$ は重み行列である．このようにして，各パラメータのゲーティングネットワークは，エキスパートの異なる混合パターンを学習でき，正誤反応と所要時間の関係を捉えることができる．

次に，各エキスパートの出力はゲーティングネットワークにより重み付けされ，各パラメータについて加重和が計算される．

$$\boldsymbol{f}^{\hat{\theta}}(\boldsymbol{x}) = \sum_{e=1}^{E} g_e^{\hat{\theta}}(\boldsymbol{x})\ \boldsymbol{f}_e(\boldsymbol{x}), \tag{23}$$

$$\boldsymbol{f}^{\hat{\zeta}}(\boldsymbol{x}) = \sum_{e=1}^{E} g_e^{\hat{\zeta}}(\boldsymbol{x})\ \boldsymbol{f}_e(\boldsymbol{x}), \tag{24}$$

$$\boldsymbol{f}^{\hat{\beta}}(\boldsymbol{x}) = \sum_{e=1}^{E} g_e^{\hat{\beta}}(\boldsymbol{x})\ \boldsymbol{f}_e(\boldsymbol{x}), \tag{25}$$

$$\boldsymbol{f}^{\hat{\lambda}}(\boldsymbol{x}) = \sum_{e=1}^{E} g_e^{\hat{\lambda}}(\boldsymbol{x})\ \boldsymbol{f}_e(\boldsymbol{x}). \tag{26}$$

重み付けされた特徴表現 $\boldsymbol{f}^{\hat{\theta}}, \boldsymbol{f}^{\hat{\zeta}}, \boldsymbol{f}^{\hat{\beta}}, \boldsymbol{f}^{\hat{\lambda}}$ は，対応する 4 つのフィードフォワードニューラルネットワークの入力として用いられる．$\boldsymbol{f}^{\hat{\theta}}$ と $\boldsymbol{f}^{\hat{\zeta}}$ が入力されるネットワークは受検者ネットワークを構成し，受検者の能力と回答速度を推定する．各ネットワークは M 層の隠れ層 ($m = \{1, 2, ..., M\}$) から構成され，受検者固有の潜在的な特徴を捉える．同様に，$\boldsymbol{f}^{\hat{\beta}}$ と $\boldsymbol{f}^{\hat{\lambda}}$ が入力されるネットワークは項目ネットワークを構成し，項目の困難度と時間困難度を推定する．

受検者ネットワークおよび項目ネットワークは以下のように定義される．

$$\boldsymbol{H}_m^{\hat{\theta}} = \begin{cases} \mathrm{ReLU}(\boldsymbol{W}_1^{\hat{\theta}}\, \boldsymbol{f}^{\hat{\theta}}(\boldsymbol{x}) + \boldsymbol{b}_1^{\hat{\theta}}), & m = 1, \\ \mathrm{ReLU}(\boldsymbol{W}_m^{\hat{\theta}}\, \boldsymbol{H}_{m-1}^{\hat{\theta}} + \boldsymbol{b}_m^{\hat{\theta}}), & m = 2, \ldots, M, \end{cases} \tag{27}$$

$$\boldsymbol{H}_m^{\hat{\zeta}} = \begin{cases} \mathrm{ReLU}(\boldsymbol{W}_1^{\hat{\zeta}}\, \boldsymbol{f}^{\hat{\zeta}}(\boldsymbol{x}) + \boldsymbol{b}_1^{\hat{\zeta}}), & m = 1, \\ \mathrm{ReLU}(\boldsymbol{W}_m^{\hat{\zeta}}\, \boldsymbol{H}_{m-1}^{\hat{\zeta}} + \boldsymbol{b}_m^{\hat{\zeta}}), & m = 2, \ldots, M, \end{cases} \tag{28}$$

$$\boldsymbol{H}_m^{\hat{\beta}} = \begin{cases} \mathrm{ReLU}(\boldsymbol{W}_1^{\hat{\beta}}\, \boldsymbol{f}^{\hat{\beta}}(\boldsymbol{x}) + \boldsymbol{b}_1^{\hat{\beta}}), & m = 1, \\ \mathrm{ReLU}(\boldsymbol{W}_m^{\hat{\beta}}\, \boldsymbol{H}_{m-1}^{\hat{\beta}} + \boldsymbol{b}_m^{\hat{\beta}}), & m = 2, \ldots, M, \end{cases} \tag{29}$$

$$\boldsymbol{H}_m^{\hat{\lambda}} = \begin{cases} \mathrm{ReLU}(\boldsymbol{W}_1^{\hat{\lambda}}\, \boldsymbol{f}^{\hat{\lambda}}(\boldsymbol{x}) + \boldsymbol{b}_1^{\hat{\lambda}}), & m = 1, \\ \mathrm{ReLU}(\boldsymbol{W}_m^{\hat{\lambda}}\, \boldsymbol{H}_{m-1}^{\hat{\lambda}} + \boldsymbol{b}_m^{\hat{\lambda}}), & m = 2, \ldots, M, \end{cases} \tag{30}$$

ここで，$\boldsymbol{W}_m^{\hat{\theta}}$, $\boldsymbol{W}_m^{\hat{\zeta}}$, $\boldsymbol{W}_m^{\hat{\beta}}$, および $\boldsymbol{W}_m^{\hat{\lambda}}$ はそれぞれ $\hat{\theta}$, $\hat{\zeta}$, $\hat{\beta}$, $\hat{\lambda}$ に対応するネットワークの m 番目の隠れ層の重み行列を表す．同様に，$\boldsymbol{b}_m^{\hat{\theta}}$, $\boldsymbol{b}_m^{\hat{\zeta}}$, $\boldsymbol{b}_m^{\hat{\beta}}$, および $\boldsymbol{b}_m^{\hat{\lambda}}$ は対応するバイアスベクトルを表す．

最後に，4 つの解釈可能なパラメータは以下のように推定される．

$$\hat{\theta}_i = \mathrm{ReLU}(\boldsymbol{W}^{\hat{\theta}} \boldsymbol{H}_M^{\hat{\theta}} + b^{\hat{\theta}}), \tag{31}$$

$$\hat{\zeta}_i = \mathrm{ReLU}(\boldsymbol{W}^{\hat{\zeta}} \boldsymbol{H}_M^{\hat{\zeta}} + b^{\hat{\zeta}}), \tag{32}$$

$$\hat{\beta}_j = \mathrm{ReLU}(\boldsymbol{W}^{\hat{\beta}} \boldsymbol{H}_M^{\hat{\beta}} + b^{\hat{\beta}}), \tag{33}$$

$$\hat{\lambda}_j = \mathrm{ReLU}(\boldsymbol{W}^{\hat{\lambda}} \boldsymbol{H}_M^{\hat{\lambda}} + b^{\hat{\lambda}}). \tag{34}$$

ここで，$\boldsymbol{W}^{\hat{\theta}}$, $\boldsymbol{W}^{\hat{\zeta}}$, $\boldsymbol{W}^{\hat{\beta}}$, および $\boldsymbol{W}^{\hat{\lambda}}$ はそれぞれ $\hat{\theta}$, $\hat{\zeta}$, $\hat{\beta}$, $\hat{\lambda}$ に対応する出力層の重みベクトルを表す．また，$b^{\hat{\theta}}$, $b^{\hat{\zeta}}$, $b^{\hat{\beta}}$, および $b^{\hat{\lambda}}$ は対応するバイアス項を表す．受検者ネットワークと項目ネットワークは，LNIRT モデルの解釈性を維持しつつ，受検者固有の潜在特性と項目固有のパラメータを同時に学習できる．

推定されたパラメータを用いて，提案手法は受検者 i の項目 j に対する正誤反応 $\hat{u}_{ij}$ と対数所要時間 $\ln \hat{t}_{ij}$ を予測する．提案手法における正誤反応の予測値は，Deep-IRT [10] と同様に以下のロジスティック定式化に従って導出される．

$$\hat{u}_{ij} = \frac{1}{1 + \exp\{-(\hat{\theta}_i - \hat{\beta}_j)\}} \tag{35}$$

次に，予測対数所要時間は以下のように定義される．

$$\ln \hat{t}_{ij} = -\hat{\zeta}_i + \hat{\lambda}_j \tag{36}$$

この定式化は，Becker ら [41] によって提案された事後期待値推定量に基づき，LNIRT における所要時間モデル (LNRT) から導出されている．

提案手法の全体の損失関数は，以下の正誤反応予測に対するバイナリークロスエントロピー (BCE) $\mathcal{L}^{(\mathrm{c/w})}$ と所要時間予測に対する平均二乗誤差 (MSE) $\mathcal{L}^{(\mathrm{rt})}$ の加重和として定義される．

$$\mathcal{L}^{(\mathrm{c/w})} = -\sum_{i=1}^{I} \sum_{j \in A_i} \Big\{ u_{ij} \ln \hat{u}_{ij} + (1 - u_{ij}) \ln(1 - \hat{u}_{ij}) \Big\}, \tag{37}$$

$$\mathcal{L}^{(\mathrm{rt})} = \frac{1}{\sum_i |A_i|} \sum_{i=1}^{I} \sum_{j \in A_i} \big(\ln t_{ij} - \ln \hat{t}_{ij} \big)^2, \tag{38}$$

$$\mathcal{L}_{\mathrm{total}} = \alpha \, \mathcal{L}^{(\mathrm{c/w})} + (1 - \alpha) \, \mathcal{L}^{(\mathrm{rt})}, \quad 0 \leq \alpha \leq 1. \tag{39}$$

ここで，u_{ij} と $\hat{u}_{ij}$ は受検者 i の項目 j への真の正誤反応と予測正誤反応を表し，$A_i \subseteq \{1, 2, \ldots, J\}$ は受検者 i が回答した項目の集合を表す．また，$\ln t_{ij}$ と $\ln \hat{t}_{ij}$ は受検者 i の項目 j への真の対数所要時間と予測対数所要時間を表す．α は，2 つの予測タスクのバランスを調整するためのハイパーパラメータである．

4　評価実験

本章では，実データを用いて提案手法と先行手法 (IRT，LNRT [31]，LNIRT [23]，Deep-IRT [10]，MMoE [32]) の予測精度を比較する．

4.1　データセット

本実験では，実データとして以下に示す CBT による試験データおよびオンライン学習システムによる学習ログデータを使用する．

1. UEC [42]: 電気通信大学で実施された CBT 形式の試験データ．
2. S-LME [43]: ある教育工学企業によって開発された独自の学習システムから収集されたデータセット．
3. Assistments: 教育データマイニング研究において広く用いられているデータセットであり，ASSIST2009 [44] および ASSIST2017 [45] から構成される．
4. Statics2011 [46]: Open Learning Initiative (OLI) プラットフォーム上の静力学コースから収集されたデータセット．
5. Slepemapy [47]: オンライン地理学習プラットフォームから収集された大規模な教育データセット．

表 1 は，各データセットにおける受検者数，項目数，正答率，および項目あたりの平均所要時間 (秒) を示す．なお，本実験では，信頼性の低い回答データを除外するために，前処理として以下の条件に該当する回答データを除外した．

- 正答率が 0 または 1 の項目に対する回答データ
- 所要時間が 3600 秒以上の回答データ
- IRT により予測された正誤反応と実際の正誤反応が異なり，かつ所要時間が該当項目における所要時間分布の上位 1% または下位 1% に属する回答データ

4.2　実験設定

本実験では，各データセットにおいて 5 分割交差検証を行い，提案手法と先行手法における所要時間および正誤反応の予測精度を比較した．予測性能を評価するために，所要時間予測には二乗平均平方根誤差 (RMSE)，平均絶対誤差 (MAE)，および決定係数 (R^2)

表 1: 各データセットの概要

データセット	受検者数	項目数	正答率	項目あたりの平均所要時間 (秒)
UEC	741	204	0.63	157.24
S-LME	6073	2671	0.89	378.85
ASSIST2017	1709	3162	0.40	29.75
ASSIST2009	8039	6651	0.59	57.95
Statics2011	333	300	0.76	26.11
Slepemapy	18563	2894	0.61	34.57

を，正誤反応予測には予測正解率 (ACC)，AUC スコア (AUC)，および F1 スコア (F1) の 6 つの指標を採用した．

IRT，LNRT，および LNIRT については，サンプリング数を 50000，バーンイン数を 10000 としてマルコフ連鎖モンテカルロ法 (MCMC) によりパラメータ推定を行った．ただし，パラメータ推定は 24 時間を上限として行い，24 時間経過した場合は取得したサンプルの先頭 20% をバーンインとして破棄した．IRT は Python の PyMC ライブラリ [48] を用いて実装し，LNRT と LNIRT は Fox ら [49, 50] による R パッケージを用いて推定した．

Deep-IRT は所要時間に関するパラメータをモデル化しておらず，所要時間を直接予測できない．そのため，本実験では受検者ネットワークが回答速度パラメータを学習し，項目ネットワークが時間困難度パラメータを学習する Deep-IRT の拡張版を構築した．これらのパラメータは，提案手法で用いられる所要時間の予測式 (36) に代入され，予測所要時間が算出される．なお，所要時間予測のための Deep-IRT では，損失関数には平均二乗誤差 (MSE) を用いて学習を行った．

Deep-IRT，MMoE および提案手法の学習では，各隠れ層のノード数を 16 とした．MMoE および提案手法におけるエキスパートネットワークおよびタワーネットワークの層数はそれぞれ 3 層とした ($R = M = 3$)．また，両手法で用いられる損失関数の重みパラメータ α とエキスパート数 E はグリッドサーチを行い，最適な値を決定した．具体的には，各データセットに対して，α は {0.1,0.2,0.3,0.4,0.5}，E は {1,2,3,4,5} から，検証データにおける全体損失 (式 (39)) が最小となる組み合わせを選択した．

正誤反応予測を行う全てのモデルにおいて，閾値として 0.5 を使用した．すなわち，正答の予測確率が 0.5 以上であれば正答 (1) と分類し，0.5 未満の場合は誤答 (0) と分類

した．

4.3 実験結果

表 2 は，全てのデータセットにおける所要時間予測の RMSE，MAE，および R^2 を示している．提案手法は，確率モデルである LNRT や LNIRT と比較して，より小さい RMSE と MAE，およびより大きい R^2 を示した．これらの確率モデルは，回答速度の潜在変数に対して固定された統計分布を仮定している．これに対し，提案手法はそのような分布の仮定に依存せず，観測データから直接回答速度を学習するため予測精度が向上したと考えられる．

また，提案手法は Deep-IRT よりも高い平均予測精度を示した．この結果は，正誤反応と所要時間の両タスクを相補的に学習することの利点を示している．具体的には，エキス

表 2: 各データセットに対する所要時間の予測精度

データセット	評価指標	LNRT	LNIRT	Deep-IRT	MMoE	Proposed
UEC	RMSE	141.2 (4.0)	140.9 (6.1)	107.2 (5.3)	107.2 (5.1)	**107.0** (**4.5**)
	MAE	83.7 (1.9)	83.4 (2.1)	**62.6** (**1.4**)	62.7 (1.3)	62.7 (1.3)
	R^2	−0.1 (0.1)	−0.1 (0.1)	**0.4** (**0.0**)	**0.4** (**0.0**)	**0.4** (**0.0**)
S-LME	RMSE	500.0 (0.0)	366.8 (2.4)	274.3 (2.9)	272.8 (2.7)	**270.3** (**3.4**)
	MAE	282.6 (4.6)	192.2 (0.7)	118.9 (0.8)	118.4 (0.8)	**117.7** (**0.8**)
	R^2	−33.5 (33.3)	−0.4 (0.0)	0.2 (0.0)	0.2 (0.0)	**0.3** (**0.0**)
ASSIST2009	RMSE	500.0 (0.0)	127.7 (25.1)	87.5 (1.3)	87.2 (1.2)	**86.9** (**3.3**)
	MAE	84.9 (1.3)	47.5 (0.1)	29.5 (0.3)	29.5 (0.3)	**29.4** (**0.2**)
	R^2	−65.8 (20.5)	−1.0 (0.8)	**0.1** (**0.0**)	**0.1** (**0.0**)	**0.1** (**0.0**)
ASSIST2017	RMSE	500.0 (0.0)	500.0 (0.0)	72.3 (1.0)	**72.2** (**1.0**)	72.9 (1.1)
	MAE	389.7 (140.1)	500.0 (0.0)	34.2 (0.2)	**34.0** (**0.2**)	**34.0** (**0.2**)
	R^2	−100.0 (0.0)	−100.0 (0.0)	**0.2** (**0.0**)	**0.2** (**0.0**)	**0.2** (**0.0**)
Statics2011	RMSE	121.5 (11.4)	90.1 (41.4)	60.5 (3.0)	60.4 (3.0)	**60.2** (**2.8**)
	MAE	35.2 (0.9)	26.0 (1.6)	19.7 (0.6)	19.7 (0.6)	**19.6** (**0.6**)
	R^2	−2.5 (1.1)	−1.2 (2.1)	0.2 (0.0)	**0.2** (**0.0**)	**0.2** (**0.0**)
Slepemapy	RMSE	64.5 (1.5)	61.2 (1.1)	**59.5** (**1.1**)	**59.5** (**1.1**)	**59.5** (**1.1**)
	MAE	13.7 (0.1)	13.0 (0.1)	**8.1** (**0.1**)	**8.1** (**0.1**)	8.2 (0.2)
	R^2	−0.2 (0.0)	−0.1 (0.0)	**0.0** (**0.0**)	**0.0** (**0.0**)	**0.0** (**0.0**)
Average	RMSE	304.5	214.5	110.2	109.9	**109.5**
	MAE	148.3	143.7	45.5	45.4	**45.3**
	R^2	−33.7	−17.1	**0.2**	**0.2**	**0.2**

表 3: 各データセットに対する正誤反応の予測精度

データセット	評価指標	IRT	LNIRT	Deep-IRT	MMoE	Proposed
UEC	ACC	74.76 (0.39)	75.34 (1.08)	78.31 (0.46)	78.77 (0.53)	**78.80** (**0.55**)
	AUC	79.32 (0.45)	77.02 (1.11)	83.88 (0.39)	84.31 (0.27)	**84.35** (**0.23**)
	F1	80.77 (0.46)	72.26 (0.95)	83.60 (0.44)	83.81 (0.57)	**83.87** (**0.59**)
S-LME	ACC	85.68 (0.21)	89.95 (0.05)	90.32 (0.09)	90.39 (0.12)	**90.51** (**0.07**)
	AUC	66.10 (0.71)	83.43 (0.32)	87.78 (0.05)	88.02 (0.07)	**88.27** (**0.16**)
	F1	92.13 (0.12)	94.52 (0.03)	94.71 (0.05)	94.75 (0.07)	**94.82** (**0.04**)
ASSIST2009	ACC	65.98 (0.29)	73.48 (0.20)	76.18 (0.21)	76.33 (0.20)	**76.35** (**0.13**)
	AUC	63.01 (0.53)	77.06 (0.26)	80.67 (0.12)	80.88 (0.13)	**80.90** (**0.10**)
	F1	76.40 (0.24)	81.06 (0.21)	83.01 (0.22)	83.24 (0.16)	**83.25** (**0.10**)
ASSIST2017	ACC	**70.48** (**0.22**)	62.21 (0.35)	70.22 (0.24)	70.30 (0.20)	70.35 (0.17)
	AUC	**77.50** (**0.09**)	66.44 (0.47)	77.07 (0.07)	77.25 (0.05)	77.26 (0.06)
	F1	72.67 (0.20)	63.01 (0.36)	**73.34** (**0.33**)	73.29 (0.25)	72.02 (0.22)
Statics2011	ACC	**76.76** (**0.26**)	74.54 (0.15)	74.39 (0.60)	76.24 (0.37)	76.51 (0.35)
	AUC	**84.30** (**0.32**)	80.36 (0.18)	81.48 (0.47)	83.69 (0.42)	84.07 (0.46)
	F1	73.71 (0.35)	71.39 (0.55)	70.94 (1.21)	73.44 (0.52)	**73.76** (**0.44**)
Slepemapy	ACC	72.73 (0.63)	61.44 (0.68)	72.97 (0.08)	72.94 (0.06)	**73.08** (**0.06**)
	AUC	68.85 (1.99)	55.78 (0.22)	69.92 (0.18)	69.97 (0.20)	**70.19** (**0.20**)
	F1	**83.10** (**0.60**)	72.62 (0.75)	82.97 (0.05)	82.94 (0.05)	**83.10** (**0.08**)
Average	ACC	75.49	73.44	77.99	78.35	**78.45**
	AUC	73.17	71.29	79.34	79.82	**79.97**
	F1	81.22	77.32	82.77	**83.18**	83.09

パートネットワークを通じて正誤反応と所要時間の特徴を同時に学習することで，提案手法は両タスクの予測性能を向上させる．

最後に，提案手法は MMoE モデルよりも優れた性能を示した．その理由は，LNIRT に関する先行研究 [23] で示唆されているように，MMoE が回答速度のような解釈可能な潜在特性を明示的に推定しないためである．このことが，予測精度の低下につながっていると考えられる．対照的に，提案手法は解釈可能な受検者ネットワークと項目ネットワークを通じて，データから重要な潜在特性を推定する．これらのネットワークは予測精度を向上させるだけでなく，e-testing において重要となる教育的な解釈も支援可能である．

表 3 は，全てのデータセットにおける項目の正誤反応予測の ACC，AUC，および F1 を示している．

提案手法は，多くのデータセットにおいて IRT および LNIRT を上回り，より高い Accuracy，AUC，および F1 スコアを達成した．正誤反応予測の結果は，表 2 で見られ

た傾向と一致している．IRT と LNIRT が潜在特性に対する分布の仮定に依存しているのに対し，提案手法はデータから直接これらの特性を学習するため，受検者と項目の相互作用をより正確にモデル化することが可能になる．

さらに，提案手法は，Deep-IRT よりも高い平均予測精度を達成した．特定のデータセットにおいては，Deep-IRT が提案手法よりもわずかに高い予測精度を示すこともあったが，その差は小さかった．これらの結果は，項目の正誤反応と所要時間を同時に学習することが予測精度を向上させることを示している．

最後に，提案手法は MMoE よりも多くの指標において優れた性能を示した．その要因は，所要時間予測の場合と同様に，MMoE が受検者の能力のような解釈可能な潜在特性を明示的に推定しないためである．提案手法は，解釈可能な受検者ネットワークと項目ネットワークを通じてデータから重要な潜在特性を推定することで，予測精度を向上させた．

4.4 受検者ネットワークおよび項目ネットワークの解釈性分析

提案手法は，受検者ネットワークおよび項目ネットワークを導入することで， LNIRT の潜在変数を解釈可能な形で推定する．この解釈性を評価するために，LNIRT および提案手法により推定されたパラメータ間の相関を分析した．具体的には，複数のデータセットにおいて，LNIRT と提案手法により推定された受検者の能力および回答速度，項目の困難度および時間困難度について，Pearson の積率相関係数と Spearman の順位相関係数を算出した．その結果を表 4 に示す．

表 4 より，項目ネットワークから推定された項目の困難度パラメータは，多くのデータセットで LNIRT による推定値と高い正の相関を示した．また，時間困難度パラメータについても，一部のデータセットで高い正の相関を示した．これらの結果は，提案手法が，特に困難度に関して，LNIRT と概ね整合的な項目パラメータの解釈性を提供していることを示唆している．

一方で，受検者ネットワークから推定された受検者の能力および回答速度パラメータは，LNIRT による推定値と相関が低い傾向が見られた．これは，LNIRT が能力や回答速度といった受検者の潜在変数に対して統計的分布を仮定するのに対し，提案手法はそのような仮定に依存せず，観測データからこれらの変数を直接学習するためであると考えられる．この仮定の影響を分析するために，本節では，提案手法と LNIRT により推定された受検者の能力および回答速度パラメータの分布を比較する．

図 3 と図 4 は，ASSIST2009 および Slepemapy データセットにおいて，提案手法と LNIRT により推定された受検者の能力および回答速度パラメータのヒストグラムを示す．

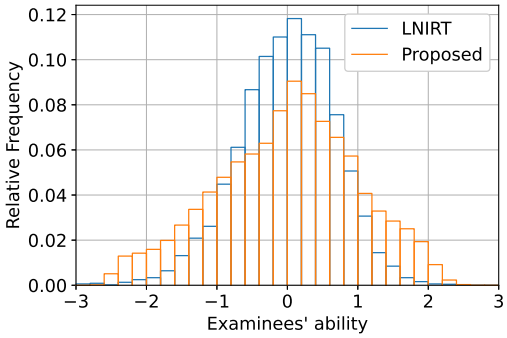
表 4: 提案手法と LNIRT のパラメータ推定値に対する Pearson 相関係数および Spearman 相関係数

データセット	評価指標	能力	困難度	回答速度	時間困難度
UEC	Pearson	0.3461	0.7550	0.1123	0.2061
	Spearman	0.3294	0.7667	0.1098	0.2516
S-LME	Pearson	0.3751	0.5733	0.1823	0.0861
	Spearman	0.4060	0.5934	0.2301	0.0750
ASSIST2009	Pearson	0.5261	0.3427	0.2509	0.4100
	Spearman	0.5321	0.4237	0.2999	0.4071
ASSIST2017	Pearson	0.0094	0.6123	−0.0156	0.3893
	Spearman	0.0081	0.5821	−0.0087	0.3560
Statics2011	Pearson	0.0057	0.8853	0.0237	0.6193
	Spearman	0.0004	0.9042	0.0197	0.6347
Slepemapy	Pearson	0.0097	0.4925	0.0131	0.2269
	Spearman	0.0117	0.4762	0.0115	0.2121

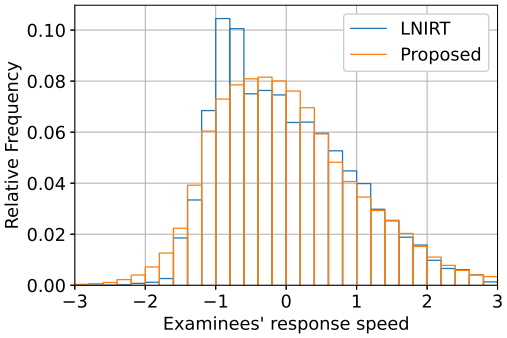
横軸は推定値を，縦軸は相対度数を表す．なお，提案手法の各パラメータの推定値は平均 0，分散 1 の分布に標準化した．

図 3 より，このデータセットでは，LNIRT と提案手法の両方が，能力および回答速度の分布を標準正規分布に類似した形状として推定していることが確認できる．さらに，この結果と合わせて，提案手法は ASSIST2009 データセットにおいて，受検者の能力および回答速度で比較的高い相関を示している (表 4)．この結果は，真の能力分布が標準正規分布に従う場合，Deep-IRT が単峰性の分布を適切に推定できると報告した Tsutsumi ら [9] の結果と整合的である．したがって，ASSIST2009 データセットにおける高い相関は，両手法が類似したパラメータ分布を推定したことに起因すると解釈できる．

一方で，ASSIST2009 データセットとは対照的に，Slepemapy データセット (図 4) では分布形状に顕著な差異が見られた．LNIRT はモデルの制約上，単峰性の分布を推定するのに対し，提案手法は能力および回答速度の多峰性を示唆する分布を捉えている．さらに，この結果と合わせて，提案手法は Slepemapy データセットにおいて，受検者の能力および回答速度で比較的低い相関を示している (表 4)．これは，提案手法が正規分布の仮定に縛られず，データに内在する複雑な構造をより柔軟に学習できている可能性を示して

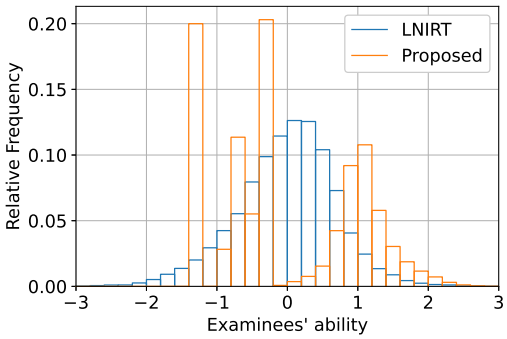


(a) 受検者の能力推定値

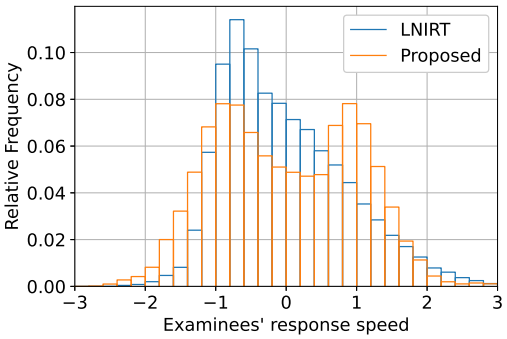


(b) 受検者の回答速度推定値

図 3: LNIRT および提案手法による受検者の能力および回答速度推定値のヒストグラム (ASSIST2009)



(a) 受検者の能力推定値



(b) 受検者の回答速度推定値

図 4: LNIRT および提案手法による受検者の能力および回答速度推定値のヒストグラム (Slepemapy)

いる．その結果，提案手法により推定される柔軟な分布は，未知項目に対する項目正誤反応および項目反応時間のより正確な予測に寄与している可能性がある．なお，推定された受検者の能力は，その分布が潜在尺度に固定されないため，IRT スコアとして用いる際には注意が必要である．

4.5 ガンマ分布を用いた所要時間分布の分析

本節では，データセットに依存しない予測精度の傾向を調査するために，所要時間の分布が予測精度にどのように影響するかを分析する．具体的には，Ueno ら [51] の方法に基づき，各項目の所要時間分布をガンマ分布により近似し，推定されたパラメータに基づいて項目を 4 つのタイプに分類する．次に，各タイプの項目に対する所要時間予測精度を比較し，所要時間分布が予測精度に与える影響を評価する．

4.5.1 ガンマ分布モデル

Ueno ら [51] では，エントロピー最大化原理により，単純な思考プロセスの所要時間が従う分布として以下の指数分布が導出される．

$$f(t) = \frac{1}{\tau} \cdot \exp\left(-\frac{t}{\tau}\right) \tag{40}$$

なお，τ は項目の実質平均所要時間を表し，最小時間 t_0 と平均所要時間 $\bar{t}$ を用いて $\tau = \bar{t} - t_0$ と定義される．

ここで，回答プロセスは，この単純な思考プロセスが α_γ 回繰り返されることで構成されると仮定する．これより，式 (40) の分布の α_γ 回の畳み込み積分を行う．これにより，項目の所要時間分布として以下のガンマ分布が得られる．

$$f(t) = \frac{t^{\alpha_\gamma - 1} \exp(-\frac{t}{\beta_\gamma})}{\beta_\gamma^{\alpha_\gamma} (\alpha_\gamma - 1)!} \tag{41}$$

ただし，$\alpha_\gamma \beta_\gamma = \tau$ である．また，モデルに含まれる二つのパラメータ α_γ および β_γ は，それぞれ以下のように解釈できる．

- α_γ：項目が要求する受検者の思考の「複合度」
- β_γ：単純な思考プロセスに要する所要時間

なお，これら 2 つのパラメータは，モーメント法 (例えば，[52, 53]) により以下のように推定される．

$$\hat{\alpha}_\gamma = \frac{\hat{\tau}^2}{\hat{\sigma}^2}, \quad \hat{\beta}_\gamma = \frac{\hat{\sigma}^2}{\hat{\tau}} \tag{42}$$

ここで，$\hat{\alpha}_\gamma, \hat{\beta}_\gamma, \hat{\tau}$ はパラメータ $\alpha_\gamma, \beta_\gamma, \tau$ の推定値を表し，$\hat{\sigma}^2$ は所要時間データの分散推定値を表す．

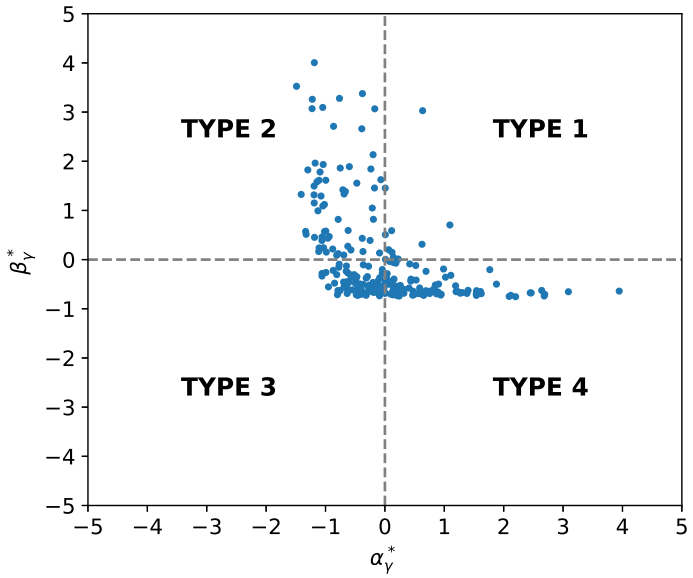


図 5: Statics2011 におけるガンマ分布のパラメータ α^*_γ および β^*_γ に基づく 4 タイプへの項目分類

4.5.2 項目のガンマ分布パラメータに基づく分類

図 5 は，Statics2011 データセットの全項目に対してガンマ分布のパラメータ推定値 $\hat{\alpha}_\gamma$ および $\hat{\beta}_\gamma$ を標準化し (それぞれ $\alpha^*_\gamma, \beta^*_\gamma$ と示す)，各項目ごとにパラメータ推定値をプロットした図である．Ueno ら [51] によると，これらのパラメータ値に基づいて，項目は以下の 4 つのタイプに分類できる：

- **TYPE 1:** $\alpha^*_\gamma > 0$ かつ $\beta^*_\gamma > 0$ (第一象限)．比較的難易度の高い思考プロセスを多数回繰り返すことで回答される項目であり，所要時間の平均と分散が大きいことを示す．
- **TYPE 2:** $\alpha^*_\gamma < 0$ かつ $\beta^*_\gamma > 0$ (第二象限)．比較的難しい思考プロセスを数回繰り返すことで回答される項目であり，平均と分散が中程度であることを示す．
- **TYPE 3:** $\alpha^*_\gamma < 0$ かつ $\beta^*_\gamma < 0$ (第三象限)．比較的簡単な思考プロセスを数回繰り返すことで回答される項目であり，平均と分散が小さいことを示す．
- **TYPE 4:** $\alpha^*_\gamma > 0$ かつ $\beta^*_\gamma < 0$ (第四象限)．比較的簡単な思考プロセスを多数回繰り返すことで回答される項目であり，平均と分散が中程度であることを示す．

4 つのタイプ間での所要時間分布の形状の違いを説明するために，図 6 に代表的な項目

の経験的累積分布とガンマ累積分布を比較した結果を示す．ガンマ累積分布は，以下の式で表される．

$$F(t) = \begin{cases} 0 & (t < t_0) \\ \int_{t_0}^{t} f(t)dt & (t \geq t_0) \end{cases} \tag{43}$$

各プロットにおいて，横軸は秒単位の実際の所要時間を表し，縦軸は 0 から 1 の範囲の累積確率を表す．階段状の線はデータから直接導出された経験的累積分布関数を示し，滑らかな曲線は当てはめられたガンマ累積分布関数を示す．これらの累積分布はよく一致しており，ガンマ分布が各項目タイプの所要時間分布の適切な近似を提供していることを示している．

図 6 において，**TYPE 1** の項目は，より長くばらつきのある所要時間と整合する右に厚い裾を持つ分布を示しているのに対し，**TYPE 3** の項目は，より短くばらつきの少ない所要時間を反映して，原点付近で急激な立ち上がりを見せている．**TYPE 2** および **TYPE 4** の項目はこれらの中間に位置しており，提案された分類が所要時間分布の形状における体系的な違いを捉えていることを示している．

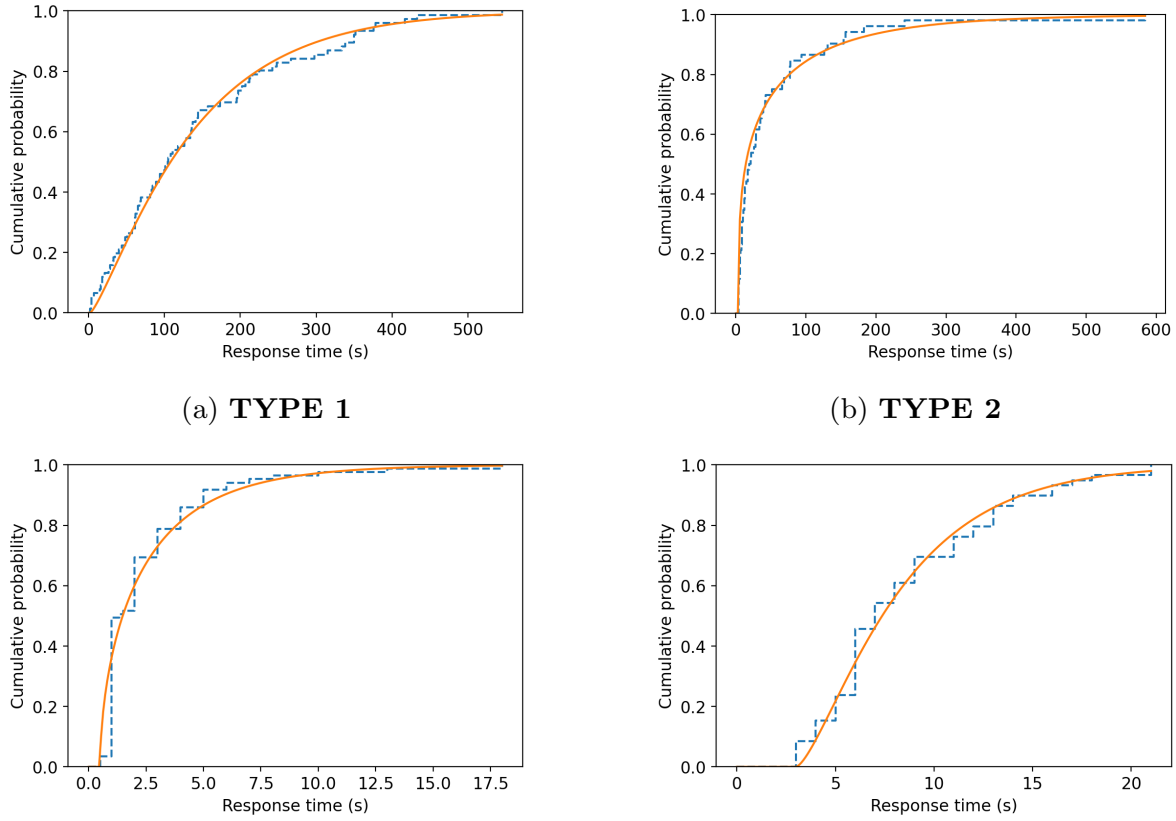
4.5.3 各タイプに基づく予測精度の比較

表 5 は，ガンマ分布のパラメータに基づく項目分類に基づき，各データセットにおける項目タイプごとの正誤反応 (ACC，AUC，および F1) と所要時間 (RMSE，MAE，および R^2) の予測精度を示す．

表 5 より，提案手法は全ての項目タイプにおいて LNIRT よりも所要時間の予測精度が改善した．特に，パラメータ β_γ の値が大きく，所要時間の平均と分散が大きい **TYPE 1** および **TYPE 2** の項目において，改善が顕著であった．これは，提案手法が受検者の回答速度の分布をより柔軟に捉え，正誤反応と所要時間の関係をデータから直接学習できる点が，所要時間の変動性が高い項目に対する予測精度の向上に寄与したと考えられる．

対照的に，所要時間の変動性が低い **TYPE 3** および **TYPE 4** の項目については，提案手法と LNIRT における所要時間の予測精度の差は相対的に小さいが，提案手法は一貫して LNIRT を上回った．

所要時間の予測精度は，提案手法と LNIRT ともに **TYPE 1** と **TYPE 2** で相対的に低く，**TYPE 3** と **TYPE 4** で高い傾向が見られた．また，**TYPE 2** は **TYPE 1** より，**TYPE 3** は **TYPE 4** よりも予測精度が高かった．この結果は，所要時間分布の立ち上がりが早い (α_γ が小さい) 項目ほど受検者の所要時間が最小所要時間付近に集中し，予測誤差が小さくなりやすいことを示唆している．以上より，β_γ が大きい項目では予測が難しくなる傾向がある一方，α_γ が小さい項目では予測が相対的に容易であるという傾



(a) **TYPE 1**

(b) **TYPE 2**

(c) **TYPE 3**

(d) **TYPE 4**

図 6: Statics2011 における 4 つの項目タイプに対する所要時間の経験的累積分布とガンマ分布累積分布の比較

向が示された．

正誤反応でも，提案手法は全ての項目タイプで LNIRT よりも予測精度を改善した．特に，複数の評価指標で提案手法と LNIRT ともに **TYPE 1** と **TYPE 4** の予測精度は低く，**TYPE 2** と **TYPE 3** は高い傾向が見られた．これは，項目タイプの α_γ の大きさが正誤反応予測精度に影響していると考えられる．α_γ が小さい項目は，思考プロセスの複合度が低く，知識の有無や基礎能力が正誤に直結しやすい．そのため，反応確率の推定が比較的単純化され，結果として予測精度が高くなった可能性がある．一方で，α_γ が大きい項目は，回答に必要な思考プロセスが比較的多く，同じ正答に至るまでの過程が多様化しやすい．その結果，正誤を受検者の能力や項目の困難度のみでは説明しにくく，予測精度の低下につながったと解釈できる．したがって，α_γ が小さい項目タイプでは反応予測精度が高くなる一方，α_γ が大きい項目タイプでは，潜在特性以外の要因による影響が

表 5: ガンマ分布パラメータタイプごとの各データセットにおける予測精度

データセット	タスク	評価指標	LNIRT				Proposed			
			TYPE 1	**TYPE 2**	**TYPE 3**	**TYPE 4**	**TYPE 1**	**TYPE 2**	**TYPE 3**	**TYPE 4**
UEC	所要時間	RMSE	245.00	171.90	52.82	80.30	**176.11**	**134.73**	**44.59**	**67.14**
		MAE	168.66	114.57	36.68	56.25	**127.56**	**81.56**	**29.01**	**44.31**
		R^2	−0.93	−0.33	0.10	0.05	**0.00**	**0.18**	**0.36**	**0.33**
	正誤反応	ACC	63.71	74.07	83.62	74.53	**67.83**	**78.54**	**86.00**	**77.91**
		AUC	68.03	79.73	72.97	73.00	**74.43**	**85.09**	**81.46**	**80.42**
		F1	59.25	76.37	90.25	82.23	**59.31**	**80.16**	**91.75**	**84.49**
S-LME		RMSE	500.00	308.26	160.76	258.80	**442.90**	**265.18**	**131.57**	**177.63**
		MAE	451.77	147.70	81.31	159.27	**266.43**	**99.93**	**51.61**	**89.60**
		R^2	−0.96	−0.30	−0.34	−0.61	**0.08**	**0.04**	**0.10**	**0.24**
	正誤反応	ACC	84.55	92.14	91.23	89.99	**86.10**	**92.41**	**91.63**	**90.31**
		AUC	83.86	83.04	80.33	83.80	**87.66**	**88.64**	**86.50**	**88.04**
		F1	90.97	95.81	95.32	94.53	**91.80**	**95.95**	**95.54**	**94.71**
ASSIST2009	所要時間	RMSE	200.71	209.54	55.82	97.39	**137.97**	**143.92**	**42.17**	**50.67**
		MAE	135.07	66.24	29.46	49.11	**83.03**	**45.11**	**18.15**	**26.82**
		R^2	−0.96	−1.01	−0.47	−1.74	**0.07**	**0.05**	**0.16**	**0.26**
	正誤反応	ACC	69.64	73.59	75.90	70.35	**72.79**	**76.19**	**78.46**	**73.87**
		AUC	76.78	79.12	76.54	73.99	**80.24**	**82.25**	**80.55**	**78.63**
		F1	69.27	79.75	84.26	77.84	**72.22**	**81.77**	**86.04**	**80.78**
ASSIST2017	所要時間	RMSE	176.65	500.00	500.00	500.00	**107.77**	**101.90**	**44.70**	**52.88**
		MAE	121.16	500.00	500.00	500.00	**64.14**	**43.88**	**21.80**	**31.07**
		R^2	−1.52	−100.00	−100.00	−100.00	**0.06**	**0.06**	**0.12**	**0.18**
	正誤反応	ACC	59.21	61.81	64.77	60.84	**69.08**	**69.96**	**71.58**	**69.85**
		AUC	62.43	65.75	68.75	64.55	**75.57**	**76.92**	**78.17**	**76.50**
		F1	45.37	60.56	70.13	59.92	**63.96**	**70.53**	**75.91**	**70.57**
Statics2011	所要時間	RMSE	150.20	182.94	20.67	25.26	**121.53**	**105.97**	**17.32**	**18.54**
		MAE	86.45	55.10	10.44	14.12	**64.16**	**43.72**	**7.85**	**9.45**
		R^2	−0.49	−1.80	−0.30	−0.40	**0.02**	**0.06**	**0.09**	**0.24**
	正誤反応	ACC	92.14	80.00	68.61	74.19	92.14	**81.74**	**70.92**	**76.22**
		AUC	53.00	70.78	74.29	79.23	**71.97**	**79.70**	**78.08**	**82.47**
		F1	**32.10**	35.93	72.29	77.42	28.57	**42.87**	**74.59**	**79.16**
Slepemapy	所要時間	RMSE	–	106.32	30.96	18.65	–	**104.84**	**28.25**	**11.71**
		MAE	–	16.83	11.75	10.89	–	**12.00**	**7.05**	**5.85**
		R^2	–	−0.03	−0.19	−1.36	–	**0.00**	**0.01**	**0.08**
	正誤反応	ACC	–	62.73	62.36	58.48	–	**74.08**	**73.68**	**71.00**
		AUC	–	55.45	55.64	55.83	–	**69.99**	**70.18**	**69.96**
		F1	–	74.24	73.78	68.55	–	**84.05**	**83.65**	**81.02**
Average	所要時間	RMSE	251.04	225.85	126.02	155.03	**195.58**	**135.68**	**52.70**	**65.99**
		MAE	181.84	132.21	98.79	120.73	**114.06**	**49.04**	**21.29**	**34.33**
		R^2	−0.90	−14.82	−14.48	−14.97	**0.04**	**0.06**	**0.13**	**0.20**
	正誤反応	ACC	70.06	74.72	74.97	69.20	**76.34**	**79.89**	**79.37**	**75.90**
		AUC	66.38	70.08	69.69	69.30	**76.12**	**79.43**	**78.50**	**77.95**
		F1	59.24	72.90	81.90	75.06	**66.06**	**78.27**	**85.44**	**81.80**

相対的に増え，予測精度が低下しやすいという傾向が示された．

本節では，α_γ (思考プロセスの複合度) と β_γ (単一思考プロセスの所要時間) に基づく項目分類を用いて予測精度を分析した．その結果，提案手法は正誤反応および所要時間の両方で一貫して LNIRT よりも高い予測精度を示した．特に，所要時間予測では，β_γ が大きい項目タイプにおいて改善が顕著であった．また，正誤反応予測では，α_γ が小さい項目タイプで高精度となる傾向が見られた．以上より，実運用においては，β_γ が大きい項目タイプを多く含むデータセットに対して提案手法による利得が大きいことが示唆される．さらに，項目タイプ依存の性能差が見られたことから，提案手法に項目タイプを考慮した機構を導入することも今後の課題として挙げられる．例えば，提案手法が採用するパラメータ固有のゲーティングネットワークの入力に項目タイプを加えることで，項目タイプも考慮したエキスパート選択を可能にし，さらなる性能向上につながる可能性がある．

5　むすび

本研究は，MMoE に基づき，項目および受検者の解釈可能なパラメータを学習し，正誤反応と所要時間を同時に予測する新たな Multi-task Deep-IRT を提案した．提案手法は，受検者ネットワークおよび項目ネットワークを導入することで，受検者の能力および回答速度，項目の困難度および時間困難度といったパラメータを解釈可能な形でデータから直接学習できる．

数値実験の結果，提案手法は正誤反応予測の精度を低下させることなく，従来手法と比較して所要時間の予測精度を向上させることを示した．さらに，提案手法は，LNIRT と同様の項目パラメータの解釈性を維持しつつ，受検者の能力および回答速度に対しては，より複雑な分布構造を捉えることができることが示唆された．

所要時間予測は，カンニングなどの異常行動の検知 [14–16]，アダプティブ・ラーニングへの応用 [22]，受検者モデリングの改善 [18, 19]，および時間制約を組み込んだ CAT 手法への導入 [20, 21] など，幅広く利用されている．特に，自動平行テスト構成手法 [6, 7] を用いた最先端の CAT 手法 [54, 55] は，一般的に回答精度を重視しており，所要時間については明示的に扱っていない．今後の展望として，提案手法をこうした手法に組み込むことが挙げられる．これにより，受検者の能力と回答速度の同時モデリングが可能となり，テストの効率性と公平性を向上させられる可能性がある．

謝辞

本研究を進めるにあたり，丁寧なご指導，ご鞭撻を賜りました指導教員の植野真臣教授に深く感謝の意を表します．そして，ゼミや日常の議論を通じて多くの示唆や知識をいただいた渕本壱真准教授にも心より御礼申し上げます．最後に，日頃から研究生活を支えてくださった研究室の皆様に感謝申し上げます．

参考文献

[1] Koun-Tem Sun, Yu-Jen Chen, Shu-Yen Tsai, and Chien-Fen Cheng. Creating IRT-based parallel test forms using the genetic algorithm method. *Applied Measurement in Education*, 2(21):141–161, 2008.

[2] Pokpong Songmuang and Maomi Ueno. Bees algorithm for construction of multiple test forms in e-testing. *IEEE Transactions on Learning Technologies*, 4:209–221, 2011.

[3] Takatoshi Ishii, Pokpong Songmuang, and Maomi Ueno. Maximum clique algorithm for uniform test forms. *The 16th International Conference on Artificial Intelligence in Education*, pages 451–462, 2013.

[4] Takatoshi Ishii, Pokpong Songmuang, and Maomi Ueno. Maximum clique algorithm and its approximation for uniform test form assembly. *IEEE Trans. Learn. Technol.*, 7(1):83–95, 2014.

[5] Takatoshi Ishii and Maomi Ueno. Algorithm for uniform test assembly using a maximum clique problem and integer programming. In *Artificial Intelligence in Education*, pages 102–112. Springer International Publishing, 2017.

[6] Kazuma Fuchimoto, Takatoshi Ishii, and Maomi Ueno. Hybrid maximum clique algorithm using parallel integer programming for uniform test assembly. *IEEE Transactions on Learning Technologies*, 15(2):252–264, 2022.

[7] Kazuma Fuchimoto, Shin-ichi Minato, and Maomi Ueno. Automated test assembly using zero-suppressed binary decision diagrams. *IEEE Access*, 2023.

[8] Chun-Kit Yeung. Deep-irt: Make deep learning based knowledge tracing explainable using item response theory. *Proceedings of the 12th International Conference on Educational Data Mining*, 2019.

[9] Emiko Tsutsumi, Ryo Kinoshita, and Maomi Ueno. Deep item response theory as a novel test theory based on deep learning. *Electronics*, 10(9):1020, April 2021.

[10] Emiko Tsutsumi, Ryo Kinoshita, and Maomi Ueno. Deep-irt with independent student and item networks. *International Educational Data Mining Society*, 2021.

[11] Emiko Tsutsumi, Yiming Guo, and Maomi Ueno. Deepirt with a hypernetwork to optimize the degree of forgetting of past data. In *Proceedings of the 15th International Conference on Educational Data Mining*, page 543, 2022.

[12] Emiko Tsutsumi, Yiming Guo, Ryo Kinoshita, and Maomi Ueno. Deep knowledge tracing incorporating a hypernetwork with independent student and item networks. *IEEE Transactions on Learning Technologies*, 2023.

[13] Emiko Tsutsumi, Tetsurou Nishio, and Maomi Ueno. Deep-irt with a temporal convolutional network for reflecting students' long-term history of ability data. In *International Conference on Artificial Intelligence in Education*, pages 250–264. Springer, 2024.

[14] Chun Wang, Gongjun Xu, Zhuoran Shang, and Nathan Kuncel. Detecting aberrant behavior and item preknowledge: A comparison of mixture modeling method and residual method. *J. Educ. Behav. Stat.*, 43(4):469–501, August 2018.

[15] Jose A Ruiperez-Valiente, Pedro J Munoz-Merino, Giora Alexandron, and David E Pritchard. Using machine learning to detect 'multiple-account' cheating and analyze the influence of student and problem features. *IEEE Trans. Learn. Technol.*, 12(1):112–122, January 2019.

[16] Kaiwen Man and Jeffrey R Harring. Assessing preknowledge cheating via innovative measures: A multiple-group analysis of jointly modeling item responses, response times, and visual fixation counts. *Educ. Psychol. Meas.*, 81(3):441–465, June 2021.

[17] Chansoon Lee, Kylie Gorney, and Jianshen Chen. Using item scores and response times to detect item compromise in computerized adaptive testing. *Educational and Psychological Measurement*, 0(0):00131644251368335, 2025.

[18] Wei Chu and Philip I Pavlik Jr. The predictiveness of pfa is improved by incorporating the learner's correct response time fluctuation. *International Educational Data Mining Society*, 2023.

[19] Rohit Murali, Cristina Conati, and David Poole. A comparison of real-time user classification methods using interaction data for open-ended learning. In Caitlin Mills, Giora Alexandron, Davide Taibi, Giosuè Lo Bosco, and Luc Paquette, editors, *Proceedings of the 18th International Conference on Educational Data Mining*, pages 7–18, Palermo, Italy, July 2025. International Educational Data Mining Society.

[20] Wim J van der Linden. Predictive control of speededness in adaptive testing. *Appl. Psychol. Meas.*, 33(1):25–41, January 2009.

[21] Wim J van der Linden and Xinhui Xiong. Speededness and adaptive testing.

Journal of Educational and Behavioral Statistics, 38(4):418–438, 2013.

[22] Everett Mettler, Christine Massey, and Philip Kellman. Improving adaptive learning technology through the use of response times. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33, 2011.

[23] Wim J van der Linden. A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, 72(3):287–308, September 2007.

[24] D Thissen. Timed testing: An approach using item response theory. *New horizons in testing*, pages 179–203, 1983.

[25] N D Verhelst, H H F M Verstralen, and M G H Jansen. A logistic model for time-limit tests. In *Handbook of Modern Item Response Theory*, pages 169–185. Springer New York, New York, NY, 1997.

[26] Edward E Roskam. Models for speed and time-limit tests. In *Handbook of Modern Item Response Theory*, pages 187–208. Springer New York, New York, NY, 1997.

[27] H Scheiblechner. Specifically objective stochastic latency mechanisms. *J. Math. Psychol.*, 1979.

[28] KK Tatsuoka and MM Tatsuoka. A model for incorporating response-time data in scoring achievement tests. In *Proceedings of the 1979 computerized adaptive testing conference*, pages 236–256. University of Minnesota, Department of Psychology, Psychometric Methods ⋯, 1980.

[29] Eric Maris. Additive and multiplicative models for gamma distributed random variables, and their application as psychometric models for response times. *Psychometrika*, 58(3):445–469, September 1993.

[30] Jeffrey N Rouder, Jun Lu, Paul Speckman, Dongchu Sun, and Yi Jiang. A hierarchical model for estimating response time distributions. *Psychon. Bull. Rev.*, 12(2):195–223, April 2005.

[31] Wim J van der Linden. A lognormal model for response times on test items. *J. Educ. Behav. Stat.*, 31(2):181–204, June 2006.

[32] Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H Chi. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pages 1930–1939, New York, NY, USA, July 2018. Association for Computing Machinery.

[33] F.M. Lord and M.R. Novick. *Statistical theories of mental test scores*. Addison-

Wesley Pub. Co., 1968.

[34] 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.

[35] Alan E Gelfand and Adrian F M Smith. Sampling-based approaches to calculating marginal densities. *J. Am. Stat. Assoc.*, 85(410):398–409, June 1990.

[36] Rich Caruana. Multitask learning. *Machine learning*, 28:41–75, 1997.

[37] Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3994–4003. IEEE, June 2016.

[38] Yong Zhang, Yuqi Xin, Zhi-Wei Liu, Ming Chi, and Guijun Ma. Health status assessment and remaining useful life prediction of aero-engine based on BiGRU and MMoE. *Reliab. Eng. Syst. Saf.*, 220(108263):108263, April 2022.

[39] Yanping Chen, Lele Ren, Hong Xia, Zhongmin Wang, Cong Gao, and Fengwei Wang. A compound fault diagnosis method based on multi-task learning with multi-gate mixture-of-experts. In *2022 14th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*, pages 281–285. IEEE, January 2022.

[40] Tong Guan, Jiaheng Peng, and Jun Liang. Spatial-temporal graph multi-gate mixture-of-expert model for traffic prediction. In *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, pages 36–41. IEEE, September 2023.

[41] Benjamin Becker, Sebastian Weirich, Frank Goldhammer, and Dries Debeer. Controlling the speededness of assembled test forms: A generalization to the three - parameter lognormal response time model. *J. Educ. Meas.*, April 2023.

[42] 電気通信大学大学院情報理工学研究科 情報・ネットワーク工学専攻 植野真臣研究室. 文部科学省大学入学者選抜改革推進委託事業（個別大学の入学者選抜等における cbt の活用）. http://www.ai.lab.uec.ac.jp/cbt/.

[43] 駿台. 駿台だからこそ提供できる！ ICT 学習のご案内. https://www2.sundai.ac.jp/ict/.（参照 2026-01-13）.

[44] Zachary A Pardos and Neil T Heffernan. Kt-idem: Introducing item difficulty to the knowledge tracing model. *User Modeling and User-Adapted Interaction*, 21(1-2):99–135, 2011.

[45] Xiaoxiao Xiong, Siyuan Zhao, Erik G Van Inwegen, and Joseph E Beck. Going deeper with deep knowledge tracing. In *Proceedings of the 9th International Conference on Educational Data Mining*, pages 545–550, 2016.

[46] Steven Ritter and Joseph A. Beck. Knowledge tracing with the Statics2011 dataset. In *Proceedings of the Fourth International Conference on Educational Data Mining (EDM 2011)*, pages 405–406. International Educational Data Mining Society, 2011.

[47] Tomáš Žabka and Radek Pelánek. Data collection for knowledge tracing: Slepemapy.cz geography learning system. In *Proceedings of the Eighth International Conference on Educational Data Mining (EDM 2015)*, pages 367–368. International Educational Data Mining Society, 2015.

[48] John Salvatier, Thomas V Wiecki, and Christopher Fonnesbeck. Probabilistic programming in python using pymc3. *PeerJ Computer Science*, 2:e55, 2016.

[49] JP Fox, K Klotzke, and R Klein Entink. *LNIRT: log-normal response time item response theory models*, 2019. `R package version 0.5.1. Available at` https://cran.r-project.org/package=LNIRT.

[50] Jean-Paul Fox, Konrad Klotzke, and Ahmet Salih Simsek. R-package lnirt for joint modeling of response accuracy and times. *PeerJ Computer Science*, 9:e1232, 2023.

[51] Maomi UENO and Keizo NAGAOKA. On-line data-analysis of e-learning response time using gamma distribution. *Educational Technology Research*, 29(1-2):65–73, 2006.

[52] Herbert CS Thom. A note on the gamma distribution. *Monthly weather review*, 86(4):117–122, 1958.

[53] E Webb Stacy and G Arthur Mihram. Parameter estimation for a generalized gamma distribution. *Technometrics*, 7(3):349–358, 1965.

[54] Wakaba Kishida, Kazuma Fuchimoto, Yoshimitsu Miyazawa, and Maomi Ueno. Item difficulty constrained uniform adaptive testing. In *International Conference on Artificial Intelligence in Education*, pages 568–573. Springer, 2023.

[55] Maomi Ueno, Kazuma Fuchimoto, Wakaba Kishida, and Yoshimitsu Miyazawa. Computerized adaptive testing to balance exposure bias and measurement accuracy using zero-suppressed binary decision diagrams. *IEEE Access*, 13:33883–33903, 2025.