学習者回答予測モデルからの設問の正答者数予測分布推定

江原遥 [1]
[1] 東京学芸大学　教育学部
ehara@u-gakugei.ac.jp

概要

本稿では，問題文から問題文の難しさを考慮して学習者が正答できるかを判定する学習者回答予測タスクを扱う．BERT などの大規模言語モデルを用いる場合，学習者ごとに異なった結果を出す判別ができない問題があり，学習者を表すトークンを問題文に付与してこの問題を解決する手法を筆者が 2022 年に提案した．本稿では，この手法をさらに拡張し，個人化判別対応の微調整済み言語モデルから，問題の難しさ等の性質を「正答者数予測分布」として抽出する手法を提案する．

1　はじめに

学習支援システムにおいて，学習者が項目に回答できるかどうかを予測する事は，学習者に合った水準の項目（設問）の提示など，適応的学習支援を行うための基本的なタスクである．学習者が項目に回答した履歴のデータがあれば，教育心理学などで能力や難しさのモデル化に多用される項目反応理論 (Item Response Theory，以下 IRT) [1] を用いることで，学習者の能力と項目の難しさを推定し，学習者の反応予測を行う事ができる．しかし，IRT に基づくモデルは通常，学習者の回答パターンにのみ依存し，項目（設問）が自然文で書かれていても文意を理解しない．自然言語処理においては，近年，Transformer モデルに代表される深層言語モデルが自然文理解で高い性能を示している [2] ため，設問文の理解に，これらの深層言語モデルを用いたい．しかし，これらの言語モデルは，通常，言語のみをモデル化するため，学習者ごとに異なった判定を行うことができず，学習者反応の予測に用いることが難しい問題があった．

この問題に対し，筆者は，設問文を考慮した学習者反応の予測問題に適用する簡便な方法を提案した [3]．この手法では，事前学習済の深層言語モデル Bidirectional Encoder Representation of Tranformers (BERT)[2] に，学習者 ID を表す語を新語として追加し，設問文の文頭に学習者 ID を表す語を置くことで，「この学習者 ID の被験者が次の設問文で表される問題に正答できるか否か」を予測するように，テスト結果データセット上で，微調整 (fine-tune) している．この手法は，自然文で記述されている設問に対して，複数の学習者が正答/誤答が明瞭にわかる形式（多肢選択式など）で回答する試験結果データであれば，幅広く適用することが原理的に可能であるため，汎用性が高い．教育の目的では，さらに，予測だけでなく，設問の難しさ（困難度）や，設問が良問である度合い（識別力）も微調整済モデルから取得したい．こうした設問に関する値を**既に学習者の回答がわかっている設定で**取得できる統計的手法としては，項目反応理論（Item Response Theory, IRT）の 2 パラメータモデルが知られている．しかし，**一部の学習者の回答結果が不明であり予測しなければならない設定で**，予測性能が高いだけでなく，さらに，こうした設問文の解釈に重要な値を深層言語モデルから取得する手法はわかっていなかった．

本稿では，微調整済の BERT による学習者回答予測モデルで，高い予測精度を持ったうえ，設問の難しさなど教育上重要な情報を取得できる正答者数分布を推定する手法を提案する．提案手法から抽出した設問の性質は，**評価データ中の学習者の回答を与えた設定で** IRT を用いて推定した設問の困難度・識別力と統計的有意に相関した．提案手法により，高い予測性能に加え，IRT のような高い解釈性を持つ微調整済 BERT を構築できることが示された．

2　関連研究

本研究で扱うのはテストなどの設問文とその回答データであり，学習者が限られた時間で回答できるものである．一方，長期にわたり，どの学習者がどの設問に正答/誤答しそうかという**時系列回答データ**から，問題の難しさに関する埋め込みを作成するタスクは知識追跡（Knowledge Tracing）という名前

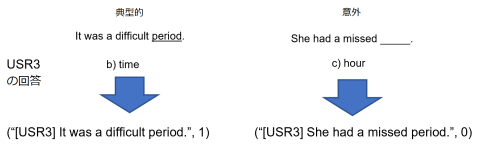

図 1　学習者トークンの導入．

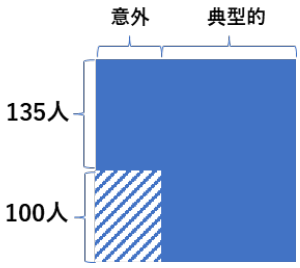

図 2　実験設定．青く塗られた部分がパラメタ推定に使われる訓練データ．斜線部が性能比較に用いられるテストデータ．

で知られており，データマイニング分野などで研究されている [4, 5, 6, 7]．しかし，これらの中では，設問文のテキスト情報は利用されていない．その理由は，知識追跡タスクの評価に標準的に用いられているデータセットが数学などいわゆる理系分野の問題であるため，設問文の言語的知識よりも，過去の回答データの方が設問の難しさを推定するのに有効な設定であるためと推察される．設問文のテキスト情報を設問間関係に変換して知識追跡に利用する研究はあるものの [8]，知識追跡は，長期の時系列回答データが対象であるため，本研究とは設定も目的も異なる．そのほか，近年，BERT を用いた教育応用が提案されているが [9, 10, 11]，これらの研究では学習者回答予測については扱われていない．

3　予測性能評価

本研究は，設問文から直接，設問の難しさを推定する研究であるため，１）一文程度の短い分量で，２）文中の語の意味を捉えることが回答の正誤に大きくかかわるデータセットで評価することが，結果がわかりやすく望ましい．このため，頻出英単語の典型的な語義と，意外な語義の設問に，それぞれ多肢選択式で答えてもらったデータセット [3] を用いて性能評価を行った．[3] のデータセットでは，英語母語話者に問題として成立していることを確認してもらったうえで，クラウドソーシングサービス Lancers[1] から，2021 年 1 月に収集した．より詳細は [3] を参照されたい．

これにより，対応する問題は 12 問となる．

1）　https://lancers.co.jp/

表 1　図 2 斜線部の予測精度 (accuracy)．

手法	精度
IRT (能力 - 235 人から推定した典型的な語義の困難度)	0.544
IRT (能力 - 135 人から推定した意外な語義の困難度)	0.644
[3] (bert-large-cased)	0.674 (**)
[3] (bert-base-cased)	**0.688** (**)
[3] (bert-base-uncased)	0.655
[3] (roberta-base)	0.681 (**)
[3] (albert-base-cased)	0.671 (*)

Transformer モデルを個人化判別に対応させる手法は，自然言語処理の言語教育応用の目的では著者の知る限り知られていない．ただし，Transformer モデルに特殊なトークン（語）を加えて微調整を行い，様々な問題設定に対応させる手法は知られており，ライブラリ上で特殊なトークンを加える機能が用意されている．本研究では，この機能を利用することで，学習者に対応するトークン（**学習者トークン**）を作り，これを入力系列の最初に置くことによって判別を行う手法を提案する（図 1）．例えば，学習者 ID が 3 番の学習者を表すトークン “[USR3]” を導入し，“[USR3] It was a difficult period.” が入力であれば，3 番の学習者が “It was a difficult period.” という文から成る設問に正答するか否かを予測する問題に帰着させる．入力文はそのままで，入力文の前に，単純に学習者トークンが挿入されている点に注意されたい．導入するトークン数は学習者数と同数である．Transformer では各トークンに対して，その語としての機能を表現する**単語埋め込みベクトル**があるので，学習者トークンに対しても埋め込みベクトルが作られる．

重要な点として，提案手法では，文中のどの語についての設問であるかという情報や，誤答選択肢の情報は与えていない．すなわち，提案手法の判別器は，表 2 のどの単語に下線が引かれているかや，表 2 や表 3 の正答以外の選択肢の情報を用いない．提案手法は，単純に正解となる文を入力とし，これを学習者が理解できるか否かを判別する判別器を構成している，と解釈できる．これにより，提案手法は，表 2 と表 3 という仔細の異なる 2 種類の多肢選択式の問題に対応できる．このように，提案手法の適用範囲を広くとることができる．今回の設定では，入力文が短文であり，学習者が 1 語でもわからなければ正答できない設問で構成されているため，語義を知っている事と正解となる文を理解できるか否かは，同一視できる．

Transformer モデルのその他の実験設定については多用される設定とした．判別には，**transformers** ラ

イブラリの **AutoModelForSequenceClassification** を用いた．微調整の訓練には Adam 法 [12] を用い，バッチサイズは 32 とした．

Transformer モデルを用いた結果を，表 1 に示す．* は IRT の最高性能と比較して Wilcoxon 検定で統計的有意であることを表し，**は $p < 0.01$，*は $p < 0.05$ を表す．また提案手法の（）内は用いた事前学習済モデル名である．表 1 では，まず，学習者トークンを導入した提案手法が，IRT を用いた従来手法より高い性能を達成していることが分かる．この実験結果は，設問文の意味を考慮する事で，IRT より高精度な判別が行えることを示している．

4 設問の難しさや識別力の抽出法

ここまでは微調整済の BERT モデルから学習者の能力値を抽出する方法であったが，さらに，設問の難しさや識別力に相当する値を抽出する方法を提案する．方法の概略を示す．BERT は被験者が設問文が指定されれば，その被験者がその設問に正答できるかどうかだけでなく，その確率値も予測として出力できる．ある設問に着目し，全被験者がその設問を解いた時の正答できる確率を BERT に出力させ，ここからその設問の正答者数の確率分布を計算する．被験者間の独立性を仮定すると，数学的には，成功確率が互いに異なる独立なベルヌーイ分布の和の分布であるポアソン 2 項分布を計算する事に相当する．この時，その設問の正答者数の確率分布の平均を設問の難易度，分散を識別力のような設問の良さと解釈する事が可能になる．

ここでは，被験者数を N 人とし，学習者の添字を n とする（厳密には，被験者の中から特定の被験者を選び N と J が異なる設定もあり得るので，違う文字でおいた）．項目数を I 個とし，項目の添字を i とする．学習データ上で予測器を微調整した後，予測器は学習者 n が項目 i に正しく回答する確率を出力することができる．この確率を $BERTProb(n,i)$ と表記する．簡単のために，ここからは設問 i に焦点を当てる．$BERTProb(n,i)$ を使って，N 人のうち，質問 i に正答する者の確率分布を求めたい．そこで $BERTProb(n,i)$ の確率で 1，そうでなければ 0 となるベルヌーイ分布に従う確率変数 A_n を $A_n \sim Bernoulli(BERTProb(n,i))$ と定義する．ここで，簡単のため，これらの確率変数 $\{A_1,\ldots,A_n\}$ は互いに独立であると仮定する．学習者について和をとり，項目 i の全 N 人の中での正答者数の確率分布は次のように書ける．

$$A_i = \sum_{n=1}^{N} A_n \tag{1}$$

式 1 は互いに独立なベルヌーイ分布の和であり，ポアソン 2 項分布と呼ばれる[2)]．この分布の計算は，動的計画法を用いて計算可能である．[13, 14] ではポアソン 2 項分布の計算を全く違うタスクに対して行う中で詳述しているので，計算アルゴリズムの詳細はこちらを参照されたい．

A_i は確率分布なので，平均と分散を計算できる．A_i は，全 N 人のうち，項目 i の正答者数である事に注意すると，A_i の平均は，問題 i の難易度を表していると解釈できる．また，A_i の分散は，問題 i の正解者数を予測のためのエラーバーと解釈できる．同じような難しさの設問の中では，分散が最も小さい，つまり，正答者数の予測がつきやすい問題が良問と考えられる．A_i の分散は，項目反応理論における「識別力」に似た性質を持つ指標である．項目反応理論の識別力は，項目が能力の高い被験者と低い被験者を識別する力を表す．直感的には，能力が本当は高い被験者が間違えてしまうような確率の少ない「良問」である度合いを表す．A_i の分散も，項目反応理論の識別力のように良問である度合を表すが，項目反応理論はモデルが固定されているのに対し，A_i の分散は予測器 $BERTProb(n,i)$ の確率値さえわかればどのような予測器を用いても計算できるので，深層転移学習のような複雑な手法を用いた場合でも計算できる．

横軸に A_i の分散，縦軸に A_i の平均値をとることでリスク・リターンプロットを作成できる．まず，どの程度の難しさの設問を選びたいかを決めて縦軸の値に注目し，次に同程度の難しさの問題の中で横軸の値が最も小さいもの（最も左にあるもの）を選ぶことで，特定の難易度の良問を選択可能である．この最も左にある点を結んだ線を「効率的フロンティア」という [14]．

図 3 と図 4 に，ある項目（設問）i について，受験者数がそれぞれ 5 人，235 人である場合の分布を描いた．（5 人については，ランダムに受験者を選んだ）図 3 から，受験者数が少ないときでも，非対称な分布の形が計算できている事が分かる．また，図 6 に，リスク-リターンプロットを描いた．各点は前述の 12 問の設問であり，破線は効率的フロン

2) https://en.wikipedia.org/wiki/Poisson_binomial_distribution

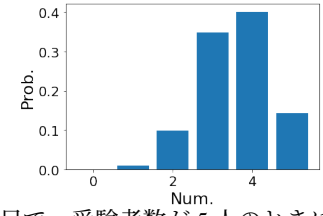

図 3　ある項目で，受験者数が 5 人のときに予測される正答者数の分布．

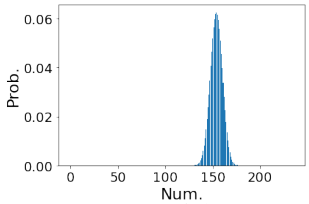

図 4　図 3 と同じ項目で，受験者数が 235 人であるときに予測される正答者数の分布．

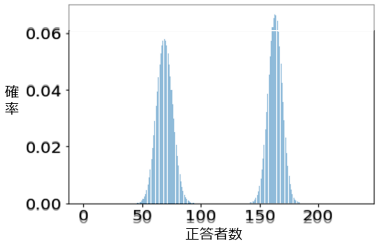

図 5　受験者数が 235 人で 2 問の予測される正答者数の分布を重ねたもの．

ティアである．全 12 問のうち，効率的フロンティア上の問題を選ぶことで，3 問の「良問」をさらに選び出せている事が分かる．また，図 5 に，難しい問題（正答者数が少ない問題）と易しい問題のグラフを重ねた．わずかではあるが，難しい問題では，釣り鐘型の横幅（分散）が大きいことが見て取れる．

今回，提案手法は予測される正答者数の分布の平均を設問の難しさとして，標準偏差を設問の良問度合い（設問の難しさ推定のしやすさ）として出力できる．こうした値は，IRT においても，それぞれ，困難度，識別力という名前で知られている．図 7，図 8 に，提案手法による値と，テストデータ中の値を与えたうえで IRT が推定した困難度・識別力の値の（簡単のため）負値を図示する．相関係数は，図 7 では 0.78 ($p < 0.01$)，図 8 では 0.62 ($p < 0.05$) であり，どちらも統計的有意な相関がみられた．

5　おわりに

本研究では，設問文を考慮して，学習者が所与の設問に正答/誤答するかを予測する学習者回答予測のタスクにおいて，高い予測性能を保ちつつ，設問の難しさや良問度合いなどの教育上重要な情報を取得できる手法を提案した．具体的には，深層言語モデル BERT の微調整段階で，学習者を表す単語を導入することにより学習者ごとに異なる出力を行う既存手法 [3] を拡張し，微調整済 BERT の予測確率値から，予測される正答者数分布を求める手法を提案した．これにより，平均を各設問の難易度とみなせ，その分散を各設問の良問度合いとみなせる事を示した．[3] により，微調整済 BERT の学習者を表す単語の埋め込みベクトルから学習者の能力値を取り出す方法も提案している．本研究により，IRT と同様，微調整済 BERT から学習者の能力値と設問の難易度の両方の情報を解釈することが可能になった．

今後の課題としては，設問の文ベクトルと難易度や識別力の関係性を求めることなどがあげられる．

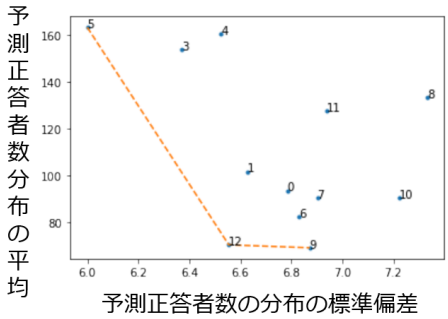

図 6　各設問について，受験者の総数が 235 である場合のリスク（横軸，各設問の予測される正解者数の分布の分散）とリターン（縦軸，各設問の予測される正解者数の分布の平均）をプロットしたものである．各点は設問を表し，各点の番号は設問番号である．縦軸が同程度の値であれば，分散が小さい設問（図中左側の設問）が良問である．

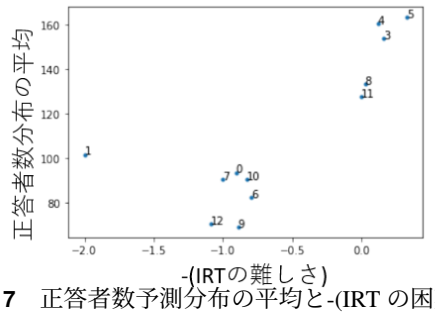

図 7　正答者数予測分布の平均と-(IRT の困難度).

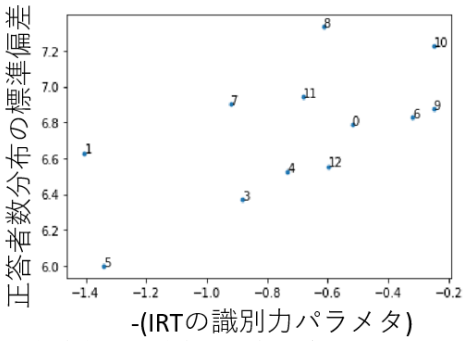

図 8　正答者数予測分布の標準偏差と-(IRT の識別力).

謝辞

本研究は，科学技術振興機構 ACT-X 研究費 (JPMJAX2006) の支援を受けた．

参考文献

[1] Frank B. Baker. **Item Response Theory : Parameter Estimation Techniques, Second Edition**. CRC Press, July 2004.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In **Proc. of NAACL**, 2019.

[3] Yo Ehara. No meaning left unlearned: Predicting learners' knowledge of atypical meanings of words from vocabulary tests for their typical meanings. In **Proc. of Educational Data Mining (short paper)**, 2022.

[4] Aritra Ghosh, Neil Heffernan, and Andrew S. Lan. Context-Aware Attentive Knowledge Tracing. In **Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining**, KDD '20, pp. 2330–2339, New York, NY, USA, 2020. Association for Computing Machinery.

[5] Shuanghong Shen, Qi Liu, Enhong Chen, Zhenya Huang, Wei Huang, Yu Yin, Yu Su, and Shijin Wang. Learning Process-consistent Knowledge Tracing. In **Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining**, KDD '21, pp. 1452–1460, New York, NY, USA, 2021. Association for Computing Machinery.

[6] Shuanghong Shen, Zhenya Huang, Qi Liu, Yu Su, Shijin Wang, and Enhong Chen. Assessing Student's Dynamic Knowledge State by Exploring the Question Difficulty Effect. In **Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval**, SIGIR '22, pp. 427–437, New York, NY, USA, 2022. Association for Computing Machinery.

[7] Ghodai Abdelrahman and Qing Wang. Deep Graph Memory Networks for Forgetting-Robust Knowledge Tracing. **IEEE Transactions on Knowledge and Data Engineering**, pp. 1–13, 2022. Conference Name: IEEE Transactions on Knowledge and Data Engineering.

[8] Shalini Pandey and Jaideep Srivastava. RKT: Relation-Aware Self-Attention for Knowledge Tracing. In **Proceedings of the 29th ACM International Conference on Information & Knowledge Management**, CIKM '20, pp. 1205–1214, New York, NY, USA, 2020. Association for Computing Machinery.

[9] Jia Tracy Shen, Michiharu Yamashita, Ethan Prihar, Neil Heffernan, Xintao Wu, Sean McGrew, and Dongwon Lee. Classifying math knowledge components via task-adaptive pre-trained bert. In **Proc. of AIED**, pp. 408–419, 2021.

[10] Lele Sha, Mladen Rakovic, Alexander Whitelock-Wainwright, David Carroll, Victoria M Yew, Dragan Gasevic, and Guanliang Chen. Assessing algorithmic fairness in automatic classifiers of educational forum posts. In **Proc. of AIED**, pp. 381–394, 2021.

[11] Shiting Xu, Guowei Xu, Peilei Jia, Wenbiao Ding, Zhongqin Wu, and Zitao Liu. Automatic task requirements writing evaluation via machine reading comprehension. In **Proc. of AIED**, pp. 446–458. Springer, 2021.

[12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In **Proc. of ICLR**, 2015.

[13] Yo Ehara. Lurat: a lightweight unsupervised automatic readability assessment toolkit for second language learners. In **Proc. of ICTAI**, pp. 806–814, 2021.

[14] Yo Ehara. Selecting reading texts suitable for incidental vocabulary learning by considering the estimated distribution of acquired vocabulary. In **Proc. of Educational Data Mining (poster paper)**, 2022.

表 2　実際の設問例.

It was a difficult period.
a) question
b) time
c) thing to do
d) book

表 3　意外な意味を問う設問例.

She had a missed ------.
a) time
b) period
c) hour
d) duration

A　データセットの詳細 [3]

[3] で用いたデータセットについて詳述する.

この 2 つの工夫を施した実際の設問例が表 3 である．“period” には通常の「期間」の他に「生理」という意味があり，これを問うている．学習者は，70 問の通常の用例の語彙テストの前に，表 3 のような設問を 13 問解くように求められる．ただし，先に解く表 3 の形式の選択肢が，表 2 の形式の問題に影響していないかどうかを後で確認できるよう，意外な語義ではあるが，通常の語義の設問群の側に対応する設問がない設問を 1 問設けた.

B　IRT による学習者回答予測

語の意外と思われる語義の難しさを典型的な語義の難しさで代替してしまうと，学習者が設問に正答/誤答するかを IRT で予測する際，どの程度の悪影響があるのだろうか？これを調べるために，次の実験を行った．まず，235 人の学習者を 135 人と 100 人に分ける（図 2）．意外と思われる語義の設問群（12 問）のパラメタについては前者の 135 人の学習者反応だけから，典型的な語義の設問群（70 問）のパラメタについては 235 人全員の学習者反応で推定する．この推定の際には，後者の 100 人× 12 問，計 1,200 件の回答データは用いていないことに注意されたい．項目反応理論では，推定された学習者 θ_j の能力値 θ_j, 語義の困難度 d_i を用い，$\theta_j > d_i$ であれば学習者 j が設問 i に正答，そうでなければ誤答と判定する．設問 i の困難度パラメタとして，意外と思われる語義の 12 問の困難度パラメタを直接用いた場合と，対応する語の典型的な語義の困難度パラメタで代替した場合で，この 1,200 件の回答データの予測精度を比較した．予測精度 (accuracy) の結果を表 1 に記す．その結果，直接用いた場合の予測精度は 64.4%，典型的な語義の困難度で代替した場合は 54.4% と，10 ポイントの差が出た．この差は，Wilcoxon 検定で $p < 0.01$ で有意であった．この結果から，学習者反応の予測における，語の語義ごとに困難度を推定することの重要性がわかる．より直接的に言い換えれば，この結果は，語の意外な用例の難しさを，語の典型的な用例の難しさで置き換えると，学習者回答予測の精度が著しく低下することを示唆している.

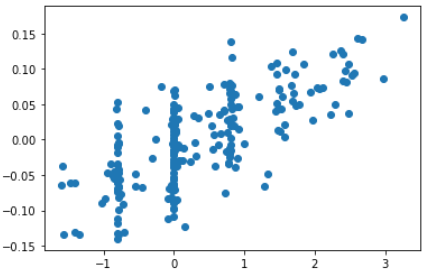

図 9　IRT の能力パラメタ（横軸，pyirt によって算出）と，学習者トークンの単語埋め込みベクトルの第一主成分得点（縦軸）.

C　能力値抽出

[3] では，次の手順で微調整済モデルからの能力値抽出に成功している．微調整後の **bert-large-cased** の場合の学習者トークンに対する単語埋め込みベクトルのみを集めた．すなわち，学習者の人数分の単語埋め込みベクトルの集合がある．このベクトル集合に対して主成分分析を行い，その第一主成分得点と IRT の能力値パラメタを比較した（図 9）．各点は学習者を表す．IRT の能力値パラメタの算出には，Python の **pyirt** ライブラリを用いた．両者は相関係数 0.72 という強い相関を示した（$p < 0.01$）．これにより，提案手法を用いた場合でも，能力値は学習者トークンの第一主成分得点として容易に抽出できることが分かった．これにより，提案手法は文意を考慮することにより IRT より高い精度を達成しながら，IRT と同様に「能力値を取り出せる」という高い解釈性を持つことが示された.

図 9 では，縦に筋が入っているように見える部分がある．これは，pyirt の内部で使われている IRT のパラメタ推定アルゴリズムの性質で，横軸の学習者の能力値パラメタの推定の際，能力に大きな差がない能力値パラメタは 1 つの値にまとめられる性質があるため，横軸が同じ値を取る学習者が存在するためである.