電気通信大学情報理工学域
I 類 (情報系) 情報数理工学プログラム卒業論文

問題テキストの特徴量を補助情報として正誤反応と所要時間を同時に予測する Multi-Task Deep-IRT

2026 年 1 月 29 日

コンピュータサイエンスプログラム

学籍番号 2210044

石山　創一郎

指導教員 植野　真臣

令和 6 年度 情報数理工学プログラム卒業論文概要

令和 2 年度　入学	学籍番号　2210044
指導教員　植野　真臣	氏名　石山　創一郎
題目	問題テキストの特徴量を補助情報として正誤反応と 所要時間を同時に予測する Multi-Task Deep-IRT

概要

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．所要時間を予測する手法として，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．一方で，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．最も予測精度が高いモデルとして石川らのモデル (Multi-Task Deep-IRT:MTDIRT) がある．MTDIRT は，MMoE に Deep-IRT の受検者ネットワークと項目ネットワークを組み合わせることにより，解釈性を維持しつつ予測精度を改善させた．しかし、MTDIRT は問題テキストの情報を考慮していないという課題がある．一般に，問題文の長さ，文や数式の複雑さなどは正誤反応および所要時間の双方に影響しうる．そこで，本論では，問題テキストの特徴量を補助情報とした MTDIRT の拡張手法を提案する．提案手法は，MTDIRT に問題テキスト特徴量を入力とする問題テキストエキスパートネットワークを追加する．さらに，ゲートネットワークを拡張することで，モデルが受検者から得られた特徴量と問題テキストから得られた特徴量の使用割合を学習できる．これにより，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，提案手法のゲート重みの分析を行い，提案手法における問題テキスト特徴量の使用率を評価する．

1　まえがき

近年，CBT（Computer Based Testing）や CAT（Computerized Adaptive Testing）に代表される e-testing は，教育評価を実施するための枠組みとして広く普及している [1–7]．e-testing の利点の一つは，出題される項目集合が異なるテストであっても，項目反応理論（Item Response Theory；IRT）に基づく推定により，受検者の得点（能力）を同一尺度上で同程度の精度で測定できる点にある [1–7]．IRT は回答データから潜在能力を推定することで，異なる項目で構成されたテストに対しても等質な評価を可能にする．

一方で，IRT の予測精度には理論的な限界が存在する．IRT では，一般的に受検者の能力が標準正規分布に従うと仮定するが，実データにおける能力分布はこの仮定から外れる場合がある．この場合，IRT の推定が理論的に最適である保証がなく，予測精度が低下する．

この課題に対処するために，Tsutsumi ら [8–11] は，深層学習を用いた IRT の拡張として Deep-IRT を提案した．深層学習は統計的分布の仮定に依存せず，IRT のような確率モデルより高い予測性能を示すことが多い．Deep-IRT は，独立な受検者の能力を推定する受検者ネットワークと項目の難易度を推定する項目ネットワークにより構成される．これにより，解釈性を維持しつつ，実データに対する頑健性と IRT を上回る予測精度を実現した．

一方で，時間制約下での公平性を考えると，e-testing では正誤反応の予測に加えて所要時間の予測も重要となる．所要時間予測は，カンニング等の不正行為検知 [12–14] や，CAT の試験時間を制御する手法 [15, 16] など，e-testing における様々な課題に応用されている。

この目的のために，van der Linden [17] は階層ベイズモデルに基づき，正誤反応と所要時間を同時に扱う Log-normal Response Time IRT（LNIRT）を提案した．LNIRT は IRT を基にした手法の中でも高い予測性能を示すことが報告されている [18–24]．しかし，LNIRT は，従来の IRT と同様に受検者の能力や回答速度に標準正規分布を仮定しており，予測精度に限界がある．

この問題に対処するために石川ら [25] は，正誤反応と所要時間を同時に予測する Multi-Task Deep-IRT(MTDIRT) を提案しており，最も高精度な所要時間予測精度を達成している．MTDIRT は，Multi-gate Mixture of Experts(MMoE) [26] を用いた深層学習モデルである。MMoE は，入力特徴量から各タスクの出力をデータ駆動で直接学習するため，LNIRT のような統計的分布仮定に依存しない。MMoE では，複数のエキスパートネットワークがタスク間で共有される表現を学習し，タスク固有のタワーネットワークが各タスク独自の表現へ変換する。これにより，MMoE は関連タスク間の相補的な情報を活用でき，予測精度を向上できる。さらに，MTDIRT では，LNIRT のような解釈可能なパラメータを学習する受検者ネットワークと項目ネットワークを MMoE に組み込む．受検者ネットワークでは，受検者の能力と回答速度を表現するパラメータを推定する．一方，項目ネットワークでは，項目の困難度と時間強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，回答速度と時間強度パラメータは所要時間の予測に使用される．これにより，MTDIRT は学習評価に必要な解釈性を保ちつつ，高い予測精度を実現した．

一方で，MTDIRT は受検者の項目への回答データに基づく特徴量のみを学習に用いており，項目の問題テキストがもつ情報を利用していない．一般的に，問題テキストの長さや含まれる語彙の難易度，文や数式の構造の複雑さといった言語的要因は，受検者の正誤反応および解答に要する時間の双方に影響する．例えば，文章が長い問題は時間内容の理解のために回答に必要な所要時間が増大する．また，難しい語彙が含まれる項目では，内容理解が困難になり，誤答受検者が多くなる場合がある．このように，問題テキストの特徴量は正誤反応および所要時間予測に寄与すると考えられる．

そこで, 本研究では, 問題テキストから得られる特徴量を補助情報として, 各受検者の各問題項目への正

誤反応と所要時間を同時に予測する Multi-task Deep-IRT with Exprerts for Text features(MTDIRT-ETF) を提案する．

提案手法は，MTDIRT に問題テキストから得られた特徴量を入力とし，所要時間と正誤反応に共通するテキストの特徴量を学習するエキスパートネットワーク（以下：問題テキストエキスパートネットワークと呼ぶ.）を追加する．この問題テキストエキスパートネットワークは従来のエキスパートネットワークとは独立に設計され，問題テキストの多様な特徴を抽出する．テキスト特徴量には，MathBERT [27] から得られた特徴量と古典的特徴量を用いる．MathBERT とは，数式を含む文章を用いて事前学習された大規模言語モデルであり，問題テキストを入力することで 768 次元の意味・文脈を表現する特徴量が得られる．古典的特徴量は，問題テキストの構造的な統計量（単語数や品詞比率など）である．提案手法では，受検者ネットワークおよび項目ネットワークで推定する 各潜在変数・項目パラメータごとにゲートネットワークを設ける．各ゲートネットワークは，従来手法で用いる特徴量と問題テキストから得られる特徴量を入力とし，従来の特徴量を入力とするエキスパートネットワークと問題テキストエキスパートネットワークの出力に対する重みを学習する．これにより，各潜在変数・項目パラメータの推定における両特徴量の寄与がデータ駆動的に決定される．これにより，提案手法はテキスト特徴量を補助情報として正誤反応および所要時間を予測することで，予測精度を向上できる．

本研究では，提案手法の有効性を示すために，従来手法との比較実験を行った．その結果，正誤反応，所要時間の予測において，提案手法は比較手法を上回った．

2　項目反応理論

項目反応理論 (Item Response Theory; IRT) は数理モデルを用いたテスト理論のひとつであり，近年，CBT や CAT に代表される e-testing の普及とともに教育， 心理学など多様な分野で使用されている [28].

2.1　２母数ロジスティックモデル

IRT[] では、受検者 $i \in \{1, \cdots, I\}$ の項目 $j \in \{1, \cdots, J\}$ に対する正誤反応 u_{ij} を以下のように表す.

$$u_{ij} = \begin{cases} 1 & \text{項目 } j \text{ に受検者 } i \text{ が正答} \\ 0 & \text{それ以外} \end{cases} \tag{1}$$

項目反応理論の一般的なモデルの一つとして，2 母数ロジスティックモデル (2-Parameter Logistic Model; 2PLM) がある．2PLM は，能力値が $\theta \in [-\infty, \infty]$ のある受検者 i が項目 j に正答する確率を以下のように示す.

$$p(u_{ij} = 1|\theta_i) = \frac{1}{1 + \exp\left(-a_j(\theta_i - b_j)\right)} \tag{2}$$

ここで， $a_j \in [0, \infty]$, $b_j \in [-\infty, \infty]$ はそれぞれ項目 j の識別力，困難度を示すパラメータである.

2.2　所要時間予測のための対数正規分布モデル

Van der Linden は，所要時間が対数正規分布に従う確率変数であるとしたモデル (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.

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

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，受検者の項目への所要時間を高精度に予測できる手法は，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) である．LNIRT は，受検者に関するパラメータと項目に関するパラメータ間の依存関係をモデル化することで，所要時間と正誤反応の依存関係を考慮できる．まず，第一層目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、これらの潜在変数が受検者および項目の集合全体で多変量正規分布に従うと仮定される。

第一層において、LNIRT は受検者と項目の各組合せに対して、正誤反応予測のための 2PLM および所要時間の対数正規分布モデル (Log-normal Response Time Theory:LNRT) [29] を定義する。LNRT では、所要時間 t_{ij} が対数正規分布に従うと仮定する。このとき、受検者 i の項目 j への所要時間 t_{ij} の確率密度関数は以下のように与えられる。

$$f(t_{ij;\zeta_i,\phi_j,\lambda_j}) = \frac{\phi_j}{t_{ij}\sqrt{2\pi}} \exp\{-\frac{1}{2}[\phi_j(\ln t_{ij} - (\lambda_j - \zeta_i))]^2\} \tag{4}$$

ここで、$\zeta \in (-\infty, \infty)$ は受検者 i の回答速度を表す潜在変数であり、$\lambda_j \in (-\infty, \infty)$ と $\phi \in (0, \infty)$ は

項目 j の時間強度および時間識別力を表すパラメータである。

第二層において、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、時間強度) は、それぞれの母集団分布からサンプリングされると仮定される。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。

$$(\theta_i, \zeta_i) \sim \mathcal{N}_2(\boldsymbol{\mu}_{(\theta_i,\zeta_i)}, \boldsymbol{\Sigma}_{(\theta_i,\zeta_i)}), \tag{5}$$

$$\boldsymbol{\mu}_{(\theta_i,\zeta_i)} = (\mu_\theta, \mu_\zeta), \tag{6}$$

$$\boldsymbol{\Sigma}_{(\theta_i,\zeta_i)} = \begin{pmatrix} \sigma_\theta^2 & \sigma_{\theta\zeta} \\ \sigma_{\zeta\theta} & \sigma_\zeta^2 \end{pmatrix} \tag{7}$$

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、$\rho_\theta^2, \rho_\zeta^2$ 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 $\rho_{\theta\zeta}, \rho_{\zeta\theta}$ 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),(5),(6) 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。

$$(a_j, b_j, \phi_j, \lambda_j) \sim \mathcal{N}_4(\boldsymbol{\mu}_{(a_j,b_j,\phi_j,\lambda_j)}, \Sigma_{(a_j,b_j,\phi_j,\lambda_j)}) \tag{8}$$

$$\mu_{(a_j,b_j,\phi_j,\lambda_j)} = (\mu_a, \mu_b, \mu_\phi, \mu_\lambda), \tag{9}$$

$$\Sigma_{(a_j,b_j,\phi_j,\lambda_j)} = \begin{pmatrix} \sigma_a^2 & \sigma_{a,b} & \sigma_{a,\phi} & \sigma_{a,\lambda} \\ \sigma_{b,a} & \sigma_b^2 & \sigma_{b,\phi} & \sigma_{b,\lambda} \\ \sigma_{\phi,a} & \sigma_{\phi,b} & \sigma_\phi^2 & \sigma_{\phi,\lambda} \\ \sigma_{\lambda,a} & \sigma_{\lambda,b} & \sigma_{\lambda,\phi} & \sigma_\lambda^2 \end{pmatrix} \tag{10}$$

ここで、$\mu_a, \mu_b, \mu_\phi, \mu_\lambda$ はそれぞれ項目全体における識別力、困難度、時間識別力、時間強度の平均を表し、共分散行列 $\Sigma_{(a_j,b_j,\phi_j,\lambda_j)}$ は、これらの項目パラメータの分散と項目間の二変量正規分散を表す。

LNIRT では，第一層および第二層における受検者潜在変数と項目パラメータはギブスサンプリング [30] 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定義され、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。

$$\boldsymbol{\Sigma}_{(\theta_i,\zeta_i)} \sim \text{Inv-Wishart}_{\nu_{(\theta_i,\zeta_i)}}(V^{-1}_{(\theta_i,\zeta_i)}), \tag{11}$$

$$\boldsymbol{\mu}_{(\theta_i,\zeta_i)} \mid \boldsymbol{\Sigma}_{(\theta_i,\zeta_i)} \sim \mathcal{N}(\boldsymbol{\mu}_{0,(\theta_i,\zeta_i)}, \boldsymbol{\Sigma}_{(\theta_i,\zeta_i)}/\kappa_{0,(\theta_i,\zeta_i)}), \tag{12}$$

$$\boldsymbol{\Sigma}_{(a_j,b_j,\phi_j,\lambda_j)} \sim \text{Inv-Wishart}_{\nu_{(a_j,b_j,\phi_j,\lambda_j)}}(V^{-1}_{(a_j,b_j,\phi_j,\lambda_j)}), \tag{13}$$

$$\boldsymbol{\mu}_{(a_j,b_j,\phi_j,\lambda_j)} \mid \boldsymbol{\Sigma}_{(a_j,b_j,\phi_j,\lambda_j)} \sim \mathcal{N}(\boldsymbol{\mu}_{0,(a_j,b_j,\phi_j,\lambda_j)}, \boldsymbol{\Sigma}_{(a_j,b_j,\phi_j,\lambda_j)}/\kappa_{0,(a_j,b_j,\phi_j,\lambda_j)}), \tag{14}$$

ここで、$\mu_{(\theta_i,\zeta_i)}$,$\mu_{(a_j,b_j,\phi_j,\lambda_j)}$ は自由度,$V_{((\theta_1,\zeta_i))}$,$V_{(a_j,b_j,\phi_j,\lambda_j)}$ は逆ウィシャート事前分布のスケール行列を表す。$\mu_{0,(\theta_i,\zeta_i)}$ および $\mu_{0,(a_j,b_j,\phi_j,\lambda_j)}$ 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、推定手順は [] に示されている。

以上により，LNIRT は所要時間と正誤反応の依存関係を適切にモデル化でき，従来手法よりも受検者の項目への所要時間と正誤反応の予測精度を向上させた.

3　正誤反応と所要時間を同時に予測する深層学習モデル

しかし，LNIRT は，従来の IRT と同様に受検者の能力や回答速度に標準正規分布を仮定しており，予測精度に限界がある．この問題を解消するために石川らは MMoE を用いた深層学習モデル (Multi-Task Deep-IRT:MTDIRT) を提案した。MTDIRT は、入力特徴量から各タスクの出力をデータ駆動で直接学習し、正誤反応と所要時間を同時に予測する。

3.1　Multi-gate Mixture of Experts

Ma ら [26] は，多タスク学習におけるタスク間関係を明示的にモデル化する手法 (Multi-gate Mixture-of-Experts:MMoE) を提案した．

MMoE のモデル図を図 1 に示す．

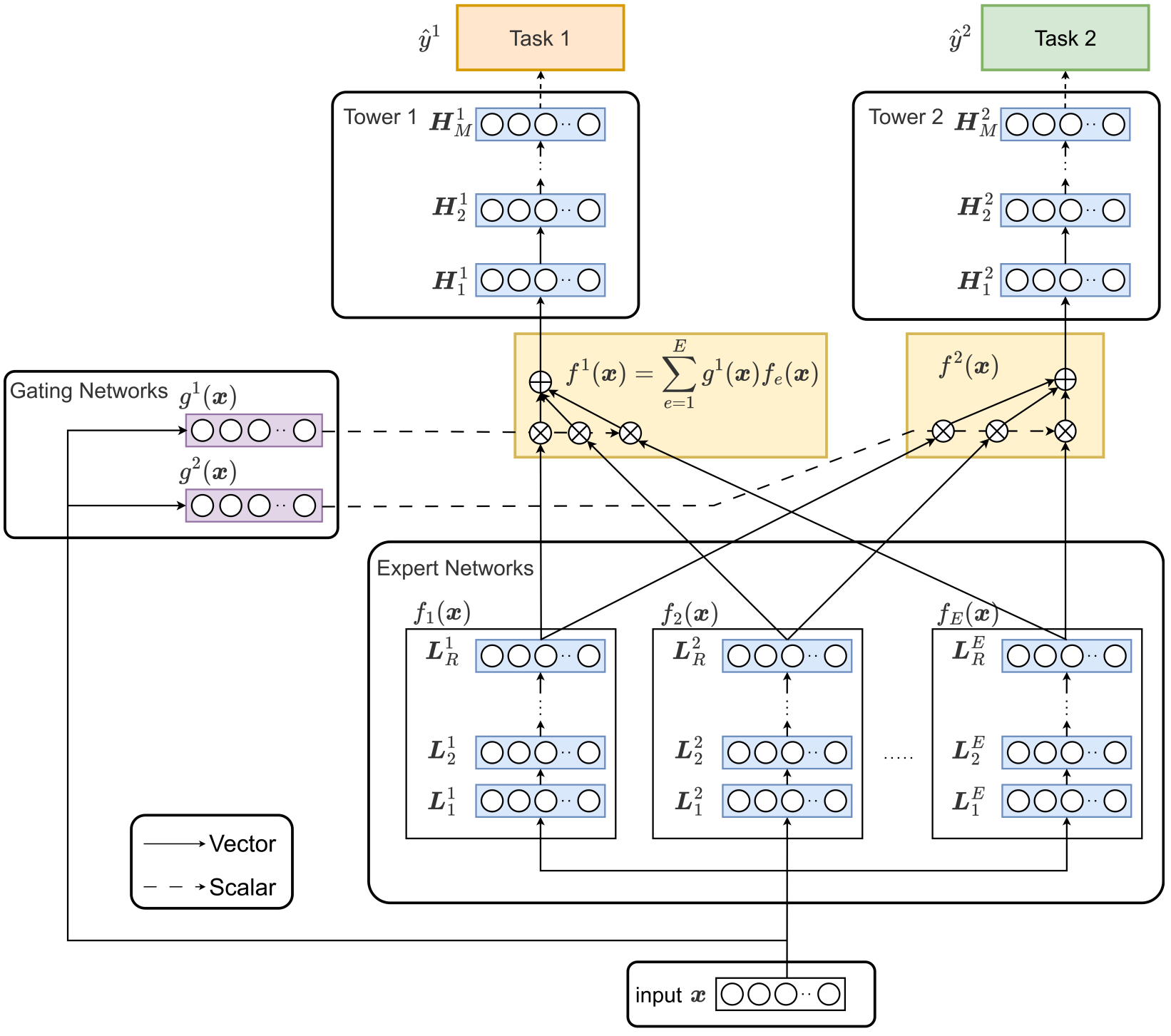

図 1　MMoE のモデル図

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 次元の特徴量ベクトル $x \in R^d$ を入力として、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が、エキスパートネットワークのそれぞれの出力に異なる重みを割り当てる。これによ

り、各タスクに対する共通特徴の重要度を学習する。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に用いられ、予測タスク固有の特徴を学習する。MMoE は、タスク間の共通の特徴を学習するために、E 個 ($e = 1, 2, \cdots, E$) のエキスパートネットワークを持つ。各エキスパートネットワークは、R 個 ($r = 1, 2, \cdots, R$) の隠れ層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 (Feed-forward Neural Network:FNN) である。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 L_r^e は、ReLU 活性化関数と線形変換を用いて以下のように計算される。

$$L_r^e = \begin{cases} \mathrm{ReLU}(W_r^e x + b_r^e) & r = 1, \\ \mathrm{ReLU}(W_r^e L_{r-1}^e + b_r^e) & r = 2, \cdots, R \end{cases} \tag{15}$$

ここで、W_r^e は 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、b_r^e は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である。

$$f_e(x) = L_R^e \tag{16}$$

次に、各タスク $t \in \{1, 2, \cdots, T\}$ に対して、タスク固有のゲーティングネットワークは、各エキスパートの特徴表現 $f_e(x)$ の重要度を学習する。

$$g_e^t(x) = softmax(W_e^t x) \tag{17}$$

$$f^t(x) = \Sigma_{e=1}^{E} g_e^t(x) f_e(x) \tag{18}$$

ここで、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 $f^t(x)$ は、M 個 ($m = 1, 2, \cdots, M$) の隠れ層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 H_m^t は、以下のように計算される。

$$H_m^t = \begin{cases} \mathrm{ReLU}(W_m^t f^t(x) + b_1^t) & m = 1, \\ \mathrm{ReLU}(W_m^t H_{m-1}^t + b_m^t) & m = 2, \cdots, M \end{cases} \tag{19}$$

ここで、W_m^t はタスク t の m 番目の隠れ層の重み行列、b_m^t はバイアスベクトルを表す。最後に、タスク固有の予測値 $\hat{y}^t$ は、最終層の出力 H_M^t にタスク依存の出力活性化関数 $\psi^t(\cdot)$ を用いて得られる。

$$\hat{y}^t = \psi^t(H_M^t) \tag{20}$$

MMoE では、各タスク t に対してタスク固有の予測値 $\hat{y}^t$ と真のラベル y^t との差を測定する損失関数 $L^{(t)}(\hat{y}^t, y^t)$ が定義される。損失関数はタスクの予測目的に応じて、例えば、分類タスクにはバイナリークロスエントロピー (Binary Cross Entropy:BCE)、回帰タスクには平均二乗誤差 (Mean Squared Error:MSE) などが設定される。MMoE 全体の損失は、タスクすべての損失の加重和として以下のように定義される。

$$L_{total} = \Sigma_{t=1}^{T} \alpha_t L^{(t)}(\hat{y}^t, y^t), \qquad s.t. \qquad \Sigma_{t=1}^{T} \alpha_t = 1, \alpha_t \geq 0 \tag{21}$$

ここで、α_t はタスク t における全体の損失に対する割合を決定するハイパーパラメータである。

以上のように，MMoE はエキスパート層をタスク間で共有しつつ，タスクごとにゲートを分離することで，タスク関連性の強弱を自動的に学習し，多タスク学習における柔軟な表現獲得と安定した性能向上を実現した．しかし、MMoE は高い予測精度を実現する一方で解釈性が低いという問題がある。

3.2 MTDIRT

この問題に対処するために、石川らは，MMoE を基盤として，タワーネットワーク部にタワーネットワークに解釈可能パラメータネットワーク (受検者ネットワークと項目ネットワーク) を組み込んだ Multi-Task Deep-IRT(MTDIRT) を提案した．MTDIRT は解釈性を保ちながら受検者の正誤反応と所要時間を同時に高い精度で予測することができる。モデル図を図 2 に示す.

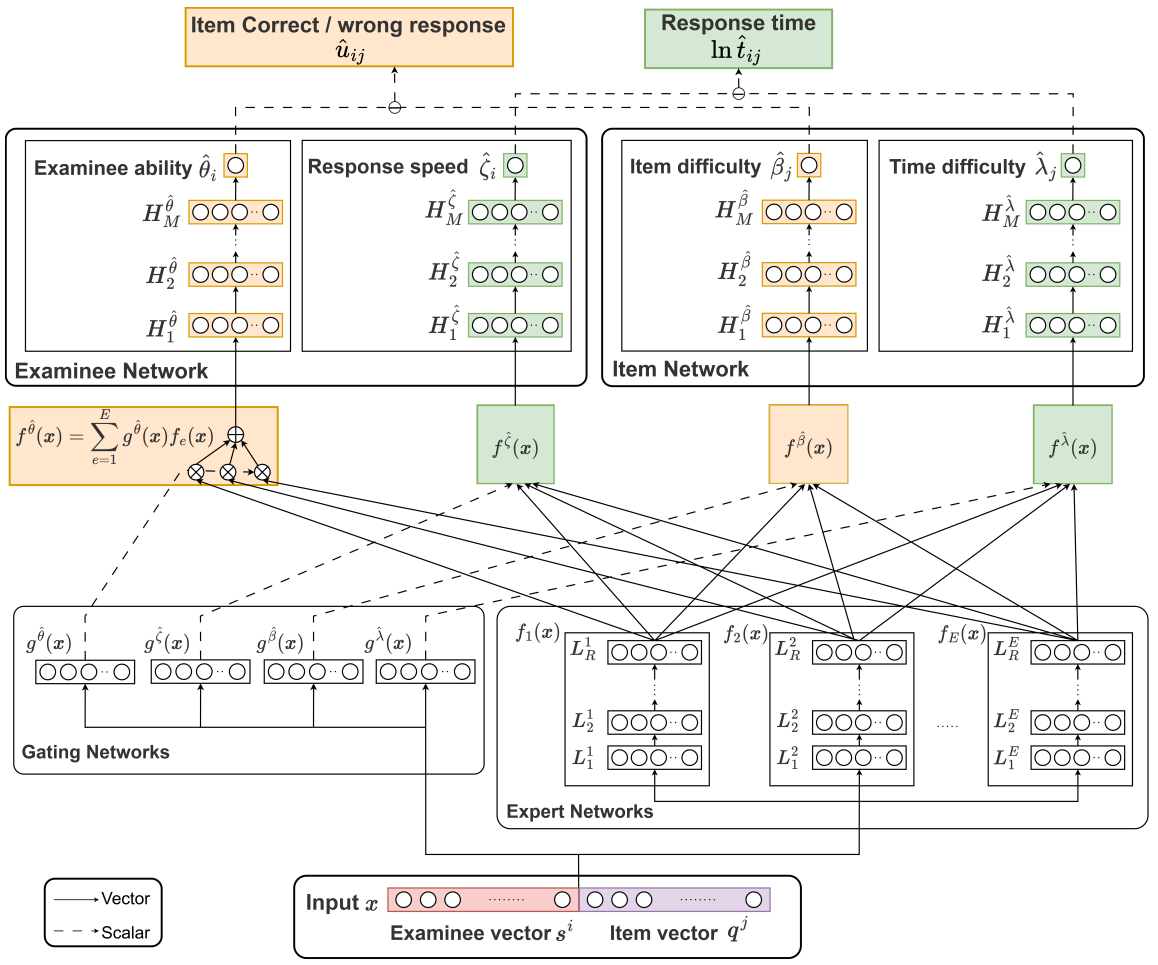

図 2　MTDIRT のモデル図

図 2 のとおり、入力ベクトル $x = (s^i, q^j)$ から 4 つの解釈可能パラメータ (受検者能力 $\hat{\theta}_i$、回答速度 $\hat{\zeta}_i$、項目困難度 $\hat{\beta}_j$、時間強度 $\hat{\lambda}_j$) を推定する。ここで $s^i = \{s^i_1, s^i_2, \cdots, s^i_I\}$ は受検者 i を表す one-hot ベクトルであり、$i' = i$ 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 $q^j = \{q^j_1, q^j_2, \cdots, q^j_J\}$ は，項目 j を表す one-hot ベクトルである．

エキスパートは E 個の多層 FNN とし、それぞれの出力をパラメータごとのゲート $(\hat{\theta}, \hat{\zeta}, \hat{\beta}, \hat{\lambda})$ で重み付けし、パラメータ別の特徴表現 $f^{\hat{\theta}}, f^{\hat{\zeta}}, f^{\hat{\beta}}, f^{\hat{\lambda}}$ を得る。これらを、受検者ネットワーク $(\hat{\theta}, \hat{\zeta},)$ と項目ネットワーク $(\hat{\beta}, \hat{\lambda})$ に入力し、最終的な予測値

$$\hat{\theta}_i = \mathrm{ReLU}(W^{\hat{\theta}} H^{\hat{\theta}}_M + b^{\hat{\theta}}), \qquad \hat{\zeta}_i = \mathrm{ReLU}(W^{\hat{\zeta}} H^{\hat{\zeta}}_M + b^{\hat{\zeta}}) \tag{22}$$

$$\hat{\beta}_j = \mathrm{ReLU}(W^{\hat{\beta}} H^{\hat{\beta}}_M + b^{\hat{\beta}}), \qquad \hat{\lambda}_j = \mathrm{ReLU}(W^{\hat{\lambda}} H^{\hat{\lambda}}_M + b^{\hat{\lambda}}) \tag{23}$$

を得る。推定されたパラメータを用いて、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．正誤反応の予測値は，Deep-IRT 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．

$$\hat{u}_{ij} = \frac{1}{1 + \exp\{-(\hat{\theta}_i - \hat{\beta}_j)\}}, \tag{24}$$

次に、予測所要時間の対数は以下のように定義される．

$$\ln \hat{t}_{ij} = -\hat{\zeta}_i + \hat{\lambda}_j \tag{25}$$

この定式化は、Becker らの事後期待値推定量に基づき、LNIRT における所要時間モデルから導出されている。

全体的な損失は正誤のバイナリークロスエントロピー (BCE)$L^{(c/w)}$ と所要時間予測に対する平均二乗誤差 (MSE)$L^{(rt)}$ の加重和とする：

$$L^{(c/w)} = -\Sigma_{i=1}^{I} \Sigma_{j \in A_i} \{u_{ij} \ln \hat{u}_{ij} + (1 - u_{ij}) \ln(1 - \hat{u}_{ij})\} \tag{26}$$

$$L^{(rt)} = \frac{1}{\Sigma_i |A_i|} \Sigma_{i=1}^{I} \Sigma_{j \in A_i} (\ln t_{ij} - \ln \hat{t}_{ij})^2 \tag{27}$$

$$L_{total} = \alpha L^{(c/w)} + (1 - \alpha) L^{(rt)}, \qquad 0 \leq \alpha \leq 1. \tag{28}$$

このように，MTDIRT は， MMoE により正誤反応と所要時間の関係性を考慮した上で LNIRT として解釈可能なパラメータを予測できる．その結果，MTDIRT は高い精度で所要時間を予測できる．

4　提案手法

従来の MTDIRT は，受検者から得られた回答データを主入力として学習を行う枠組みであるが，項目が持つ問題テキストそのものの情報を直接には利用していない．一般に，問題文の長さ，含まれる語彙の難易度，文や数式の構造の複雑さといった言語的要因は，正誤反応および解答所要時間の双方に影響しうる．例えば，文章が長い問題では内容理解に要する時間が増大し，難解な語彙や複雑な表現を含む項目では，理解が困難となり誤答が増える可能性がある．したがって，問題テキストから得られる特徴量を MTDIRT へ統合することは，正誤および所要時間予測の精度向上に寄与すると考えられる．

本研究では，上記の課題を踏まえ，MathBERT により抽出したテキスト特徴量を MTDIRT へ導入する拡張手法を提案する．MathBERT は BERT [31] を基盤とし，数学領域テキストで事前学習されたモデルであり，文章の文脈に加えて数式・数学記号を含む表現に適している．提案手法の概要を図 3 に示す．

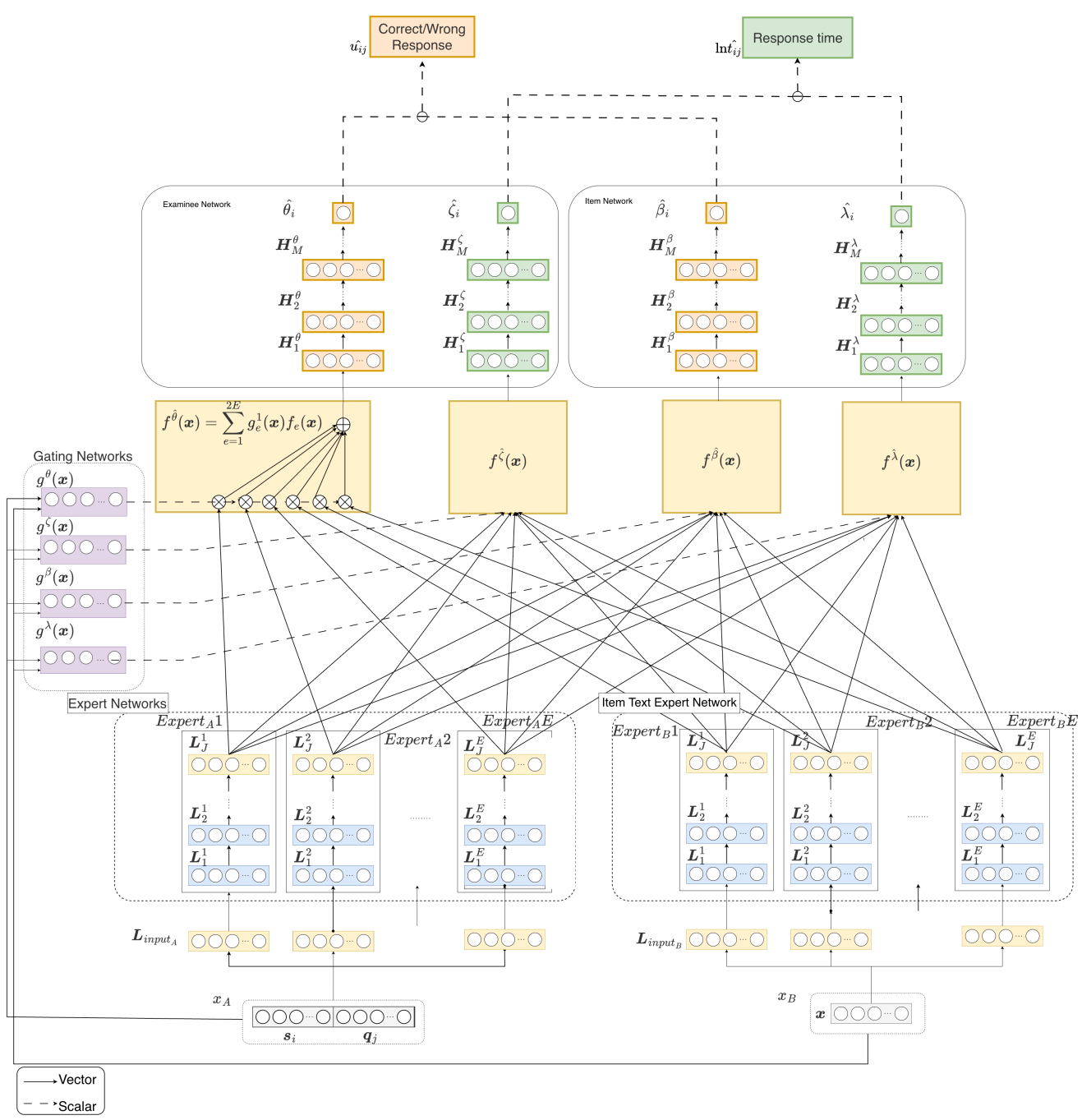

図 3　提案手法のモデル図

各タスクにおいて，ゲートが受検者由来情報と問題テキスト由来情報の寄与率をデータ駆動的に調整できる点が提案手法の要点である．

提案手法では，(i) 受検者の回答データ由来の特徴量と，(ii) 問題テキスト由来の特徴量を併用する．受検者 i と項目 j の ID に由来する特徴量を $\boldsymbol{x}_A(i,j)$，項目 j の問題テキストに由来する特徴量を $\boldsymbol{x}_B(j)$

とする．ゲート入力は両者の結合

$$\boldsymbol{x}(i,j) = [\boldsymbol{x}_A(i,j); \boldsymbol{x}'_B(j)] \tag{29}$$

で与える．ここで $\boldsymbol{x}'_B(j)$ は，後述する次元圧縮後のテキスト特徴量である（本研究では 64 次元）．

項目 j の問題テキストを文単位に分割し，文 s を MathBERT へ入力してトークン表現 $\boldsymbol{h}_{j,s,t} \in \mathbb{R}^d$（$t$ はトークン位置）を得る．可変長のトークン列・文集合を固定長ベクトルへ集約するため，本研究では階層的注意機構を用い，(1) 文内トークン表現から文ベクトルを得る「トークン注意」，(2) 文ベクトル集合から文章ベクトルを得る「文注意」を適用する．

まずトークン注意により文ベクトル $\boldsymbol{u}_{j,s} \in \mathbb{R}^d$ を

$$\boldsymbol{r}_{j,s,t} = \tanh(\boldsymbol{W}_w \boldsymbol{h}_{j,s,t}), \tag{30}$$

$$\alpha_{j,s,t} = \frac{\exp(\boldsymbol{v}_w^\top \boldsymbol{r}_{j,s,t})}{\sum_{t'} \exp(\boldsymbol{v}_w^\top \boldsymbol{r}_{j,s,t'})}, \tag{31}$$

$$\boldsymbol{u}_{j,s} = \sum_t \alpha_{j,s,t} \boldsymbol{h}_{j,s,t} \tag{32}$$

として計算する．続いて文注意により，項目 j 全体の文章ベクトル $\boldsymbol{z}_j \in \mathbb{R}^d$ を

$$\boldsymbol{q}_{j,s} = \tanh(\boldsymbol{W}_s \boldsymbol{u}_{j,s}), \tag{33}$$

$$\beta_{j,s} = \frac{\exp(\boldsymbol{v}_s^\top \boldsymbol{q}_{j,s})}{\sum_{s'} \exp(\boldsymbol{v}_s^\top \boldsymbol{q}_{j,s'})}, \tag{34}$$

$$\boldsymbol{z}_j = \sum_s \beta_{j,s} \boldsymbol{u}_{j,s} \tag{35}$$

として得る．ここで $\alpha_{j,s,t}$ は文 s 内のトークン重要度（トークン注意重み），$\beta_{j,s}$ は項目 j における文 s の重要度（文注意重み）を表す．

MathBERT 埋め込みに加えて，問題文の表層的・統計的性質を表す古典的特徴量ベクトル $\boldsymbol{c}_j$ を導入する．本研究で用いる古典的特徴量を表 1 に示す．スケール差の影響を避けるため，各次元は以下のように定義される Min-Max 正規化により $[0,1]$ へ変換する：

$$\tilde{x}_{j,k} = \frac{x_{j,k} - \min_{j'} x_{j',k}}{\max_{j'} x_{j',k} - \min_{j'} x_{j',k} + \epsilon}, \qquad 0 \le \tilde{x}_{j,k} \le 1. \tag{36}$$

ここで，$x_{j,k}$ は項目 j に対する k 次元目の特徴量値，$\min_{j'} x_{j',k}$ および $\max_{j'} x_{j',k}$ は，それぞれ同一データセット内の全項目 j' における k 次元目の特徴量の最小値・最大値を表す．また，$\epsilon > 0$ は分母が 0 となることを避けるための微小定数である．

表 1　古典的テキスト特徴量（特徴量名と定義）

特徴量名	定義
`sentence_count`	項目 j の文数 S_j
`POS ratios`	品詞比率（名詞 NN・代名詞 PRP・動詞 VB・形容詞 JJ の割合）
`num_symbols`	数字・数学記号（$+, -, =, \sqrt{\ }, \Sigma, \pi, \int, \times, \div$ 等）の出現数
`unk_ratio`	トークナイザの未知語トークン（`[UNK]`）の割合
`flesch_reading_ease`	Flesch Reading Ease スコア
`word_count_scaled`	単語数 W_j の正規化値（式 (36) に従う）

以上より，項目 j のテキスト特徴量は

$$\boldsymbol{x}_B(j) = [\boldsymbol{z}_j; \boldsymbol{c}_j] \tag{37}$$

として与える（$[;]$ はベクトル結合）．

入力の 2 系統化に対応して，エキスパートネットワークも 2 系統に分離する．すなわち，$\boldsymbol{x}_A(i,j)$ を

入力とする受検者エキスパート群（E 個）と，$\boldsymbol{x}'_B(j)$ を入力とする問題テキスト由来エキスパート群（E 個）を独立に設計し，合計 $2E$ 個のエキスパートを用いる．

しかし，BERT 由来の表現は高次元であり，そのまま利用すると計算量増大や過学習の懸念がある．そこで本研究では，テキスト特徴量 $\boldsymbol{x}_B(j)$ を線形変換により低次元へ圧縮したものをモデル入力として用いる．次元数はグリッドサーチにより決定した．

それに加えて、受検者エキスパート群と問題テキストエキスパート群を合わせて $2E$ 個のエキスパートを用意し，タスク k ごとにゲートを設ける．ゲートは結合入力 $\boldsymbol{x}(i,j)$（式 (29)）から各エキスパートの重みを出力する：

$$\boldsymbol{g}^k(\boldsymbol{x}) = \mathrm{softmax}(\boldsymbol{W}_{g^k}\boldsymbol{x}) \in \mathbb{R}^{2E}, \tag{38}$$

$$\boldsymbol{f}^k(\boldsymbol{x}) = \sum_{e=1}^{2E} g_e^k(\boldsymbol{x})\, \boldsymbol{f}_e(\cdot) \in \mathbb{R}^h. \tag{39}$$

ここで $\boldsymbol{f}_e(\cdot)$ はエキスパート e の出力であり，受検者エキスパートは $\boldsymbol{x}_A(i,j)$ を，問題テキストエキスパートは $\boldsymbol{x}'_B(j)$ を入力として処理する．これにより各タスクにおいて，受検者による特徴量とテキスト特徴量をどの程度利用するかをゲートが自動的に学習でき，予測精度の向上が期待される．

5　評価実験

提案手法の有効性を示すために，実データセットを用いた実験により従来手法と比較する．従来手法には，IRT [1–7]，LNRT [29]，LNIRT [18–24]，石川らの Multi-task Deep-IRT(MTDIRT) と MTDIRT を片方のタスクのみ予測するように変更した Single-task Deep-IRT(STDIRT)，さらに、MTDIRT の入力に提案手法と同様の問題テキスト特徴量を追加したモデル（以降 MTDIRT-TF と呼ぶ）を用いた．

5.1　データセット

本実験では，以下の実データを使用する．

1. UEC: 電気通信大学 (UEC) で実施された CBT 形式の試験データである．2023 年，2024 年，および 2025 年のデータ (数学) である．

表 2 は，各データセットにおける受検者数，項目数，正答率，平均所要時間をまとめたものである．

表 2　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	平均所要時間
UEC2023	666	99	0.68	170.26
UEC2024	734	140	0.69	185.02
UEC2025	741	203	0.63	162.57

5.2　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受検者の項目への所要時間と正誤反応の予測精度を比較した。評価指標として、所要時間予測には二乗平均平方根誤差（RMSE)、正誤反応予測には ROC 曲線下面積（AUC）を採用した。なお、受検者 i の予測所要時間の RMSE は以下のように計算される。

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i-\hat{y}_i\right)^2} \tag{40}$$

ここで、y_i は所要時間の真値，$\hat{y}_i$ が所要時間の推定値，N をデータの総数としている。

IRT，LNRT，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．

MTDIRT、STDIRT、および提案手法の学習では、各隠れ層のノード数を 16、隠れ層の数を 6 とした。また、各タスクの損失関数の重み α とエキスパートの数 E は，各データセットごとに，検証データの loss が最も低くなるように $\alpha = \{0.1, \cdots, 0.5\}$，E=$\{1,2,3,4,5\}$ からグリッドサーチを行い選択した。

なお，問題テキスト特徴量を用いる提案手法，MTDIRT-TF については，α,E のグリッドサーチに加えてテキスト特徴量アブレーション実験を行った．具体的には，各年度のデータごとに独立して，まずは作成した全特徴量を投入して学習・評価を行い，その後は「一度に一種類だけ特徴量を外す」という操作を全特徴量について実施した．最も指標が改善した特徴量を削除と判定し，残りの特徴量集合で

同じ手順を繰り返し，性能がこれ以上良くならなくなる時点で最終の採用特徴量集合を確定した．

正誤反応予測をするすべてのモデルにおける閾値を 0.5 に設定した．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、それ以外の場合を誤答 (0) とした。

5.3 実験結果

本節では，提案手法を IRT, LNRT, LNIRT, MTDIRT, Single-Task Deep-IRT(STDIRT), MTDIRT-TF と推定精度を比較した結果を示す．

表 3 は、各データセットにおける予測所要時間の RMSE を示す．提案手法は，従来手法より小さい RMSE を示した．これは，問題テキスト特徴量を入力に追加したことで，モデルがより多くの情報を予測に活用できたためと考えられる．さらに，提案手法は MTDIRT-TF よりも予測精度を大きく改善させた．この結果は、受検者特徴量と問題テキスト特徴量を異なるエキスパートネットワークに入力し，それぞれ特徴を学習することでモデルが適切な割合で各特徴量を利用できたためと考えられる．

表 3　予測所要時間の RMSE 平均 (標準偏差)

データセット	LNRT	LNIRT	STDIRT	MTDIRT	MTDIRT-TF	Proposed
UEC2023	185.91(10.79)	185.9(10.79)	120.41(4.49)	120.47(4.51)	125.35(3.99)	**119.67(4.26)**
UEC2024	253.0(19.75)	253.0(19.74)	168.32(6.51)	168.57(7.01)	171.64(6.67)	**167.50(6.51)**
UEC2025	145.99(6.11)	145.99(6.11)	106.47(5.38)	109.39(5.66)	111.39(7.03)	**106.46(5.02)**
平均	194.96	194.96	131.73	132.81	136.12	**131.21**

表 4 は，各データセットにおける予測正誤反応の AUC を示す．提案手法は，従来手法よりも平均で大きな AUC を示した．所要時間予測と同様に，テキスト特徴量を適切に学習に用いることで改善が見られたと考えられる．MTDIRT-TF と比較すると，平均で大きな改善が見られた．この結果についても，問題テキストエキスパートネットワークを追加したことによって正誤反応予測に必要なエキスパートに適切な重み付けをできた結果だと考えられる．

表 4　予測正誤反応の AUC(%) 平均 (標準偏差)

データセット	IRT	LNIRT	STDIRT	MTDIRT	MTDIRT-TF	Proposed
UEC2023	75.60(0.97)	71.36(1.14)	82.39(0.68)	82.23(0.61)	82.01(0.50)	**82.42(0.68)**
UEC2024	77.12(0.65)	73.05(0.24)	80.58(0.42)	**80.66(0.37)**	80.33(0.48)	80.64(0.46)
UEC2025	79.32(0.45)	77.02(1.11)	84.06(0.28)	84.22(0.29)	84.10(0.32)	**84.30(0.31)**
平均	77.34	74.57	82.34	82.37	82.14	**82.45**

提案手法は、MTDIRT が用いる正誤反応・所要時間に加えて，問題文から得られるテキスト特徴量を入力として取り込み，さらに問題テキスト専用のエキスパートネットワークを導入した．これにより，MMoE のゲートがタスクごとに受検者特徴量と問題テキスト特徴量を選択的に活用できるようになり，正誤反応予測，所要時間予測において MTDIRT よりも予測性能を向上した．

5.4 各タスクに割り当てられたゲート重みの分析

本節では，提案手法により推定された各潜在特性 (能力，回答速度，困難度，時間困難度) について，受検者特徴量と問題テキスト特徴量のどちらがより推定に寄与したかを分析するために、受験者情報および問題テキスト情報のエキスパートネットワークに対するゲートネットワークの重みを集計し比較した．

具体的には，評価時の実験ログから各タスクに割り振られた重みの値を抽出し，各年度ごとの 5 回分の平均ゲート重みを算出した．その結果を表 5 に示す．

表 5　年度別平均ゲート重み 平均 (標準偏差)

年度	エキスパート	能力	回答速度	困難度	時間困難度
2023	受検者	0.005(0.007)	0.876(0.049)	0.004(0.006)	0.002(0.002)
	テキスト	0.995(0.007)	0.124(0.049)	0.996(0.006)	0.998(0.002)
2024	受検者	0.028(0.053)	0.615(0.308)	0.005(0.006)	0.393(0.320)
	テキスト	0.972(0.053)	0.385(0.308)	0.995(0.006)	0.607(0.320)
2025	受検者	0.00018(0.00009)	0.0048(0.0021)	0.00070(0.00026)	0.0034(0.0011)
	テキスト	0.99982(0.00009)	0.995(0.002)	0.99930(0.00026)	0.9966(0.0011)

まず 2023 年では，能力・困難度・時間困難度 の重みがほぼテキスト側に割り当てられており，能力・困難度・強度に関する予測は主に問題テキスト特徴量で説明されていることが示唆される．一方で，速度 回答速度 のみは受検者側の寄与が大きい．

2024 年では，能力 と 困難度 は引き続きテキスト側が割合のほとんどを占めており，能力・困難度の予測における問題テキスト特徴量の重要性は維持されている．他方で 回答速度 は受検者側の寄与が大きく，時間困難度 は受検者とテキストの双方が一定程度寄与する構造となった．さらに 回答速度 と 時間困難度 の標準偏差が比較的大きいことから，fold（データ分割）によりどちらのエキスパートを重視するかが変動しやすく，速度・強度系のモデリングはデータ条件に敏感である可能性がある．

2025 年では，能力・回答速度・困難度・時間困難度のすべてでテキスト側の重みが極めて大きく，受検者側の寄与はほぼゼロに近い．また標準偏差も小さいため，問題テキスト特徴量を重視する傾向が fold をまたいで安定していると解釈できる．

まとめると，能力と困難度 は年度を通じて一貫してテキスト側の寄与が支配的であり，正誤の予測にはテキスト特徴量が重要という傾向が確認できる．一方，所要時間予測に使用される回答速度 や 時間困難度 は，年度により受検者寄与の大きさが変わり，特に 2024 年では分割によるデータ依存性も強い．

6　むすび

本研究では，Multi-Task Deep-IRT を拡張し，補助情報としての問題テキスト特徴量の追加を提案した．提案手法は，テキスト補助情報を追加入力として与えることでより多くのデータを学習できる．また，ゲートネットワークを追加したことによって，モデルが適切にテキスト特徴量を学習に使用できる．

数値実験の結果，提案手法はテキスト情報を学習に利用しつつ正誤反応，所要時間の両タスクにおいて予測精度を向上させることを示した．さらに、提案手法は、MTDIRT と比較して問題テキスト特徴量を優先的に利用することによって推定精度を向上することができた．MTDIRT-TF との比較では，問題テキストエキスパートネットワークの重要性を示すことができた．

一方で，問題テキスト特徴量の使用率が極端に偏っている年度も見受けられるため原因の追求が重要である．今後は，オープンソースデータセットでの実験，ゲート使用率を調整する機構の追加などを検討する．

参考文献

[1] Cheng, T., Sun, K.-T., Chen, Y.-J., Tsai, S.-Y. and Cheng, C.-F.: Creating IRT-Based Parallel Test Forms Using the Genetic Algorithm Method, *Applied Measurement in Education*, Vol. 21, pp. 1–41 (2008).

[2] Songmuang, P. and Ueno, M.: Bees Algorithm for Construction of Multiple Test Forms in E-Testing, Vol. 4, No. 3 (2011).

[3] Ishii, T., Songmuang, P. and Ueno, M.: Maximum Clique Algorithm for Uniform Test Forms Assembly, Vol. 7926, pp. 451–462 (2013).

[4] Ishii, T., Songmuang, P. and Ueno, M.: Maximum Clique Algorithm and Its Approximation for Uniform Test Form Assembly, *IEEE Transactions on Learning Technologies*, Vol. 7, No. 1, pp. 83–95 (2014).

[5] Ishii, T. and Ueno, M.: Algorithm for Uniform Test Assembly Using a Maximum Clique Problem and Integer Programming, pp. 102–112 (2017).

[6] Fuchimoto, K., Ishii, T. and Ueno, M.: Hybrid Maximum Clique Algorithm Using Parallel Integer Programming for Uniform Test Assembly, *IEEE Trans. Learn. Technol.*, Vol. 15, No. 2, p. 252–264 (2022).

[7] Fuchimoto, K., Minato, S.-i. and Ueno, M.: Automated Test Assmbly using Zero-suppressed Binary Decision DiagramsZero-suppressed Binary Decision Diagrams を用いた自動テスト構成, *Transactions of the Japanese Society for Artificial Intelligence*, Vol. 37 (2022).

[8] Tsutsumi, E., Kinoshita, R. and Ueno, M.: Deep Item Response Theory as a Novel Test Theory Based on Deep Learning, *Electronics*, Vol. 10, No. 9 (2021).

[9] Tsutsumi, E., Kinoshita, R. and Ueno, M.: Deep-IRT with　independent student and item networks, *Educational Data Mining* (2021).

[10] Tsutsumi, E., Guo, Y., Kinoshita, R. and Ueno, M.: Deep Knowledge Tracing Incorporating a Hypernetwork With Independent Student and Item Networks, *IEEE Trans. Learn. Technol.*, Vol. 17, p. 951–965 (2024).

[11] Tsutsumi, E., Nishio, T. and Ueno, M.: Deep-IRT with a Temporal Convolutional Network for Reflecting Students' Long-Term History of Ability Data, pp. 250–264 (2024).

[12] Wang, C., Xu, G., Shang, Z. and Kuncel, N.: Detecting Aberrant Behavior and Item Preknowledge: A Comparison of Mixture Modeling Method and Residual Method, *Journal of Educational and Behavioral Statistics*, Vol. 43, No. 4, pp. 469–501 (2018).

[13] Ruipérez-Valiente, J. A., Merino, P., Alexandron, G. and Pritchard, D.: Using Machine Learning to Detect 'Multiple-Account' Cheating and Analyze the Influence of Student and Problem Features, *IEEE Transactions on Learning Technologies*, Vol. PP, pp. 1–1 (2017).

[14] Man, K. and Harring, J. R.: Assessing Preknowledge Cheating via Innovative Measures: A Multiple-Group Analysis of Jointly Modeling Item Responses, Response Times, and Visual Fixation Counts, *Educational and Psychological Measurement*, Vol. 81, No. 3, pp. 441–465 (2021).

[15] van der Linden, W. J.: Predictive Control of Speededness in Adaptive Testing, *Applied Psychological Measurement*, Vol. 33, No. 1, pp. 25–41 (2009).

[16] {van der Linden}, W. and Xiong, X.: Speededness and adaptive testing, *Journal of educational*

and behavioral statistics, Vol. 38, No. 4, pp. 418–438 (2013).

[17] Linden, W.: A Hierarchical Framework for Modeling Speed and Accuracy on Test Items, *Psychometrika*, Vol. 72, pp. 287–308 (2007).

[18] : 9 - Timed Testing: An Approach Using Item Response Theory, *New Horizons in Testing* (WEISS, D. J., ed.), Academic Press, San Diego, pp. 179–203 (1983).

[19] Verhelst, N. D., Verstralen, H. H. F. M. and Jansen, M. G. H.: *A Logistic Model for Time-Limit Tests*, pp. 169–185, Springer New York (1997).

[20] Roskam, E. E.: *Models for Speed and Time-Limit Tests*, pp. 187–208, Springer New York (1997).

[21] : Specifically objective stochastic latency mechanisms, *Journal of Mathematical Psychology*, Vol. 19, No. 1, pp. 18–38 (1979).

[22] Maris, E.: Additive and Multiplicative Models for Gamma Distributed Random Variables, and their Application as Psychometric Models for Response Times, *Psychometrika*, Vol. 58, No. 3, p. 445–469 (1993).

[23] Rouder, J. N., Jun, L., Paul, S., DongChu, S. and Yi, J.: A hierarchical model for estimating response time distributions, Vol. 12, No. 2 (2005).

[24] Linden, W.: A Lognormal Model for Response Times on Test Items, *Journal of Educational and Behavioral Statistics - J EDUC BEHAV STAT*, Vol. 31, pp. 181–204 (2006).

[25] 石川文弥，渕本壱真，植野真臣
：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-IRT，人工知能学会全国大会論文集 第 39 回 (2025)，一般社団法人 人工知能学会，pp. 4P3GS1004–4P3GS1004 (2025).

[26] Ma, J., Zhao, Z., Yi, X., Chen, J., Hong, L. and Chi, E. H.: Modeling task relationships in multi-task learning with multi-gate mixture-of-experts, *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1930–1939 (2018).

[27] Peng, S., Yuan, K., Gao, L. and Tang, Z.: Mathbert: A pre-trained model for mathematical formula understanding, *arXiv preprint arXiv:2105.00377* (2021).

[28] Van der Linden, W. J. and van der Linden, W.: *Handbook of item response theory*, Vol. 1, CRC press New York (2016).

[29] Van der Linden, W. J.: A lognormal model for response times on test items, *Journal of Educational and Behavioral Statistics*, Vol. 31, No. 2, pp. 181–204 (2006).

[30] Gelfand, A. E. and Smith, A. F. M.: Sampling-Based Approaches to Calculating Marginal Densities, *Journal of the American Statistical Association*, Vol. 85, pp. 398–409 (1990).

[31] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding, *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186 (2019).