

# Multi-Task DNN による所要時間を考慮した適応型テスト

2025 年 1 月 31 日

情報数理工学プログラム

学籍番号 2010130

岡田 幸征

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 2010130 |
| 指導教員 植野 真臣                           | 氏名 岡田 幸征     |
| 題目 Multi-Task DNN による所要時間を考慮した適応型テスト |              |

## 概要

適応型テストは、逐次的に推定した受検者の能力から情報量が最大となる問題を出題するテストである。これにより、従来型の試験よりも少ない問題数で高精度に測定できる。しかし、従来の適応型テストは、各問題への所要時間を考慮しておらず、試験時間が増加したり、制限時間内で能力推定値の不安定になるといった問題がある。この問題を解決するため、所要時間を予測する確率的手法を用いた適応型テストが提案されてきた。しかし、近年では、確率的手法よりも深層学習手法の方が予測精度が高いことが知られている。そこで、本研究では、逐次的に推定した能力から最先端手法の Multi-Task DNN を用いて予測された所要時間を考慮した3つの項目選択手法、(1) 単位時間当たりの情報量が最大となる項目選択手法、(2) 制限時間を考慮した項目選択手法、(3) (1) と (2) を組み合わせた項目選択手法、を提案する。(1) では単位時間あたりの能力推定精度向上が期待され、(2) では、制限時間内で精度の高い能力推定値が得られることが期待される。(3) では、制限時間を考慮した上で、単位時間当たりの情報量が最大となる項目選択を行うことで、能力推定値のさらなる向上が期待できる。提案手法の有効性は実データを用いたシミュレーションにより示した。

# 1 まえがき

近年、e テスティングと呼ばれるコンピュータ上で実施されるテスト形式が普及している。この形式のテストでは、異なる問題項目から構成されているにも関わらず、受検者の能力値を同一尺度で測定することができる。e テスティングの一つとして、適応型テスト (Computerized Adaptive Testing: CAT) が知られている。適応型テストは、受験者の回答が得られるたびに能力を推定し、その能力推定値に応じて最適な項目を出題する。これにより、能力値の推定精度を減少させることなく、出題項目数や受験時間を軽減できるという利点がある。この利点から、適応型テストは、Synthetic Personality Inventory(SPI)[1] や Global Test of English Communication(GTEC)[2] など、さまざまな試験で実用化されている。

適応型テストにおける項目選択では、能力推定値に対する情報量が最大となる項目を選択する MI 法 (Maximum Information) が最もよく用いられている。[3] しかし、MI 法は受験者の回答にかかる所要時間を考慮していないために、主に 2 つの問題点がある。1 つ目は、受験者間のテスト終了時間に大きな差が生じ、試験の公平性や運用効率が損なわれる可能性がある。アイテムバンク内の項目に、回答所要時間に大きなばらつきがあると、受験者のテスト終了時間に大きな差が生じることになる。特に難易度の高い項目ほど、所要時間が長くなる傾向があるため [4]、高い能力値を持つ受検者に対して、テスト全体の所要時間が長くなる傾向がある。2 つ目は、制限時間が設定された場合、所要時間を考慮しない適応型テストでは、受験者の能力推定精度が低くなる可能性がある。特に、制限時間内に全ての項目を解き終えることができないと、未回答の項目が生じ、能力の推定誤差が増加する。また、所要時間の長い項目セットが与えられた受験者は、所要時間の短い受験者と比較して、制限時間内に回答を終わらせるために推測を行う可能性が高まり、結果的に能力推定が不安定になることが報告されている [5]。

1 つ目の問題点を解決するために、Fan らは、適応型テストの項目選択に単位時間あたりの情報量が最大となる項目を選択する MIT (Maximum Information per Time Unit) 法を提案した。MIT 法は、各受検者の各項目への所要時間が、対数正規分布に従う変数と仮定したモデル [6] を用いて、単位時間当たりの情報量を算出する。MIT 法は、MI 法と比較して、同じ項目数を出題した場合でも、予測精度を維持しつつテスト全体の所要時間を短縮させた [7]。次に、2 つ目の問題点を解決するために、シャドウテストアプローチ [8] を用いた制限時間の制約のある適応型テストが提案されている。シャドウテストアプローチでは、項目選択時にすでに出題した項目を含み、制約を満たしながら、情報量が最大となる項目集合 (シャドウテスト) を整数計画法を用いて構成する。さらに、この

シャドウテストの中から最適な項目を出題することで、試験終了時に出題した項目集合が試験の制約を満たし、かつ情報量が最大となることが保証される。これにより、制約条件がある場合でも、推定精度を維持できることが示されている。Linden は、シャドウテストアプローチに、制限時間の制約を追加することで、受検者が制限時間内に回答を終えることができる項目を出題した。[9] その結果、受検者が制限時間内に提出した項目に回答でき、能力値の推定誤差を維持できた。これらの手法では、より適切な項目選択のために、所要時間の予測精度が重要となる。これらの手法に用いられている所要時間予測手法は、確率モデルに基づくアプローチが主に採用されている。その一方で、近年では、教育評価の分野で、確率モデルよりも予測精度の高い深層学習モデルが複数提案されている [10, 11, 12, 13]。石川らは、Multi-gate Mixture-of-Experts (MMoE) [14] というマルチタスク学習を用いて、正誤反応と所要時間を同時に学習する深層学習モデルを提案した（以降、Multi-Task DNN と呼ぶ）[15]。このモデルでは、正誤反応と所要時間の依存関係を同時に学習することで、所要時間の予測精度を大きく向上させた。

そこで、本論文では、Multi-Task DNN を用いて、所要時間を考慮した適応型テストの項目選択手法を 3 つ提案する。まず 1 つ目の手法では、Multi-Task DNN を用いて、受検者の能力推定値から各項目への所要時間を予測し、その予測所要時間に基づいて単位時間当たりの情報量が最大となるように項目を選択する。これにより、従来手法よりも高精度な所要時間の予測を可能とし、能力推定精度を維持しながらテストの所要時間をさらに短縮することが期待される。2 つ目に、Multi-Task DNN を用いて、所要時間を考慮したシャドウテストアプローチに基づく、制限時間のある適応型テストの項目選択手法を提案する。本手法では、Multi-Task DNN による出題項目の所要時間を予測し、それが制限時間に収まるよう制約式として整数計画法に与え、シャドウテストを構成する。また、受検者によって回答速度が異なるため、出題項目数を固定せず、出題項目数に上限を設けず、制限時間を最大限に活用を行う。Multi-Task DNN を用いることで、所要時間を高精度に予測し、より正確な項目選択を実現する。これにより、制限時間内に回答可能でありながら、予測精度をさらに向上させることが期待される。3 つ目に、1 つ目の単位時間当たりの情報量が最大となる項目選択と 2 つ目の制限時間を考慮した手法を組み合わせた項目選択手法を提案する。この手法では、まず短時間で高精度な能力推定値を得られ、それに基づいてより正確な情報量や所要時間を得る。これらの値から制限時間内で能力推定値が最大となる項目をより正確に選択することができる。

提案手法の有効性は、実データを用いたシミュレーション実験により示される。Multi-Task DNN による所要時間を考慮した単位時間当たりの情報量が最大となる項目選択手法は、従来手法と比較して、予測精度を保ったまま、回答時間を短縮することができた。また、Multi-Task DNN による所要時間を考慮した制限時間のある適応型テストでは、従

来手法と比較して、制限時間内で出題項目数を増加させ、予測精度を向上させた。加えて、制限時間のある適応型テストでは、情報量が最大となる項目より単位時間あたりの情報量が最大となる項目を出題するほうが、予測精度が高くなることも示された。

## 2 適応型テスト

### 2.1 項目反応理論

適応型テストは、項目反応理論 (Item Response Theory: IRT) に基づき、受験者の能力値を効率的に推定する。項目反応理論は、数理モデルを用いたテスト理論のひとつであり、近年コンピュータテストの普及とともに多様な評価場面で活用されている [16, 17, 18, 19, 20]。IRT は異なる項目への受験者の反応を同一尺度で測定可能であるという特徴がある。本論文では、項目反応理論の中でも最も広く利用されている 2 母数ロジスティックモデル (2PLM: 2-Parameter Logistic Model) を用いる。2PLM では、能力値  $\theta \in (-\infty, \infty)$  の受験者が項目  $i \in \{1, \dots, N\}$  に正答する確率を以下の式で表す。

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

ここで、 $u_i$  は受験者が項目  $i$  に正答するとき 1、誤答するとき 0 をとる変数である。また、 $a_i \in [0, \infty)$  は、項目  $i$  の識別力と呼ばれ、どのくらいの精度で受験者の能力値を識別できるかを表す。 $b_i \in (-\infty, \infty)$  は、項目  $i$  の難易度パラメータと呼ばれ、その項目の難しさを表す。

#### 2.1.1 フィッシャー情報量

2PLM では、能力値  $\theta$  の受験者に対して項目  $i$  のフィッシャー情報量を以下の式で表す。

$$I_i(\theta) = \frac{\left[ \frac{\partial}{\partial \theta} p(u_i = 1|\theta) \right]^2}{p(u_i = 1|\theta)(1 - p(u_i = 1|\theta))}$$

このフィッシャー情報量は、その項目が能力値  $\theta$  付近でその能力値をどのくらいよく識別できるかを表す。従来の適応型テストでは、このフィッシャー情報量が高い項目を逐次的に出題することで効率の良い能力値の推定を実現している。本論文では、情報量はフィッシャー情報量を指す。なお、テストを構成する項目集合  $T$  の情報量の総和をテスト情報量とよび、次のように示す。

$$I(\theta) = \sum_{i \in T} I_i(\theta)$$

また、テスト情報量の逆数が受験者の能力推定値の漸近分散に一致することが知られている．[21] つまり、テスト情報量を最大化することで、能力値の測定誤差を最小化することができる。

### 2.1.2 能力値 $\theta$ の推定

受検者の能力値の推定には EAP 推定 (Expected a posteriori) を用いる [22]． $m - 1$  項目までの反応データのベクトル  $\mathbf{u}_{m-1}$  と能力値  $\theta$  の事前分布  $f(\theta)$  を用いて、能力値の事後分布  $f(\theta|\mathbf{u}_{m-1})$  は次のように表される．

$$f(\theta|\mathbf{u}_{m-1}) = \frac{L(\mathbf{u}_{m-1}|\theta)f(\theta)}{\int_{-\infty}^{\infty} L(\mathbf{u}_{m-1}|\theta)f(\theta)d\theta}$$

ここで、 $L(\mathbf{u}_{m-1}|\theta)$  は能力値を  $\theta$  としたときに反応  $\mathbf{u}_{m-1}$  を返す尤度である．EAP 推定は事後分布  $f(\theta|\mathbf{u}_{m-1})$  の  $\theta$  に関する期待値を能力推定値とする．EAP 推定を用いて能力推定値  $\hat{\theta}$  は次式より求められる．

$$\hat{\theta} = \int_{-\infty}^{\infty} \theta f(\theta|\mathbf{u}_{m-1})d\theta$$

## 2.2 古典的な適応型テスト

適応型テストは、受検者の能力値を逐次的に推定し、その能力推定値に応じて情報量が最大の項目を出題する．これにより、適応型テストは、少ない項目数で受検者の能力値を高精度に測定することができる．情報量の最大化を目的とする MI 法 (情報量最大化法) は、適応型テストにおける最も広く利用されている項目選択戦略の一つである．[3] MI 法は、次のように  $t+1$  番目に出題すべき項目を選択する．

$$\max_l \left\{ I_l \left( \hat{\theta}^{(t)} \right) \right\}$$

ここで、 $\hat{\theta}^{(t)}$  は  $t$  問回答した時点での能力推定値であり、 $I_l \left( \hat{\theta}^{(t)} \right)$  は、能力推定値  $\hat{\theta}^{(t)}$  に基づく項目  $l$  の情報量である．具体的には、MI 法は、項目のパラメータが既知のアイテムバンクを用いて、次のアルゴリズムにしたがって項目選択する．

1. 能力推定値を  $\hat{\theta} = 0$  に初期化する．
2. 能力推定値  $\hat{\theta}$  を所与として情報量が最大となる項目をアイテムバンクから選択し受検者に出題する．
3. 受験者の回答履歴から受検者の能力推定値  $\hat{\theta}$  を推定する．

4. テスト終了条件まで手順 (2), (3) を繰り返す.

この手順に従って、適応型テストは受検者の能力を逐次的に推定し、情報量の高い項目を受検者に出題することで、少ない項目数で高精度な能力値の推定を実現する. MI 法は、項目数の効率が優れている一方で、時間効率については考慮されない. MI 法により選択された項目は、多くの情報量を持つが、その一方で回答に長い時間がかかる可能性がある. 通常テストには、時間制限がある場合が多く、受験者と試験監督者の長時間の拘束にも、コストがかかるため、時間効率の観点での評価することが重要となる.

## 2.3 適応型テストのためのシャドウテストアプローチ

制限時間などテストの制約のある適応型テストでは、シャドウテストアプローチが効果的に活用される. [9] テストの制約を満たすためには、項目出題時ではなく、テスト終了時まで制約を満たす必要がある. シャドウテストアプローチでは、テスト終了時に制約を満たすことを保証するために、出題される項目に対してのみ制約を課すのではなく、項目集合に対して制約を課すことで簡易的に仮想的なテスト（シャドウテスト）を作成する. シャドウテストとは、(1) テストの制約（例えばカテゴリごとの出題項目数）をすべて満たし、(2) 受検者に出題した項目をすべて含め、(3) 能力推定値に対してフィッシャー情報量の総和が最大になるように構成された項目集合である. 具体的には以下のアルゴリズムに従ってシャドウテストの中から項目を選択する.

1. 受検者の能力推定値を  $\hat{\theta} = 0$  に初期化する.
2. 次の整数計画法を用いてシャドウテストを構成する.

$$\begin{aligned} \text{variables} \quad & x_i = \begin{cases} 1 & \text{項目 } i \text{ がシャドウテストに含まれる} \\ 0 & \text{それ以外} \end{cases} \\ \text{maximize} \quad & \sum_{i=1}^N I_i(\hat{\theta}) x_i \\ \text{subject to} \quad & \sum_{i=1}^N x_i = n \quad (\text{テストの項目数}) \\ & (\text{その他テストの制約が含まれる}) \end{aligned}$$

3. シャドウテストから項目選択基準に従って最適な項目を受検者に出題する.
4. 受検者の能力値  $\hat{\theta}$  を推定する.
5. テストの制約を満たす間、手順 (2) から (4) を繰り返す.

このようにして、シャドウテストアプローチは、テスト終了時に出題した項目集合が、

制約を満たしながら情報量を最大化することができる。

### 3 予測所要時間を用いた適応型テスト

#### 3.1 単位時間あたりの情報量最大化する項目選択手法 (MIT)

古典的な適応型テストの問題点に対処するために, Fan らは [7], 単位時間あたりの情報量を最大化する項目を選択する手法 (以降, MIT と呼ぶ.) を提案した. MIT は, 次のように  $t + 1$  番目に出題すべき項目を選択する.

$$\max_{l \in I} \left\{ \frac{I_l(\hat{\theta}^{(t)})}{T_l} \right\}, (\text{未出題項目 } I)$$

ここで,  $\hat{\theta}^l$  は  $t$  問回答時点の受験者の能力推定値であり,  $I_l(\hat{\theta}^{(t)})$  は,  $t$  問目回答時点の問題  $l$  の情報量を表す.

#### 3.2 LNIRT

所要時間を考慮した適応型テストでは, IRT と LNIRT を組み合わせた階層モデルが用いられる. LNIRT では, 1 層目で IRT と LNIRT を独立して定義する. LNIRT は, 受験者の項目への所要時間を対数正規分布に従う変数としてモデル化する. 具体的には, 受験者  $i$  の項目  $j$  に対する所要時間の対数  $\ln t_{ij}$  が以下の正規分布に従うと仮定する.

$$\ln t_{ij} = \lambda_j - \phi_j \zeta_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim \mathcal{N}(0, \sigma_{\epsilon_j}^2)$$

ここで,  $\lambda_j \in [0, \infty]$  は項目  $j$  の回答に必要な時間量 (対数スケール) を表す所要時間パラメータである.  $\zeta_i \in [-\infty, \infty]$  は受験者  $i$  の回答速度を表すスピードパラメータ,  $\phi_j \in [0, \infty]$  は回答速度が所要時間に及ぼす効果を示す時間識別力パラメータである. また,  $\epsilon_{ij}$  は平均 0, 項目固有の残差分散  $\sigma_{\epsilon_j}^2$  の正規分布に従う誤差項であり, 予測所要時間の不確かさを表す.

次に, 2 層目では, IRT と LNIRT の受験者に関するパラメータと項目に関するパラメータの母集団分布を定義する. 受験者の能力と回答速度を表すパラメータ  $\rho = (\theta_i, \zeta_i)$  の事前分布は, 以下の 2 変量正規分布が定義される:

$$(\theta_i, \zeta_i) \sim \mathcal{N}_2(\mu_\rho, \Sigma_\rho),$$

$$\mu_\rho = \begin{pmatrix} \mu_\theta \\ \mu_\zeta \end{pmatrix}, \quad \Sigma_\rho = \begin{pmatrix} \sigma_\theta^2 & \sigma_{\theta\zeta} \\ \sigma_{\theta\zeta} & \sigma_\zeta^2 \end{pmatrix}.$$

$$(\theta_i, \zeta_i) \sim \mathcal{N}_2(\mu_\rho, \Sigma_\rho), \quad \mu_\rho = \begin{pmatrix} \mu_\theta \\ \mu_\zeta \end{pmatrix}, \quad \Sigma_\rho = \begin{pmatrix} \sigma_\theta^2 & \sigma_{\theta\zeta} \\ \sigma_{\theta\zeta} & \sigma_\zeta^2 \end{pmatrix}.$$

$$E \left[ T_l \mid \hat{\tau}^{(t)} \right] = \exp \left( \beta_l - \hat{\tau}^{(t)} + \frac{1}{2} \alpha_l^2 \right)$$

### 3.3 Multi-Task DNN を用いた所要時間予測モデル

最も高精度に所要時間の予測が可能なモデルとして、石川らの深層学習手法である Multi-Task DNN をもとにしたモデルがある [15]。このモデルでは、正誤反応と所要時間を2つの出力に持ち、これらの依存関係を学習する。一般的に、受験者が時間をかけて試行錯誤を繰り返すことで正答確率は高くなると期待される。そのため、正誤反応と所要時間の関係には依存関係があると考えられる。このモデルでは、正誤反応と所要時間を同時に学習することで、この依存関係を捉え、結果的に正誤反応、所要時間において予測精度が向上した。このアプローチによって、両者の相乗効果が得られることが示されている。図は、石川らの Multi-Task DNN の構造を示している。Multi-Task DNN は、各項目を表す one-hot ベクトルに加え、項目反応理論による受験者の能力推定値  $\theta$  を入力とし、出力は正誤反応 (正答ならば1, 誤答なら0) と所要時間となっている。

入力ベクトル  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  は入力層  $L_{\text{input}}$  を通り独立した E 個のフィードフォワードネットワーク層に入力される。

$$\begin{aligned} L_{\text{input}} &= \text{leaky\_relu}(W_{L_{\text{input}}} \mathbf{x} + b_{L_{\text{input}}}), \\ L_i^{e1} &= \text{leaky\_relu}(W_{L_i^e} L_{\text{input}} + b_{L_i^{e1}}), \\ L_j^e &= \text{leaky\_relu}(W_{L_j^e} L_{j-1}^e + b_{L_j^e}), \end{aligned}$$

$$\text{leaky\_relu}(z) = \begin{cases} z & \text{if } z > 0 \\ \gamma z & \text{if } z \leq 0 \end{cases}$$

ここで  $L_j^e$  は e 番目のフィードフォワードネットワーク層の第 j 層の出力を表す。また  $\mathbf{W}$  はニューラルネットワークの重みパラメータ、 $\mathbf{b}$  はバイアスパラメータ、 $\gamma$  はチューニングパラメータである。

その後、以下のゲート付きネットワークにより、各下層ニューラルネットワークの結果にタスク k の重みが割り当てられる。

$$\begin{aligned}
g^k(\mathbf{x}) &= \text{softmax}(\mathbf{W}_{g_k} \mathbf{x} + \mathbf{b}_{g_k}), \\
f^k(\mathbf{x}) &= \sum_{e=1}^E g^k(\mathbf{x}) L^e, \\
H_1^k &= \text{leaky\_relu}(\mathbf{W}_{H_1^k} f^k(\mathbf{x}) + \mathbf{b}_{H_1^k}), \\
H_m^k &= \text{leaky\_relu}(\mathbf{W}_{H_m^k} H_{m-1}^k + \mathbf{b}_{H_m^k}), \\
(m &= \{2, 3, \dots, M\}).
\end{aligned}$$

最後に、タスク固有の予測値は以下から得られる

$$y^k = \text{leaky\_relu}(\mathbf{W}_{y_k} H_M^k + \mathbf{b}_{y_k}).$$

ただし、 $k=1$  は正誤反応の予測タスク、 $k=2$  は所要時間の予測タスクである。

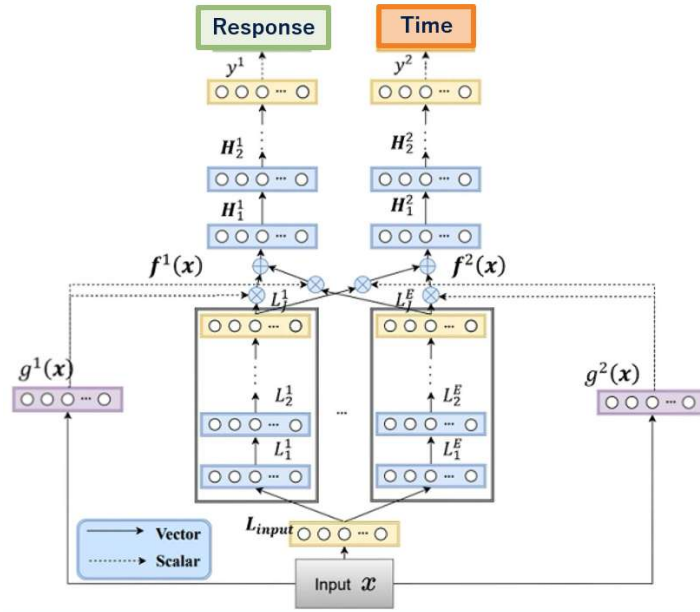


図 1: 所要時間のための Multi-Task DNN

## 4 Multi-Task DNN による予測所要時間を考慮した適応型テストの項目選択手法

本研究では、所要時間の予測精度において従来の確率モデルを上回る性能を示す Multi-Task DNN を活用し、3つの項目選択手法、(1) 単位時間当たりの情報量が最大となる項

目選択手法、(2) 制限時間を考慮した項目選択手法、(3) (1) と (2) を組み合わせた項目選択手法、を提案する。これらの手法に Multi-Task DNN の予測所要時間を用いることで、より正確な項目選択ができる。具体的には、(1) では能力推定精度を維持しつつ所要時間を削減し、(2) では制限時間内で能力推定精度を向上させることが期待される。(3) では、(1) により短時間で精度の高い能力推定値を得ることで、早期に項目の情報量や予測時間が正確になり、それらを基に (2) で制限時間内でより正確な項目を選択できる。このアプローチにより、制限時間内で能力値推定精度の向上が実現できると期待される。提案手法では、事前データを用いて学習した Multi-Task DNN を用い、受験者の逐次推定される能力値を所与として、各項目への所要時間を予測する。

#### 4.1 Multi-Task DNN を用いた単位時間あたりの情報量が最大となる項目選択手法

本節では、Multi-Task DNN[15] による各項目への予測所要時間を用いて、単位時間あたりの情報量が最大となる項目を選択する適応型テストの項目選択手法を提案する。Multi-Task DNN を活用することで、所要時間の予測精度の向上を実現し、受験者の各項目への予測時間を用いて単位時間あたりの情報量を最適化した項目選択が可能となる。この結果、従来の手法に比べて単位時間あたり能力推定精度の向上が期待される。具体的なアルゴリズムは、以下の通りである。

1. 受験者の能力推定値を  $\hat{\theta} = 0$  に初期化する。
2. Multi-Task DNN を用い、能力推定値  $\hat{\theta}$  を所与として、各項目への所要時間を以下のように予測する。ここで、 $\mathbf{z}$  は項目を表す onehot-vector であり、 $u_i, t_i$  は、それぞれ項目  $i$  の正誤反応、所要時間である。

$$\text{入力: } \begin{matrix} \mathbf{z} \\ \hat{\theta} \end{matrix} = \{z_1, z_2, \dots, z_N\},$$

$$\text{出力: } \begin{matrix} u_i \\ t_i \end{matrix} = \begin{cases} 1 & \text{正答のとき} \\ 0 & \text{それ以外,} \end{cases}$$

3. 能力推定値を所与として、情報量  $\hat{\theta}$  が最大となる項目を、未出題項目集合  $P$  から次のように選択する。

$$i^* = \arg \max_{i \in P} \frac{I_i(\hat{\theta})}{t_i}$$

4. 受験者の能力値  $\hat{\theta}$  を推定し、更新する。
5. テストの制約を満たす間、手順 (2) から (4) を繰り返す。

この手法では、制限時間を考慮していないという問題点がある．そのため、制限時間内に回答が終わらない場合、未回答の項目が生じる可能性が高まり、能力推定精度が低下する可能性がある．

## 4.2 Multi-Task DNN を用いた制限時間のある適応型テストの項目選択手法

本節では、Multi-Task DNN[15] による予測所要時間を用いて、制限時間のある適応型テストの項目選択手法を提案する．各項目への所要時間は、受験者の能力推定値  $\hat{\theta}$  を所与として、Multi-Task DNN により予測する．この予測所要時間を基に、シャドウテストアプローチ [8] に従って、予測所要時間が制限時間内に収まるという制約を満たしながら、テスト情報量が最大となるシャドウテストを構成する．次に、構成されたシャドウテストの中から情報量が最も高い項目を出題する．この手法により、制限時間を考慮して、最大のテスト情報量を持つ項目選択が可能となる．また、Multi-Task DNN の予測所要時間を活用することで、高い予測精度を実現し、より正確な項目選択が可能となる．この結果、制限時間内で能力値の推定精度の向上が期待できる．具体的なアルゴリズムは、以下の通りである．

1. 受験者の能力推定値を  $\hat{\theta} = 0$  に初期化する．
2. Multi-Task DNN を用いて、各項目への所要時間を以下のように予測する．ここで、 $\mathbf{z}$  は項目を表す onehot-vector であり、 $u_i, t_i$  は、それぞれ項目  $i$  の正誤反応、所要時間である．

$$\text{入力: } \begin{matrix} \mathbf{z} \\ \hat{\theta} \end{matrix} = \{z_1, z_2, \dots, z_N\},$$

$$\text{出力: } \begin{matrix} u_i \\ t_i \end{matrix} = \begin{cases} 1 & \text{正答のとき} \\ 0 & \text{それ以外,} \end{cases}$$

3. 予測した各項目への所要時間  $t_i$  を用いて、テストの所要時間が制限時間内に収まるように以下の制約式をテストの制約に追加する．ただし、テストの制限時間を  $T$

とする.

$$\begin{aligned}
&\text{variables} \quad x_i = \begin{cases} 1 & \text{項目 } i \text{ がシャドウテストに含まれる} \\ 0 & \text{それ以外} \end{cases} \\
&\text{maximize} \quad \sum_{i=1}^N I_i(\hat{\theta}) x_i \\
&\text{subject to} \quad \sum_{i=1}^n t_i x_i \leq T
\end{aligned}$$

4. 整数計画法を用いてシャドウテスト  $S$  を構成する.
5. 能力推定値  $\hat{\theta}$  を所与として, 情報量  $I_i(\hat{\theta})$  が最大となる項目を, シャドウテスト  $S$  の中から次のように選択する.

$$i^* = \arg \max_{i \in S} I_i(\hat{\theta})$$

6. 受検者の能力値  $\hat{\theta}$  を推定し, 更新する.
7. テストの制約を満たす間, 手順 (2) から (6) を繰り返す.

しかし, この手法では, シャドウテスト内の情報量が最大となる項目を優先して出題することで, 所要時間の長い項目が選ばれる傾向がある. そのため, 能力推定値の精度が低いテストの序盤にも所要時間の長い問題項目が出題される可能性が高く, 測定精度の効率が悪いかもしれない.

### 4.3 Multi-Task DNN を用いた制限時間のある単位時間あたりの情報量が最大となる項目選択手法

本節では, 4.1 および 4.2 の 2 つの手法を組み合わせた適応型テストの項目選択手法を提案する. 制限時間を考慮した上で, 単位時間あたりの情報量が最大となる項目選択を行うことで, 精度の高い能力推定値が得られると期待できる. 提案手法の具体的なアルゴリズムは以下のとおりである.

1. 受検者の能力推定値を  $\hat{\theta} = 0$  に初期化する.
2. Multi-Task DNN を用いて, 各項目への所要時間を以下のように予測する. ここで,  $\mathbf{z}$  は項目を表す onehot-vector であり,  $u_i, t_i$  は, それぞれ項目  $i$  の正誤反応,

所要時間である.

$$\text{入力: } \begin{matrix} \mathbf{z} = \{z_1, z_2, \dots, z_N\}, \\ \hat{\theta} \end{matrix}$$

$$\text{出力: } \begin{matrix} u_i = \begin{cases} 1 & \text{正答のとき} \\ 0 & \text{それ以外,} \end{cases} \\ t_i \end{matrix}$$

3. 予測した各項目への所要時間  $t_i$  を用いて, テストの所要時間が制限時間内に収まるように以下の制約式をテストの制約に追加する. ただし, テストの制限時間を  $T$  とする.

$$\begin{aligned} \text{variables } x_i &= \begin{cases} 1 & \text{項目 } i \text{ がシャドウテストに含まれる} \\ 0 & \text{それ以外} \end{cases} \\ \text{maximize } & \sum_{i=1}^N I_i(\hat{\theta}) x_i \\ \text{subject to } & \sum_{i=1}^n t_i x_i \leq T \end{aligned}$$

4. 整数計画法を用いてシャドウテスト  $S$  を構成する.  
 5. 能力推定値  $\hat{\theta}$  を所与として, 情報量  $I_i(\hat{\theta})$  が最大となる項目を, シャドウテスト  $S$  の中から次のように選択する.

$$i^* = \arg \max_{i \in S} \frac{I_i(\hat{\theta})}{t_i}$$

6. 受検者の能力値  $\hat{\theta}$  を推定し, 更新する.  
 7. テストの制約を満たす間, 手順 (2) から (6) を繰り返す.

## 5 評価実験

提案手法の有効性を示すため, 実データを用いて従来手法と比較する.

### 5.1 Multi-Task DNN を用いた単位時間あたりの情報量が最大となる項目選択手法 (MIT with MTDNN)

本節では, まず逐次的に能力推定値に対して情報量最大の項目を出題する適応型テスト (以降, CAT と呼ぶ.), MIT と本手法 (MIT with MTDNN と呼ぶ) を比較する. 実験の手順は以下の通りである.

1. 受験者の能力推定値を  $\hat{\theta} = 0$  に初期化する.
2. 受験者の能力推定値  $\hat{\theta}$  を所与として, 各手法のアルゴリズムに従い, アイテムバンクから出題項目を選択する.
3. 出題項目への反応データを所与として, 受験者の能力推定値  $\hat{\theta}$  を推定する.
4. テスト終了条件まで手順 2, 3 を繰り返す.

受験者に出題した項目から, 出題項目数, 予測所要時間の誤差, テストの所要時間と情報量を求めた. また, 予測時間の誤差は, 出題した項目数  $N$ ,  $n$  問目出題時点での未出題項目集合  $I_n$ , 項目  $i$  に対する実際の応答時間  $t_i$ , 予測時間を  $\hat{t}_{i,n}$  として, 以下の式で表される.

$$\frac{1}{N} \sum_{n=1}^N \sqrt{\frac{1}{|I_n|} \sum_{i \in I_n} (t_i - \hat{t}_{i,n})^2}$$

実験には, 2023 年度に電気通信大学で実施された入学生を対象とした学力調査試験の数学の結果を実験データとして用いた. アイテムバンクの詳細は表 1 の通りである.

表 1: アイテムバンクの詳細

| 問題数 | 識別力       |      |      | 難易度         |        |      |
|-----|-----------|------|------|-------------|--------|------|
|     | Range     | Mean | SD   | Range       | Mean   | SD   |
| 100 | 0.37~2.07 | 1.12 | 0.15 | - 1.34~0.79 | - 0.50 | 0.25 |

実施された試験では, 666 人の受験者がこのアイテムバンクから構成されたテストを受験した. このデータから, 正誤反応データおよび回答時間データが利用された. ただし, 各受験者が回答した項目が異なるため, 未回答の項目については, Multi-Task DNN を用いて反応データを補完した. 具体的には, 実際の反応データを基に学習した Multi-Task DNN により, 受験者の能力推定値を所与として, 正誤反応および回答時間を予測し, それらの予測値を補完データとして利用した. テスト終了条件は, 先行研究 [7] と同様にテストの長さとした. 本実験ではテストの長さを 10, 20, 30, 40 項目と 10 刻みでテストの長さを変えた. 各手法による所要時間の予測精度 (time.RMSE), テストの所要時間 (秒) の平均 (Avg.time), テスト情報量の平均 (Avg.info) を表 2 に示す.

表 2: テスト結果

| テストの長さ | 評価指標      | CAT                | MIT                | MIT with MTDNN          |
|--------|-----------|--------------------|--------------------|-------------------------|
| 10     | time.RMSE | -                  | 197.49 (220.89)    | <b>64.44</b> (3.41)     |
|        | Avg.time  | 1404.77 (497.96)   | 868.55 (317.45)    | <b>798.23</b> (329.14)  |
|        | Avg.info  | 0.47 (0.22)        | <b>0.49</b> (0.22) | <b>0.49</b> (0.22)      |
| 20     | time.RMSE | -                  | 258.79 (276.49)    | <b>66.91</b> (4.43)     |
|        | Avg.time  | 3370.22 (656.08)   | 2403.55 (496.04)   | <b>2005.77</b> (423.73) |
|        | Avg.info  | <b>0.83</b> (0.34) | 0.82 (0.34)        | 0.82 (0.34)             |
| 30     | time.RMSE | -                  | 291.43 (308.33)    | <b>67.93</b> (4.43)     |
|        | Avg.time  | 5125.89 (642.60)   | 4224.94 (549.29)   | <b>3832.32</b> (571.34) |
|        | Avg.info  | <b>1.36</b> (0.54) | 1.32 (0.40)        | 1.32 (0.39)             |

実験結果から、CAT と MIT のテスト所要時間の平均を比較すると、先行研究で示された通り、所要時間を考慮した MIT は CAT に比べて大幅に短縮することができた。次に、予測時間の RMSE を比較すると MIT よりも提案手法（表中の MIT with MTDNN）の方が小さい。これにより、単位時間あたりの情報量がより正確に得られる。また、MIT と提案手法のテスト所要時間の平均を比較すると、提案手法は MIT よりもさらに小さく抑えることができた。最後に、各手法のテスト情報量の平均を比較すると、CAT が最も高い値を示したが、MIT および提案手法ともほぼ同程度であった。このことから、提案手法は従来の手法と比較して、単位時間あたりの情報量を正確に算出でき、テスト情報量を維持しつつ、所要時間を短縮することができた。

## 5.2 Multi-Task DNN を用いた制限時間のある適応型テストの項目選択手法 (STA with MTDNN)

本節では、制限時間を考慮しない適応型テスト（以降、CAT と呼ぶ.），STA 法 [9] と本手法（STA with MTDNN）を比較する。実験手順は前の実験と同様である。終了条件は、制限時間とした。本実験では、制限時間を、30 分、40 分、50 分、60 分とした。実験結果は、表 3 に示す。

表 3: テスト結果

| 制限時間 | 評価指標      | CAT             | STA             | STA with MTDNN  | MIT+STA with MTDNN  |
|------|-----------|-----------------|-----------------|-----------------|---------------------|
| 30 分 | time.RMSE | 0.73 (0.21)     | 230.47 (243.17) | 67.33 (4.14)    | <b>64.56</b> (4.09) |
|      | Avg.time  | 1630.69 (82.17) | 1684.03 (25.43) | 1763.65 (12.95) | 1760.06 (13.65)     |
|      | Avg.item  | 11.44 (3.07)    | 16.16 (2.52)    | 18.03 (3.07)    | <b>20.91</b> (3.07) |
|      | Avg.info  | 0.56 (0.32)     | 0.76 (0.33)     | 0.81 (0.34)     | <b>0.84</b> (0.33)  |
| 40 分 | time.RMSE | 0.76 (0.20)     | 247.44 (260.78) | 68.11 (4.32)    | <b>65.02</b> (4.71) |
|      | Avg.time  | 2237.51 (66.82) | 2282.79 (24.42) | 2366.50 (11.61) | 2361.95 (10.75)     |
|      | Avg.item  | 14.35 (2.29)    | 18.95 (2.80)    | 21.21 (3.02)    | <b>25.57</b> (3.39) |
|      | Avg.info  | 0.68 (0.34)     | 0.82 (0.35)     | 0.89 (0.37)     | <b>1.05</b> (0.43)  |
| 50 分 | time.RMSE | 0.80 (0.17)     | 263.61 (276.52) | 68.63 (4.57)    | <b>65.61</b> (5.01) |
|      | Avg.time  | 2849.51 (42.44) | 2877.90 (20.73) | 2962.71 (12.49) | 2960.80 (12.90)     |
|      | Avg.item  | 17.82 (2.26)    | 22.47 (2.36)    | 24.45 (2.52)    | <b>29.64</b> (3.18) |
|      | Avg.info  | 0.80 (0.33)     | 0.83 (0.35)     | 1.05 (0.43)     | <b>1.19</b> (0.47)  |
| 60 分 | time.RMSE | 0.81 (0.17)     | 276.03 (291.27) | 68.63 (4.57)    | <b>65.61</b> (5.38) |
|      | Avg.time  | 3461.00 (46.51) | 3476.79 (17.35) | 3560.09 (15.85) | 3555.48 (15.07)     |
|      | Avg.item  | 21.37 (3.10)    | 25.89 (2.44)    | 27.61 (2.87)    | <b>33.22</b> (3.10) |
|      | Avg.info  | 0.94 (0.38)     | 1.11 (0.44)     | 1.19 (0.50)     | <b>1.29</b> (0.45)  |

実験結果から、テスト情報量を従来の手法である CAT, STA と比較すると、提案手法 (STA with MTDNN) が最も大きくすることができた。平均テスト所要時間に注目すると、提案手法が最も長くなっている。これにより、制限時間を十分に活用することができ、出題項目数を多くすることができたと推測できる。一方で、制限時間を考慮しない CAT は、平均テスト所要時間が短く、出題項目数が最も少ない。これは、CAT は制限時間の近くで、所要時間の長い項目を出題してしまい、制限時間を超過して受験者の回答が得られなかった可能性が挙げられる。これに対し、提案手法は、制限時間が制限時間が短くなった後半でも、所要時間の短い項目を出題することにより、より多くの項目を出題することができたと推測される。このことから、提案手法は従来の手法と比較して、制限時間内でより多くの項目を出題し、テスト情報量を大きくすることができた。

### 5.3 Multi-Task DNN を用いた制限時間のある単位時間あたりの情報量が最大となる項目選択手法 (MIT + STA with MTDNN)

本節では、前節の手法 (STA with MTDNN) と本手法 (以降, MIT+STA with MTDNN と呼ぶ) を比較する。結果は表 3 に示されている。MIT+STA with MTDNN は STA with MTDNN に比べて出題項目数が増加し、テスト情報量をさらに高めることに成功した。これは、単位時間当たりの情報量が最大となる項目選択 (MIT) を取り入れたことが影響していると考えられる。このアプローチにより、短時間で高精度な能力推定値が得られ、それに基づいて項目の情報量や所要時間が正確に算出される。その結果、制限時間という制約の中で、情報量を最大化するための最適化が正確に行われたと推測される。

## 6 むすび

本研究では、Multi-Task DNN による受検者の所要時間を考慮した適応型テストの項目選択手法を 3 つ提案した。実データを用いたシミュレーション実験により、1 つ目の Multi-Task DNN を用いた単位時間当たりの情報量最大となる項目選択手法では、従来手法 (MIT) より、予測精度を維持したまま、所要時間を短縮することができた。2 つ目の Multi-Task DNN を用いた制限時間のある適応型テストの項目選択手法では、従来手法 (STA 法) よりも制限時間内で出題項目数を多くし、能力値推定精度を向上させることができた。最後に、1 つ目と 2 つ目の手法を組み合わせた項目選択手法を検討し、制限時間内でより多くの項目を出題し、能力値推定精度を一層向上させることができた。適応型テストは、特定の項目が過度にテストに用いられ、露出数が偏ってしまう問題がある。適応型テストを繰り返し運用するためには、露出数の偏りを抑える必要がある。また項目の

露出数と能力値推定精度はトレードオフの関係にあり，それを制御可能な 2 段階適応型等質テスト [23] が提案されている．そのため今後は，2 段階適応型等質テストに本手法の適応を行いたい．

## 参考文献

- [1] Recruit. Synthetic personality inventory.
- [2] 株式会社ベネッセコーポレーション. Global test of english communication, 2014.
- [3] Wim J van der Linden. Some alternatives to sympson-hetter item-exposure control in computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, 28(3):249–265, 2003.
- [4] Wim J. Van Der Linden. Conceptual issues in response-time modeling. *Journal of Educational Measurement*, 46(3):247–272, 2009.
- [5] Brent Bridgeman and Frederick Cline. Effects of differentially time-consuming tests on computer-adaptive test scores. *Journal of Educational Measurement*, 41(2):137–148, 2004.
- [6] Wim J. van der Linden. Some alternatives to sympson-hetter item-exposure control in computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, 28(3):249–265, 2003.
- [7] Zhewen Fan, Chun Wang, Hua-Hua Chang, and Jeffrey Douglas. Utilizing response time distributions for item selection in cat. *Journal of Educational and Behavioral Statistics*, 37(5):655–670, 2012.
- [8] Wim J. van der Linden and Lynda M. Reese. A model for optimal constrained adaptive testing. *Applied Psychological Measurement*, 22(3):259–270, 09 1998.
- [9] Wim J. van der Linden. Predictive control of speededness in adaptive testing. *Applied Psychological Measurement*, 33(1):25–41, 2009.
- [10] Chun-Kit Yeung. Deep-irt: Make deep learning based knowledge tracing explainable using item response theory, 2019.
- [11] Emiko Tsutsumi, Ryo Kinoshita, and Maomi Ueno. Deep-irt with independent student and item networks. In *Educational Data Mining*, 2021.
- [12] Emiko Tsutsumi, Ryo Kinoshita, and Maomi Ueno. Deep item response theory as a novel test theory based on deep learning. *Electronics*, 10(9), 2021.
- [13] 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*, 17:951–965, 2024.
- [14] 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, 2018. Association for Computing Machinery.
- [15] 石川 文弥, 淵本 壱真, and 植野 真臣. 問題項目への所要時間を予測する multi-task deep gaussian process. **人工知能学会研究会資料 先進的学習科学と工学研究会**, 101:01–06, 2024.
  - [16] F M Lord. *Applications of item response theory to practical testing problems*. Routledge, London, England, November 2012.
  - [17] Frederic M Lord and Melvin R Novick. *Statistical Theories of Mental Test Scores*. IAP, November 2008.
  - [18] F B Baker and S H Kim. *Item response theory: Parameter estimation techniques*. CRC Press, 2004.
  - [19] W J Van der Linden et al. *Handbook of item response theory, volume one: Models*. Chapman and Hall/CRC, 2016.
  - [20] W J Van der Linden et al. *Handbook of item response theory, volume two: Statistical Tools*. Chapman and Hall/CRC, 2016.
  - [21] 植野真臣 and 永岡慶三. *e テスティング*. 培風館, 2009.
  - [22] R. Darrell Bock and Robert J. Mislevy. Adaptive eap estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, 6(4):431–444, 1982.
  - [23] Maomi Ueno and Yoshimitsu Miyazawa. *Two-Stage Uniform Adaptive Testing to Balance Measurement Accuracy and Item Exposure*, pages 626–632. 07 2022.