On the Limitations of Human-Computer Agreement in Automated Essay Scoring

Afrizal Doewes, Mykola Pechenizkiy
Eindhoven University of Technology
{a.doewes, m.pechenizkiy}@tue.nl

ABSTRACT

Scoring essays is generally an exhausting and time-consuming task for teachers. Automated Essay Scoring (AES) facilitates the scoring process to be faster and more consistent. The most logical way to assess the performance of an automated scorer is by measuring the score agreement with the human raters. However, we provide empirical evidence that a well-performing essay scorer from the quantitative evaluation point of view are still too risky to be deployed. We propose several input scenarios to evaluate the reliability and the validity of the system, such as off-topic essays, gibberish, and paraphrased answers. We demonstrate that automated scoring models with high human-computer agreement fail to perform well on two out of three test scenarios. We also discuss the strategies to improve the performance of the system.

Keywords

Automated Essay Scoring, Testing Scenarios, Reliability and Validity

1. INTRODUCTION

Automated Essay Scoring (AES) system is a computer software designed as a tool to facilitate the evaluation of student essays. Theoretically, AES systems work faster, reduce cost in term of evaluator's time, and eliminate concerns about rater consistency. The most logical way to assess the performance of an automated scorer is by measuring the score agreement with the human raters. The score agreement rate must exceed a specific threshold value to be considered as having a good performance. Consequently, most studies have focused on increasing the level of agreement between human and computer scoring. However, the process of establishing reliability should not stop with the calculation of inter-coder reliability, because automated scoring poses some distinctive validity challenges such as the potential to misrepresent the construct of interest, vulnerability to cheating, impact on examinee behavior, and users' interpretation on score and use of scores [1]. Bennet and Bejar [2] have argued that reliability scores are limited in their reliance on human ratings for evaluating the performance of automated scoring primarily because human graders are fallible. Humans raters may experience fatigue and have problems with scoring consistency across time. Reliability calculations alone are therefore not adequate as the current trend for establishing validity [3]. A well-performing essay scorer from the quantitative evaluation perspective is too risky to be deployed before evaluating the system's reliability and validity.

The initial attempt to discuss validity issues regarding automated scoring in a larger context of a validity argument for the assessment was made by Clauser et al. [4]. They presented several outlines of the potential validity threats that automated scoring would introduce to the overall interpretation and use. Enright and Quinlan [5] discussed how the evidence for a scoring process that uses both human and e-rater scoring is relevant to validity argument. They described an e-rater model which was proposed to score one of the two writing tasks on the TOEFL-iBT writing section. Automated scorer was investigated as a tool to complement to human judgement on essays written by English language learners.

Several criticisms for Automated Essay Scoring (AES) system were highlighted in [6]. They argued that there were limited studies on how effective automated writing evaluation was used in writing classes as a pedagogical tool. In their study, the students gave negative reactions towards the automated assessment. One of the problems was that the scoring system favored lengthiness; higher scores were awarded to longer essays. It also overemphasized the use of transition words, which increased the score of an essay immediately. Moreover, it ignored coherence and content development as an essay could achieve a high score by having four or five paragraphs with relevant keywords, although it had major coherence problems and illogical ideas. Another concern is described in [7]. Specifically, knowing how the system evaluates an essay may be a reason why students can fool the system into assigning a higher score than what is warranted. They concluded that the system was not ready yet as the only scorer, especially for high-stakes testing, without the help of expert human raters.

Most researchers agree that human - automated score agreement still serves as the standard baseline for measuring the quality of machine score prediction. However, there is an inherent limitation with this measurement because the agreement rate is usually derived only from the data used for training and testing the machine learning model. The aim of this paper is to highlight some limitations of standard performance metrics used to evaluate automated essay scoring model, using several input scenarios to evaluate the reliability and the validity of the system, such as off-topic essays, gibberish, and paraphrased answers. We show empirical evidence that a well-performing automated essay scorer, with high agreement rate between human-machine, is not necessarily ready for deployment for operational use, since it fails to perform well on two out of three test scenarios. In addition, we also discuss some strategies to improve the performance of the system. This paper begins with the explanation of the quantitative performance acceptance criteria for an automated scoring model from [1]. Then, we present the experiment settings, including the training algorithm and the essay features, for creating the model. Afterwards, we discuss the experiment results, model performance analysis, reliability and validity evaluation and the strategies for improvement, and finally, we conclude our work.

2. SCORE AGREEMENT EVALUATION

According to Williamson et al. in [1], the following are some of the acceptance criteria used for evaluation of automated scoring with respect to human scores when automated scoring is intended to be used in conjunction with human scoring:

1. Agreement of scores between human raters and computer

Agreement between human scores and automated scores has been a long-established measure of the performance of automated scoring. This is to measure whether the agreement satisfies a predefined threshold. The quadratic weighted kappa (QWK) between automated and human scoring must be at least .70 (rounded normally). It is important to note that the performance of automated scorer will rely on the quality of the human scoring. Therefore, the interrater agreement among human raters must first be reliable.

2. Degradation from human-human score agreement

The human-automated scoring agreement cannot be more than .10 lower than the human-human agreement. This standard prevents the case in which automated essay scoring may reach the .70 threshold but still be notably deficient in comparison with human scoring. In addition, it does not rule out the cases in which the automated–human agreement has been slightly less than the .70 threshold, but very close to a borderline performance for human scoring. For example, a human-computer QWK of .69 and human-human QWK of .71. Such model is approved for operational use based on being highly similar to human scoring. Moreover, Williamson et al. stated that it is relatively usual to observe automated-human agreements that are higher than the human-human agreements for tasks that predominantly target linguistic writing quality (e.g. GRE Issue and TOEFL Independent tasks).

3. Standardized mean score difference between human and automated scores.

Another measure for association of automated scores with human scores is that the standardized mean score difference (standardized on the distribution of human scores) between the human and computer cannot exceed .15. The standardized difference of the mean is formalized as follows:

$$\overline{Z} = \frac{[\overline{X}_{AS} - \overline{X}_H]}{\sqrt{\frac{SD_{AS}^2 - SD_H^2}{2}}} \quad (1)$$

where $\overline{X}_{AS}$ is the mean of the automated score, $\overline{X}_H$ is the mean of the human score, SD_{AS}^2 is the variance of the automated score, and SD_H^2 is the variance of the human score.

3. EXPERIMENTS

3.1 DATASET

We used the Automated Student Assessment Prize (ASAP) dataset[1], hosted by the Kaggle platform, as our experiment data. ASAP is the most widely used dataset to evaluate the performance of AES systems [8]. All the essays provided are already human graded. ASAP dataset consists of eight prompts, with varying score ranges for each prompt. Table 1 highlights the topics of each prompt in ASAP dataset.

Table 1 Prompts in ASAP Dataset

ASAP Dataset	Topics
Prompt 1	The effects computers have on people
Prompt 2	Censorship in the libraries
Prompt 3	Respond to an extract about how the features of a setting affected a cyclist
Prompt 4	Explain why an extract from *Winter Hibiscus* by Minfong Ho was concluded in the way the author did.
Prompt 5	Describe the mood created by the author in an extract from *Narciso Rodriguez* by Narciso Rodriguez
Prompt 6	The difficulties faced by the builders of the Empire State Building in allowing dirigibles to dock there
Prompt 7	Write a story about patience
Prompt 8	The benefits of laughter

3.2 FEATURES EXTRACTION

Each essay is transformed into a 780 dimension of features vector. We extract the essay features into two categories: 12 interpretable features, and 768 dimension of Sentence-BERT vector representation. Table 2 contains the essay features we used to train the scoring model.

Table 2 Essay Features

Type	Description
Interpretable essay features (12 features)	Answer Length (Character counts) Word count Average word length Count of "good" POS n-grams Number of overlapping tokens with the prompt Number of overlapping tokens (including synonyms) with the prompt Number of punctuations Spelling errors Unique words count Prompt – answer similarity score (SBERT representation) Prompt – answer similarity score (BOW representation) Language Errors
Sentence-BERT features (768 dim)	The encoding of the essay using Sentence-BERT pretrained model

3.2.1 Interpretable Essay Features

Six out of the twelve interpretable essay features are extracted from EASE (Enhanced AI Scoring Engine) library[2], written by one of the winners in ASAP Kaggle competition. This features set have been proven to be robust [9]. EASE generates 414-length features. However, we exclude most of the features generated by EASE library that are mostly Bag-of-Words vectors. The other six features extracted from the text are the number of punctuations, the number of spelling errors, unique words count, similarity scores between

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

[2] https://github.com/edx/ease

answer and prompt using S-BERT and BOW (Bag-of-Words) vector representations, and the number of language errors.

The grammar feature is measured by the number of good n-grams in the essay. EASE library extracts the essay text into its POS-tags and compares them with a list of valid POS-tag combinations in English. Good POS n-grams are defined as the ones that separate high- from low-scoring essays, determined using the Fisher test [10]. Moreover, we count the number of language errors in an answer using Language Tool Python library[3]. Mechanics in a language include aspects such as the usage of punctuation and the number of spelling errors found in the answer.

The average word length and long words count are used by Mahana et al. to estimate language fluency and dexterity [11]. Larkey also used the number of long words to indicate the complexity of term usage [12]. Unique words count feature is useful to estimate the richness of vocabulary in the answer.

The relevance factor of an answer combines two features from EASE library, which are related to the degree of tokens overlap between the prompt and the answer, including their synonyms. Two additional features are the cosine similarity measurement between the answer and the prompt, both using the Sentence-BERT and the BOW representation.

3.2.2 Sentence-BERT representation

Sentence-BERT, introduced by Reimers and Gurevych (2019), is a modification of pretrained BERT network using Siamese and triplet network [13]. It converts a text into a 768-dimension feature vectors and produces semantically meaningful sentence embedding. The embedding result can then be compared using cosine-similarity.

3.3 MODEL TRAINING

We train the regression models using Gradient Boosting algorithms, with 80% training data and 20% testing data (using 5-fold cross-validation). We use Quadratic Weighted Kappa (QWK) [14] score as the evaluation metric, which measures the agreement between system predicted scores and human-annotated scores. QWK is the standard evaluation metric to measure the performance of an AES system [1]. Although ASAP dataset has 8 prompts, we trained 9 models in total. The reason is that prompt 2 was scored in two different domains (Writing Application and Language Conventions). Therefore, we must create two separate predictions for this essay prompt. To train all nine models, we used different hyperparameters for each model.

4. RESULTS

4.1 Model Performance Evaluation

In this subsection, we evaluated the model performance based on the quantitative evaluation criteria as discussed in Section 2. Furthermore, we also analyzed the distribution of overall holistic scores assigned by human raters and computer.

4.1.1 Score Agreement Evaluation

We conducted the quantitative evaluation of our models based on the acceptance criteria by Williamson et al. in [1], and the results are shown in Table 3. The table describes the performance measurements for the scoring model of each prompt. The first column is the QWK score, which measures the human-computer agreement. The human-human agreement score in the second column is used to calculate the degradation value. The last column contains the standardized mean score difference between human and automated scores.

Table 3 Model Performance Evaluation for ASAP Dataset

Dataset	QWK Score	Human Agreement	Degradation	$\overline{Z}$
1	0.7826	0.72095	-0.06165	0.0056
2_dom1	0.6731	0.81413	0.14103	0.0007
2_dom2	0.6715	0.80175	0.13025	0.0394
3	0.6887	0.76923	0.08053	0.0272
4	0.7736	0.85113	0.07753	0.0094
5	0.8065	0.7527	-0.0538	0.0229
6	0.7985	0.77649	-0.02201	0.0102
7	0.7771	0.72148	-0.05562	0.0023
8	0.6668	0.62911	-0.03769	0.0147

Based on the above results, we concluded that five models (prompt 1, 4, 5, 6, and 7) satisfy the quantitative evaluation criteria defined as a standard in [1]. Next, we continued our analysis and the reliability and validity tests on only these well-performing models and ignored the other underperforming models (prompt 2_dom1, 2_dom2, 3, and 8).

4.1.2 Distribution of overall holistic scores

We investigated the distribution of the overall holistic scores assigned by human raters and the automated scorer. It is important to understand the distribution of the scores, especially in relation to the decision of an exam. For this purpose, we presented the decision into three categories: passing, borderline, and failing. Table 4 shows the rubric score and resolved score range in ASAP dataset, in which our model passed the quantitative performance evaluation in the previous subsection. The resolved score is the final score after combining the rubric scores from two human raters. Each prompt has a different score resolution. In some prompts, if there was a difference between scorer 1 and scorer 2, the final score was always the higher of the two. In another prompt, the final score was the sum of scores from rater 1 and rater 2 if their scores are adjacent. If non-adjacent, an expert scorer will determine the final score.

Table 4 ASAP Rubric and Resolved Score Range

Dataset	Rubric score	Resolved Score	Borderline Score
Prompt 1	1 - 6	2 - 12	6-8
Prompt 4	0 - 3	0 - 3	1
Prompt 5	0 - 4	0 - 4	2
Prompt 6	0 - 4	0 - 4	2
Prompt 7	0 - 12	0 - 24	12

To categorize the exam results, we created a borderline score from the resolved score. It is not necessarily the exact middle score because some datasets do not have it. We considered scores below the borderline as failing and scores above the borderline as passing.

[3] https://github.com/jxmorris12/language_tool_python

Table 5 Frequency Comparison for Passing, Borderline, and Failing (in %)

	ASAP 1			ASAP 4			ASAP 5			ASAP 6			ASAP 7		
	HR1	**HR2**	**AES**	**HR1**	**HR2**	**AES**	**HR1**	**HR2**	**AES**	**HR1**	**HR2**	**AES**	**HR1**	**HR2**	**AES**
Passing	35.2	35.5	50.1	39.6	40	45.2	37.8	39.2	45.9	59.4	59.1	65	74	72.8	82
Borderline	62.7	62.5	48.8	42.7	41.9	42.3	38.5	36	38.7	25.7	25.9	25.6	8.9	9.8	4.9
Failing	2.1	2	1.1	17.7	18.1	12.5	23.7	24.8	15.4	14.9	15	9.4	17.1	17.4	13.1
Total	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**

As can be seen from Table 5, the use of automated scorer will put the students at an advantage, compared to when the exam only employs human raters. In all datasets, we have the same overview. The AES models assigned scores with much higher passing rates than both human raters. And as far as the failure rates are concerned, the scores from AES indicate very low failing rates, compared to human raters. This finding supports the result in [15], with a different dataset for the experiment.

4.2 Evaluating and Improving the Model

We examined the performance of the model using three scenarios: gibberish answer, paraphrased answer, and off-topic answer. We discuss the evaluation results and strategies to improve the system in the following sections.

4.2.1 Off-topic essay

One way to validate the use of an automated essay scoring system is by checking its performance against off-topic answers. For this study, we use ASAP dataset which has eight sub datasets (prompts). To simulate the experiment, for each model, we used the answers in the other seven prompts as the off-topic essays. We randomly sampled 50 essays from each dataset, resulting in a total of 350 off-topic essays. Using 5-fold cross-validation, we measured the accuracy of the model in predicting the score of the off-topic essays, which we assume should get 0 (zero), due to the complete irrelevance with the corresponding prompt.

Table 6 Accuracy of off-topic detection

Dataset	Training Data	Accuracy (%)	QWK Score
Prompt 1 (2 – 12)	Original	0%	0.7826
	Original + 350 off-topic	55.4%	0.7031
Prompt 4 (0 – 3)	Original	4.3%	0.7736
	Original + 350 off-topic	91.4%	0.7697
Prompt 5 (0 - 4)	Original	0.6%	0.8065
	Original + 350 off-topic	88%	0.7951
Prompt 6 (0 – 4)	Original	5.80%	0.799
	Original + 350 off-topic	97%	0.787
Prompt 7 (0 - 24)	Original	0%	0.7771
	Original + 350 off-topic	45.7%	0.7225

We also investigated the change of value of the QWK scores after retraining the model. The motivation is, we want to avoid that the retraining process degrades the performance of the original model. The new model should still perform well in predicting the original essay set.

Table 6 describes the experiment results of the retraining process, the improvement in accuracies, and the change in QWK scores. The models trained with only the original dataset performed with very low accuracies. The highest result is by prompt 6, with 5.8% of the off-topic essays are correctly given 0 scores. Prompt 1 and prompt 7 are the worst with no correct prediction at all. It means all off-topic essays are graded with scores greater than 0.

We can observe the effect of including the off-topic essays in the training data. The model performance for prompt (4, 5, and 6) drastically increase. For prompt 6, the accuracy on predicting the unseen data of the off-topic essays (test set) reached 97%, with only a slight decrease on the QWK score. There are moderate improvements for prompt 1 and especially prompt 7. We assume this is due to a larger score range, which for prompt 7 is 0 – 24. If we are being less strict about the score 0 (zero) policy for off-topic essays, for example by categorizing score between 0 – 3 as failed score, we obtain a much better accuracy, which is 81.7%. Meanwhile, for prompt 1, the lowest resolved score is 2. However, we trained the model to give off-topic essays 0 score prediction. If we create a score range 0 – 2 as failed category, the accuracy increases to 93%. Because we have many score predictions for the off-topic answers ranging from 0 to 2 by prompt 1.

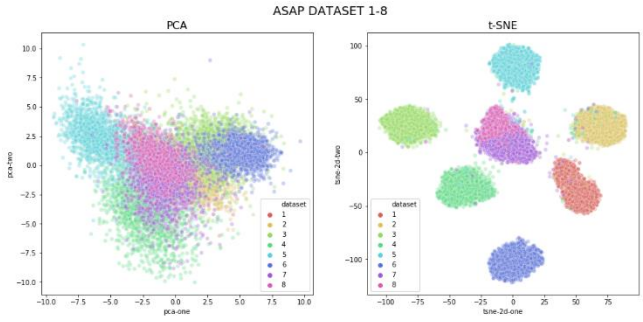

Figure 1 SBERT representation for 8 prompts in ASAP dataset

We conclude that the solution for detecting off-topic essays is relatively simple. Without using additional features for detecting off-topic essays, we found out that the SBERT features are very helpful to be used as features for training a scoring model. Using PCA and t-SNE, we plotted the SBERT vector representation of the essays in all eight of ASAP dataset prompts. Figure 1 shows that using t-SNE, all prompts are almost perfectly separated, although we can see both dataset 7 and dataset 8 are close to each other in the middle of the plot. We assume that it is caused by their similar prompt topics. If we check Table 1 for the description of topics in ASAP dataset, in prompt 7 the students are asked to write a story about patience, while prompt 8 discusses about the benefits of laughter. Both topics are arguably more closely related to each other than when we compare them with the topics of prompt 1 – prompt 6.

4.2.2 *Gibberish*

For the next input scenario, we want to avoid that the system receives invalid answers such as gibberish, and undeservedly returns scores other than zero. Ideally, any gibberish answer must get the score zero. However, using our model, we tested several gibberish as the answers, and the scores are not zero. We provide some examples of the inputs as shown in Table 7. In this table, we show the examples of wrong predictions by the scoring model that was trained using ASAP dataset prompt 6. Nevertheless, we can also observe similar problems in the other models.

Table 7 Examples of Wrong Predictions by Prompt 6

Answer	Score (0-4)
asdafafdf adjhgladghad	1
Eyoqtuwrpituauoyeqo ngbambgagadhkq3124 31794613 hbfka df	1
orkesbroh	1

To analyze the reason, we conducted local model interpretation, which means that we are interested in understanding which variable, or combination of variables, determines the specific prediction. We use SHAP (SHapley Additive exPlanations) values to help in determining the most predictive variables in a single prediction [16]. In AES, the system output is a real number. Each variable contribution will either increase or decrease the output value. One implementation of SHAP libraries is TreeExplainer, a faster library for obtaining SHAP values for tree ensemble methods [17].

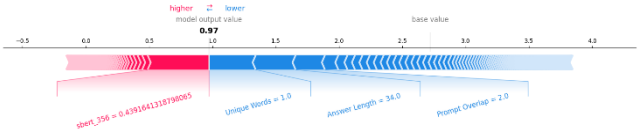

Figure 2 Local interpretation for a single prediction

From Figure 2, it seems that some of the most important interpretable features (number of unique words, answer length, and prompt overlap) have the correct effects to the prediction, they play a role in decreasing the score. However, there is one feature from the SBERT vector representation that helps the score of gibberish answers to increase, i.e. sbert_356. Most of the wrong gibberish predictions have a similar explanation to the one shown in Figure 2. SBERT vector is not interpretable, therefore we cannot explain and analyze the reason of this peculiar model behavior. We propose two solutions to this problem as follows:

1. Retrain model using gibberish data, with label score zero.

We created 200 gibberish essays and transformed them into a 780-dimension vector representation of text, and then include them in the training and testing data. Table 8 shows the performance comparison of three model training scenarios. We can observe a large improvement on the accuracy of the model to detect gibberish answers, and to punish them with the scores zero. For example, by prompt 1, the accuracy increases from 0% to 92.8% by adding only 100 gibberish data to the model training. By adding 100 more data, the accuracy only improves by a little less than 2%. After changing the training data by adding gibberish for the training process, we want to make sure that the main performance metrics (QWK score) on the original data is not being sacrificed. The results show that in all prompts, the addition of gibberish data to the training phase did not harm the performance of the models. The QWK scores decreased by a very small margin, and they still have the human-computer agreement score above the required threshold. Even in prompt 7, the final QWK score increased with the addition of gibberish to the training set.

Table 8 Accuracy of Gibberish Detection

Dataset	Training Data	Accuracy (%)	QWK Score
Prompt 1	Original	0	0.7826
	Original + 100 gibberish	92.8	0.7768
	Original + 200 gibberish	94.4	0.7683
Prompt 4	Original	18.6	0.7736
	Original + 100 gibberish	97.8	0.779
	Original + 200 gibberish	98.6	0.7749
Prompt 5	Original	3.5	0.8065
	Original + 100 gibberish	98	0.7986
	Original + 200 gibberish	98.6	0.8049
Prompt 6	Original	18.2	0.799
	Original + 100 gibberish	96	0.794
	Original + 200 gibberish	97.9	0.782
Prompt 7	Original	0	0.7771
	Original + 100 gibberish	68	0.7798
	Original + 200 gibberish	73.4	0.781

2. Use rule-based mechanism.

This is arguably a simpler solution, without the need to involve any additional model retraining process and could be a more generalizable solution. The system is configured to automatically give score zero for possible gibberish answer, this can be done automatically for example with a valid English word detection library. If none of the token in the answer is valid English vocabulary, we can consider the answer as gibberish. The main drawback is possibly the added processing time for the program to validate each word in the answer, depending on how large the vocabulary is.

4.2.3 *Paraphrased Answer*

To further evaluate the performance of the system, we investigated the reliability of the system by testing whether the model consistently gives the same score for the same answer. For this experiment, we generated paraphrased answers of all answers in the dataset. And we examine whether the model would predict the same score for each paraphrased answer. We utilized an online paraphrasing tool[4] to generate the paraphrased version of the answer.

We use Quadratic Weighted Kappa (QWK) to compute the agreement between the original test set prediction and the

[4] https://spinbot.com/

paraphrased test set prediction. Based on the scores in Table 9, it is evident that the agreements for all datasets are high. Therefore, we conclude that the models perform consistently in predicting the scores of paraphrased answers.

Table 9 Agreement of Prediction between Original and Paraphrased Answers

Dataset	QWK
Prompt 1	0.8109
Prompt 4	0.8674
Prompt 5	0.8645
Prompt 6	0.846
Prompt 7	0.9411

The highest agreement score is achieved by the model of prompt 7, which shows a near perfect agreement with QWK score of 0.9411. Although not as high as the result of prompt 7, the QWK scores in prompt 1, 4, 5, and 6 are considered as very high agreement.

5. CONCLUSION

The purpose of this research is to highlight the limitations of the current performance measurement standard for automated essay scoring. A quantitatively well-performing model with high human – automated score agreement rate, is not necessarily ready for deployment in the real-world usage. We demonstrated that such models still possess some performance concerns against varying input scenarios. We showed empirical evidence that those models have some difficulties, proven by very low accuracies, in detecting off-topic essays and gibberish. We also proposed and proved several strategies that can successfully improve the performance of the system. In another scenario, for consistency testing, the models already performed quite well for predicting paraphrased answers, judged from high agreement results with the predictions on the original answer. While we are aware that there remain more validity questions to be studied, this research can serve as additional techniques towards a better holistic evaluation framework for AES.

6. References

[1] D. M. Williamson, X. Xi and F. J. Breyer, "A Framework for Evaluation and Use of Automated Scoring," *Educational Measurement: Issues and Practice*, vol. 31, pp. 2-13, 2012.

[2] R. E. Bennett and I. I. Bejar, "Validity and automad scoring: It's not only the scoring," *Educational Measurement: Issues and Practice*, vol. 17, p. 9–17, 1998.

[3] Y. Attali and J. Burstein, "Automated essay scoring with e-rater® V. 2," *The Journal of Technology, Learning and Assessment*, vol. 4, 2006.

[4] B. E. Clauser, M. T. Kane and D. B. Swanson, "Validity Issues for Performance-Based Tests Scored With Computer-Automated Scoring Systems," *Applied Measurement in Education*, vol. 15, pp. 413-432, 2002.

[5] M. K. Enright and T. Quinlan, "Complementing human judgment of essays written by English language learners with e-rater® scoring," *Language Testing*, vol. 27, pp. 317-334, 2010.

[6] C.-F. E. Chen and W.-Y. E. C. Cheng, "Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes," *Language Learning & Technology*, vol. 12, p. 94–112, 2008.

[7] D. E. Powers, J. C. Burstein, M. Chodorow, M. E. Fowles and K. Kukich, "Stumping E-Rater: Challenging the validity of automated essay scoring," *ETS Research Report Series*, vol. 2001, p. i–44, 2001.

[8] J. Liu, Y. Xu and Y. Zhu, "Automated essay scoring based on two-stage learning," *arXiv preprint arXiv:1901.07744*, 2019.

[9] P. Phandi, K. M. A. Chai and H. T. 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*, 2015.

[10] R. A. Fisher, "On the interpretation of χ 2 from contingency tables, and the calculation of P," *Journal of the Royal Statistical Society*, vol. 85, pp. 87-94, 1922.

[11] M. Mahana, M. Johns and A. Apte, "Automated essay grading using machine learning," *Mach. Learn. Session, Stanford University*, 2012.

[12] L. S. Larkey, "Automatic essay grading using text categorization techniques," in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, 1998.

[13] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[14] J. Cohen, "Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit.," *Psychological bulletin*, vol. 70, p. 213, 1968.

[15] J. Wang and M. S. Brown, "Automated essay scoring versus human scoring: A comparative study.," *Journal of Technology, Learning, and Assessment*, vol. 6, p. n2, 2007.

[16] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett, Eds., Curran Associates, Inc., 2017, p. 4765–4774.

[17] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal and S.-I. Lee, "From local explanations to global understanding with explainable AI for trees," *Nature Machine Intelligence*, vol. 2, pp. 2522-5839, 2020.