




Robust Neural Automated Essay Scoring Using Item Response Theory

Masaki Uto^(✉)  and Masashi Okano

The University of Electro-Communications, Tokyo, Japan
uto@ai.lab.uec.ac.jp

Abstract. Automated essay scoring (AES) is the task of automatically assigning scores to essays as an alternative to human grading. Conventional AES methods typically rely on manually tuned features, which are laborious to effectively develop. To obviate the need for feature engineering, many deep neural network (DNN)-based AES models have been proposed and have achieved state-of-the-art accuracy. DNN-AES models require training on a large dataset of graded essays. However, assigned grades in such datasets are known to be strongly biased due to effects of rater bias when grading is conducted by assigning a few raters in a rater set to each essay. Performance of DNN models rapidly drops when such biased data are used for model training. In the fields of educational and psychological measurement, item response theory (IRT) models that can estimate essay scores while considering effects of rater characteristics have recently been proposed. This study therefore proposes a new DNN-AES framework that integrates IRT models to deal with rater bias within training data. To our knowledge, this is a first attempt at addressing rating bias effects in training data, which is a crucial but overlooked problem.

Keywords: Deep neural networks · Item response theory · Automated essay scoring · Rater bias

1 Introduction

In various assessment fields, essay-writing tests have attracted much attention as a way to measure practical and higher-order abilities such as logical thinking, critical reasoning, and creative thinking [1, 4, 13, 18, 33, 35]. In essay-writing tests, examinees write essays about a given topic, and human raters grade those essays based on a scoring rubric. However, grading can be an expensive and time-consuming process when there are many examinees [13, 16]. In addition, human grading is not always sufficiently accurate even when a rubric is used because assigned scores depend strongly on rater characteristics such as strictness and inconsistency [9, 11, 15, 26, 31, 43]. Automated essay scoring (AES), which utilizes natural language processing (NLP) and machine learning techniques to automatically grade essays, is one approach toward resolving this problem.

Many AES methods have been developed over the past decades, and can generally be classified as *feature-engineering* or *automatic feature extraction* approaches [13, 16].

The feature-engineering approach predicts scores using manually tuned features such as essay length and number of spelling errors (e.g., [3, 5, 22, 28]). Advantages of this approach include interpretability and explainability. However, these approaches generally require extensive feature redesigns to achieve high prediction accuracy.

To obviate the need for feature engineering, automatic feature extraction based on deep neural networks (DNNs) has recently attracted attention. Many DNN-AES models have been proposed in the last few years (e.g., [2, 6, 10, 14, 23, 24, 27, 37, 47]) and have achieved state-of-the-art accuracy. This approach requires a large dataset of essays graded by human raters as training data. Essay grading tasks are generally shared among many raters, assigning a few raters to each essay to lower assessment burdens. However, assigned scores are known to be strongly biased due to the effects of rater characteristics [8, 15, 26, 31, 34, 39, 40]. Performance of DNN models rapidly drops when biased data are used for model training, because the resulting model reflects bias effects [3, 12, 17]. This problem has been generally overlooked or ignored, but it is a significant issue affecting all AES methods using supervised machine learning models, including DNN, and because cost concerns make it generally difficult to remove rater bias in practical testing situations.

In the fields of educational and psychological measurement, statistical models for estimating essay scores while considering rater characteristic effects have recently been proposed. Specifically, they are formulated as item response theory (IRT) models that incorporate parameters representing rater characteristics [9, 29, 30, 38, 42–45]. Such models have been applied to various performance tests, including essay writing. Previous studies have reported that they can provide reliable scores by removing adverse effects of rater bias (e.g., [38, 39, 41, 42, 44]).

This study therefore proposes a new DNN-AES framework that integrates IRT models to deal with rater bias in training data. Specifically, we propose a two-stage architecture that stacks an IRT model over a conventional DNN-AES model. In our framework, the IRT model is first applied to raw rating data to estimate reliable scores that remove effects of rater bias. Then, the DNN-AES model is trained using the IRT-based scores. Since the IRT-based scores are theoretically free from rater bias, the DNN-AES model will not reflect bias effects. Our framework is simple and easily applied to various conventional AES models. Moreover, this framework is highly suited to educational contexts and to low- and medium-stakes tests, because preparing high-quality training data in such situations is generally difficult. To our knowledge, this study is a first attempt at mitigating rater bias effects in DNN-AES models.

2 Data

We assume the training dataset consists of essays written by J examinees and essay scores assigned by R raters. Let e_j be an essay by examinee $j \in \mathcal{J} =$

$\{1, \dots, J\}$ and let U_{jr} represent a categorical score $k \in \mathcal{K} = \{1, \dots, K\}$ assigned by rater $r \in \mathcal{R} = \{1, \dots, R\}$ to e_j . The score data can then be defined as $\mathbf{U} = \{U_{jr} \in \mathcal{K} \cup \{-1\} \mid j \in \mathcal{J}, r \in \mathcal{R}\}$, with $U_{jr} = -1$ denoting missing data. Missing data occur because only a few graders in \mathcal{R} can practically grade each essay e_j to reduce assessment workload. Furthermore, letting $\mathcal{V} = \{1, \dots, V\}$ be a vocabulary list for essay collection $\mathbf{E} = \{e_j \mid j \in \mathcal{J}\}$, essay $e_j \in \mathbf{E}$ is definable as a list of vocabulary words $e_j = \{\mathbf{w}_{jt} \in \mathcal{V} \mid t = \{1, \dots, N_j\}\}$, where \mathbf{w}_{jt} is a one-hot representation of the t -th word in e_j , and N_j is the number of words in e_j . This study aimed at training DNN-AES models using this training data.

3 Neural Automated Essay Scoring Models

This section briefly introduces the DNN-AES models used in this study. Although many models have been proposed in the last few years, we apply the most popular model that uses convolution neural networks (CNN) with long short-term memory (LSTM) [2], and an advanced model based on bidirectional encoder representations from transformers (BERT) [7].

3.1 CNN-LSTM-Based Model

A CNN-LSTM-based model [2] proposed in 2016 was the first DNN-AES model. Figure 1(a) shows the model architecture. This model calculates a score for a given essay, which is defined as a sequence of one-hot word vectors, through the following multi-layered neural networks.

Lookup table layer: This layer transforms each word in a given essay into a D -dimensional word-embedding representation, in which words with the same meaning have similar representations. Specifically, letting \mathbf{A} be a $D \times V$ -dimensional embeddings matrix, the embedding representation corresponding to $\mathbf{w}_{jt} \in e_j$ is calculable as the dot-product $\mathbf{A} \cdot \mathbf{w}_{jt}$.

Convolution layer: This layer extracts n-gram level features using CNN from the sequence of word embedding vectors. These features capture local textual dependencies among n-gram words. Zero padding is applied to outputs from this layer to preserve the word length. This is an optional layer, often omitted in current studies.

Recurrent layer: This layer is a LSTM network that outputs a vector at each timestep to capture long-distance dependencies of the words. A single-layer unidirectional LSTM is generally used, but bidirectional or multilayered LSTMs are also often used.

Pooling layer: This layer transforms outputs of the recurrent layer $\mathcal{H} = \{\mathbf{h}_{j1}, \mathbf{h}_{j2}, \dots, \mathbf{h}_{jN_j}\}$ into a fixed-length vector. Mean-over-time (MoT) pooling, which calculates an average vector $\mathbf{M}_j = \frac{1}{N_j} \sum_{t=1}^{N_j} \mathbf{h}_{jt}$, is generally used because it tends to provide stable accuracy. Other frequently used pooling methods include the last pool, which uses the last output of the recurrent layer \mathbf{h}_{jN_j} , and a pooling-with-attention mechanism.

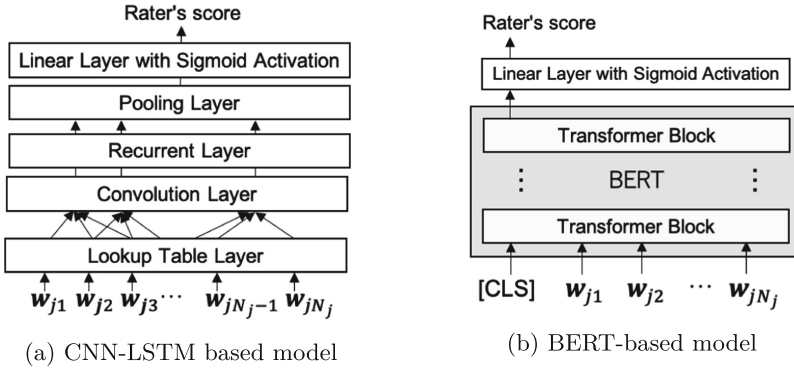


Fig. 1. Architectures of DNN-AES models.

Linear layer with sigmoid activation: This layer projects pooling-layer output to a scalar value in the range $[0, 1]$ by utilizing the sigmoid function as $\sigma(\mathbf{W}\mathbf{M}_j + b)$, where \mathbf{W} is a weight matrix and b is a bias. Model training is conducted by normalizing gold-standard scores to $[0, 1]$, but the predicted scores are rescaled to the original score range in the prediction phase.

3.2 BERT-Based Model

BERT, a pretrained language model released by the Google AI Language team, has achieved state-of-the-art results in various NLP tasks [7]. BERT has been applied to AES [32] and automated short-answer grading (SAG) [19, 21, 36] since 2019, and provides good accuracy.

BERT is defined as a multilayer bidirectional transformer network [46]. Transformers are a neural network architecture designed to handle ordered sequences of data using an attention mechanism. Specifically, transformers consist of multiple layers (called *transformer blocks*), each containing a multi-head self-attention and a position-wise fully connected feed-forward network. See Ref. [46] for details of this architecture.

BERT is trained in *pretraining* and *fine-tuning* steps. Pretraining is conducted on huge amounts of unlabeled text data over two tasks, *masked language modeling* and *next-sentence prediction*, the former predicting the identities of words that have been masked out of the input text and the latter predicting whether two given sentences are adjacent.

Using BERT for a target NLP task, including AES, requires fine-tuning (retraining), which is conducted from a task-specific supervised dataset after initializing model parameters to pretrained values. When using BERT for AES, input essays require preprocessing, namely adding a special token (“CLS”) to the beginning of each input. BERT output corresponding to this token is used as the aggregate sequence representation [7]. We can thus score an essay by

inputting its representation to a *linear layer with sigmoid activation*, as illustrated in Fig. 1(b).

3.3 Problems in Model Training

Training of CNN-LSTM-based AES models and fine-tuning of BERT-based AES models are conducted using large datasets of essays by graded human raters. For model training, the mean-squared error (MSE) between predicted and gold-standard scores is used as the loss function. Specifically, letting y_j be the gold-standard score for essay e_j and letting \hat{y}_j be the predicted score, the MSE loss function is defined as $\frac{1}{J} \sum_{j=1}^J (y_j - \hat{y}_j)^2$.

The gold-standard score y_j is a score for essay e_j assigned by a human rater in a set of raters \mathcal{R} . When multiple raters grade each essay, the gold-standard score should be determined by selecting one score or by calculating an average or total score. In any case, such scores depend strongly on rater characteristics, as discussed in Sect. 1. The accuracy of a DNN model drops when such biased data are used for model training, because the trained model inherits bias effects [3, 12, 17]. In educational and psychological measurement research, item response theory (IRT) models that can estimate essay scores while considering effects of rater characteristics have recently been proposed [9, 29, 30, 38, 42–44]. The main goal of this study is to train AES models using IRT-based unbiased scores. The next section introduces the IRT models.

4 Item Response Theory Models with Rater Parameters

IRT [20] is a test theory based on mathematical models. IRT represents the probability of an examinee response to a test item as a function of latent examinee ability and item characteristics such as difficulty and discrimination. IRT is widely used for educational testing because it offers many benefits. For example, IRT can estimate examinee ability considering effects of item characteristics. Also, the abilities of examinees responding to different test items can be measured on the same scale, and missing response data can be easily handled.

Traditional IRT models are applicable to two-way data (examinees \times test items), consisting of examinee test item scores. For example, the generalized partial credit model (GPCM) [25], a representative polytomous IRT model, defines the probability that examinee j receives score k for test item i as

$$P_{ijk} = \frac{\exp \sum_{m=1}^k [\alpha_i(\theta_j - \beta_i - d_{im})]}{\sum_{l=1}^K \exp \sum_{m=1}^l [\alpha_i(\theta_j - \beta_i - d_{im})]}, \quad (1)$$

where θ_j is the latent ability of examinee j , α_i is a discrimination parameter for item i , β_i is a difficulty parameter for item i , and d_{ik} is a step difficulty parameter denoting difficulty of transition between scores $k - 1$ and k in the item. Here, $d_{i1} = 0$, and $\sum_{k=2}^K d_{ik} = 0$ is given for model identification.

However, conventional GPCM ignores rater factors, so it is not applicable to rating data given by multiple raters as assumed in this study. Extension models

that incorporate parameters representing rater characteristics have been proposed to resolve this difficulty [29,30,38,42–45]. This study introduces a state-of-the-art model [44,45] that is most robust for a large variety of raters. This model defines the probability that rater r assigns score k to examinee j 's essay for a test item (e.g., an essay task) i as

$$P_{ijrk} = \frac{\exp \sum_{m=1}^k [\alpha_r \alpha_i (\theta_j - \beta_r - \beta_i - d_{rm})]}{\sum_{l=1}^K \exp \sum_{m=1}^l [\alpha_r \alpha_i (\theta_j - \beta_r - \beta_i - d_{rm})]}, \quad (2)$$

where α_r is the consistency of rater r , β_r is the strictness of rater r , and d_{rk} is the severity of rater r within category k . For model identification, we assume $\sum_{i=1}^I \log \alpha_i = 0$, $\sum_{i=1}^I \beta_i = 0$, $d_{r1} = 0$, and $\sum_{k=2}^K d_{rk} = 0$.

This study applies this IRT model to rating data \mathbf{U} in training data. Note that DNN-AES models are trained for each essay task. Therefore, rating data \mathbf{U} are defined as two-way data (examinees \times raters). When the number of tasks is fixed to one in the model, the above model identification constraints make α_i and β_i ignorable, so Eq. (2) becomes

$$P_{jrk} = \frac{\exp \sum_{m=1}^k [\alpha_r (\theta_j - \beta_r - d_{rm})]}{\sum_{l=1}^K \exp \sum_{m=1}^l [\alpha_r (\theta_j - \beta_r - d_{rm})]}. \quad (3)$$

This equation is consistent with conventional GPCM, regarding use of item parameters as the rater parameters. Note that θ_j in Eq. (3) represents not only the ability of examinee j but also the latent unbiased scores for essay e_j , because only one essay is associated with each examinee. This model thus provides essay scores with rater bias effects removed.

5 Proposed Method

We propose a DNN-AES framework that uses IRT-based unbiased scores $\boldsymbol{\theta} = \{\theta_j \mid j \in \mathcal{J}\}$ to deal with rater bias in training data.

Figure 2 shows the architectures of the proposed method. As that figure shows, the proposed method is defined by stacking an IRT model over a conventional DNN-AES model. Training of our models occurs in two steps:

1. Estimate the IRT scores $\boldsymbol{\theta}$ from the rating data \mathbf{U} .
2. Train AES models using the IRT scores $\boldsymbol{\theta}$ as the gold-standard scores. Specifically, the MSE loss function for training is defined as $\frac{1}{J} \sum_{j=1}^J (\theta_j - \hat{\theta}_j)^2$, where $\hat{\theta}_j$ represents the AES's predicted score for essay e_j . Since scores $\boldsymbol{\theta}$ are estimated while considering rater bias effects, a trained model will not reflect bias effects. Note that the gold-standard scores must be rescaled to the range $[0, 1]$ for training because sigmoid activation is used in the output layer. In IRT, 99.7% of θ_j fall within the range $[-3, 3]$ because a standard normal distribution is generally assumed. We therefore apply a linear transformation from the range $[-3, 3]$ to $[0, 1]$ after rounding the scores lower than -3 to -3 , and those higher than 3 to 3 .

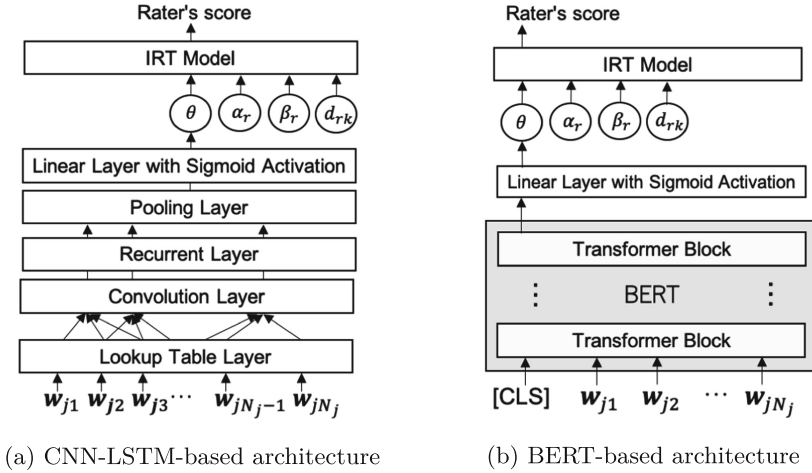


Fig. 2. Proposed architectures.

Note that the increase in training time for the proposed method compared with a conventional method is the time for IRT parameter estimation.

In the testing phase, the score for new essay $e_{j'}$ is predicted in two steps:

1. Predict the IRT score $\theta_{j'}$ from a trained AES model, and rescale it to the range $[-3,3]$.
2. Calculate the expected score $\hat{U}_{j'}$, which corresponds to an unbiased original-scaled score of $e_{j'}$ [39], as

$$\hat{U}_{j'} = \frac{1}{R} \sum_{r=1}^R \sum_{k=1}^K k \cdot P_{j'rk}. \tag{4}$$

6 Experiments

This section describes evaluation of the effectiveness of the proposed method through actual data experiments.

6.1 Actual Data

These experiments used the Automated Student Assessment Prize (ASAP) dataset, which is widely used as benchmark data in AES studies. This dataset consists of essays on eight topics, originally written by students from grades 7 to 10. There are 12,978 essays, averaging 1,622 essays per topic. However, this dataset cannot be directly used to evaluate the proposed method, because despite its essays having been graded by multiple raters, it contains no rater identifiers.

We therefore employed other raters and asked them to grade essays in the ASAP dataset. We used essay data for the fifth ASAP topic, because the number

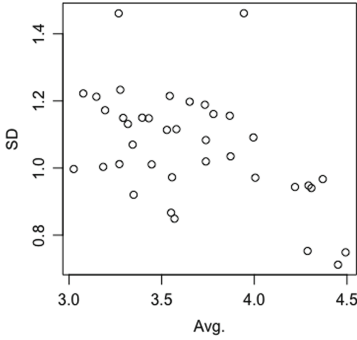


Fig. 3. Score statistics (average and SD) for each rater.

Table 1. Category usage rates.

Rater ID	Rating category				
	1	2	3	4	5
1	4%	23%	27%	28%	17%
2	4%	3%	36%	48%	10%
3	2%	6%	7%	32%	54%
4	2%	4%	10%	22%	62%
5	3%	20%	35%	30%	12%
6	6%	16%	33%	25%	21%
7	3%	22%	41%	23%	12%
8	12%	8%	11%	10%	58%
9	1%	11%	33%	43%	12%
10	9%	24%	28%	23%	17%

of essays in that topic is relatively large ($n = 1805$). We recruited 38 native English speakers as raters through Amazon Mechanical Turk and assigned four raters to each essay. Each rater graded around 195 essays. The assessment rubric used the same five rating categories as ASAP. Average Pearson’s correlation between the collected rating scores and the original ASAP scores was 0.675.

To confirm any differences in rater characteristics, we plotted averaged score values and standard deviations (SD) for each rater, as shown in Fig. 3. In that figure, each plot represents a rater, and horizontal and vertical axes respectively show the average and SD values. In addition, Table 1 shows appearance rates in the five rating categories for 10 representative raters. The figure and table show extreme differences in grading characteristics among the raters, suggesting that consideration of rater bias is required.

6.2 Experimental Procedures

This subsection shows that the proposed method can provide more robust scores than can conventional AES models, even when the rater grading each essay in the training data changes. The experimental procedures, which are similar to those used in previous studies examining IRT scoring robustness [39–42], were as follows:

1. We estimated IRT parameters by the Markov chain Monte Carlo (MCMC) algorithm [30, 42] using all rating data.
2. We created a dataset consisting of (essay, score) pairs by randomly selecting one score for each essay from among the scores assigned by multiple raters. We repeated this data generation 10 times. Hereafter, the m -th generated dataset is represented as U'_m .
3. From each dataset U'_m , we estimated IRT scores θ (referred to as θ_m) given the rater parameters obtained in Step 1, and then created a dataset U''_m comprising essays and θ_m values.

Table 2. Evaluations of prediction robustness.

	Kappa		Weighted Kappa		RMSE		Correlation	
	Prop.	Conv.	Prop.	Conv.	Prop.	Conv.	Prop.	Conv.
CNN+LSTM (MoT)	0.749	0.624	0.778	0.727	0.191	0.301	0.937	0.931
CNN+LSTM (Last)	0.696	0.459	0.701	0.551	0.212	0.400	0.829	0.783
LSTM (MoT)	0.831	0.697	0.845	0.779	0.142	0.237	0.965	0.958
LSTM (Last)	0.612	0.371	0.624	0.514	0.300	0.518	0.804	0.775
BERT	0.790	0.629	0.808	0.743	0.159	0.311	0.960	0.935

- Using each dataset \mathbf{U}''_m , we conducted five-fold cross validation to train AES models and to obtain predicted scores $\hat{\theta}_m$ for all essays.
- We calculated metrics for agreement between the expected scores calculated by Eq. (4) given $\hat{\theta}_m$ and those calculated given $\hat{\theta}_{m'}$ for all unique $m, m' \in \{1, \dots, 10\}$ pairs (${}_{10}C_2 = 45$ pairs in total). As agreement metrics, we used Cohen’s kappa, weighted kappa, root mean squared error (RMSE), and Pearson correlation coefficient.
- We calculated average metric values obtained from the 45 pairs.

High kappa and correlation and low RMSE values obtained from the experiment indicate that score predictions are more robust for different raters.

We conducted a similar experiment using conventional DNN-AES models without the IRT model. Specifically, using each dataset \mathbf{U}'_m , we predicted essay scores from a DNN-AES model through five-fold cross validation procedures as in Step 4. We then calculated the four agreement metrics among the predicted scores obtained from different datasets \mathbf{U}'_m , and averaged them.

These experiments were conducted with several DNN-AES models. Specifically, we examined CNN-LSTM models using MoT pooling or last pooling, those models without a CNN layer, and the BERT model. These models were implemented in Python with the Keras library. For the BERT model, we used the *base*-sized pretrained model. The hyperparameters and dropout settings were determined following Refs. [2, 7, 46].

6.3 Experimental Results

Table 2 shows the results, which indicate that the proposed method sufficiently improves agreement metrics as compared to the conventional models in all cases. The results indicate that the proposed method provides stable scores when the rater allocation for each essay in training data is changed, thus demonstrating that it is highly robust against rater bias. Note that the values in Table 2 are not comparable with the results of previous AES studies because our experiment and previous experiments evaluated different aspects of AES performance.

In addition, as in previous AES studies, we evaluated score (θ) prediction accuracy of the proposed method through five-fold cross-validation. We mea-

Table 3. Prediction accuracy for IRT score θ by the proposed method

	MAE	RMSE	Correlation	R^2
CNN+LSTM (MoT)	0.431	0.546	0.719	0.499
CNN+LSTM (Last)	0.580	0.717	0.417	0.161
LSTM (MoT)	0.408	0.519	0.749	0.557
LSTM (Last)	0.509	0.640	0.584	0.340
BERT	0.400	0.511	0.763	0.562

sured accuracy using mean absolute error (MAE), RMSE, the correlation coefficient, and the coefficient of determination (R^2), because θ is a continuous variable. Table 3 shows the results, which indicate that the CNN-LSTM and LSTM models with MoT pooling achieved higher performance than did those with last pooling. The table also shows that the CNN did not effectively improve accuracy. These tendencies are consistent with a previous study [2]. In addition, the BERT provided the highest accuracy, which is also consistent with current NLP studies.

Tables 2 and 3 show that the score prediction robustness in Table 2 tends to increase with score prediction accuracy. This might be because scores in low-performance DNN-AES models are strongly biased not only by rater characteristics, but also by prediction errors arising from the model itself. With increasing accuracy of DNN-AES models, rater bias effects as a percentage of overall error increases, suggesting that the impact of the proposed method increases.

7 Conclusion

We showed that DNN-AES model performance strongly depends on the characteristics of raters grading essays in training data. To resolve this problem, we proposed a new DNN-AES framework that integrates IRT models. Specifically, we formulated our method as a two-stage architecture that stacks the IRT model over a conventional DNN-AES model. Through experiments using an actual dataset, we demonstrated that the proposed method can provide more robust essay scores than can conventional DNN-AES models. The proposed method is simple but powerful, and is easily applicable to any AES model. As described in the Introduction, our method is also highly suited to situations where high-quality training data are hard to prepare, including educational contexts.

In future studies, we expect to evaluate effectiveness of the proposed method using various datasets. Although this study mainly focused on robustness against rater bias, the proposed method might also improve prediction accuracy for each rater's score. In future studies, the accuracy should be evaluated. Our method is defined as a two-stage procedure for separately training IRT models and DNN-AES models. However, conducting end-to-end optimization would further improve the performance. This extension is another topic for future study.

Acknowledgment. This work was supported by JSPS KAKENHI 17H04726 and 17K20024.

References

1. Abosalem, Y.: Beyond translation: adapting a performance-task-based assessment of critical thinking ability for use in Rwanda. *Int. J. Secondary Educ.* **4**(1), 1–11 (2016)
2. Alikaniotis, D., Yannakoudakis, H., Rei, M.: Automatic text scoring using neural networks. In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 715–725 (2016)
3. Amorim, E., Caçado, M., Veloso, A.: Automated essay scoring in the presence of biased ratings. In: *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 229–237 (2018)
4. Bernardin, H.J., Thomason, S., Buckley, M.R., Kane, J.S.: Rater rating-level bias and accuracy in performance appraisals: the impact of rater personality, performance management competence, and rater accountability. *Hum. Resour. Manag.* **55**(2), 321–340 (2016)
5. Dascalu, M., Westera, W., Ruseti, S., Trausan-Matu, S., Kurvers, H.: *ReaderBench* learns Dutch: building a comprehensive automated essay scoring system for Dutch language. In: André, E., Baker, R., Hu, X., Rodrigo, M.M.T., du Boulay, B. (eds.) *AIED 2017*. LNCS (LNAI), vol. 10331, pp. 52–63. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61425-0_5
6. Dasgupta, T., Naskar, A., Dey, L., Saha, R.: Augmenting textual qualitative features in deep convolution recurrent neural network for automatic essay scoring. In: *Proceedings of the Workshop on Natural Language Processing Techniques for Educational Applications*, Association for Computational Linguistics, pp. 93–102 (2018)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186 (2019)
8. Eckes, T.: Examining rater effects in TestDaF writing and speaking performance assessments: a many-facet Rasch analysis. *Lang. Assess. Q.* **2**(3), 197–221 (2005)
9. Eckes, T.: *Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluating Rater-Mediated Assessments*. Peter Lang Publication Inc., New York (2015)
10. Farag, Y., Yannakoudakis, H., Briscoe, T.: Neural automated essay scoring and coherence modeling for adversarially crafted input. In: *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 263–271 (2018)
11. Hua, C., Wind, S.A.: Exploring the psychometric properties of the mind-map scoring rubric. *Behaviormetrika* **46**(1), 73–99 (2018). <https://doi.org/10.1007/s41237-018-0062-z>
12. Huang, J., Qu, L., Jia, R., Zhao, B.: O2U-Net: a simple noisy label detection approach for deep neural networks. In: *Proceedings of the IEEE International Conference on Computer Vision* (2019)
13. Hussein, M.A., Hassan, H.A., Nassef, M.: Automated language essay scoring systems: a literature review. *PeerJ Comput. Sci.* **5**, e208 (2019)

14. Jin, C., He, B., Hui, K., Sun, L.: TDNN: a two-stage deep neural network for prompt-independent automated essay scoring. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics, pp. 1088–1097 (2018)
15. Kassim, N.L.A.: Judging behaviour and rater errors: an application of the many-facet Rasch model. *GEMA Online J. Lang. Stud.* **11**(3), 179–197 (2011)
16. Ke, Z., Ng, V.: Automated essay scoring: a survey of the state of the art. In: Proceedings of the International Joint Conference on Artificial Intelligence, pp. 6300–6308 (2019)
17. Li, S., et al.: Coupled-view deep classifier learning from multiple noisy annotators. In: Proceedings of the Association for the Advancement of Artificial Intelligence (2020)
18. Liu, O.L., Frankel, L., Roohr, K.C.: Assessing critical thinking in higher education: current state and directions for next-generation assessment. *ETS Res. Rep. Ser.* **1**, 1–23 (2014)
19. Liu, T., Ding, W., Wang, Z., Tang, J., Huang, G.Y., Liu, Z.: Automatic short answer grading via multiway attention networks. In: Isotani, S., Millán, E., Ogan, A., Hastings, P., McLaren, B., Luckin, R. (eds.) AIED 2019. LNCS (LNAI), vol. 11626, pp. 169–173. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-23207-8_32
20. Lord, F.: Applications of Item Response Theory to Practical Testing Problems. Erlbaum Associates, Mahwah (1980)
21. Lun, J., Zhu, J., Tang, Y., Yang, M.: Multiple data augmentation strategies for improving performance on automatic short answer scoring. In: Proceedings of the Association for the Advancement of Artificial Intelligence (2020)
22. Shermis, M.D., Burstein, J.C.: Automated Essay Scoring: A Cross-disciplinary Perspective. Taylor & Francis, Abingdon (2016)
23. Mesgar, M., Strube, M.: A neural local coherence model for text quality assessment. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 4328–4339 (2018)
24. Mim, F.S., Inoue, N., Reisert, P., Ouchi, H., Inui, K.: Unsupervised learning of discourse-aware text representation for essay scoring. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pp. 378–385 (2019)
25. Muraki, E.: A generalized partial credit model. In: van der Linden, W.J., Hambleton, R.K. (eds.) Handbook of Modern Item Response Theory, pp. 153–164. Springer, Heidelberg (1997). https://doi.org/10.1007/978-1-4757-2691-6_9
26. Myford, C.M., Wolfe, E.W.: Detecting and measuring rater effects using many-facet Rasch measurement: part I. *J. Appl. Measur.* **4**, 386–422 (2003)
27. Nadeem, F., Nguyen, H., Liu, Y., Ostendorf, M.: Automated essay scoring with discourse-aware neural models. In: Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics, pp. 484–493 (2019)
28. Nguyen, H.V., Litman, D.J.: Argument mining for improving the automated scoring of persuasive essays. In: Proceedings of the Association for the Advancement of Artificial Intelligence, pp. 5892–5899 (2018)
29. Patz, R.J., Junker, B.W., Johnson, M.S., Mariano, L.T.: The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *J. Educ. Behav. Stat.* **27**(4), 341–384 (2002)
30. Patz, R.J., Junker, B.: Applications and extensions of MCMC in IRT: multiple item types, missing data, and rated responses. *J. Educ. Behav. Stat.* **24**(4), 342–366 (1999)

31. Rahman, A.A., Ahmad, J., Yasin, R.M., Hanafi, N.M.: Investigating central tendency in competency assessment of design electronic circuit: analysis using many facet Rasch measurement (MFRM). *Int. J. Inf. Educ. Technol.* **7**(7), 525–528 (2017)
32. Rodriguez, P.U., Jafari, A., Ormerod, C.M.: Language models and automated essay scoring. *arXiv, cs.CL* (2019)
33. Rosen, Y., Tager, M.: Making student thinking visible through a concept map in computer-based assessment of critical thinking. *J. Educ. Comput. Res.* **50**(2), 249–270 (2014)
34. Saal, F., Downey, R., Lahey, M.: Rating the ratings: assessing the psychometric quality of rating data. *Psychol. Bull.* **88**(2), 413–428 (1980)
35. Schendel, R., Tolmie, A.: Assessment techniques and students' higher-order thinking skills. *Assess. Eval. High. Educ.* **42**(5), 673–689 (2017)
36. Sung, C., Dhamecha, T.I., Mukhi, N.: Improving short answer grading using transformer-based pre-training. In: Isotani, S., Millán, E., Ogan, A., Hastings, P., McLaren, B., Luckin, R. (eds.) *AIED 2019. LNCS (LNAI)*, vol. 11625, pp. 469–481. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-23204-7_39
37. Taghipour, K., Ng, H.T.: A neural approach to automated essay scoring. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1882–1891 (2016)
38. Ueno, M., Okamoto, T.: Item response theory for peer assessment. In: *Proceedings of the IEEE International Conference on Advanced Learning Technologies*, pp. 554–558 (2008)
39. Uto, M.: Rater-effect IRT model integrating supervised LDA for accurate measurement of essay writing ability. In: Isotani, S., Millán, E., Ogan, A., Hastings, P., McLaren, B., Luckin, R. (eds.) *AIED 2019. LNCS (LNAI)*, vol. 11625, pp. 494–506. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-23204-7_41
40. Uto, M., Thien, N.D., Ueno, M.: Group optimization to maximize peer assessment accuracy using item response theory. In: André, E., Baker, R., Hu, X., Rodrigo, M.M.T., du Boulay, B. (eds.) *AIED 2017. LNCS (LNAI)*, vol. 10331, pp. 393–405. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61425-0_33
41. Uto, M., Duc Thien, N., Ueno, M.: Group optimization to maximize peer assessment accuracy using item response theory and integer programming. *IEEE Trans. Learn. Technol.* **13**(1), 91–106 (2020)
42. Uto, M., Ueno, M.: Item response theory for peer assessment. *IEEE Trans. Learn. Technol.* **9**(2), 157–170 (2016)
43. Uto, M., Ueno, M.: Empirical comparison of item response theory models with rater's parameters. *Heliyon* **4**(5), 1–32 (2018). Elsevier
44. Uto, M., Ueno, M.: Item response theory without restriction of equal interval scale for rater's score. In: Penstein Rosé, C., et al. (eds.) *AIED 2018. LNCS (LNAI)*, vol. 10948, pp. 363–368. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93846-2_68
45. Uto, M., Ueno, M.: A generalized many-facet Rasch model and its Bayesian estimation using Hamiltonian Monte Carlo. *Behaviormetrika* **47**, 1–28 (2020). <https://doi.org/10.1007/s41237-020-00115-7>
46. Vaswani, A., et al.: Attention is all you need. In: *Proceedings of the International Conference on Advances in Neural Information Processing Systems*, pp. 5998–6008 (2017)
47. Wang, Y., Wei, Z., Zhou, Y., Huang, X.: Automatic essay scoring incorporating rating schema via reinforcement learning. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 791–797 (2018)