

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	岸田 若葉	学籍番号	2331047
論 文 題 目	項目解答時間における Deep-IRT		
要 旨 e-テストの普及に伴い、新たに受検者の解答時間データを収集できるようになり、その活用が求められている。解答時間のモデルに関する研究には、主に項目反応理論(Item Response Theory: IRT) が用いられている。IRT とは確率モデルを用いて受検者の能力と問題項目特性から正誤反応を予測するテスト理論である。代表的な解答時間の分析のための数理モデルである階層モデルは、IRT を基に受検者および項目の時間特性パラメータを推定するため、解釈性をもつ。これにより、受検者ごとに異なる項目セットを解いた場合でも、受検者・項目の時間特性をそれぞれ同一尺度上で比較できるという利点がある。しかし、一般に確率モデル手法よりも深層学習手法の方が予測精度は高いことが知られており、階層モデルによる解答時間の予測精度は限定的であるかもしれない。 一方、アダプティブ・ラーニング分野では、深層学習と IRT を組み合わせることにより、高精度な正誤反応予測とパラメータ解釈性を両立する Deep-IRT モデルが開発されている。Tsutsumi らの Deep-IRT は深層学習による高精度な正誤反応予測に加え、学習者の能力値を示す学習者ネットワークと項目の難易度を示す項目ネットワークの二つの独立なネットワークにより、IRT と同様に解釈できるパラメータ推定を実現している。 そこで、本論では、解釈性を有しつつ高精度に解答時間を予測するため、Tsutsumi らの Deep-IRT を基に、受検者および項目に関する時間特性パラメータをもった項目解答時間における Deep-IRT を新たに提案する。具体的には、独立した二つの受検者ネットワークと項目ネットワークからそれぞれ算出した、受検者の解答速度パラメータと項目の所要時間パラメータから解答時間を予測する。評価実験では、提案手法が推定するパラメータが受検者および項目の時間特性として解釈できることを示す。さらに、提案手法と先行研究の未出題項目への解答時間予測精度を比較し、提案手法の有効性を示す。最後に、提案手法を用いて学力試験の実データに対し時間特性分析を実施する。			

令和6年度 修士論文

項目解答時間における Deep-IRT

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 情報数理工学プログラム

学籍番号 2331047

岸田 若葉

主任指導教員: 植野 真臣

指導教員: 宇都 雅輝

提出年月日: 2025年2月13日

目次

1	まえがき	2
2	項目反応理論	4
3	解答時間に関する項目反応理論	5
3.1	解答時間を考慮した項目反応理論	5
3.2	項目への解答時間のみをモデル化した手法	5
3.3	反応と解答時間を同時に表現する階層モデル	6
3.4	EBLUP による解答時間予測手法	7
4	提案手法	8
5	実データへの適応	11
5.1	UEC 検定データ	11
5.2	パラメータ推定	12
5.3	解答時間予測の評価	15
5.4	所要時間パラメータと項目難易度の分析	18
5.5	解答速度パラメータと能力値の分析	20
6	むすび	21

1 まえがき

近年、e-ラーニングシステムやe-テストシステムの普及に伴い、膨大な教育ビッグデータを収集できるようになった。教育ビッグデータには学習者の解答履歴データや解答時間データなどが含まれており、これらのデータの活用方法の研究が急務となっている。本論では問題項目への解答時間データに注目し、新しい数理モデル「項目解答時間における Deep-IRT」を提案する。

解答時間のモデルに関する研究には、主に項目反応理論 (Item Response Theory: IRT) が用いられている。ここで、IRT とは確率モデルを用いて受検者の能力と問題項目特性から正誤反応を予測できるテスト理論であり、CBT (Computer Based Testing) で広く用いられている [1]。IRT の解答時間モデルには3つのアプローチがある。

1. 従来の正誤反応における IRT に解答時間に関するパラメータを追加し、解答時間を考慮した反応予測を行う手法 [2-4]
2. 反応とは独立に項目への解答時間のみをモデル化する手法 [5, 6].
3. 反応と解答時間を同時にモデル化する手法 [7-9]

特に代表的な手法として、van der Linden らが提案した対数正規分布モデル [5] と Fox らが提案した階層モデル [7, 8] が挙げられる。van der Linden らの対数正規分布モデルは、過去に取り組んだ項目への解答時間を目的変数とすることで、IRT における受検者の能力パラメータと難易度パラメータを、それぞれ対数正規分布で表される解答速度パラメータと所要時間パラメータとして解釈するモデルである。Fox らの階層モデルは、受検者の能力値と解答速度、項目の難易度と時間特性は相互に影響すると仮定されたモデルである。具体的には、1 層目では IRT に基づいた反応予測モデルと解答時間モデルをそれぞれ構成し、2 層目では 1 層目で用いたパラメータ群をそれぞれ受検者と項目に関連するものに分類し、独立に多変量正規分布を仮定した。

これらのモデルは、受検者の解答速度パラメータと項目の所要時間パラメータをもち、受検者・項目ごとに異なる時間特性を表現できる。これにより、受検者ごとに異なる項目セットを解いた場合でも、受検者・項目の時間特性をそれぞれ同一尺度上で比較できる。

解答時間の予測精度向上のために、最新の研究では、EBLUP (Empirical Best Linear Unbiased Predictor) という確率モデルによる機械学習を用いた手法が提案されている [10]。この手法は未出題項目への解答時間を予測することに特化しており、短時間で高精度に解答時間を予測できる。しかし、受検者および項目の時間特性を表現するパラメー

タをもたないため、解釈性が低い。

一方、近年、アダプティブ・ラーニング分野では深層学習手法と IRT を組み合わせることにより、高精度な正誤反応予測とパラメータの解釈性を両立する Deep-IRT モデルが開発されている [11–17]。Tsutsumi らの Deep-IRT の特徴は、深層学習による高精度な反応予測に加え、学習者の能力値を推定する学習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の独立な二つのニューラルネットワークでパラメータ推定を行うことで、従来の IRT で推定される学習者の能力値と項目の難易度と同等のパラメータ解釈性を示すことである [11–17]。

本論では、Deep-IRT モデル [13] にインスパイアされ、解釈性を有しつつ高精度に解答時間を予測することを目的に、受検者および項目に関する時間特性パラメータを持った項目解答時間における Deep-IRT を新たに提案する。具体的には、独立した二つの受検者ネットワークと項目ネットワークからそれぞれ算出した、受検者の解答速度パラメータと項目の所要時間パラメータから解答時間を予測する。評価実験では、提案手法で推定した受検者の解答速度パラメータおよび項目の所要時間パラメータと、実際の解答時間データを比較し、パラメータの解釈性を示す。さらに、提案手法と先行研究の未出題項目への解答時間予測精度を比較し、提案手法の有効性を示す。最後に、提案手法を用いて学力試験の実データに対し時間特性分析を実施する。

2 項目反応理論

本章では、最もよく知られたテスト理論である項目反応理論について述べる。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、受検者の項目への正答確率をモデル化したものである [1]。これにより、異なる項目への受検者の反応を同一尺度上で評価できる。

よく用いられる IRT モデルとして 3 母数ロジスティックモデル (3-Parameter Logistic Model: 3PLM) が知られている。3PLM では、多肢選択式の問題において能力値の低い受検者が偶然正答する可能性を考慮して、能力値 $\theta \in (-\infty, \infty)$ の受検者が項目 $i \in \{1, \dots, N\}$ に正答する確率を次の式で示す。

$$p(u_i = 1|\theta) = c_i + \frac{1 - c_i}{1 + \exp(-1.7a_i(\theta - b_i))} \quad (1)$$

ここで、 u_i は受検者が項目 i に正答するとき 1、それ以外の場合は 0 である変数である。また、 $a_i \in [0, \infty)$, $b_i \in (-\infty, \infty)$, $c_i \in [0, 1]$ はそれぞれ項目 i の識別力パラメータ、難易度パラメータ、当て推量パラメータである。問題形式が自由記述形式であるなど、受検者が偶然正答する可能性を考慮しない場合 ($c_i = 0$)、2 母数ロジスティックモデル (2-Parameter Logistic Model: 2PLM) と呼ぶ。

他の IRT モデルとして、大問項目を扱う一般化部分採点モデル (Generalized Partial Credit Model:GPCM) がある。GPCM では能力値 θ の受検者が k 個の小問に正答する確率を次の式で示す。

$$p(k|\theta) = \frac{\exp(\sum_{m=1}^k \alpha_i(\theta - \beta_{im}))}{\sum_{l=1}^K [\exp(\sum_{m=1}^l \alpha_i(\theta - \beta_{im}))]} \quad (2)$$

ここで、 β_{im} は項目 i の小問正答数が $m - 1$ から m に遷移する難易度を示す。

3 解答時間に関する項目反応理論

本章では、IRT を基に発展した解答時間に関するモデルを紹介する。

3.1 解答時間を考慮した項目反応理論

Roskam(1987,1997) では、IRT に解答時間に関するパラメータを追加し、解答時間を考慮した反応予測を行う手法を提案している [2,3]。具体的には、解答時間が増加すると正答確率が低くなるように、IRT に解答時間データを追加した [2,3]。

$$p_i(\theta_j) = \frac{1}{1 + \exp(-\theta_j + \ln t_{ij} + b_i)} \quad (3)$$

ここで、 θ_j は受検者 j の能力値、 b_i は項目 i の難易度、 t_{ij} は項目への反応（正答・誤答）にかかわらず、受検者 j が項目 i の解答に要した時間である。この手法は解答時間を考慮しているものの、受検者固有の解答速度や項目ごとの時間特性は表現していない。また、解答時間そのものを解答速度とみなすならば、全ての受検者が同一の項目群を受験する必要があると指摘されている [8]。

3.2 項目への解答時間のみをモデル化した手法

Roskam(1987,1997) の問題を解決するために、van der Linden(2006) により受検者の解答時間を対数正規分布を用いてモデリングする手法が提案されている [5]。過去に取り組んだ項目への解答時間を目的変数とすることで、IRT における受検者の能力パラメータと難易度パラメータを、それぞれ対数正規分布に従う解答速度パラメータと所要時間パラメータとして解釈できる。受検者の解答時間分布は次式で定義される。

$$f(t_{ij}; \tau_j, \alpha_i, \beta_i) = \frac{\alpha_i}{t_{ij}\sqrt{2\pi}} \left\{ -\frac{1}{2} [\alpha_i(\ln t_{ij} - (\beta_i - \tau_j))]^2 \right\} \quad (4)$$

ここで、 τ_j は受検者 j の解答速度パラメータ、 α_i 、 β_i はそれぞれ項目 i の時間識別力パラメータと所要時間パラメータである。このモデルは項目ごとに時間識別力パラメータと所要時間パラメータを推定することで、項目の時間特性を表現できる。これにより、受検者ごとに異なる項目セットを解いた場合でも、受検者・項目の時間特性をそれぞれ同一尺度上で比較できるという利点をもつ。一方で、受検者の反応と解答時間の関係を表現できていないという問題があった。

3.3 反応と解答時間を同時に表現する階層モデル

Fox らは、受検者の能力値と解答速度、項目の難易度と時間特性は相互に影響すると考え、受検者の反応と解答時間の関係を同時に表現する手法として階層モデル手法を提案した [7, 8]. 1 層目では IRT に基づいた反応予測モデルと解答時間モデルをそれぞれ独立に考える. 具体的には、反応データ $\mathbf{u}$ には (1) 式より定義される IRT, 解答時間データ $\mathbf{t}$ は (4) 式より定義される対数正規分布を用いてそれぞれモデリングする.

$$\mathbf{u} \sim f(u_{ij}; \theta_j, a_i, b_i, c_i), \mathbf{t} \sim f(t_{ij}; \tau_j, \alpha_i, \beta_i) \quad (5)$$

ここで、受検者 j に関するパラメータ群を $\xi_j = (\theta_j, \tau_j)$, 項目 i に関するパラメータ群を $\psi_i = (a_i, b_i, c_i, \alpha_i, \beta_i)$ と表記すると、1 層目は次式で表すことができる.

$$f(\mathbf{u}_j, \mathbf{t}_j; \xi_j, \psi_i) = \prod_{i=1}^I f(u_{ij}; \theta_j, a_i, b_i, c_i) f(t_{ij}; \tau_j, \alpha_i, \beta_i) \quad (6)$$

次に、2 層目では受検者 j に関するパラメータ群 ξ_j , 項目 i に関するパラメータ群 ψ_i の関係性をそれぞれ独立に多変量正規分布を用いてモデリングする. パラメータ群 ξ_j の関係性を表現するモデル $\mathcal{P}$ は次式より定義される.

$$f(\xi_j; \mu_{\mathcal{P}}, \Sigma_{\mathcal{P}}) = \frac{|\Sigma_{\mathcal{P}}^{-1}|^{1/2}}{2\pi} \exp\left[-\frac{1}{2}(\xi_j - \mu_{\mathcal{P}})^T \Sigma_{\mathcal{P}}^{-1}(\xi_j - \mu_{\mathcal{P}})\right] \quad (7)$$

ここで、

$$\mu_{\mathcal{P}} = (\mu_{\theta}, \mu_{\tau}), \quad (8)$$

$$\Sigma_{\mathcal{P}} = \begin{pmatrix} \sigma_{\theta}^2 & \sigma_{\theta\tau} \\ \sigma_{\theta\tau} & \sigma_{\tau}^2 \end{pmatrix}. \quad (9)$$

である. パラメータ群 ψ_i の関係性を表現するモデル $\mathcal{I}$ は次式より定義される.

$$f(\psi_j; \mu_{\mathcal{I}}, \Sigma_{\mathcal{I}}) = \frac{|\Sigma_{\mathcal{I}}^{-1}|^{1/2}}{(2\pi)^{5/2}} \exp\left[-\frac{1}{2}(\psi_i - \mu_{\mathcal{I}})^T \Sigma_{\mathcal{I}}^{-1}(\psi_i - \mu_{\mathcal{I}})\right] \quad (10)$$

ここで、

$$\mu_{\mathcal{I}} = (\mu_a, \mu_b, \mu_c, \mu_{\alpha}, \mu_{\beta}), \quad (11)$$

$$\Sigma_{\mathcal{P}} = \begin{pmatrix} \sigma_a^2 & \sigma_{ab} & \sigma_{ac} & \sigma_{a\alpha} & \sigma_{a\beta} \\ \sigma_{ba} & \sigma_b^2 & \sigma_{bc} & \sigma_{b\alpha} & \sigma_{b\beta} \\ \sigma_{ca} & \sigma_{cb} & \sigma_c^2 & \sigma_{c\alpha} & \sigma_{c\beta} \\ \sigma_{\alpha a} & \sigma_{\alpha b} & \sigma_{\alpha c} & \sigma_{\alpha}^2 & \sigma_{\alpha\beta} \\ \sigma_{\beta a} & \sigma_{\beta b} & \sigma_{\beta c} & \sigma_{\beta\alpha} & \sigma_{\beta}^2 \end{pmatrix}. \quad (12)$$

である。したがって、モデル全体の確率分布は次式より表される。

$$f(\mathbf{u}, \mathbf{t}; \boldsymbol{\xi}, \boldsymbol{\psi}) = \prod_{j=1}^J \prod_{i=1}^I f(\mathbf{u}_j, \mathbf{t}_j; \boldsymbol{\xi}_j, \boldsymbol{\psi}_i) f(\boldsymbol{\xi}_j; \boldsymbol{\mu}_P, \boldsymbol{\Sigma}_P) f(\boldsymbol{\psi}_j; \boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I) \quad (13)$$

この手法は、受検者・項目ごとに異なる時間特性を表現できることに加え、2層目で反応と解答時間の関係をモデルに組み込んでいる。

3.4 EBLUP による解答時間予測手法

解答時間の予測精度向上のために、Peng(2024) は EBLUP(Empirical Best Linear Unbiased Predictor) という確率モデルによる機械学習アプローチを用いて解答時間のみを予測する手法を提案した。Peng(2024) では線形混合効果モデル (Linear Mixed-Effects Model:LMM) を用いて、受検者 j の項目 i への解答時間の対数 T_{ij} を次式で求める。

$$T_{ij} = \mu + X_i \beta + b_i + e_{ij} \quad (14)$$

ここで、 μ は従属変数 T_{ij} の期待値、 $X_i \in \mathbb{R}^{1 \times p}$ は受検者 j の項目 i への正答率、 $\beta \in \mathbb{R}^{p \times 1}$ はその係数、 $b_i \sim N(0, \sigma_b^2)$ は項目 i のランダム効果、 $e_{ij} \sim N(0, \sigma_e^2)$ はランダムエラーである。EBLUP では従属変数 T_i を制限付き最尤法を用いて次式より推定する。

$$\hat{T}_i^{EBLUP}(\sigma_b^2, \sigma_e^2) = (1 - \hat{\gamma}_i)[\hat{\mu} + \bar{X}_i \hat{\beta}] + \hat{\gamma}_i \bar{T}_i \quad (15)$$

ただし、 $\bar{X}_i$ 、 $\bar{T}_i$ はそれぞれ項目 i に関する平均正答率、平均解答時間であり、 $\hat{\gamma} = \hat{\sigma}_b^2(\hat{\sigma}_b^2 + \hat{\sigma}_e^2 n_i^{-1})^{-1}$ である。ここで、 T_i は項目 i に受検者の 95% が解答を終える時間と定義されている。EBLUP による最終的な推定値は、次式より求められる。

$$\hat{T}_i^{EBLUP'} = \hat{T}_i^{EBLUP} + z \times \hat{\sigma}_{T_i} \quad (16)$$

ここで、 z は十分割交差検証を用いて最適化されるハイパーパラメータ、 $\hat{\sigma}_{T_i}$ は T_i の標準偏差である。

EBLUP を用いた解答時間予測は、特に項目ごとの解答数が少ない場合に有効であることが示されている [10]。また、他の手法と比較して、解答速度パラメータや所要時間パラメータなどを推定する必要がないため、短時間で高精度に解答時間を予測できる。しかし、受検者および項目に関する時間特性を表現するパラメータをもたないため、解釈性が低く、時間特性分析を実施できない。

4 提案手法

近年、アダプティブ・ラーニング分野では深層学習手法と IRT を組み合わせることにより、高精度な反応予測とパラメータの解釈性を両立する Deep-IRT モデルが提案されている [11–17]. 一般的に深層学習を用いた予測手法は、高い予測精度を示す一方、学習者の能力値や項目の難易度のような解釈できるパラメータをもたない. Tsutsumi らの Deep-IRT [11–17] では学習者の能力値を表現する学習者ネットワークと項目の難易度を表す項目ネットワークの二つの独立な多層のニューラルネットワークで構成することにより、IRT と同様に解釈できるパラメータ推定を実現している. さらに、従来の IRT では受検者間の関係性を推定していなかったが、Tsutsumi らの Deep-IRT [11–17] では受検者間の関係性を考慮した予測を行うことができるため、従来の IRT よりも高い反応予測精度を示す.

そこで本研究では、時間特性と解釈できるパラメータを推定しつつ高精度に解答時間を予測するために、Tsutsumi らの Deep-IRT [13] を基に項目解答時間のための新たな Deep-IRT を提案する. 提案手法は、受検者の解答速度を表す受検者ネットワークと項目への解答所要時間を表す項目ネットワークの二つの独立なネットワークによって構成される. それぞれのネットワークの出力から受検者の項目への反応時間を予測する. 提案手法の構造を図 1 に示す.

受検者ネットワークでは j 番目の受検者を表現する、 j 番目の要素のみが 1、その他の要素が 0 の one-hot vector $\mathbf{s}_j \in \mathbb{R}^J$ を入力とし、次のように n 層のニューラルネットワークを構成する.

$$\boldsymbol{\tau}_1^{(j)} = \tanh(\mathbf{W}^{(\tau_1)} \mathbf{s}_j + \mathbf{b}^{(\tau_1)}), \quad (17)$$

$$\boldsymbol{\tau}_k^{(j)} = \tanh(\mathbf{W}^{(\tau_k)} \boldsymbol{\tau}_{k-1}^{(j)} + \mathbf{b}^{(\tau_k)}), \quad (18)$$

$$\tau_n^{(j)} = \mathbf{W}^{(\tau_n)} \boldsymbol{\tau}_{n-1}^{(j)} + b^{(\tau_n)}. \quad (19)$$

ここで、 $\tau_n^{(j)}$ を受検者 j の解答速度パラメータとみなす. 本研究では、項目ネットワークへの入力手法として次の 2 つを検証する.

1. i 番目の項目を表現する、 i 番目の要素のみが 1、その他の要素が 0 の one-hot vector $\mathbf{q}_i \in \mathbb{R}^I$ を入力とする.
2. 本研究で用いるデータセットでは、受検者の反応 $\mathbf{u}$ によって解答時間の分布が大きく異なるため、正答データと誤答データを区別する. 具体的には次式により定義

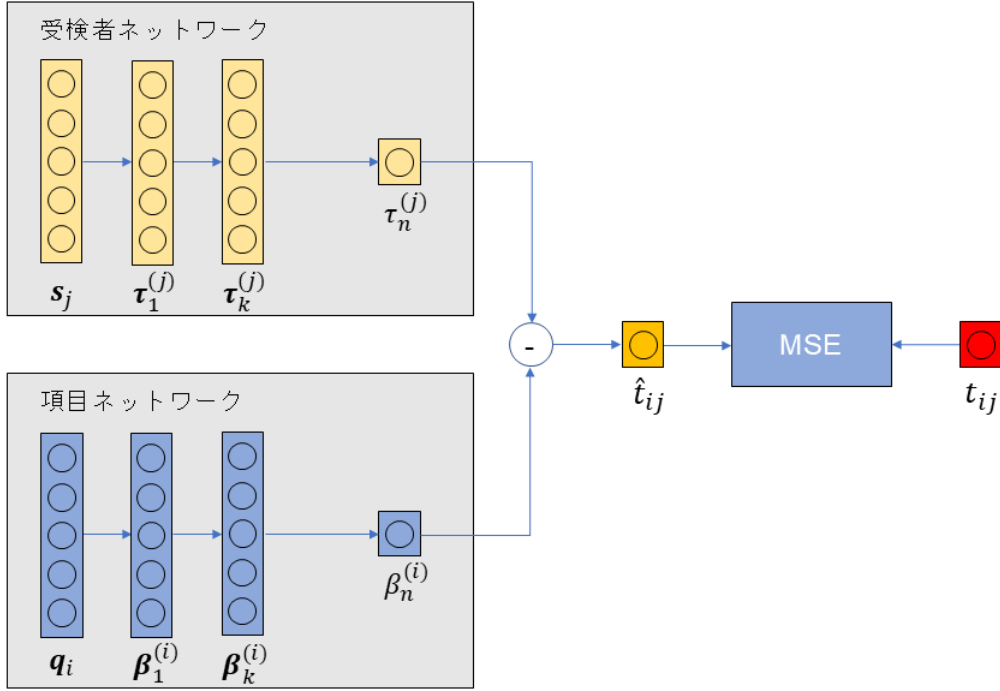


図 1: モデル図

されたベクトル $\mathbf{q}_i \in \mathbb{R}^{2I}$ を入力する.

$$\mathbf{q}_i = \begin{cases} [\mathbf{c}_i, \mathbf{0}] & (u_{ij} = 1) \\ [\mathbf{0}, \mathbf{c}_i] & (u_{ij} = 0) \end{cases} \quad (20)$$

ここで, $\mathbf{c}_i$ は i 番目の要素のみが 1, その他の要素が 0 の one-hot vector である. また, u_{ij} は受検者の反応を表し, 受検者 j が項目 i に正答したとき 1, 誤答したとき 0 である.

このベクトル $\mathbf{q}_i$ を入力とし, 次のような n 層のニューラルネットワークを構成する.

$$\beta_1^{(i)} = \tanh(\mathbf{W}^{(\beta_1)} \mathbf{q}_i + \mathbf{b}^{(\beta_1)}), \quad (21)$$

$$\beta_k^{(i)} = \tanh(\mathbf{W}^{(\beta_k)} \beta_{k-1}^{(i)} + \mathbf{b}^{(\beta_k)}), \quad (22)$$

$$\beta_n^{(i)} = \mathbf{W}^{(\beta_n)} \beta_{n-1}^{(i)} + \mathbf{b}^{(\beta_n)}. \quad (23)$$

ここで出力された $\beta_n^{(i)}$ を項目 j の所要時間パラメータとみなす. 次に, 受検者の解答速度パラメータと項目の所要時間パラメータの差より受検者 i の項目 j への解答時間 $\hat{t}_{ij}$ を

算出する．

$$\hat{t}_{ij} = W(-\tau_n^{(j)} + \beta_n^{(i)}) + b \quad (24)$$

本手法では，平均二乗誤差 (Mean Squared Error: MSE) を損失関数 $loss$ とし，誤算逆伝播法を用いてパラメータの学習を行う．実際の解答時間 t_{ij} と予測解答時間 $\hat{t}_{ij}$ から次の式で MSE を計算する．

$$loss = \frac{1}{I \times J} \sum_I \sum_J (t_{ij} - \hat{t}_{ij})^2 \quad (25)$$

この手法は，独立した 2 つのネットワークにより受検者の解答速度パラメータと項目の所要時間パラメータを推定することで，受検者ごとに異なる項目セットを解いた場合でも，受検者および項目の時間特性を同一尺度上で比較できる．この利点から，提案手法により推定された解答速度パラメータおよび所要時間パラメータを用いて，実際の試験結果に対し時間特性分析を実施できる．さらに，IRT に深層学習手法を組み合わせることで，高精度に解答時間を予測できる．

5 実データへの適応

本章では，提案手法が推定するパラメータが受検者および項目の時間特性として解釈できることを示す．そして，実際に行われた学力試験である UEC 検定 [18] のデータを用いて，提案手法により解答時間を予測し，他の先行研究と予測精度を比較する．さらに，UEC 検定データに対して提案手法を用いて受検者および項目に関する時間特性分析を行う．

5.1 UEC 検定データ

本節では，本研究で利用した UEC 検定データの概要を説明する．UEC 検定は，電気通信大学の入学者を対象に行われた基礎学力試験である．試験は CBT 形式で実施されており，CBT システムによって受検者の各項目への反応（正答・誤答）および解答時間が記録される．項目は自由記述式，多肢選択式，大問式の 3 つの問題形式で構成されている．自由記述式はキーボードを用いて枠内に解答を入力する形式，多肢選択式は 2 つ以上の選択肢から正しい選択枝を選択する形式，大問式は複数の自由記述式および多肢選択式の項目から構成される形式である．

表 1: 問題形式と推定に用いた IRT の対応

問題形式	IRT モデル
自由記述式	2PLM
多肢選択式	3PLM
大問式	GPCM

表 2: 科目（データセット）ごとの基礎統計

科目（データセット）	情報	数学	物理
受検者数	62	666	317
項目数	284	100	134
平均正答率	0.57	0.71	0.43
平均解答時間	61.34	160.74	74.93
項目あたりの平均解答数	23.10	204.70	126.70

UEC 検定は、受検者の所属するクラスごとに出題された項目セットが異なる。また、不正行為防止のため、受検者ごとに異なる順序で項目が出題されており、試験中受検者は自由に解答順序を変更できる。試験終了後に、収集した受検者の反応データから IRT を用いて受検者の能力値および項目のパラメータを推定し、パラメータ推定値を受検者にフィードバックする。このとき、表 1 に従って問題形式ごとに異なる IRT モデルを用いてパラメータ推定が行われる。

また、UEC 検定の実施科目は情報、数学、物理であり、それぞれの試験結果の基礎統計を表 2 にまとめた。以降、情報、数学、物理の試験にて収集された反応および解答時間データをそれぞれデータセット情報、数学、物理と呼ぶ。ただし、各項目の解答時間データについて、正答データの最小解答時間よりも短い時間で誤答したデータは、受検者が正常に解答に取り組んでいないとみなして削除した。

5.2 パラメータ推定

本節では、提案手法が推定する受検者の解答速度パラメータおよび項目の所要時間パラメータが、それぞれ受検者および項目の時間特性と解釈できることを示す。

まず、データセット情報、数学、物理について提案手法を用いて解答速度パラメータ、所要時間パラメータを推定した。解答速度パラメータ、所要時間パラメータと平均解答時間の関係性を示す。図 2a, 2b, 2c は各データセットから推定した解答速度パラメータと受検者ごとの平均解答時間の散布図である。ただし、受検者ごとの平均解答時間の上位下位 1% にあたるデータは削除した。図 2a, 2b, 2c より、解答速度パラメータが大きくなればなるほど、平均解答時間が小さくなる傾向があることがわかる。つまり、平均解答時間が短い受検者ほど解答速度パラメータは大きい値で推定される傾向がある。

また、所要時間パラメータと項目ごとの平均解答時間の散布図を図 3a, 3b, 3c に示す。ただし、図 2a, 2b, 2c と同様に、項目ごとの平均解答時間の上位下位 1% にあたるデータは削除した。図 3a, 3b, 3c より、所要時間パラメータが大きくなればなるほど、平均解答時間が大きくなる傾向があることがわかる。つまり、平均解答時間が長い項目ほど所要時間パラメータは大きな値で推定される傾向がある。

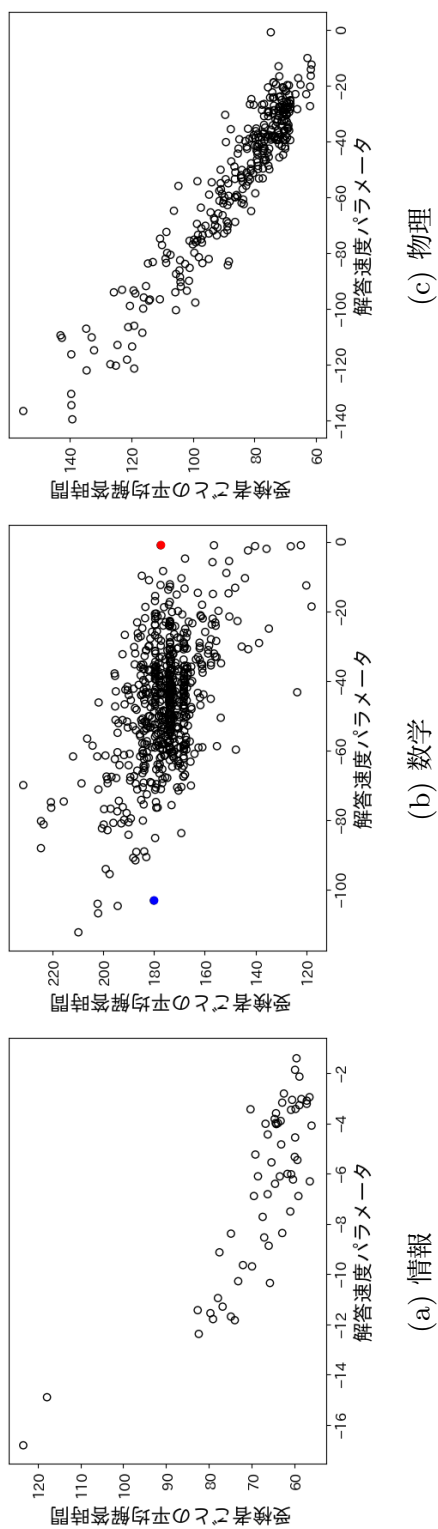


図 2: 解答速度パラメータと受検者ごとの平均解答時間の散布図

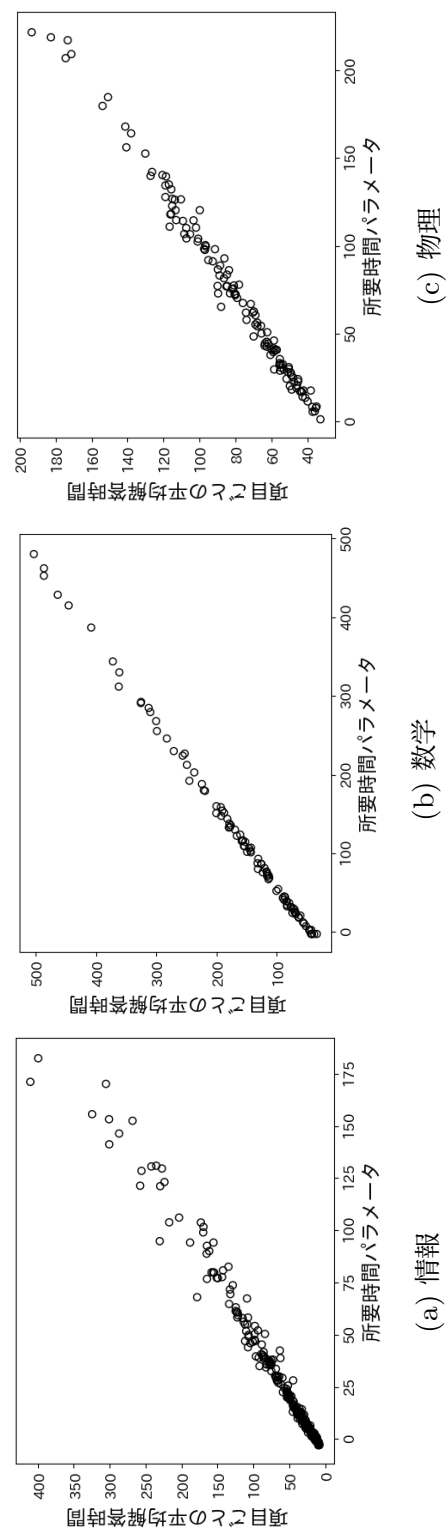


図 3: 所要時間パラメータと項目ごとの平均解答時間の散布図

式 (24) は重み W の要素が正である場合、解答速度パラメータが大きくなればなるほど解答時間が短くなり、所要時間パラメータが大きくなればなるほど解答時間が長くなるという関係性を示している。散布図より、この式 (24) の示す通り解答速度パラメータと所要時間パラメータが推定されていることがわかる。

特に、図 3a, 3b, 3c はそれぞれ相関係数が 0.94, 0.99, 0.98 であり、項目ごとの平均解答時間と所要時間パラメータの間には非常に強い相関がある。この理由は、同一項目に対する受検者間の解答時間の分散が小さかったからであると考えられる。

一方、図 2a, 2b, 2c はそれぞれ相関係数が -0.80 , -0.51 , -0.94 であり、受検者ごとの平均解答時間と解答速度パラメータの間の相関は、項目ごとの平均解答時間と所要時間パラメータ間の相関よりも弱い。実際に、平均解答時間が同等であるが解答速度パラメータが大きく異なる受検者が存在する。その一例として、図 2b 中の赤いプロットと青いプロットで示された受検者に注目する。赤いプロットを受検者 A、青いプロットを受検者

表 3: データセット数学より着目した 2 点について

受検者名	平均解答時間	解答速度パラメータ
受検者 A	177.30	-0.68
受検者 B	180.00	-102.95

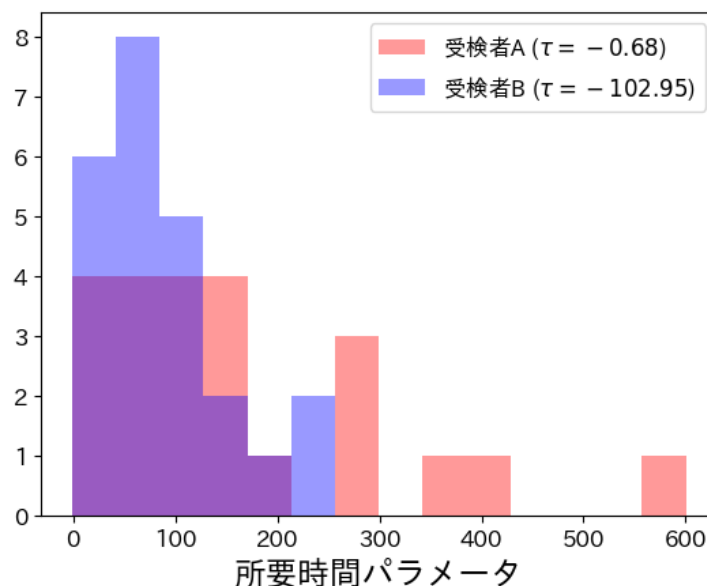


図 4: 受検者 A, B への出題項目の所要時間パラメータのヒストグラム

Bとして、受検者 A, B の平均解答時間と解答速度パラメータの値を表 3 に示す。そして、受検者 A, B への出題項目の所要時間パラメータのヒストグラムを図 4 に示す。図 4 より、受検者 A は所要時間パラメータが小さい項目から大きい項目まで出題されていることがわかる。一方、受検者 B は所要時間パラメータが小さい項目に偏って出題されている。すなわち、受検者 A は受検者 B よりも長い解答時間を要する項目が多く出題されている。それにも関わらず、受検者 A, B の平均解答時間は同等であるため、受検者 A の解答速度は受検者 B よりも速い。よって、受検者 A の解答速度パラメータは受検者 B より大きな値で推定されたと考えられる。このことから、提案手法が項目ごとに異なる時間特性を考慮しながら解答時間パラメータを推定したことがわかる。

一部の先行研究では、受検者の解答時間そのものを解答速度として扱っている [2, 3]。それに対し van der Linden は、解答時間そのものを受検者の時間特性として扱う場合、全受験者が同一の項目セットを解く必要があると主張している [8]。しかし、実際の CBT 試験では全受験者に対し同一の項目セットが出題されることは稀であり、そのような状況は常に保証できない。一方、提案手法は項目ごとに異なる時間特性を考慮しながら解答速度パラメータを推定しているため、異なる項目セットを解いた場合でも、受検者・項目の時間特性をそれぞれ同一尺度上で比較できる。これは、提案手法が独立な 2 つのネットワークを用いて解答速度パラメータと所要時間パラメータを推定していることに起因する。よって、提案手法の推定する解答速度パラメータと所要時間パラメータはそれぞれ受検者および項目の時間特性を表現していると解釈できる。

5.3 解答時間予測の評価

本節では、提案手法と解答時間予測を行う先行研究（階層モデル [7, 8], EBLUP [10]) を用いて受検者の項目への解答時間を予測し、精度比較を行う。提案手法は、項目ネットワークへの入力に反応データのエンベディングを用いた場合と用いない場合について検証した。解答時間の予測精度は次式より定義される RMSE(平均二乗偏差) を用いて評価した。

$$RMSE = \sqrt{\frac{1}{I \times J} \sum_I \sum_J (t_{ij} - \hat{t}_{ij})^2} \quad (26)$$

本実験では、UEC 検定のデータセット情報、数学、物理を用いた。ただし、各項目の解答データについて、正答者の最小解答時間よりも短い時間で誤答したデータは受検者が正常に解答に取り組んでいないとみなして削除した。データセットは訓練データ、検証データ、評価データに分割し、訓練データ、検証データより学習したモデルにより、評価デー

タの解答時間予測を行う 10 分割交差検証法を用いた。各手法のパラメータ推定は次のように行った。

階層モデルは、パラメータ推定で用いる MCMC を以下の 2 通りの設定で実験した。尚、初期のサンプリング値は不安定であるため、burn-in の回数より前のサンプルは削除した。

1. サンプリング:1000 回, burn-in: 100 回
2. サンプリング:50000 回, burn-in:20000 回

EBLUP は訓練データ、検証データ、評価データ全てに各項目 1 つ以上の反応および解答時間データが必要であるため、条件を満たすように設定した。また、EBLUP は項目ごとの平均解答時間を予測する手法であるため、本実験では項目に対する平均解答時間の予測値を、その項目に対する各受検者の予測解答時間として扱う。EBLUP におけるハイパーパラメータ γ のチューニングは先行研究 [10] と同様に行った。

提案手法のニューラルネットワークの層数 n は、データセットに合わせてチューニングを行った。提案手法の受検者および項目ネットワークの層数 n を変化させた場合の解答時間の予測精度を図 5 に示す。これらのグラフより、データセット情報では層数 $n = 4$

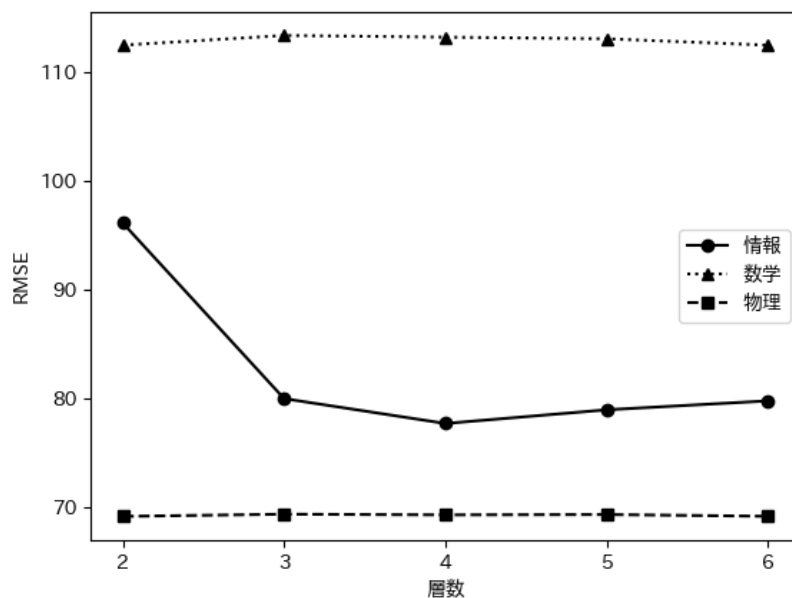


図 5: 層数の変化に伴う提案手法の予測精度の推移

表 4: 実験結果 (RMSE)

データセット	階層モデル (1000)	階層モデル (50000)	EBLUP	Proposal	Proposal + embedding
情報	2013.29	807.89	134.56	77.70	78.38
数学	163.60	162.88	184.45	113.23	113.10
物理	190.41	165.32	125.04	69.30	69.10

で予測精度が最も高く、数学と物理では層数による変化はほとんど見られなかった。したがって、以降の実験は、すべてのデータセットにおいて層数 $n = 4$ で行う。

各手法を用いた解答時間予測精度を表 4 に示す。尚、階層モデルの末尾の括弧内の値は MCMC のサンプリング回数を示す。結果から、全てのデータセットにおいて提案手法が最も高い予測精度を示した。提案手法について、一部のデータセットにおいては、項目ネットワークへの入力に正誤反応データのエンベディングを用いた場合のほうが用いない場合よりも精度がやや良いことがわかる。

階層モデルは、データセット情報、物理では最も予測精度が低い。さらに、データセット情報では、他手法と比較して予測精度が大幅に悪化している。MCMC のサンプリング回数を増加させたとき、全てのデータセットで精度が向上している。しかし、データセット数学、物理では大幅な精度向上はみられず、MCMC のサンプリング回数を 50000 回としたとき十分に値が収束していると考えられる。また、サンプリング回数の増加に伴い情報データセットでは精度が大きく向上しているものの、依然として他の比較手法よりも精度が大幅に低い。そのため、データセット情報は、MCMC のサンプリング回数を増加するとさらに精度が向上するかもしれないが、それに伴いパラメータ推定に要する時間が長くなるため、階層モデルによる解答時間予測は効率が悪いといえる。

EBLUP ではデータセット情報、物理において階層モデルよりも高い予測精度を示した。この理由として、表 2 から分かるようにデータセット情報、物理はデータセット数学に比べて項目あたりのデータ数が小さいことが挙げられる。EBLUP は特に項目への解答数が少ないときに有効な手法であるため、データセット情報、物理における予測精度が向上したと考えられる [10]。

5.4 所要時間パラメータと項目難易度の分析

ここでは、問題形式ごとに、各項目について提案手法で推定した所要時間パラメータとIRTで推定した難易度パラメータの関係性を分析する。項目の難易度パラメータを推定する手法はIRTの他にもある[11–13, 19]が、本研究では実際に受検者へのフィードバックに用いた、IRTで推定した難易度パラメータと比較する。また、1つの項目に対して複数の難易度パラメータをもつGPCMは、1つの項目に対して1つの難易度パラメータをもつ2PLMおよび3PLMと同様には比較できないため、GPCMを用いてパラメータ推定を行った大問式については本分析の対象としない。各データセットについて、自由記述式項目と多肢選択式項目の所要時間パラメータと難易度パラメータの散布図をそれぞれ図6a, 6b, 6c, 図7a, 7b, 7cに示す。尚、問題形式間での傾向を比較するために同一データセットの散布図は所要時間パラメータと難易度パラメータの範囲を揃えている。また、図中の直線は回帰直線を表す。これらのグラフから、データセットによって相関の強さに差はあるものの、自由記述式項目と多肢選択式項目の両方において、所要時間パラメータの増加に伴って難易度パラメータも増加することがわかる。一般に、難易度が高い問題ほど長い解答時間を要する傾向があり[20]、本研究で推定した所要時間パラメータと難易度パラメータの間にも同様の関係性がみられた。また、データセット情報と数学の多肢選択式項目は、自由記述式項目と比較して所要時間パラメータの小さい項目が多くなっている。これは、自由記述式と異なり多肢選択式には選択肢が用意されているため、難易度が高い場合でも受検者が長考せずに当て推量などで解答している可能性が考えられる。

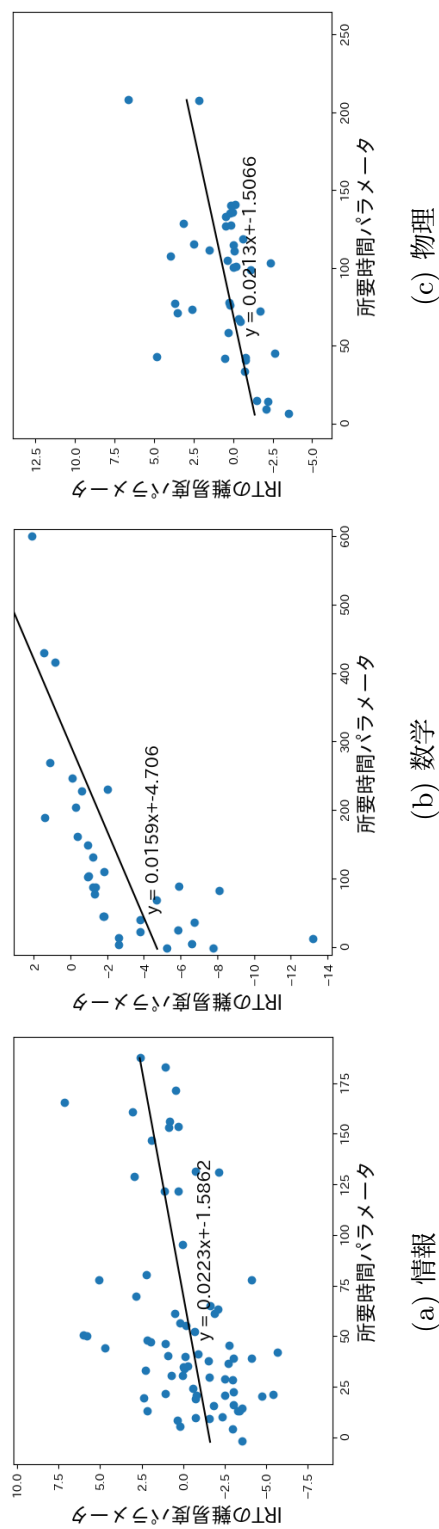


図 6: 自由記述式の項目に関する所要時間パラメータと難易度パラメータの散布図

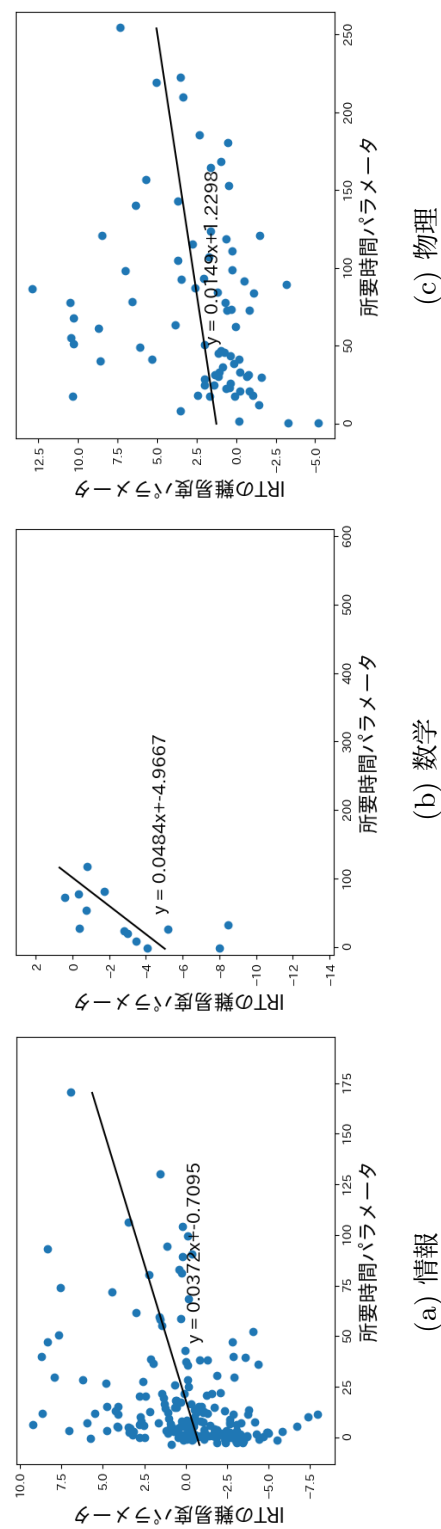


図 7: 多肢選択式の項目に関する所要時間パラメータと難易度パラメータの散布図

5.5 解答速度パラメータと能力値の分析

最後に、提案手法で推定した受検者の解答速度パラメータと IRT で推定した能力値の関係性を分析する。図 8a, 8b, 8c は各データセットより推定した解答速度パラメータと能力値の散布図である。全てのデータセットにおいて、解答速度パラメータの値は能力値によらず、同一能力値の受検者であっても解答速度は異なることがわかる。したがって、受検者の解答速度は従来の IRT で推定される能力値から判断することはできず、能力値とは独立に解答時間データから推定する必要があると言える。

能力値に従わない解答速度パラメータ推定は、適応型テストに応用することができる。現在の適応型テストでは、解答者の能力値に適した項目を出題するため、同一能力の受検者には同一難易度の項目を出題する傾向にある。そのため、制限時間下において、同一能力値の受検者でも解答速度が低い受検者が高スコアを得るためには解答速度を上げる必要がある。しかし、一般に個人の解答速度とスコアにはトレードオフの関係があると指摘されており [21]、解答速度を上げた受検者は真の能力値よりも低い能力値で推定されてしまう。そのため、受検者および項目の時間特性を考慮しながら出題することで、テストの公平性を保つことができると考えられる。

以上のように、提案手法を用いることで、従来の紙面に解答するテスト形式では捉えられなかった受検者および項目に関する時間特性を分析できた。

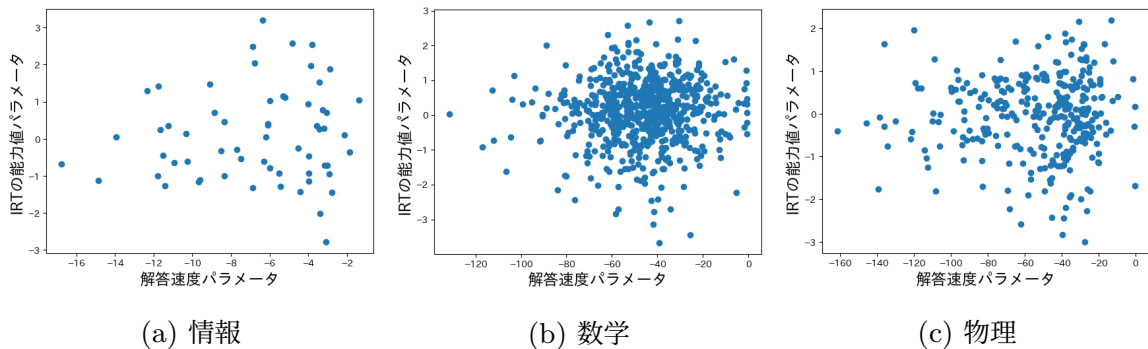


図 8: 解答速度パラメータと能力値の散布図

6 むすび

本研究では、時間特性として解釈可能なパラメータを推定しつつ高精度に解答時間を予測するために、Tsutsumi らの Deep-IRT [13] を基に項目解答時間のための新たな Deep-IRT を提案した。実データを用いた実験では、提案手法は従来手法よりも高精度に解答時間を予測できた。また、提案手法は受検者ごとに異なる項目セットを解いた場合でも、項目ごとに異なる時間特性を考慮しながら解答速度パラメータを推定できていることを示した。さらに、本研究では提案手法により推定した時間特性に関するパラメータと、IRT により推定した反応に関するパラメータを組み合わせることで実試験データの分析を行った。この分析では、反応精度と解答速度の関係性や問題形式ごとの解答時間の傾向の違いなど、従来の紙媒体で行われる試験では分析できなかった受検者および項目の時間特性を捉えることができた。

今後は、解答時間の異常検知 [22, 23] への応用を行いたい。解答時間の異常検知の先行研究 [22, 23] では、受検者の反応および解答時間データをリアルタイムに入力しながら異常値を判別している。しかし、本提案手法は受検者を長さが総受検者数の one-hot vector を用いて入力するため、未知の受検者の反応をリアルタイムに入力して解答時間を予測することが不可能である。この提案手法の問題を解決し、解答時間の異常検知への応用をすすめたい。

謝辞

本研究の遂行にあたり，指導教員の植野真臣教授，宇都雅輝准教授から丁寧かつ熱心なご指導を賜りました．ここに感謝の意を表します．また，研究に関する議論にてたくさんの示唆をいただいた堤瑛美子講師に深く感謝いたします．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] F.M. Lord and M.R. Novick, *Statistical Theories of Mental Test Scores*, Addison-Wesley, 1968.
- [2] E.E. Roskam, “Toward a psychometric theory of intelligence,” *Progress in Mathematical Psychology*, pp.151–171, 1987.
- [3] E.E. Roskam, “Models for speed and time-limit tests,” *Handbook of modern item response theory*, pp.187–208, Springer, 1997.
- [4] N.D. Verhelst, H.H. Verstralen, and M. Jansen, “A logistic model for time-limit tests,” *Handbook of modern item response theory*, pp.169–185, Springer, 1997.
- [5] W.J. van der Linden, “A lognormal model for response times on test items,” *Journal of Educational and Behavioral Statistics*, vol.31, no.2, pp.181–204, 2006.
- [6] 永岡慶三, 植野真臣, “回答所要時間における項目応答理論,” *行動計量学*, vol.18, no.2, pp.1–8, 1991.
- [7] G.J. Fox, R.K. Entink, and W.J. van derLinden, “Modeling of responses and response times with the package cirt,” *Journal of Statistical software*, vol.20, no.7, pp.1–14, 2007.
- [8] W.J. van der Linden, “A hierarchical framework for modeling speed and accuracy on test items,” *Psychometrika*, vol.72, no.3, pp.287–308, 2007.
- [9] D. Molenaar, D. Oberski, J. Vermunt, and P. De Boeck, “Hidden markov item response theory models for responses and response times,” *Multivariate behavioral research*, vol.51, no.5, pp.606–626, 2016.
- [10] L. Peng, “Predicting response time of questions using linear mixed-effects model,” *Proceedings of the 17th International Conference on Educational Data Mining*, pp.877–881, 2024.
- [11] E. Tsutsumi, R. Kinoshita, and M. Ueno, “Deep-IRT with independent student and item networks,” *Proceedings of the 14th International Conference on Educational Data Mining*, pp.510–517, 2021.
- [12] 堤瑛美子, 木下涼, 植野真臣, “独立な学習者・項目ネットワークをもつ Deep-IRT,” *電子情報通信学会論文誌 D*, vol.104, no.7, pp.596–608, 2021.
- [13] E. Tsutsumi, R. Kinoshita, and M. Ueno, “Deep item response theory as a novel test theory based on deep learning,” *Electronics*, vol.10, no.9, p.1020, 2021.

- [14] 堤瑛美子, 郭亦鳴, 植野真臣, “学習データの忘却を最適化する Hypernetwork を組み込んだ Deep-IRT,” 電子情報通信学会論文誌, vol.J106-D, no.No.2, pp.72–83, 2023.
- [15] E. Tsutsumi, Y. Guo, R. Kinoshita, and M. Ueno, “Deep knowledge tracing incorporating a hypernetwork with independent student and item networks,” IEEE Transactions on Learning Technologies, pp.951–965, 2023.
- [16] E. Tsutsumi, T. Nishio, and M. Ueno, “Deep-irt with a temporal convolutional network for reflecting students’ long-term history of ability data,” International Conference on Artificial Intelligence in EducationSpringer, pp.250–264 2024.
- [17] 西尾徹朗, 堤瑛美子, 植野真臣, “学習者の能力の時系列変化を畳み込む Temporal Convolutional Network を組み込んだ Deep-IRT,” 電子情報通信学会論文誌 D, vol.107, no.3, pp.98–110, 2024.
- [18] 植野真臣, “AI 採点システム: 1. CBT の最前線,” 情報処理, vol.64, no.5, pp.e1–e6, 2023.
- [19] 木下 涼, 植野真臣, “深層学習によるテスト理論: Item Deep Response Theory,” 電子情報通信学会論文誌, vol.J103-D, pp.314–329, 2020.
- [20] W.J. van der Linden, “Predictive control of speededness in adaptive testing,” Applied Psychological Measurement, vol.33, no.1, pp.25–41, 2009.
- [21] R.D. Luce, Response times: Their role in inferring elementary mental organization, Oxford University Press, 1991.
- [22] W.J. van der Linden and F. Guo, “Bayesian procedures for identifying aberrant response-time patterns in adaptive testing,” Psychometrika, vol.73, no.3, pp.365–384, 2008.
- [23] 植野真臣, “e ラーニングにおける所要時間データの異常値オンライン検出,” 電子情報通信学会論文誌 D, vol.90, no.1, pp.40–51, 2007.