多次元項目反応理論と深層学習に基づく複数観点同時自動採点手法

柴田拓海 [1] 宇都雅輝 [1]
[1] 電気通信大学大学院院
{shibata,uto}@ai.lab.uec.ac.jp

概要

近年，深層学習を用いた小論文自動採点手法として，全体得点と複数の評価観点別得点を同時に予測する手法が提案されている．しかし従来手法は，予測の根拠について解釈性が低いという問題があった．この問題を解決するために，本研究では，多次元項目反応理論を利用して予測根拠の解釈性を高めた複数観点同時自動採点手法を提案する．

1 はじめに

近年，小論文試験の採点をコンピュータを用いて自動化する小論文自動採点（Automated Essay Scoring；AES）手法が多数提案されている．自動採点を実現する手法は特徴量ベースと深層学習ベースの手法に大別される．これまでは，特徴量ベースの手法が一般的であったが（e.g., [1–4]），近年では深層学習を用いた自動採点モデルが多数提案されている（e.g., [5–19]）．深層学習自動採点モデルは文章の単語系列を入力として，データから自動で複雑な特徴量を学習でき，高精度を達成している．

従来の自動採点モデルの多くは，全体得点のみを予測する採点場面を想定している（e.g., [6–15]）．しかし，学習場面などで小論文試験を運用する場合，より詳細なフィードバックを受験者に行うために，論理構成力や文章表現力などの評価観点別の得点付けを行いたい場面も少なくない [16]．そこで，複数の評価観点に対応する得点を同時に予測できるモデルもいくつか提案されている（e.g., [16–19]）．

現時点では Ridley ら [19] のモデルが最高精度を達成しているが，このモデルには解釈性の観点から次のような問題がある．(1) 評価観点ごとに複雑な多層ニューラルネットワークを持つため予測根拠を解釈することが難しい．(2) 複数観点での評価では，測定対象の能力に観点間で共通性が仮定できる場合が多いが [20]，このモデルでは観点間の相関は考慮しているものの，背後にどのような能力尺度が想定されるかは解釈できない．

これらの問題を解決するために，本研究では，項目反応理論を組み込んだモデルを提案する．具体的には，評価観点の特性を考慮した多次元項目反応モデル [21] を出力層とし，それ以外を Ridley らのモデルを元にした評価観点共通のニューラルネットワークとしたモデルを開発する．提案手法の利点は以下の通りである．(1) 評価観点固有の出力層は，識別力と困難度と呼ばれる項目反応理論で一般的な 2 種類のパラメータのみで説明されるため，それらのパラメータ値に基づいて観点ごとの特性を定量的に解釈できる．(2) 多次元項目反応モデル層の能力次元数を最適化してパラメータを分析することで，複数評価観点の背後に想定される能力尺度を解釈できる．

本論文では，複数観点自動採点の研究で広く利用されるベンチマークデータセットを用いて提案手法の有効性を評価する．実験の結果，提案手法は従来手法から大きく性能を落とすことなく，妥当な解釈が可能なパラメータ値を与えたことが確認できた．また，背後に想定される能力次元数としては 1 次元が最適であることが確認できた．このことは，少なくとも本研究で使用したデータセットにおいては，観点間の関係は従来モデルのような複雑な構成でなくとも説明できることを示唆している．

2 複数観点同時自動採点モデル

提案モデルは，Ridley ら [19] が提案した複数観点同時自動採点モデルを基礎モデルとするため，本章ではこのモデルについて説明する．モデルの概念図を図 1（左）に示した．このモデルは，受験者 n の小論文を入力とし，評価観点 $m \in \mathcal{M} = \{1, 2, \ldots, M\}$ に対応する得点 $\hat{y}_n^m$ を出力する．ここで M は評価観点数を表す．また，受験者 n の小論文は単語系列と

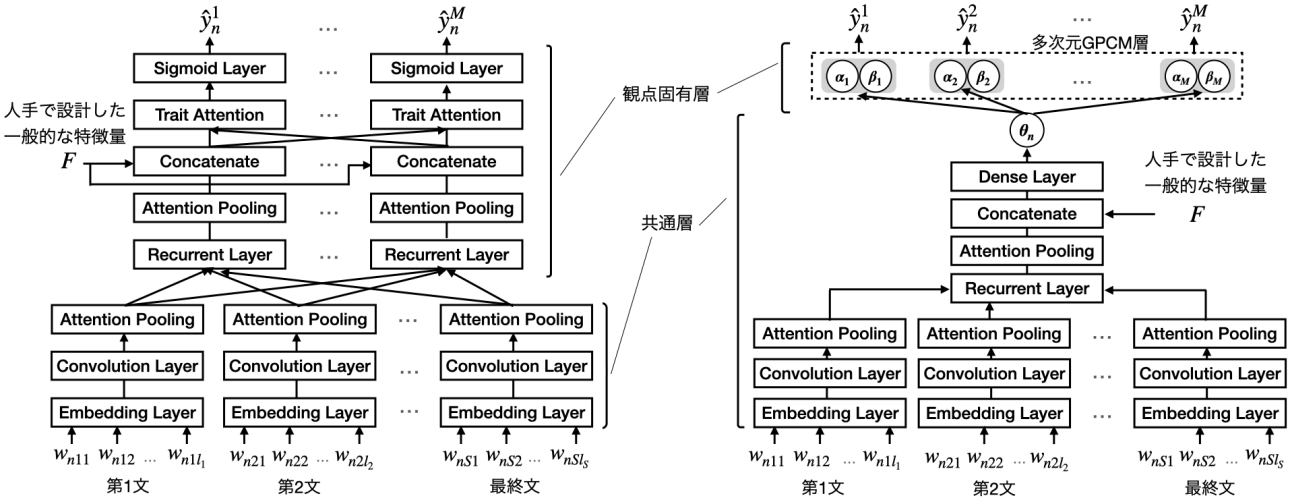

図 1　従来の複数観点同時自動採点モデル（左）と提案モデル（右）の概念図

して，$\{w_{nsl}|s \in \{1,2,\ldots,S\}, l \in \{1,2,\ldots,l_s\}\}$ と表せる．w_{nsl} は受験者 n の小論文における s 番目の文の l 番目の単語であり，S はその小論文の文数，l_s は s 番目の文の単語数である．このモデルでは，各単語 w_{nsl} をそれぞれ品詞（part-of-speech；POS）タグ p_{nsl} に変換し，POS タグ系列を入力として用いている．

得点予測は，共通層と観点固有層の二段階でデータを処理することで行われる．共通層では，文ごとに Embedding 層，Convolution 層，Attention Pooling 層 [11] が適用され，全観点に共通する文単位の分散表現の系列が得られる．

次に，観点ごとに独立に処理を行う観点固有層で，共通層で得られた分散表現の系列をもとに評価観点ごとの得点を予測する．共通層の出力系列に対して，観点ごとに Recurrent 層 [22]，Attention Pooling 層が適用される．さらに単語数や可読性，文章の複雑さなどを表す人手で設計した特徴量のベクトル F を結合（concat）することで文章単位の分散表現が得られる．次にこの結合されたベクトルに対して，評価観点間の関係を考慮するために Trait Attention [19] を適用し，得点予測のための最終的な分散表現 $\boldsymbol{c}_{nm}$ が得られる．最後に，この $\boldsymbol{c}_{nm}$ に対し，シグモイド関数を活性化関数に持つ全結合層を適用させ，受験者 n の m 番目の観点別得点 $\hat{y}_n^m$ を $\hat{y}_n^m = \sigma(W_m \boldsymbol{c}_{nm} + b_m)$ で予測する．ここで，σ はシグモイド関数，W_m は重み，b_m はバイアスを表す．なお，このモデルは得点予測にシグモイド関数を使用しているため，$\hat{y}_n^m$ は 0 から 1 の間の値をとる．実際の得点尺度がこれと異なる場合には，$\hat{y}_n^m$ を一次変換して実際の得点尺度に合わせる．

モデルの学習は，平均二乗誤差（Mean Squared Error；MSE）を損失関数として誤差逆伝播法で行われる．訓練データの小論文数が N，予測観点数が M のとき，MSE 誤差は以下のように表される．

$$\mathcal{L}_{MSE} = \frac{1}{NM} \sum_{n=1}^{N} \sum_{m=1}^{M} D(n,m)(\hat{y}_n^m - y_n^m)^2 \qquad (1)$$

ここで，y_n^m は受験者 n の小論文における m 番目の観点の真の得点を表す．また $D(n,m)$ は，n 番目の小論文における m 番目の観点に対するデータのときに 1 を，そうでなければ 0 を返す関数である．

第 1 章でも述べた通り，このモデルは評価観点ごとに固有の複雑な層を持つため予測根拠の解釈が難しい．この問題を解決するために，本研究では次章で説明する項目反応理論を用いる．

3　項目反応理論

項目反応理論（Item Response Theory；IRT）[23] は，近年のコンピュータ・テスティングの発展に伴い，様々な分野で実用化が進められている数理モデルを用いたテスト理論の一つである．IRT モデルは正誤データなどの 2 値型のデータを前提とするものが多いが，多段階の得点データに対応した IRT モデルも多数提案されている．また，一般的な IRT モデルでは，測定対象の能力に 1 次元性を仮定しているが，測定される能力に多次元性を仮定できるモデルも提案されている．

本研究では，代表的な多次元多値型 IRT モデルである多次元一般化部分採点モデル（Generalized Partial Credit Model；GPCM）[24] を使用する．ここでは，先行研究 [20,21] のように各評価観点を項目とみなして多次元 GPCM を適用する．具体的には，受験

者 n が評価観点 m において，得点 $k \in \{1, 2, \ldots, K_m\}$ を得る確率を次式で与えるモデルを適用する．

$$P_{nmk} = \frac{\exp(k\boldsymbol{\alpha}_m^T\boldsymbol{\theta}_n + \sum_{u=1}^{k}\beta_{mu})}{\sum_{v=1}^{K_m}\exp(v\boldsymbol{\alpha}_m^T\boldsymbol{\theta}_n + \sum_{u=1}^{v}\beta_{mu})} \quad (2)$$

ここで，$\boldsymbol{\theta}_n = (\theta_{n1}, \theta_{n2}, \ldots, \theta_{nd})$ は受験者 n の d 次元の能力を表すパラメータベクトルであり，ベクトルの各要素は各次元の能力値を表す．$\boldsymbol{\alpha}_m = (\alpha_{m1}, \alpha_{m2}, \ldots, \alpha_{md})$ は $\boldsymbol{\theta}_n$ に対応した評価観点 m の d 次元識別力，β_{mu} は評価観点 m においてカテゴリ $u-1$ から u に遷移する困難度を表すパラメータである．K_m は，評価観点 m における得点段階数を表す．なお，モデルの識別性のために，$\beta_{m1} = 0 : \forall m$ を所与とする．

4 提案モデル

本研究では，上記の多次元 GPCM を組み込んだ複数観点同時自動採点モデルを提案する．提案モデルの概念図を図 1（右）に示す．図 1 からわかるように，提案モデルは入力層から Concatenate 層まで評価観点数 $M = 1$ とした従来モデルと同じ構造を持ち，これらの層を用いて文章単位の分散表現 $\boldsymbol{c}_n$ を生成する．提案モデルでは，このベクトル $\boldsymbol{c}_n$ を全結合層に入力し，多次元 IRT における能力値ベクトル $\boldsymbol{\theta}_n$ に対応する値を $\boldsymbol{\theta}_n = \boldsymbol{W}\boldsymbol{c}_n + \boldsymbol{b}$ で求める．ここで，$\boldsymbol{W}$ は重み行列，$\boldsymbol{b}$ はバイアスベクトルを表す．

最後に，得られた $\boldsymbol{\theta}_n$ を用いて，多次元 GPCM 層で式（2）を計算することで，各評価観点 $m \in \mathcal{M}$ に対する得点の出力確率が得られる．得点予測の際には，確率 P_{nmk} が最大となるカテゴリ $\arg\max_k P_{nmk}$ を予測得点とする．提案モデルにおける多次元 GPCM 層は従来モデルにおける観点固有層に相当する．

損失関数には，多クラス交差エントロピー（Categorical Cross-Entropy；CCE）誤差を用いる．訓練データの小論文数が N，予測観点数が M のとき，CCE 誤差は以下のように表せる．

$$\mathcal{L}_{CCE} = -\frac{1}{NM}\sum_{n=1}^{N}\sum_{m=1}^{M}\sum_{k=1}^{K_m} y_{nmk}\log(P_{nmk}) \quad (3)$$

なお，モデルの各種ハイパーパラメータは先行研究 [19] に合わせ，最適化アルゴリズムには学習率を 0.001 に設定した RMSProp [25] を用いる．

以降では，提案モデルの解釈の方法について述べる．前章でも述べた通り，提案モデルで使用している多次元 GPCM は，評価観点の識別力 $\boldsymbol{\alpha}_m$ と困難度 $\boldsymbol{\beta}_m$，および受験者の能力 $\boldsymbol{\theta}_n$ の 3 つのパラメータで説明される．識別力 $\boldsymbol{\alpha}_m$ と困難度 $\boldsymbol{\beta}_m$ の値からは，各評価観点がどの程度受験者の能力を識別でき，どれくらいの難易度であるかが解釈可能である．さらに，能力値 $\boldsymbol{\theta}_n$ からは，その小論文を執筆した受験者の潜在的な多次元能力を読み取ることができる．

また，$\boldsymbol{\theta}_n$ の次元数を変えて，提案モデルの性能を評価することで，得点データの背後に想定される最適な能力次元数を分析できる．例えば，$\boldsymbol{\theta}_n$ に 2 次元を想定したときに最大の予測精度を示したとすると，その評価観点の背後に 2 次元的な能力尺度が想定されると解釈できる．最適な次元数でモデルパラメータの分析を行うことで，尺度の構成を解釈することが可能となる．

5 実験

本研究では，実データとして，Automated Student Assessment Prize (ASAP) と，ASAP++ [26] を用いる．ASAP は AES 研究の分野で広く使用されるデータセットである．ASAP と ASAP++には 8 つの小論文課題に関する答案が含まれており，それぞれの答案に対して全体得点と評価観点別の得点が与えられている．小論文数の課題ごとの平均は約 1622，平均単語数は 275 である．なお課題 1 と課題 2 は，Content，Organization，Word Choice，Sentence Fluency，Conventions，課題 3 から課題 6 は Content，Prompt Adherence，Language，Narrativity，課題 7 は Content，Organization，Conventions，Style，課題 8 は課題 1 と 2 に共通する評価観点に加えて Voice といった評価観点でそれぞれ得点付けされている．

5.1 得点予測精度の評価実験

ここでは，提案モデルの次元数を 1，2，3 と変化させて得点予測精度を評価する実験を行う．本実験では，モデルへの入力として，Ridley らと同様の POS タグを用いる場合と，一般的な深層学習自動採点モデルと同様に単語系列を利用する場合の両方を考える．なお単語系列を入力する場合には，Embedding 層で，50 次元の GloVe [27] による事前学習済みの単語埋め込みを利用する．

モデルの性能評価は，5 分割交差検証を用いて行う．5 分割交差検証は課題ごとに独立して実施し，エポック数は全てのモデルで 30 としている．評価関数には，2 次の重み付きカッパ係数（Quadratic Weighted Kappa；QWK）を用いる．

表 1　課題別の平均 QWK スコア

入力	モデル	課題番号									p 値		
		1	2	3	4	5	6	7	8	Avg.	提案-1dim	提案-2dim	提案-3dim
POS	従来モデル	**0.678**	0.629	0.595	0.659	**0.696**	0.673	**0.672**	**0.569**	**0.646**	1.000	1.000	1.000
	提案-1dim	0.651	0.616	0.620	**0.670**	0.682	0.685	0.619	0.480	0.628	-	1.000	1.000
	提案-2dim	0.661	0.608	0.629	**0.670**	0.679	0.675	0.620	0.445	0.623	-	-	1.000
	提案-3dim	0.636	**0.633**	**0.634**	0.656	0.685	**0.694**	0.636	0.471	0.631	-	-	-
単語	従来モデル	**0.694**	**0.664**	**0.660**	0.718	**0.704**	**0.745**	**0.728**	**0.530**	**0.680**	0.096	0.093	0.062
	提案-1dim	0.641	0.625	0.646	0.718	0.690	0.737	0.637	0.464	0.645	-	0.973	0.698
	提案-2dim	0.636	0.620	0.656	**0.721**	0.692	0.736	0.675	0.486	0.653	-	-	1.000
	提案-3dim	0.656	0.630	0.656	0.712	0.696	0.734	0.687	0.472	0.655	-	-	-

表 2　提案モデル（1 次元）を用いて推定した課題 2 の評価観点パラメータ

	識別力	困難度				
	α_{21}	β_{22}	β_{23}	β_{24}	β_{25}	β_{26}
全体得点	1.64	-4.81	-3.24	-0.07	3.51	4.96
Content	2.35	-4.74	-1.57	0.53	2.15	4.56
Organization	2.43	-4.45	-1.27	1.12	2.39	5.55
Word Choice	2.57	-4.71	-1.61	0.81	2.83	5.04
Sentence Fluency	2.25	-4.61	-2.47	0.19	2.57	4.76
Conventions	2.21	-4.22	-1.61	0.59	2.71	5.12

実験結果を表 1 に示す．表 1 では観点ごとに QWK スコアを計算し，その平均スコアを課題ごとに示している．各条件で最も精度が高い手法の結果を太字で示してある．

表 1 より，提案モデルは，次元数によって予測精度が異なることが読み取れるが，各課題に最適な次元数を持つ提案モデルは従来モデルと比較しても精度に大きな差はないことがわかる．ここで，各モデルの平均スコアに有意な差があるかを定量的に測定するため，ボンフェローニ法による多重比較検定を行った．結果を表 1 の「p 値」列に示す．表から，POS 入力，単語入力ともに，全てのモデル間で有意な差は認められなかったことが読み取れる．このことから，提案モデルは，従来モデルと比較して精度を落とさずに得点予測ができたことがわかる．

5.2　次元性の検証

本節では能力尺度の次元性を分析する．ここでは，まず課題 1 の得点データに関して因子分析を行った．その結果，因子数が 1 のときの固有値は 4.97，因子数が 2 のときの固有値は 0.41 と算出され，因子数が 1 から 2 に変化すると固有値が 1 を下回るほど大幅に減少していることがわかる．つまり，データの背後にある尺度構成は 1 次元で十分に説明できることが示唆される．なお，他の課題でも，これと同様の結果が得られた．

また，前節で示した表から，提案モデルの次元数の違いによる予測精度の差はほとんど見られなかったことがわかる．以上から，能力尺度の説明力としては 1 次元で十分であると解釈できる．

5.3　評価観点パラメータの解釈

ここでは，提案モデルで推定された評価観点パラメータの解釈について述べる．例として表 2 に課題 2 のデータにおいて，能力次元数を 1 次元としたときの評価観点パラメータの推定結果を示した．

まず表 2 に示した識別力の値から，それぞれの評価観点が受験者の能力をどの程度測定できるかを解釈できる．例えば，全体得点の識別力値は他の評価観点の値よりも小さいことが読み取れ，このことは全体得点が他の観点と比べて受験者の能力を測定する力が低いことを示している．また，困難度パラメータからは，各観点における各得点の出現分布を解釈することができる．例えば，全体得点のパラメータを所与としたときの式 (2) で表現される項目特性曲線（Item Characteristic Curve；ICC）を描くと，全体得点は得点に中心化傾向があることがわかる．

このように提案モデルでは得点予測の背後にある構造を解釈できることがわかる．

6　まとめ

本研究では，全体得点と同時に観点別得点も予測できる複数観点同時自動採点モデルに，多次元項目反応理論を組み込んだモデルを提案した．提案モデルを用いた実験から，自動採点で広く使用されるベンチマークデータである ASAP と ASAP++データセットでは，採点に多くの評価観点が用いられているにもかかわらず，その背後には少数の能力尺度しか想定されない可能性が示唆された．このことは，観点固有層として，従来モデルのような複雑な内部表現は必要でない可能性や，多次元的な能力評価が実現できていない可能性を示唆している．

今後は多様なデータセットで提案モデルの有効性を評価したい．

謝辞

本研究は JSPS 科研費 19H05663, 20K20817, 21H00898 の助成を受けたものです.

参考文献

[1] Yigal Attali and Jill Burstein. Automated essay scoring with e-rater® v.2. **The Journal of Technology, Learning and Assessment**, Vol. 4, No. 3, pp. 1–30, 2006.

[2] Hongbo Chen and Ben He. Automated essay scoring by maximizing human-machine agreement. In **Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing**, pp. 1741–1752, 2013.

[3] Peter Phandi, Kian Ming A Chai, and Hwee Tou Ng. Flexible domain adaptation for automated essay scoring using correlated linear regression. In **Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing**, pp. 431–439, 2015.

[4] Mihai Dascalu, Wim Westera, Stefan Ruseti, Stefan Trausan-Matu, and Hub Kurvers. Readerbench learns dutch: building a comprehensive automated essay scoring system for dutch language. In **International Conference on Artificial Intelligence in Education**, pp. 52–63. Springer, 2017.

[5] Masaki Uto. A review of deep-neural automated essay scoring models. **Behaviormetrika**, pp. 1–26, 2021.

[6] Robert Ridley, Liang He, Xinyu Dai, Shujian Huang, and Jiajun Chen. Prompt agnostic essay scorer: A domain generalization approach to cross-prompt automated essay scoring. arXiv, 2020.

[7] Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. Automatic text scoring using neural networks. arXiv, 2016.

[8] Kaveh Taghipour and Hwee Tou Ng. A neural approach to automated essay scoring. In **Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing**, pp. 1882–1891, 2016.

[9] Fei Dong and Yue Zhang. Automatic features for essay scoring–an empirical study. In **Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing**, pp. 1072–1077, 2016.

[10] Yi Tay, Minh C Phan, Luu Anh Tuan, and Siu Cheung Hui. Skipflow: Incorporating neural coherence features for end-to-end automatic text scoring. In **Thirty-Second AAAI Conference on Artificial Intelligence**, pp. 5948–5955, 2018.

[11] Fei Dong, Yue Zhang, and Jie Yang. Attention-based recurrent convolutional neural network for automatic essay scoring. In **Proceedings of the 21st Conference on Computational Natural Language Learning**, pp. 153–162, 2017.

[12] Youmna Farag, Helen Yannakoudakis, and Ted Briscoe. Neural automated essay scoring and coherence modeling for adversarially crafted input. arXiv, 2018.

[13] Masaki Uto, Yikuan Xie, and Maomi Ueno. Neural automated essay scoring incorporating handcrafted features. In **Proceedings of the 28th International Conference on Computational Linguistics**, pp. 6077–6088, 2020.

[14] 岡野将士, 宇都雅輝. 評価者バイアスの影響を考慮した深層学習自動採点手法. 電子情報通信学会論文誌 D, Vol. 104, No. 8, pp. 650–662, 2021.

[15] Pedro Uria Rodriguez, Amir Jafari, and Christopher M Ormerod. Language models and automated essay scoring. arXiv, 2019.

[16] Sandeep Mathias and Pushpak Bhattacharyya. Can neural networks automatically score essay traits? In **Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications**, pp. 85–91, 2020.

[17] Mohamed A. Hussein, Hesham A. Hassan, and Mohammad Nassef. A trait-based deep learning automated essay scoring system with adaptive feedback. **International Journal of Advanced Computer Science and Applications**, Vol. 11, No. 5, pp. 287–293, 2020.

[18] Farjana Sultana Mim, Naoya Inoue, Paul Reisert, Hiroki Ouchi, and Kentaro Inui. Unsupervised learning of discourse-aware text representation for essay scoring. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop**, pp. 378–385, 2019.

[19] Robert Ridley, Liang He, Xin-yu Dai, Shujian Huang, and Jiajun Chen. Automated cross-prompt scoring of essay traits. In **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 35, pp. 13745–13753, 2021.

[20] Masaki Uto. A multidimensional item response theory model for rubric-based writing assessment. In **Artificial Intelligence in Education**, pp. 420–432. Springer International Publishing, 2021.

[21] Masaki Uto. A multidimensional generalized many-facet rasch model for rubric-based performance assessment. **Behaviormetrika**, Vol. 48, No. 2, pp. 425–457, 2021.

[22] Jeffrey L Elman. Finding structure in time. **Cognitive science**, Vol. 14, No. 2, pp. 179–211, 1990.

[23] Frederic M Lord. **Applications of item response theory to practical testing problems**. Routledge, 2012.

[24] Lihua Yao and Richard D. Schwarz. A multidimensional partial credit model with associated item and test statistics: An application to mixed-format tests. **Applied Psychological Measurement**, Vol. 30, No. 6, pp. 469–492, 2006.

[25] Yann Dauphin, Harm de Vries, and Yoshua Bengio. Equilibrated adaptive learning rates for non-convex optimization. In **Advances in Neural Information Processing Systems**, Vol. 28, pp. 1504–1512. Curran Associates, Inc., 2015.

[26] Sandeep Mathias and Pushpak Bhattacharyya. ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores. Proceedings of the Eleventh International Conference on Language Resources and Evaluation, 2018.

[27] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing**, pp. 1532–1543, 2014.