# Analytic Automated Essay Scoring based on Deep Neural Networks Integrating Multidimensional Item Response Theory

**Takumi Shibata**
The University
of Electro-Communications
shibata@ai.lab.uec.ac.jp

**Masaki Uto**
The University
of Electro-Communications
uto@ai.lab.uec.ac.jp

## Abstract

Essay exams have been attracting attention as a way of measuring the higher-order abilities of examinees, but they have two major drawbacks in that grading them is expensive and raises questions about fairness. As an approach to overcome these problems, automated essay scoring (AES) is in increasing need. Many AES models based on deep neural networks have been proposed in recent years and have achieved high accuracy, but most of these models are designed to predict only a single overall score. However, to provide detailed feedback in practical situations, we often require not only the overall score but also analytic scores corresponding to various aspects of the essay. Several neural AES models that can predict both the analytic scores and the overall score have also been proposed for this very purpose. However, conventional models are designed to have complex neural architectures for each analytic score, which makes interpreting the score prediction difficult. To improve the interpretability of the prediction while maintaining scoring accuracy, we propose a new neural model for automated analytic scoring that integrates a multidimensional item response theory model, which is a popular psychometric model.

## 1 Introduction

Rapid changes in society in recent years have led to an increased need for cultivating and assessing not only knowledge and skills but also practical abilities, such as expression skills, logical thinking, and creativity (Erguvan and Aksu Dunya, 2020; Uto, 2021a). Essay exams are one of the test formats that aim to evaluate these abilities, and consequently, they have been used in various educational and assessment settings (Erguvan and Aksu Dunya, 2020; Hussein et al., 2019). However, essay exams have two considerable drawbacks in the time and monetary costs required to grade them (Taghipour and Ng, 2016). Furthermore, it is difficult to ensure consistently fair and reliable evaluation due to subjective influences on the part of the rater (Uto and Ueno, 2020; Saal et al., 1980). Automated Essay Scoring (AES) has been attracting attention as a method for resolving these difficulties (Dong and Zhang, 2016; Taghipour and Ng, 2016).

Conventional AES systems can be broadly classified into two categories (Hussein et al., 2019): those that take a feature-engineering approach and those that take a neural approach. The feature-engineering approach, which has traditionally been the greater used of the two, utilizes a statistical or machine learning model with pre-defined hand-crafted features (e.g. Attali and Burstein, 2006; Chen and He, 2013; Phandi et al., 2015; Dascalu et al., 2017; Hastings et al., 2018; Yao et al., 2019). The neural approach, on the other hand, which has become popular recently, uses deep neural networks to extract features automatically from texts (e.g. Alikaniotis et al., 2016; Taghipour and Ng, 2016; Dong and Zhang, 2016; Tay et al., 2018; Dong et al., 2017; Farag et al., 2018; Jin et al., 2018; Uto et al., 2020; Rodriguez et al., 2019; Uto et al., 2020; Ridley et al., 2020; Uto, 2021c). In this study, we focus on the neural approach because of the high accuracy it has achieved in many prior studies.

Most neural AES studies have focused on holistic scoring (Ridley et al., 2021; Ke and Ng, 2019), which provides a single overall score for each essay. However, to provide richer feedback, especially when essay exams are used for educational purposes, we often require not only the overall score but also analytic scores corresponding to various aspects of the essay, such as *content*, *organization*, and *word choice* (Hussein et al., 2020). Several AES models that can predict these analytic scores along with the overall score have recently been proposed for this purpose (Mathias and Bhattacharyya, 2020; Hussein et al., 2020; Mim et al., 2019; Ridley et al., 2021). From here on, we will refer to such



*Proceedings of the 29th International Conference on Computational Linguistics*, pages 2917–2926
October 12–17, 2022.

models as analytic AES models.

Mathias and Bhattacharyya (2020) proposed an early neural analytic AES model that took the simple approach of separately applying a conventional holistic scoring model (Dong et al., 2017) to each analytic score. Then, Hussein et al. (2020) proposed a multi-output model in which the output layers are branched by the number of analytic scores and the other layers are shared. One of the more recent models is a multi-output model proposed by Ridley et al. (2021) that has a complex deep neural architecture as the output layer for each analytic score. Although this model produces state-of-the-art accuracy, it has some problems in terms of interpretability.

1. It has a complex neural architecture for each analytic score, decreasing the interpretability of the prediction.

2. In general, analytic scores are designed to measure latent abilities in examinees that a test developer wishes to evaluate (Uto, 2021b). However, this model ignores the existence of an ability scale, further restricting the interpretability of the score prediction.

To resolve these problems, we propose to extend a conventional analytic AES model by integrating it with an item response theory (IRT) (Lord, 1980) model, a well-known psychometric model. Specifically, we extend the multi-output model of Ridley et al. (2021) by replacing the complex output layers for each analytic score with a multidimensional IRT model (Yao and Schwarz, 2006). The advantages of the proposed model are as follows.

1. The output IRT layer is explained by only three types of parameters: the discriminatory power and difficulty corresponding to each analytic score and the latent ability of each examinee. These allow us to better interpret the reasoning behind score predictions.

2. Investigating an optimal number of ability dimensions in the multidimensional IRT model layer and analyzing the estimated parameters allows us to interpret the ability scale implied by the multiple analytic scores.

In this study, we used benchmark datasets that have been widely used in analytic AES research to conduct experiments that evaluated the effectiveness of our model. They showed that our model offers reasonably interpretable parameters without significantly degrading scoring accuracy. Furthermore, an interesting finding from our experiment was that, although the benchmark dataset consisted of many analytic scores for each essay, only one or two latent abilities were measured by those multiple scores.

Note that a similar AES framework combining deep neural networks and IRT was recently proposed (Uto and Okano, 2021). However, they used IRT to improve the quality of training data by mitigating rater effects, so their research objective was completely different from the one we focus on in this study.

## 2 Conventional analytic AES model

This section introduces the conventional analytic AES model proposed by Ridley et al. (2021), which we use as a baseline model. The architecture of this model is displayed on the left side of Figure 1.

This model takes in an essay from examinee $n$ and outputs multiple analytic scores $\{\hat{y}_{nm} \mid m \in \{1, 2, \ldots, M\}\}$, where $\hat{y}_{nm}$ is the $m$-th analytic score and $M$ is the total number of analytic scores. An essay from examinee $n$ is defined as a word sequence $\{w_{nsl} \mid s \in \{1, 2, \ldots, S\}, l \in \{1, 2, \ldots, l_s\}\}$, where $w_{nsl}$ is the $l$-th word in the $s$-th sentence of examinee $n$'s essay, $S$ is the number of the sentences in the essay, and $l_s$ is the number of words in the $s$-th sentence. Note that in our paper, we regard the overall score as one of the analytic scores.

The model consists of two types of layers: a *shared layer* and an *item-specific layer*. The shared layer receives the word sequence in each sentence and produces a sentence-level distributed representation through an embedding layer, a convolutional layer, and an attention pooling layer (Dong et al., 2017). The sequence for the sentence-level distributed representation is used in the item-specific layer.

The item-specific layer consists of the same number of heads as the number of analytic scoring items, which are evaluation items corresponding to analytic scores, such as *content*, *organization*, and *word choice*. An item-specific layer for an analytic scoring item receives the sequence of the sentence-level distributed representation and produces a score value for the corresponding scoring item. Specifically, the input sequence is first processed through a recurrent neural network (RNN),

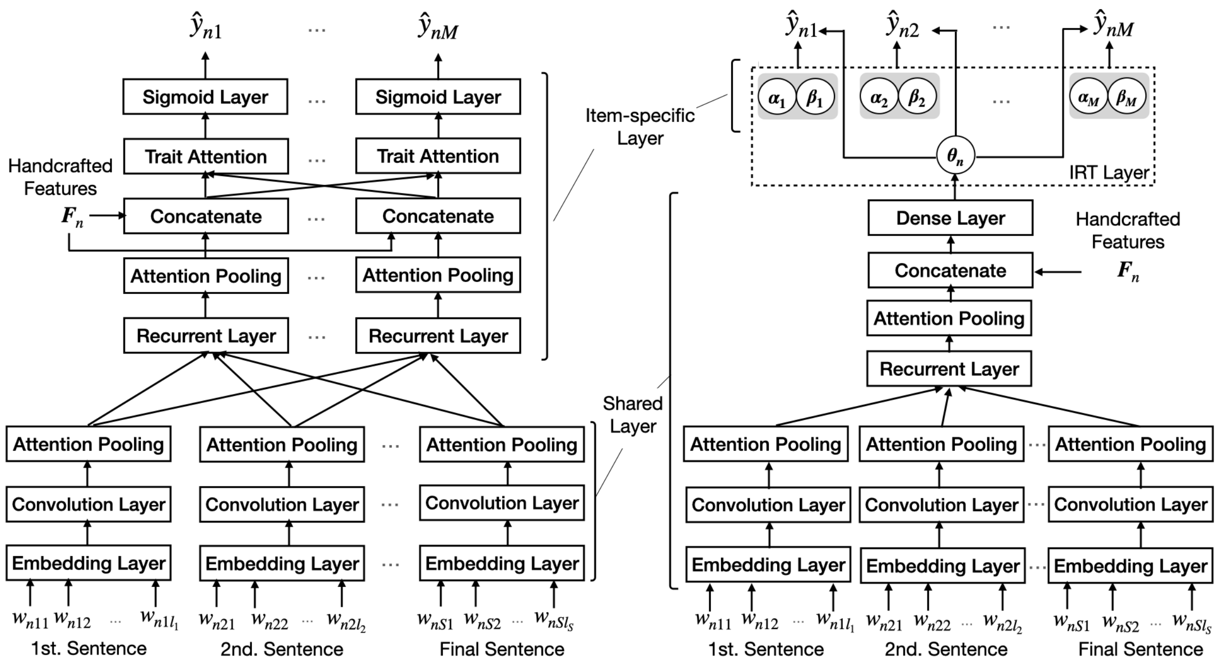


Figure 1: Architecture of a conventional analytic scoring model (left) and our model (right).

one in which the long short-term memory (Hochreiter and Schmidhuber, 1997) was used as the RNN. Then, an output sequence from the RNN layer is aggregated into a fixed-length hidden vector through an attention pooling layer (Dong et al., 2017). The hidden vector is concatenated with a manually designed feature vector $\boldsymbol{F}_n$, and the concatenated vector $\boldsymbol{h}_{nm}$ is input to the trait attention layer. For capturing the relation between analytic scoring items, the trait attention layer is defined as

$$a_{nmt} = \frac{\exp(\boldsymbol{h}_{nm} \cdot \boldsymbol{h}_{nt})}{\sum_{\substack{t=1 \\ t \neq m}}^{M} \exp(\boldsymbol{h}_{nm} \cdot \boldsymbol{h}_{nt})}, \ \forall t, \forall m, t \neq m \tag{1}$$

$$\boldsymbol{x}_{nm} = \sum_{\substack{t=1 \\ t \neq m}}^{M} a_{nmt} \boldsymbol{h}_{nt} \tag{2}$$

$$\tilde{\boldsymbol{x}}_{nm} = \mathrm{Concat}(\boldsymbol{x}_{nm}, \boldsymbol{h}_{nm}). \tag{3}$$

Finally, a linear layer with the sigmoid activation maps $\tilde{\boldsymbol{x}}_{nm}$, a trait attention output vector, to the prediction score $\hat{y}_{nm}$:

$$\hat{y}_{nm} = \sigma(\boldsymbol{W}_m \tilde{\boldsymbol{x}}_{nm} + b_m), \tag{4}$$

where $\sigma$ is the sigmoid function, $\boldsymbol{W}_m$ is a weight vector, and $b_m$ is a bias value. Note that this model uses a sigmoid function to predict scores, so $\hat{y}_{nm}$ takes values between 0 and 1. Thus, in the score prediction phase, the output scores must be linearly transformed to the original score scale.

This model is trained through a back-propagation algorithm using the Mean Squared Error (MSE) as a loss function. This is given by

$$\mathcal{L}_{MSE} = \frac{1}{NM} \sum_{n=1}^{N} \sum_{m=1}^{M} (\hat{y}_{nm} - y_{nm})^2, \tag{5}$$

where $N$ is the number of essays and $y_{nm}$ is the gold-standard score of examinee $n$ for the $m$-th analytic scoring item. The gold-standard scores $y_{nm}$ must be linearly transformed into the range between 0 and 1.

Note that Ridley et al. (2021) input the part-of-speech (POS) tags instead of the words themselves when applying the model to cross-prompt scoring tasks. However, we use word sequences as input because they are expected to be more accurate for the prompt-specific scoring tasks used in this study.

As previously mentioned, this model has a complex architecture for each analytic score, making it difficult to interpret the score prediction. Our main focus is to use IRT to increase the interpretability of score prediction.

## 3 Item Response Theory

IRT (Lord, 1980) is a popular psychometric model that has been widely used for making measurements in educational and psychological research.

Typical IRT models define the probability that an examinee will receive a certain score on a test item as a function of the examinee's latent ability and the item's characteristic parameters, such as the discrimination and difficulty parameters. Of the various existing IRT models, we employ a multidimensional generalized partial credit model (M-GPCM) (Yao and Schwarz, 2006), a representative multidimensional polytomous IRT model that can be applied to ordinal score data and can examine multidimensional latent abilities for each examinee.

If we regard IRT parameters for test items as those for analytic scoring items following the approach in previous studies (Uto, 2021b), then M-GPCM defines the probability that examinee $n$ will receive score $k$ for the $m$-th analytic scoring item as

$$P_{nmk} = \frac{\exp(k\boldsymbol{\alpha}_m^T\boldsymbol{\theta}_n + \sum_{u=1}^{k}\beta_{mu})}{\sum_{v=1}^{K_m}\exp(v\boldsymbol{\alpha}_m^T\boldsymbol{\theta}_n + \sum_{u=1}^{v}\beta_{mu})}, \quad (6)$$

where $\boldsymbol{\theta}_n = (\theta_{n1}, \theta_{n2}, \ldots, \theta_{nd})$ is the $d$-dimensional latent ability of examinee $n$, $\boldsymbol{\alpha}_m = (\alpha_{m1}, \alpha_{m2}, \ldots, \alpha_{md})$ is a $d$-dimensional discrimination parameter for analytic scoring item $m$, $\beta_{mu}$ is a step parameter denoting the difficulty of the transition between scores $u-1$ and $u$ in item $m$, and $K_m$ is the number of possible scores for the $m$-th item. Here, $\beta_{m1} = 0 : \forall m$ is assumed for model identification.

All of these model parameters, $\boldsymbol{\theta}_n$, $\boldsymbol{\alpha}_m$, and $\beta_{mu}$, can be estimated from a collection of observed scores. These parameters are clearly interpretable, as will be explained in sections 4.3 and 5.3.

# 4 Proposed Model

We propose an analytic AES model that incorporates the M-GPCM mentioned in the previous section. The architecture of this model is displayed on the right side of Figure 1.

As Figure 1 shows, our model and the conventional model share the same layers from the input to the concatenate layer. Specifically, in both models, each sentence in an essay is fed to the embedding layer, convolution layer, and attention pooling layer, and then a sequence of the sentence-level distributed representation vectors is transformed into a fixed-length vector through the recurrent layer and the attention pooling layer. Finally, the concatenate layer creates an essay-level vector $\boldsymbol{h_n}$ by combining the output from the attention pooling layer and the handcrafted feature vector $\boldsymbol{F}_n$.

The main differences between the models occur after the concatenate layer. Given the essay-level vector $\boldsymbol{h_n}$, our model obtains the latent ability vector $\boldsymbol{\theta}_n$, which is used in the subsequent M-GPCM layer, by applying a dense layer given by

$$\boldsymbol{\theta}_n = \boldsymbol{W}\boldsymbol{h}_n + \boldsymbol{b}, \quad (7)$$

where $\boldsymbol{W}$ is a weights matrix and $\boldsymbol{b}$ is a bias vector. The latent ability $\boldsymbol{\theta}_n$ is input to the M-GPCM defined in Eq. (6) to obtain the score probabilities for each analytic scoring item $m$. Our model then uses the obtained probability $P_{nmk}$ to predict the analytic scores.

## 4.1 Model Training

We train our model using the following Categorical Cross-Entropy (CCE) as a loss function:

$$\mathcal{L}_{CCE} = -\frac{1}{NM}\sum_{n=1}^{N}\sum_{m=1}^{M}\sum_{k=1}^{K_m} y_{nmk}\log(P_{nmk}). \quad (8)$$

We use this because the output IRT layer gives the probability distribution over score categories $P_{nmk}$. Note that during the training process, our model simultaneously estimates the IRT parameters, namely $\boldsymbol{\theta}_n$, $\boldsymbol{\alpha}_m$, and $\boldsymbol{\beta}_m = (\beta_{m1}, \beta_{m2}, \ldots, \beta_{mK_m})$, and the parameters in the other layers in an end-to-end manner.

The hyper-parameters in our model are the same as those in the conventional model (Ridley et al., 2021), and we use the RMSProp Optimizer (Dauphin et al., 2015) with a learning rate of 0.001. Furthermore, since IRT generally assumes a normal distribution for $\boldsymbol{\theta}_n$, we apply an L2-regularization for $\boldsymbol{\theta}_n$ so that its distribution closes to a normal distribution with mean zero.

## 4.2 Score Prediction

We have the following two options for predicting a score based on the output score probabilities $P_{nmk}$.

- Argmax score: $\arg\max_k P_{nmk}$.
- Expected score: $\sum_{k=1}^{K_m} kP_{nmk}$.

We compare these two options in the experiments discussed in section 5.2.

Table 1: Summary of the ASAP and ASAP++ dataset: *Org* refers to organization, *WC* to word choice, *SF* to sentence fluency, *Conv* to conventions, *PA* to prompt adherence, *Lang* to language, and *Narr* to narrativity.

| Prompt | Num Essays | Mean Length | Analytic Scoring Items | Score Range | |
|---|---|---|---|---|---|
| | | | | Overall | Analytic |
| 1 | 1783 | 350 | Overall, Content, Org, WC, SF, Conv | 2-12 | 1-6 |
| 2 | 1800 | 350 | Overall, Content, Org, WC, SF, Conv | 1-6 | 1-6 |
| 3 | 1726 | 150 | Overall, Content, PA, Lang, Narr | 0-3 | 0-3 |
| 4 | 1772 | 150 | Overall, Content, PA, Lang, Narr | 0-3 | 0-3 |
| 5 | 1805 | 150 | Overall, Content, PA, Lang, Narr | 0-4 | 0-4 |
| 6 | 1800 | 150 | Overall, Content, PA, Lang, Narr | 0-4 | 0-4 |
| 7 | 1569 | 250 | Overall, Content, Org, Conv, Style | 0-30 | 0-6 |
| 8 | 723 | 650 | Overall, Content, Org, WC, SF, Conv, Voice | 0-60 | 2-12 |

### 4.3 Interpretability of our model

As explained in section 3, the M-GPCM consists of three types of trainable parameters: both the discrimination parameters $\boldsymbol{\alpha}_m$ and the difficulty parameters $\boldsymbol{\beta}_m$ for each analytic scoring item and the latent examinee ability parameter $\boldsymbol{\theta}_n$.

The discrimination parameter $\boldsymbol{\alpha}_m$ provides information on how well each analytic scoring item distinguishes examinee ability, whereas the difficulty parameter $\boldsymbol{\beta}_m$ reflects how difficult examinees find each score category for the $m$-th analytic scoring item to be. The examinee ability parameter $\boldsymbol{\theta}_n$ represents the ability level of each examinee. Section 5.3 shows an example of the interpretation of these parameters.

Furthermore, our model enables us to perform an analysis of the optimal number of ability dimensions assumed under multiple analytic scores by comparing its performance with different numbers of dimensions. For example, if the score prediction performance of our model is maximized when two ability dimensions are assumed, then we can interpret this as indicating that the given analytic scoring items measure a two-dimensional latent ability of examinees. We can also interpret what each ability dimension measures by analyzing the multidimensional discrimination parameter $\boldsymbol{\alpha}_m$. Section 5.3 gives an example of how the ability dimension can be interpreted.

Our model predicts the scores by using the output IRT layer with the IRT parameters mentioned above. Thus, interpreting these parameters allows us to understand how the model determines analytic scores for a given essay.

## 5 Experiments

In this section, we discuss how the effectiveness of our model was evaluated through experiments using real-word data.

### 5.1 Real-word data

In our experiments, we used real-word data from the Automated Student Assessment Prize (ASAP)[1] and the ASAP++ (Mathias and Bhattacharyya, 2018).

The ASAP was introduced in the Kaggle competition and has since been widely used in AES research. The ASAP dataset consists of examinees' essays for eight different prompts and scores for them. Only an overall score is given for prompts 1 through 6, while some analytic scores are given in addition to the overall score for prompts 7 and 8. The ASAP++, a dataset designed to supplement ASAP, offers analytic scores for prompts 1 through 6.

Table 1 gives a summary of the ASAP with the ASAP++ dataset.

### 5.2 Evaluation of our model

Using the ASAP with the ASAP++ dataset, we evaluated the scoring accuracy of our model while varying the number of ability dimensions from 1 to 3 and compared the results to those from the conventional baseline model described in section 2. The scoring accuracy was independently evaluated for each prompt through a 5-fold cross validation using the Quadratic Weighted Kappa (QWK), which is used in AES studies. Concretely, we evaluated the QWK score for each analytic scoring item and then calculated the average QWK score for each prompt. We examined two input types in this experiment: a word sequence and a POS tag sequence. We used Glove (Pennington et al., 2014), a pre-trained word embedding, in the embedding layer for models using word sequences as inputs. Furthermore, in our model, we evaluated the two types of prediction scores, the argmax scores and the expected scores, explained in section 4.2.

Table 2 and Table 3 show the results obtained when the expected scores and the argmax scores

[1] https://www.kaggle.com/c/asap-aes

Table 2: QWK scores for our model with the expected scores and the conventional model.

| Input | Model | Prompts | | | | | | | | | $p$-value | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | Avg. | 1dim | 2dim | 3dim |
| POS | Conventional | **0.688** | **0.632** | 0.610 | **0.680** | 0.686 | **0.684** | **0.694** | 0.548 | **0.653** | 0.460 | 0.169 | 0.767 |
| | Proposed-1dim | 0.662 | 0.605 | **0.623** | 0.663 | **0.693** | 0.670 | 0.640 | 0.542 | 0.637 | - | 1.000 | 1.000 |
| | Proposed-2dim | 0.671 | 0.627 | 0.608 | 0.657 | 0.680 | 0.669 | 0.669 | **0.555** | 0.642 | - | - | 1.000 |
| | Proposed-3dim | 0.678 | 0.629 | 0.615 | 0.643 | 0.691 | 0.677 | 0.682 | 0.544 | 0.645 | - | - | - |
| Word | Conventional | **0.685** | **0.655** | **0.660** | 0.720 | **0.706** | **0.750** | 0.694 | 0.568 | **0.680** | 0.009 | 0.699 | 0.014 |
| | Proposed-1dim | 0.656 | 0.617 | 0.620 | 0.713 | 0.689 | 0.731 | 0.638 | 0.549 | 0.652 | - | 0.180 | 0.378 |
| | Proposed-2dim | 0.666 | 0.631 | 0.637 | **0.722** | 0.699 | 0.732 | **0.704** | **0.576** | 0.671 | - | - | 1.000 |
| | Proposed-3dim | 0.679 | 0.633 | 0.642 | 0.704 | 0.698 | 0.734 | 0.696 | 0.553 | 0.667 | - | - | - |

Table 3: QWK scores for our model with the argmax scores and the conventional model.

| Input | Model | Prompts | | | | | | | | | $p$-value | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | Avg. | 1dim | 2dim | 3dim |
| POS | Conventional | **0.688** | 0.632 | 0.610 | **0.680** | **0.686** | 0.684 | **0.694** | **0.548** | **0.653** | 0.253 | 0.469 | 0.420 |
| | Proposed-1dim | 0.651 | 0.616 | 0.620 | 0.670 | 0.682 | 0.685 | 0.619 | 0.480 | 0.628 | - | 0.053 | 0.755 |
| | Proposed-2dim | 0.661 | 0.608 | 0.629 | 0.670 | 0.679 | 0.675 | 0.620 | 0.445 | 0.623 | - | - | 1.000 |
| | Proposed-3dim | 0.636 | **0.633** | **0.634** | 0.656 | 0.685 | **0.694** | 0.636 | 0.471 | 0.631 | - | - | - |
| Word | Conventional | **0.685** | **0.655** | **0.660** | 0.720 | **0.706** | **0.750** | **0.694** | **0.568** | **0.680** | 0.080 | 0.100 | 0.090 |
| | Proposed-1dim | 0.641 | 0.625 | 0.646 | 0.718 | 0.690 | 0.737 | 0.637 | 0.464 | 0.645 | - | 1.000 | 1.000 |
| | Proposed-2dim | 0.636 | 0.620 | 0.656 | **0.721** | 0.692 | 0.736 | 0.675 | 0.486 | 0.653 | - | - | 1.000 |
| | Proposed-3dim | 0.656 | 0.630 | 0.656 | 0.712 | 0.696 | 0.734 | 0.687 | 0.472 | 0.655 | - | - | - |

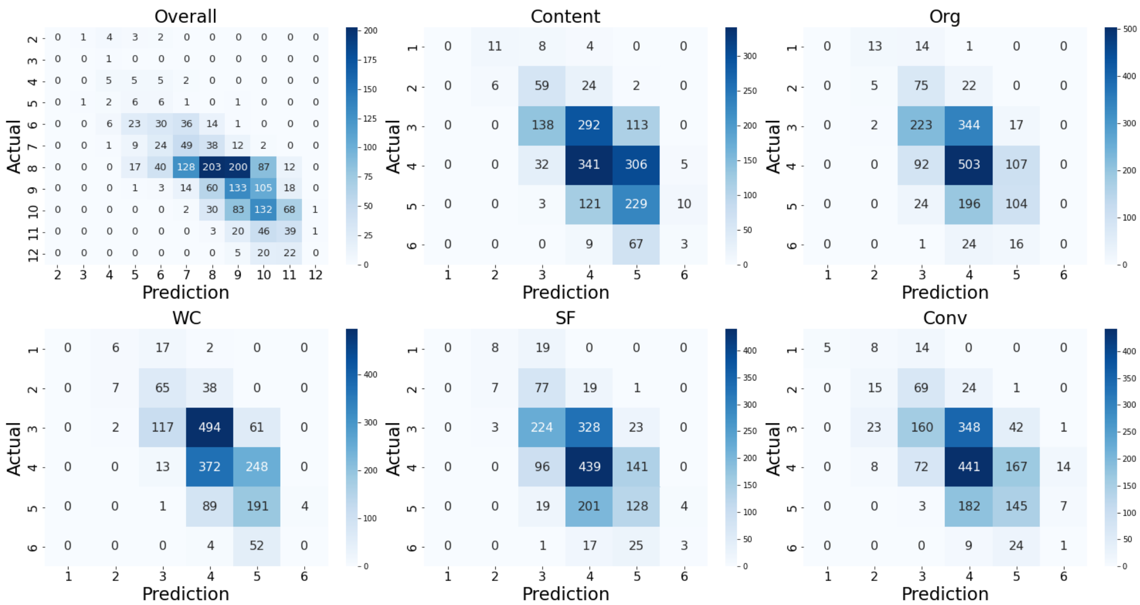


Figure 2: Confusion matrices between gold-standard scores and the expected scores from our model for prompt 1.

were used in our model, respectively. Note that the results for the conventional model are the same in both of these tables, and the highest QWK scores for each setting are shown in bold.

At first, comparing the input types suggests that the word input shows higher averaged performances in all settings. Ridley et al. (2021) used the POS tag input assuming cross-prompt tasks, as noted in section 2, whereas our experiment suggests that the word input is better for prompt-specific tasks.

Next, comparing Tables 2 and 3 shows that using the expected scores with our model tended to produce better results than when the argmax scores were used. Figure 2 shows the confusion matrices between the gold-standard scores and the expected scores given by our model for all of the analytic scoring items associated with prompt 1. According to this figure, the diagonal components of the matrices are responsive, indicating that the scores are predicted relatively well.

Finally, comparing variants of our model with different numbers of ability dimensions shows that the two- and three-dimensional models tended to outperform the one-dimensional model, although the differences are relatively small. Moreover, although the conventional model had the highest average performance, the degradations in performance of our model were small overall. We performed Bonferroni's multiple comparison test to quantitatively measure whether there were significant differences in the average QWK scores among the

Table 4: IRT parameters for analytic scoring items estimated by the one-dimensional variant of our model.

| | $\alpha_1$ | $\beta_{12}$ | $\beta_{13}$ | $\beta_{14}$ | $\beta_{15}$ | $\beta_{16}$ | $\beta_{17}$ | $\beta_{18}$ | $\beta_{19}$ | $\beta_{1\ 10}$ | $\beta_{1\ 11}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *Overall* | 1.15 | -2.61 | -3.32 | -3.00 | -3.79 | -1.48 | -2.14 | 0.85 | 0.87 | 2.47 | 2.84 |
| *Content* | 2.01 | -4.71 | -3.85 | -0.56 | 2.02 | 4.20 | - | - | - | - | - |
| *Org* | 1.88 | -4.45 | -3.59 | -0.32 | 2.25 | 4.75 | - | - | - | - | - |
| *WC* | 2.09 | -4.71 | -3.76 | 0.06 | 2.51 | 4.59 | - | - | - | - | - |
| *SF* | 2.06 | -4.74 | -3.71 | -0.36 | 2.20 | 4.82 | - | - | - | - | - |
| *Conv* | 2.01 | -4.66 | -3.56 | -0.36 | 2.29 | 5.01 | - | - | - | - | - |

Table 5: IRT parameters for analytic scoring items estimated by the two-dimensional variant of our model.

| | $\alpha_{11}$ | $\alpha_{12}$ | $\beta_{12}$ | $\beta_{13}$ | $\beta_{14}$ | $\beta_{15}$ | $\beta_{16}$ | $\beta_{17}$ | $\beta_{18}$ | $\beta_{19}$ | $\beta_{1\ 10}$ | $\beta_{1\ 11}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Overall* | 1.81 | 0.14 | -2.56 | -3.53 | -3.26 | -3.75 | -1.59 | -2.24 | 0.95 | 1.09 | 2.90 | 3.19 |
| *Content* | 1.54 | 1.38 | -4.80 | -3.98 | -0.60 | 2.04 | 4.35 | - | - | - | - | - |
| *Org* | 1.23 | 1.41 | -4.38 | -3.58 | -0.39 | 2.18 | 4.65 | - | - | - | - | - |
| *WC* | 1.38 | 1.63 | -5.10 | -3.87 | -0.01 | 2.55 | 4.73 | - | - | - | - | - |
| *SF* | 1.10 | 1.90 | -4.73 | -3.93 | -0.49 | 2.21 | 5.02 | - | - | - | - | - |
| *Conv* | 1.04 | 1.95 | -4.93 | -3.87 | -0.54 | 2.36 | 5.29 | - | - | - | - | - |

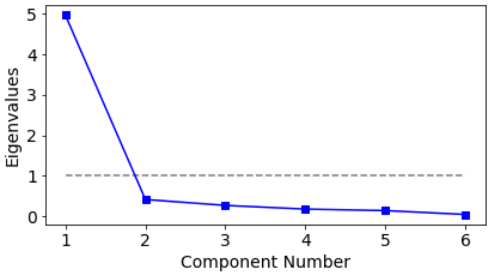


Figure 3: Scree plot for prompt 1.

models. The results are given in the $p$-value column of Table 2. The $p$-values indicate that there was no difference at the 5% significance level between the conventional model and our model with optimal dimension. This result is surprising because the scoring accuracy remains even though the item-specific layers in our model are described by significantly fewer parameters than the conventional model. Thus, we can conclude that our model does not lead to a significant decrease in the scoring accuracy.

### 5.3 Interpretation of our model

In this subsection, we explain how we interpreted the predictions from our model.

We first examined the optimal number of ability dimensions. In IRT studies, principal component analysis (PCA) is generally used for investigating the optimal number of dimensions (Nunnally and Bernstein, 1994; Bjorner et al., 2003). For this reason, Figure 3 shows the eigenvalues obtained by PCA for different numbers of dimensions in prompt 1; the horizontal axis shows the number of dimensions (component), and the vertical axis indicates the eigenvalue. A significant decrease in the eigenvalues occurs at the point where the component number is 2, suggesting that the ability dimension assumed under the data for the multiple analytic scores in prompt 1 is only explainable with a one-dimensional ability scale. Other prompts yielded the same results. Note that, as explained in the previous section, the one-dimensional model shows slightly lower QWK scores than the two- or three-dimensional models, so if prediction accuracy is a priority, then the two-dimensional model may be a better choice. Thus, we will now explain the interpretation of our model when one and two dimensions are assumed.

Tables 4 and 5 show the IRT-layer parameters for the analytic scoring items estimated with the one- and two-dimensional variants of our model, respectively. Only the results for prompt 1 are given here as an example.

The discrimination parameters provide information for interpreting how well the analytic scoring items measure examinees' abilities and what each ability dimension measures. For example, according to Table 4, the *overall* item has a lower discrimination value than the other analytic scoring items, suggesting that the *overall* item is relatively inaccurate for measuring a one-dimensional latent ability constructed by the multiple analytic scoring items. This also suggests the possibility that the ability measured by the *overall* item might differ from that of the other items, something that can be confirmed from the discrimination parameters in the two-dimensional model shown in Table 5. Specifically, the *overall* item in Table 5 has a large discrimination value in the first dimension but an extremely small value in the second dimension, whereas the other analytic scoring items have large discrimination values in the second dimension. Furthermore, taking a closer look at the other analytic scoring items, we can see that the *content* item

Table 6: Examples of examinees' latent abilities and the predicted scores estimated by our model.

| Examinee $n$ | Ability $\theta_n$ | Predicted Scores | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | *Overall* | *Content* | *Org* | *WC* | *SF* | *Conv* | Avg. |
| 4 | 0.14 | 8 | 4 | 4 | 4 | 4 | 4 | 4.67 |
| 27 | 2.21 | 11 | 6 | 5 | 6 | 5 | 5 | 6.33 |
| 31 | -2.01 | 6 | 2 | 2 | 2 | 2 | 2 | 2.67 |
| 916 | -4.19 | 2 | 1 | 1 | 1 | 1 | 1 | 1.17 |
| 1651 | 3.40 | 12 | 6 | 6 | 6 | 6 | 6 | 7.00 |

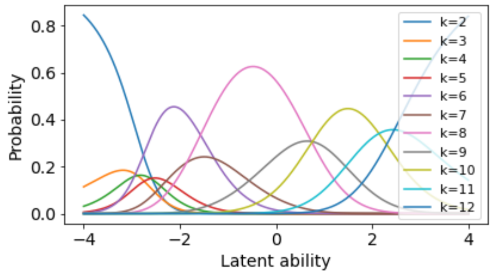

Figure 4: ICCs for the *overall* score.

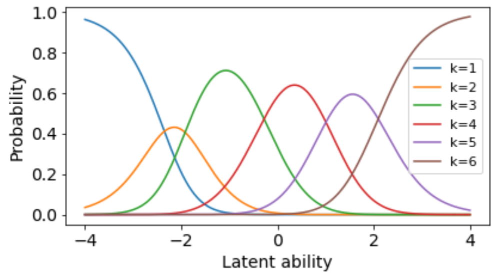

Figure 5: ICCs for the *content* score.

is like the *overall* item in that it has a higher discrimination value for the first dimension than for the second dimension, while the other items have higher discrimination values for the second dimension. These results suggest that the first ability dimension measures the overall ability relating to the skills for enriching content in an essay, while the second dimension measures the ability shared among *organization*, *word choice*, *sentence fluency*, and *convention*, which would make it a technical writing ability.

Furthermore, the difficulty parameters show how the score categories are obtained for each analytic scoring item. For instance, Figures 4 and 5 show item characteristic curves (ICC), which illustrate the probabilistic curve based on Eq. (6), for the *overall* and *content* items under the one-dimensional setting. In these figures, the horizontal axis indicates the latent ability of the examinees, and the vertical axis indicates the probability $P_{nmk}$. Note that the horizontal axis shows values for ability $\theta$ around zero because, as was explained in section 4.1, the distribution of the ability estimates follows a normal distribution with zero mean. Figures 4 and 5 show that examinees with a higher ability have a greater probability of obtaining a high score. Moreover, scores of 2, 6, 8, 10, 11, and 12 for the *overall* item are likely to be used while scores of 3, 4, 5, and 7 tend to be avoided. In the *content* item, a score of 2 tends to be avoided slightly. It is in this way that the difficulty parameters enable us to make an interpretation of how the score categories are used for the analytic scoring items. Note that although we highlighted the one-dimensional model results here, the difficulty parameters in the one- and two-dimensional models are similar and, thus, provide similar interpretations.

Our model predicts analytic scores based on these characteristics of the analytic scoring items and on estimations of the examinees' abilities. Table 6 shows examples of examinees' latent abilities and the predicted analytic scores estimated by the one-dimensional variant of our model. Table 6 indicates that our model tends to provide higher scores for essays written by examinees with higher abilities. Furthermore, comparing Table 6 with Figures 4 and 5, we can confirm that the predicted scores for the *overall* and *content* items follow the ICCs reasonably well. For example, examinee 4, who had a nearly zero value of $\theta_n$, obtained an overall score of 8 and a content score of 4. The ICCs show high response probabilities for these scores around $\theta_n = 0$.

These results demonstrate that our model enables us to interpret predictions that are based on the IRT-layer model parameters.

## 6 Conclusions

In this study, we proposed a new neural-based analytic AES model that incorporates a multidimensional IRT model. Through experiments using the well-known benchmark datasets ASAP and ASAP++, we demonstrated that, compared to the latest conventional model, our model succeeds in improving interpretability without significantly losing performance.

Our experiments also suggested that one- or two-dimensional abilities can sufficiently explain the multiple analytic scores, including the overall score. This is an important finding suggesting that the analytic scoring items in the dataset may fail to measure multiple aspects of ability. This is undesirable because the objective of analytic scoring is to evaluate multiple aspects of ability.

Future studies will be required to evaluate our model using various datasets, including other benchmark datasets. Moreover, another challenge to address in future work is to develop an extension of our model for cross-prompt tasks.

## Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers 19H05663, 20K20817, and 21H00898.

## References

Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. 2016. Automatic text scoring using neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 715–725. Association for Computational Linguistics.

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater® v.2. *The Journal of Technology, Learning and Assessment*, 4(3):1–30.

Jakob B Bjorner, Mark Kosinski, and John E Ware Jr. 2003. The feasibility of applying item response theory to measures of migraine impact: A re-analysis of three clinical studies. *Quality of Life Research*, 12(8):887–902.

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

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

Yann Dauphin, Harm de Vries, and Yoshua Bengio. 2015. Equilibrated adaptive learning rates for nonconvex optimization. In *Advances in Neural Information Processing Systems*, volume 28, pages 1504–1512. Curran Associates, Inc.

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

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

Inan Deniz Erguvan and Beyza Aksu Dunya. 2020. Analyzing rater severity in a freshman composition course using many facet rasch measurement. *Language Testing in Asia*, 10(1):1–20.

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

Peter Hastings, Simon Hughes, and M Anne Britt. 2018. Active learning for improving machine learning of student explanatory essays. In *International Conference on Artificial Intelligence in Education*, pages 140–153. Springer.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Mohamed A. Hussein, Hesham A. Hassan, and Mohammad Nassef. 2020. A trait-based deep learning automated essay scoring system with adaptive feedback. *International Journal of Advanced Computer Science and Applications*, 11(5):287–293.

Mohamed Abdellatif Hussein, Hesham Hassan, and Mohammad Nassef. 2019. Automated language essay scoring systems: A literature review. *PeerJ Computer Science*, 5:e208.

Cancan Jin, Ben He, Kai Hui, and Le Sun. 2018. Tdnn: a two-stage deep neural network for prompt-independent automated essay scoring. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1088–1097.

Zixuan Ke and Vincent Ng. 2019. Automated essay scoring: A survey of the state of the art. In *IJCAI*, volume 19, pages 6300–6308.

F.M. Lord. 1980. *Applications of item response theory to practical testing problems*. Erlbaum Associates.

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

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

Farjana Sultana Mim, Naoya Inoue, Paul Reisert, Hiroki Ouchi, and Kentaro Inui. 2019. 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*, pages 378–385.

J.C. Nunnally and I.H. Bernstein. 1994. The assessment of reliability. *Psychometric Theory*, 3:248–292.

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

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

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

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

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

Frank E Saal, Ronald G Downey, and Mary A Lahey. 1980. Rating the ratings: Assessing the psychometric quality of rating data. *Psychological bulletin*, 88(2):413.

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

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

Masaki Uto. 2021a. Accuracy of performance-test linking based on a many-facet rasch model. *Behavior Research Methods*, 53(4):1440–1454.

Masaki Uto. 2021b. A multidimensional generalized many-facet rasch model for rubric-based performance assessment. *Behaviormetrika*, 48(2):425–457.

Masaki Uto. 2021c. A review of deep-neural automated essay scoring models. *Behaviormetrika*, 48(2):1–26.

Masaki Uto and Masashi Okano. 2021. Learning automated essay scoring models using item-response-theory-based scores to decrease effects of rater biases. *IEEE Transactions on Learning Technologies*, 14(6):763–776.

Masaki Uto and Maomi Ueno. 2020. A generalized many-facet rasch model and its bayesian estimation using hamiltonian monte carlo. *Behaviormetrika*, 47(2):469–496.

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

Lihua Yao and Richard D. Schwarz. 2006. A multidimensional partial credit model with associated item and test statistics: An application to mixed-format tests. *Applied Psychological Measurement*, 30(6):469–492.

Lili Yao, Shelby J Haberman, and Mo Zhang. 2019. Prediction of writing true scores in automated scoring of essays by best linear predictors and penalized best linear predictors. *ETS Research Report Series*, 2019(1):1–27.