Modeling and Analyzing Scorer Preferences in Short-Answer Math Questions

Mengxue Zhang
UMass Amherst
mengxuezhang@umass.edu

Neil Heffernan
Worcester Polytechnic Institute
nth@wpi.edu

Andrew Lan
UMass Amherst
andrewlan@cs.umass.edu

ABSTRACT

Automated scoring of student responses to open-ended questions, including short-answer questions, has great potential to scale to a large number of responses. Recent approaches for automated scoring rely on supervised learning, i.e., training classifiers or fine-tuning language models on a small number of responses with human-provided score labels. However, since scoring is a subjective process, these human scores are noisy and can be highly variable, depending on the scorer. In this paper, we investigate a collection of models that account for the individual preferences and tendencies of each human scorer in the automated scoring task. We apply these models to a short-answer math response dataset where each response is scored (often differently) by multiple different human scorers. We conduct quantitative experiments to show that our scorer models lead to improved automated scoring accuracy. We also conduct quantitative experiments and case studies to analyze the individual preferences and tendencies of scorers. We found that scorers can be grouped into several obvious clusters, with each cluster having distinct features, and analyze them in detail.

Keywords

Automated Scoring, Scorer Models, Bias

1. INTRODUCTION

Automated scoring (AS), i.e., using algorithms to automatically score student (textual) responses to open-ended questions, has significant potential to complement and scale up human scoring, especially with an ever-increasing number of students. AS algorithms are often driven by *supervised* machine learning-based algorithms and require a small number of example responses and their score labels to train on. These algorithms mostly consist of two components: a *representation* component that use either hand-crafted features [8, 17, 21, 27, 28, 37] or language models [24, 25, 34, 36, 42] to represent the (mostly textual) content in questions, student responses, and other information, e.g., rubrics [12] and a *scoring* component that use classifiers [4, 26] to predict the score of a response from its textual representation. In different subject domains, the representation component can be quite different, from hand-crafted features and neural language model-based textual embeddings in automated essay scoring (AES) [2, 27], automatic short answer grading (ASAG) [35, 47], and reading comprehension scoring [16] to specialized representations in responses where mathematical expressions are present [6, 31, 32, 40]. On the contrary, the scoring model does not vary significantly across different subject domains, often relying on simple classifiers such as logistic regression, support vector machines, random forests, or linear projection heads in neural networks [20]. We provide a more detailed discussion on related work in Section 1.2.

One key factor that limits the accuracy of AS methods is that the scoring task is a *subjective* one; human scorers are often given a set of rubrics [1] and asked to score responses according to them. However, different individuals interpret rubrics and student responses differently, leading to significant variation in their scores. For example, inter-scorer agreement can be as quite high in NAEP reading comprehension question scoring, with a quadratic weighted Kappa (QWK) score of 0.88 [16] and quite low in open-ended math question scoring, with a Kappa score of 0.083 (see Section 3.1 for details and Table 1 for a concrete example). This variation creates a *noisy labels* problem, which is a common problem in machine learning where one often needs to acquire a large number of labels via crowdsourcing [3, 18, 19]. In educational applications such as AS, this problem is even more important since the amount of labels we have access to is often small, which amplifies the negative impact of noisy score labels. Therefore, there is a significant need to analyze the preferences and tendencies of individual scorers, to not only improve AS accuracy by providing cleaner labels to train on but also understand where the variation in scores comes from and investigate whether we can reduce it.

1.1 Contributions

In this paper, we propose a collection of models for the variation in human scorers due to their individual preferences and tendencies, from simple models that use only a few parameters to account for the bias and variance of each scorer to complex models that use a different set of neural network parameters for each scorer. We ground our work in an AS task for short-answer mathematical questions and show that by adding our model to the classification component of

AS models, we can improve AS accuracy by more than 0.02 in Kappa score and 0.01 in AUC compared to AS methods that do not account for individual scorer differences. We also conduct qualitative experiments and case studies to analyze the individual preference and tendencies of scorers. We found that scorers can be grouped into several major, obvious clusters, with each cluster having distinct features, which we explain in detail. **We emphasize that our goal is NOT to develop the most accurate AS model; instead, our goal is to show that accounting for the variation across different individual scorers can potentially improve the accuracy of any AS model.**

1.2 Related work

Noisy labels. Individual scorers often exhibit different preferences and tendencies, as found in [38]. Some of our models for scorer preference and tendency are closely related to models used in peer grading [30], where students grade each others' work, which is often deployed in settings such as massive open online courses (MOOCs) where a large number of open-ended responses make it impossible for external human scorers to score all responses. Most of these models are inspired by methods in machine learning on combining labels from human labelers with different expertise in crowdsourcing contexts [41]. These models are simple and interpretable, with the most basic version involving a single bias parameter (towards certain score labels) and a single variance parameter (across different score labels) for each scorer. On the contrary, we experiment with not only these models but also more flexible but uninterpretable models, which are compatible with using pre-trained neural language models [13, 29] in the representation component of AS models.

AS and math AS. The majority of existing ASAG and AES methods focus on non-mathematical domains [7, 9, 11, 21, 27, 37, 39]. Recently, some AS methods are developed for specific domains that contain non-textual symbols, e.g., Chemistry, Computer Science, and Physics, which exist in student responses in addition to text, achieving higher and higher AS accuracy [5, 14, 23, 33, 34]. Our work is grounded in the short-answer math question scoring setting, which is studied in prior works [5, 6, 32, 46]. The key technical challenge here is that mathematical expressions that are often contained in open-ended student responses can be difficult to parse and understand in the representation component. The authors of [5] proposed a scoring approach for short-answer math questions using sentence-BERT (SBERT)-based representation of student responses and simply ignored mathematical expressions. The authors of [6] developed an additional set of features specifically designed for mathematical expressions and used them in conjunction with the SBERT representations as input to the scoring component. The authors of [32] fine-tuned a language model, BERT [13], further pre-trained on math textbooks, as the representation component; however, this representation was found to not be highly effective in later works [46]. The authors of [46] used a sophisticated in-context meta-training approach for automated scoring by inputting not only the response that needs to be scored but also scored examples to a language model, enabling the language model to learn from examples, which results in significant improvement in AS accuracy and especially generalizability to previously unseen questions.

Another line of related work is about fairness in educational data analysis since scorer preference can be classified as a form of individual bias. Researchers have proposed methods to incorporate constraints and regularization into predictive models to improve parity and mitigate fairness issues [10, 44, 45]. On the contrary, our work does not attempt at reducing biases; our focus is only on identifying a specific source of bias, individual scorer bias, in the AS context. Therefore, the only approach we use to mitigate biases is to leverage scorer identification information and investigate its impact on AS accuracy, following prior work on using this information in predictive models [43].

2. MODEL

We now detail our models for individual scorer preference and tendency in AS tasks. For all models, we use a BERT model [13] as the corresponding representation component of the AS model, which has been shown to perform well and reach state-of-the-art performance on the short math answer AS task with an appropriate input structure [46]. Let us denote each question-response pair that needs to be scored as q_i, while the j-th scorer assigns a score $y_{i,j} \in \{1, \ldots, C\}$ where C denotes the number of possible score categories.

2.1 Baseline

Our base AS model is one that directly uses the output `[CLS]` embedding of BERT as the representation of the question-response pair $\mathbf{r}_i \in \mathbb{R}^D$, where $D = 768$ is the dimension of the embedding. We also use a linear classification head (omitting the bias terms for simplicity) with softmax output [20] for all score categories, i.e.,

$$p(y_{i,j} = c) \propto e^{(\mathbf{w}_c^T \mathbf{r}_i) + b_c},$$

where $\mathbf{w}_c$ denotes the D-dimensional parameter for each score category and $b_c \in \mathbb{R}$ is the universal bias toward each score category.

2.2 Scalar bias and variance with scorer embeddings

The first version of our model is the simplest and most interpretable: we use a scalar temperature, i.e., variance parameter for each scorer, and a scalar offset, i.e., bias parameter on each score category for each scorer, i.e.,

$$p(y_{i,j} = c) \propto e^{\alpha_j (\mathbf{w}_c^T \mathbf{r}_i + b_{c,j})}, \tag{1}$$

where $\alpha_t > 0$ is the "temperature" parameter that controls the scorer's uncertainty across categories: larger values indicate higher concentrations of the probability mass around the most likely score category, which corresponds to more consistent scoring behavior. $b_{c,j} \in \mathbb{R}$ is the "offset" parameter that controls the scorer's bias towards each score category: larger values indicate a higher probability of selecting some score category, which corresponds to more positive/negative scoring preferences.

In practice, we found that parameterizing biases with a set of *scorer embeddings* lead to better performance than simply parameterizing the biases as learnable scalars. Specifically, we introduce a high-dimensional embedding for each scorer,

Table 1: Example questions, student responses, and scores. Some scorers assign highly different scores to similar responses.

question_id	*question_body*	*response*	*scorer_id*	*score*
43737	Chris spent \$9 of the \$12 he was given for his birthday. His sister Jessie says that he has spent exactly 0.75 of the money. Chris wonders if Jessie is correct. Explain your reasoning.	Jessie is correct because 0.75 in fraction form is 3/4. 9 is 3/4 of 12, so she is right.	1	4
		Jessie is wrong.	1	0
		she is correct	1	1
		Jessie is incorrect.	2	4
		Jessie is right because if you divide 12 by 9 you get 0.75.	2	2

$\mathbf{e}_j \in \mathbb{R}^D$, and use a $C \times D$ matrix $\mathbf{S}$ to map it to a low-dimensional vector that corresponds to the bias terms for all score categories. This advantage is likely due to the fact that more model parameters make the model more flexible and more capable in capturing detailed nuances in scorer preferences and tendencies.

2.3 Content-driven scorer bias and variance

In the models above, we have set the scorer biases and variances to be scorer-dependent but not question/response-dependent, i.e., the bias and variance of a scorer stay the same across all question-response pairs. However, in practice, it is possible that these parameters depend on the actual textual content of the question and the student's response. Therefore, we extend the scorer model of Eq in Sec 2.2 into

$$\begin{aligned} &\mathbf{b}_{i,j} = f_b(\mathbf{r}_i, \mathbf{e}_j), \quad \alpha_t = f_\alpha(\mathbf{r}_i, \mathbf{e}_j), \\ &\text{where} \quad f_b(\mathbf{r}_i, \mathbf{e}_j) = \mathbf{r}_i^T \mathbf{A}_b \mathbf{e}_j, \quad f_\alpha(\mathbf{r}_i, \mathbf{e}_j) = \mathbf{r}_i^T \mathbf{A}_\alpha \mathbf{e}_j, \end{aligned}$$

where the bias $\mathbf{b}_{i,j}$ is now a $C \times 1$ vector of biases across all score categories and both question-response pair (i)-dependent and scorer (j)-dependent. f_b and f_α denote functions that map the textual representation of the question-response pair and the scorer embedding to the bias and variance parameters, which can be implemented in any way (from simple linear models to complex neural networks). In this work, we found that using bi-linear functions of the question-response pair representation $\mathbf{r}_i$ and the scorer embedding $\mathbf{e}_j$, using two $D \times D$ matrices $\mathbf{A}_b$ and $\mathbf{A}_\alpha$, results in the best AS accuracy.

2.4 Training with different losses

We explore using various different loss functions as objectives to train our AS model, which we detail below.

2.4.1 Cross-entropy

Since the AS task corresponds to a multi-category classification problem, the standard loss function that we minimize is the cross-entropy (CE) loss [20], summed over all question-response pairs and scorers, as

$$\mathcal{L}_{\text{CE}} = -\sum_{i,j} \sum_{c=1}^{C} \mathbf{1}_{y_{i,j}=c} \log p(y_{i,j} = c)$$

where $\mathbf{1}_{y_{i,j}=c}$ is the indicator function that is non-zero only if $y_{i,j} = c$. In other words, we are minimizing the negative log-likelihood of the actual score category among the category probabilities predicted by the AS model, $p(y_{i,j} = c)$.

2.4.2 Ordinal log loss

One obvious limitation of the standard CE loss is that it assumes that the categories are unordered, which works for many applications. Therefore, it penalizes all misclassifications equally. However, for AS, the score categories are naturally ordered, which means that score classification errors are not equal: if the actual score is 1 out of 5, then a misclassified score of 2 is better than 5, but they are weighted equally in the standard CE loss. Therefore, we follow the approach outlined in [15] and use an ordinal log loss (OLL), which we define as

$$\mathcal{L}_{\text{OLL}} = -\sum_{i,j} \sum_{c=1}^{C} |y_{i,j} - c| \log(1 - p(y_{i,j} = c)),$$

where we weight the misclassification likelihood, i.e., $-\log(1 - p(y_{i,j} = c))$, according to the difference between the actual score, $y_{i,j}$, and the predicted score, c. In the aforementioned example, this objective function would increase the penalty of a misclassified score of 5 by four times compared to a misclassified score of 2 when the actual score is 1, which effectively leverages the ordered nature of the score categories.

2.4.3 Mean squared error

Since the score categories are integers and can be treated as numerical values, one simple alternative to the CE loss is the mean squared error (MSE) loss, i.e.,

$$\mathcal{L}_{\text{MSE}} = \sum_{i,j} (y_{i,j} - \sum_{c=1}^{C} p(y_{i,j} = c)c)^2, \tag{2}$$

where we simply square the difference between the actual score and the expected (i.e., weighted average) score under the category probabilities predicted by the AS model.

3. QUANTITATIVE EXPERIMENTS

We now detail experiments that we conducted to validate the different scoring components of AS models and loss functions that capture scorer preferences and tendencies. Section 3.1 discusses details on the real-world student response dataset we use and the pre-processing steps. Section 3.2 details the evaluation metrics we use in our experiments. Section 3.3 details our experimental setting, and Section 3.4 details the experimental results and corresponding discussion.

Table 2: Comparing different scorer models on short-answer math scoring. The combination of content-driven scorer bias and temperature with the OLL loss outperforms other scorer models and training losses.

Bias (b) & Temperature (α)	Loss Function	AUC	RMSE	Kappa
Universal (b_c, $\alpha = 1$)	CE	0.765 ± 0.003	0.954 ± 0.014	0.614 ± 0.009
Universal (b_c, $\alpha = 1$)	MSE	0.764 ± 0.003	0.946 ± 0.018	0.615 ± 0.008
Universal (b_c, $\alpha = 1$)	OLL	0.768 ± 0.003	0.944 ± 0.015	0.617 ± 0.006
Scorer-specific ($b_{c,j}$, α_j)	CE	0.768 ± 0.005	0.928 ± 0.023	0.628 ± 0.006
Scorer-specific ($b_{c,j}$, α_j)	MSE	0.772 ± 0.005	0.926 ± 0.025	0.625 ± 0.006
Scorer-specific ($b_{c,j}$, α_j)	OLL	0.770 ± 0.003	**0.916 ± 0.013**	0.628 ± 0.004
Content-driven ($b_{c,j}(\mathbf{r}_i)$, $\alpha_j(\mathbf{r}_i)$)	CE	0.772 ± 0.003	0.923 ± 0.016	0.631 ± 0.006
Content-driven ($b_{c,j}(\mathbf{r}_i)$, $\alpha_j(\mathbf{r}_i)$)	MSE	0.774 ± 0.004	0.922 ± 0.021	0.629 ± 0.005
Content-driven ($b_{c,j}(\mathbf{r}_i)$, $\alpha_j(\mathbf{r}_i)$)	OLL	**0.779 ± 0.004**	0.924 ± 0.013	**0.641 ± 0.005**

3.1 Dataset

We use data collected from an online learning platform that has been used in prior work [5, 14], which contains student responses to open-ended, short-answer math questions, together with scores assigned by human scores. There are a total of 141,612 total student responses made by $25,069$ students to $2,042$ questions, with 891 different teachers being scorers. The set of possible score categories is from 0 (no credit) to 4 (full credit). The dataset mainly contains math word problems, where the answer could be mathematical such as numbers and equations or textual explanations, sometimes in the format of images.

We found that different scorers sometimes assign very different scores to the same response, which motivated this work. As an example, we analyze question-response pairs that are scored by more than one scorer and evaluate the Kappa score between these scorers. The *human* Kappa score is only 0.083, which means a minimal agreement between different scorers. Although there are only 523 such pairs, this case study still shows that even for the same exact response, scorers have highly different individual preferences and tendencies and may assign them highly different scores.

We also perform a series of pre-processing steps to the original dataset. For example, since some of the scorers do not score many responses, e.g., less than 100, there may not be enough information on these scorers for us to model their behavior. Therefore, we remove these scores from the dataset, which results in 203 scorers, $1,273$ questions, and $118,079$ responses. The average score is 3.152 ± 1.417. Table 1 shows some examples of data points of this dataset; each data point consists of the question statement, the student's response, the scorer's ID, and the score.

3.2 Metrics

We utilize three standard evaluation metrics for integer-valued scores that have been commonly used in the AS task [5, 14]. First, the area under the receiver operating characteristic curve (**AUC**) metric, which we adapt to the multi-category classification problem by averaging the AUC numbers over each possible score category and treating them as separate binary classification problems, following [22]. Second, we use the root mean squared error (**RMSE**) metric, which simply treats the integer-valued score categories as numbers. Third and most importantly, we use the multi-class Cohen's **Kappa** metric for ordered categories, which is often used to evaluate AS methods [1].

3.3 Experimental setting

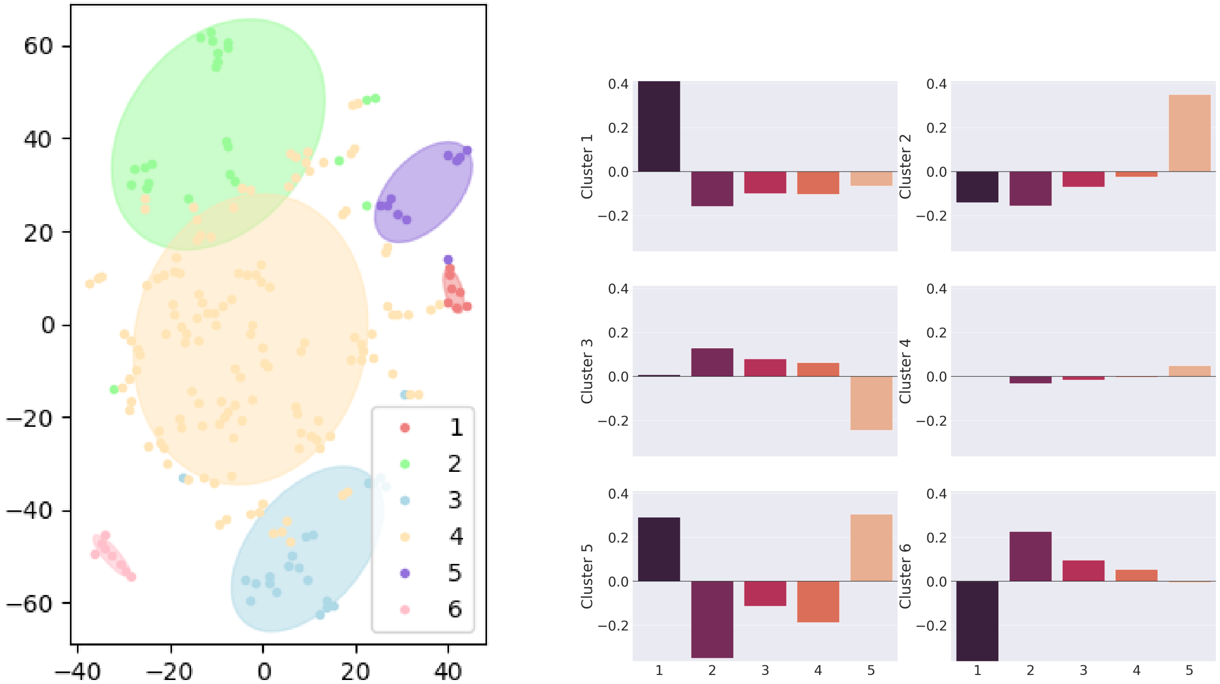
In the quantitative experiment, we focus on studying whether adding scorer information leads to improved AS accuracy. Therefore, when we are splitting a dataset into training, validation, and test sets, we ensure that every scorer is included in the training set. We divide the data points (question-response pairs, scorer ID, score) into 10 equally-sized folds for cross-validation. During training, we use 8 folds as training data, 1 fold for validation for model selection, and 1 fold for the final testing.

For a fair comparison, every model uses BERT[1] as the pre-trained model for question-response pair representation, which has been shown to result in state-of-the-art AS accuracy in prior work [46]. We emphasize that our work on **scorer models** can be added on top of **any** AS method for response representation; applying these models on other AS methods is left for future work. We use the Adam optimizer, a batch size of 16, and a learning rate of $1e-5$ for 10 training epochs on an NVIDIA RTX8000 GPU. We do not perform any hyper-parameter tuning and simply use the default settings.

3.4 Results and discussion

Table 2 shows the mean and standard deviation of each scorer model trained under each loss function. We see that generally, models with content-driven scorer biases and variances outperform scorer-specific biases and variances, which outperform the base AS model that treats each scorer the same with universal values for bias and variance. The improvement in AS accuracy is significant, up to about 0.02 in the most important metric–Kappa, for the content-driven biases and variances over the standard AS approach of not using scorer information. This observation validates the need to account for individual scorer preferences and tendencies in the highly subjective AS task. Meanwhile, since the content-driven scorer bias and variance models outperform the scorer-specific bias and variance models, we can conclude that the content of the question and response does play an important role in scorer preference.

[1] https://huggingface.co/bert-base-uncased



(a) 2-D visualization of the learned scorer embedding space

(b) Bias for each score category

Figure 1: Visualization of clustering result on scorer embedding learned via scorer-specific model. The left figure shows the 2-D visualization of scorer embedding space, and the right figure shows the average bias for each cluster

We also observe that training scorer models with the OLL loss outperform the other losse, while training with the MSE loss does not even lead to the best results on the RMSE metric. This observation suggests that taking into account the ordered nature of score categories instead of treating them as parallel ones is important to the AS task.

4. QUALITATIVE ANALYSIS

Despite the content-driven model delivering the highest AUC and Kappa results, the complexity of the information contained in its embedding space renders it difficult to interpret. Consequently, we have elected to concentrate on examining the scorer-specific model (detailed in Sec. 2.2).

4.1 Visualization of scorer embedding

Figure 1 shows a 2-D visualization of the learned scorer embedding space; We see that there are obvious clusters among all scorers. We then fit the learned scorer embeddings under a mixture-of-Gaussian model via the expectation-maximization (EM) algorithm with 6 clusters. The subfigures to each side of the main plot shows each cluster's average bias towards each score category, which are 0, 1, 2, 3, and 4 from left to right.

4.2 Features analysis based on each cluster

Cluster 1 shows a negative scoring profile, with a strong, positive bias towards the lowest score category 0 (positive $b_{c,j}$ values) and small, negative biases against higher scores, 1, 2, and 3 (negative $b_{c,j}$ values). These scorers assign 0 scores much more often than other score categories, compared to other scorers. The average score across question-response pairs is the lowest for this cluster, at 1.69. Meanwhile, this cluster has a relatively high score variance of 1.69, meaning that these scorers tend to have inconsistent behavior and assign a wide variety of score labels.

Cluster 2 shows a positive scoring profile, with a strong, positive bias towards the highest score, 4, and moderate negative biases against other scores. These scorers prefer to assign scores that are overwhelmingly higher compared to other scorers. The average score across question-response pairs is the lowest for this cluster, at 3.45. Meanwhile, this cluster has a relatively low score variance of 0.92, meaning that these scorers are consistent in scoring responses higher than other scorers.

Cluster 3 shows a conservative scoring profile, with small, positive biases towards the middling scores 1, 2, and 3 and a strong, negative bias against the top score 4. The average score across question-response pairs is 2.41 for this cluster with a variance of 1.4, which is high considering that scorers in this cluster rarely use the top score category, indicating that their scoring behavior is not highly consistent.

Cluster 4 shows an unbiased scoring profile, with a low bias towards or against any score category, with a slight preference for the top score category, 4. This cluster contains almost half of the scorers, which means that the majority of scorers are reliable (their scores depend mostly on the actual quality of the response, i.e., the $\mathbf{w}_c^T \mathbf{r}_i$ term of Eq in 2.2 rather than the bias term.

Cluster 5 shows a polarizing scoring profile, with strong, pos-

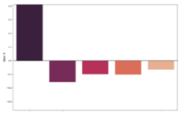
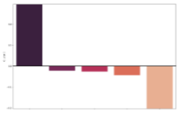
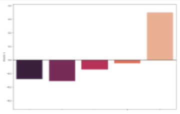
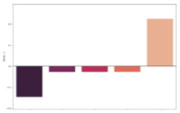
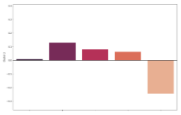
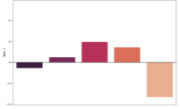
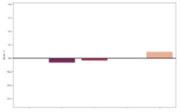
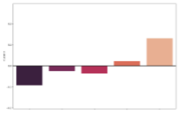
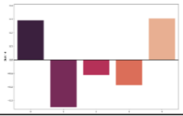
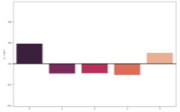
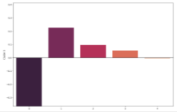
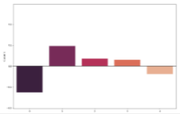
Cluster	Bias	Observed scoring distribution (normalized)	Temperature	Score	Response features		
					math tok (%)	img (%)	length
1			1.013	1.685 ± 1.644	29.13	0.101	23.06
2			1.034	3.451 ± 0.919	32.12	1.286	24.40
3			0.996	2.415 ± 1.400	23.51	1.311	36.16
4			1.033	3.074 ± 0.991	29.48	0.304	21.94
5			1.026	2.558 ± 1.806	45.18	5.271	14.35
6			1.007	2.714 ± 1.331	33.83	1.403	13.34

Figure 2: Detailed biases and variance (inverse of temperature) for each scorer profile, their observed scoring distributions, and average response features. We normalize the observed scoring distributions to zero-mean, which makes them easier to visually compare against the learned biases. *math tok (%)* is the percentage of math tokens in the response. *img (%)* is the percentage of images in the response. *length* is the number of word tokens in the response.

itive biases toward both the lowest score, 0, and the highest score, 4, while having strong, negative biases against score categories in between. Scorers in this cluster often score a response as all or nothing while using the intermediate score values sparingly. The average score across question-response pairs is 2.55 for this cluster with a variance of 1.81, the highest among all clusters, which agrees with our observation that these scorers are highly polarizing and rarely judge any response to be partially correct.

Cluster 6 shows a lenient scoring profile, with a strong, negative bias against the lowest score, 0, and a moderate, positive bias towards the next score, 1, with minimal bias across higher score categories. Scorers in this cluster tend to award students a single point for an incorrect response instead of no points at all. The average score across question-response pairs is 2.71 for this cluster with a middling variance of 1.33.

5. CONCLUSIONS AND FUTURE WORK

In this paper, We created models to account for individual scorer preferences and tendencies in short-answer math response automated scoring. Our models differ from previous work by focusing on capturing the subjective nature of scoring rather than textual content. Our models range from simple to complex, with some using bias and variance as a function of the question and response. Our experiments on a dataset with low inter-rater agreement showed that accounting for scorer preferences and tendencies improved performance by more than 0.02 in the Kappa metric. Qualitative analysis showed obvious patterns among scorers, some with biases towards certain scores. Scorer-specific settings can model scorer grading behavior very well. In other words, the scorer's grading behavior is highly controllable, and the scorer's grading behavior representation is also well-represented in the hidden space. One practical extension could be adjusting the learned scorer bias by using a different type of scorer embedding to control model grading in a different scorer style. Future work can address limitations in our analysis. Our dataset only provides scorer IDs, lacking gender, race, or location. Investigating biases with this additional information is crucial, including how teacher-student relationships or shared demographics impact biases. Our analysis also did not consider student demographic information, which is important for fairness studies. Additionally, our scorer models were only validated with a BERT-based textual representation model, so further testing is needed to determine their adaptability to traditional, feature-based automated scoring methods.

6. ACKNOWLEDGEMENTS

The authors thank the NSF (under grants 1917713, 2118706, 2202506, 2215193) for partially supporting this work.

7. REFERENCES

[1] The ed.gov national assessment of educational progress (naep) automated scoring challenge. Online: https://github.com/NAEP-AS-Challenge/info, 2021.

[2] The hewlett foundation: Automated essay scoring. Online: https://www.kaggle.com/c/asap-aes, 2021.

[3] G. Algan and I. Ulusoy. Image classification with deep learning in the presence of noisy labels: A survey. *arXiv preprint arXiv:1912.05170*, 2019.

[4] Y. Attali and J. Burstein. Automated essay scoring with e-rater® v. 2. *The Journal of Technology, Learning and Assessment*, 4:3–10, 2006.

[5] S. Baral, A. F. Botelho, J. A. Erickson, P. Benachamardi, and N. T. Heffernan. Improving automated scoring of student open responses in mathematics. *International Educational Data Mining Society*, 2021.

[6] S. Baral, K. Seetharaman, A. F. Botelho, A. Wang, G. Heineman, and N. T. Heffernan. Enhancing auto-scoring of student open responses in the presence of mathematical terms and expressions. In *International Conference on Artificial Intelligence in Education*, pages 685–690. Springer, 2022.

[7] S. Basu, C. Jacobs, and L. Vanderwende. Powergrading: a clustering approach to amplify human effort for short answer grading. *Transactions of the Association for Computational Linguistics*, 1:391–402, 2013.

[8] J. Burstein. The e-rater® scoring engine: Automated essay scoring with natural language processing. 2003.

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

[10] Y.-W. Chu, S. Hosseinalipour, E. Tenorio, L. Cruz, K. Douglas, A. Lan, and C. Brinton. Mitigating biases in student performance prediction via attention-based personalized federated learning. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 3033–3042, 2022.

[11] A. Condor, M. Litster, and Z. Pardos. Automatic short answer grading with sbert on out-of-sample questions. *International Educational Data Mining Society*, 2021.

[12] A. Condor, Z. Pardos, and M. Linn. Representing scoring rubrics as graphs for automatic short answer grading. In *International Conference on Artificial Intelligence in Education*, pages 354–365. Springer, 2022.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[14] J. A. Erickson, A. F. Botelho, S. McAteer, A. Varatharaj, and N. T. Heffernan. The automated grading of student open responses in mathematics. In *Proceedings of the International Conference on Learning Analytics & Knowledge*, page 615–624, 2020.

[15] F. C. et al. A simple log-based loss function for ordinal text classification. online: https://openreview.net/pdf?id=khB9is39GvL, 2022.

[16] N. Fernandez, A. Ghosh, N. Liu, Z. Wang, B. Choffin, R. G. Baraniuk, and A. S. Lan. Automated scoring for reading comprehension via in-context bert tuning. In *International Conference on Artificial Intelligence in Education*, page 0, 2022.

[17] P. W. Foltz, D. Laham, and T. K. Landauer. The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, 1(2):939–944, 1999.

[18] B. Frénay and M. Verleysen. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, 25(5):845–869, 2013.

[19] Github. Awesome-learning-with-label-noise, 2020.

[20] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[21] A. C. Graesser, D. S. McNamara, M. M. Louwerse, and Z. Cai. Coh-metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, 36(2):193–202, 2004.

[22] D. J. Hand and R. J. Till. A simple generalisation of the area under the roc curve for multiple class classification problems. *Machine learning*, 45(2):171–186, 2001.

[23] C. Leacock and M. Chodorow. C-rater: Automated scoring of short-answer questions. *Computers and the Humanities*, 37(4):389–405, 2003.

[24] S. Lottridge, B. Godek, A. Jafari, and M. Patel. Comparing the robustness of deep learning and classical automated scoring approaches to gaming strategies. Technical report, Cambium Assessment Inc., 2021.

[25] E. Mayfield and A. W. Black. Should you fine-tune BERT for automated essay scoring? In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 151–162, Seattle, WA, USA → Online, July 2020. Association for Computational Linguistics.

[26] D. S. McNamara, S. A. Crossley, R. D. Roscoe, L. K. Allen, and J. Dai. A hierarchical classification approach to automated essay scoring. *Assessing Writing*, 23:35–59, 2015.

[27] E. B. Page. The imminence of grading essays by computer. *The Phi Delta Kappan*, 47(5):238–243, 1966.

[28] I. Persing and V. Ng. Modeling prompt adherence in student essays. In *52nd Annual Meeting of the Association for Computational Linguistics*, pages 1534–1543, 2014.

[29] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9–20, 2019.

[30] K. Raman and T. Joachims. Methods for ordinal peer grading. In *Proc. 20th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, pages 1037–1046, Aug. 2014.

[31] A. Scarlatos and A. Lan. Tree-based representation and generation of natural and mathematical language, 2023.

[32] J. T. Shen, M. Yamashita, E. Prihar, N. Heffernan, X. Wu, B. Graff, and D. Lee. Mathbert: A pre-trained language model for general nlp tasks in mathematics education. *arXiv preprint arXiv:2106.07340*, 2021.

[33] S. Srikant and V. Aggarwal. A system to grade

computer programming skills using machine learning. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1887–1896, 2014.

[34] K. Taghipour and H. T. Ng. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891, Austin, Texas, Nov. 2016. Association for Computational Linguistics.

[35] M. Uto and Y. Uchida. Automated short-answer grading using deep neural networks and item response theory. In *International Conference on Artificial Intelligence in Education*, pages 334–339, 2020.

[36] M. Uto, Y. Xie, and M. Ueno. Neural automated essay scoring incorporating handcrafted features. In *28th Conference on Computational Linguistics*, pages 6077–6088, 2020.

[37] S. Valenti, F. Neri, and A. Cucchiarelli. An overview of current research on automated essay grading. *Journal of Information Technology Education: Research*, 2(1):319–330, 2003.

[38] J. Z. Wang, A. S. Lan, P. J. Grimaldi, and R. G. Baraniuk. A latent factor model for instructor content preference analysis. *International Educational Data Mining Society*, 2017.

[39] Z. Wang, A. Lan, A. Waters, P. Grimaldi, and R. Baraniuk. A meta-learning augmented bidirectional transformer model for automatic short answer grading. In *Proc. 12th Int. Conf. Educ. Data Mining (EDM)*, pages 1–4, 2019.

[40] Z. Wang, M. Zhang, R. G. Baraniuk, and A. S. Lan. Scientific formula retrieval via tree embeddings. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 1493–1503. IEEE, 2021.

[41] J. Whitehill, P. Ruvolo, T. Wu, J. Bergsma, and J. R. Movellan. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. Proceedings of a meeting held 7-10 December 2009, Vancouver, British Columbia, Canada*, pages 2035–2043. Curran Associates, Inc., 2009.

[42] R. Yang, J. Cao, Z. Wen, Y. Wu, and X. He. Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. *Findings of the Association for Computational Linguistics: EMNLP*, 2020:1560–1569, 2020.

[43] R. Yu, H. Lee, and R. F. Kizilcec. Should college dropout prediction models include protected attributes? In *8th ACM Conference on Learning@ Scale*, pages 91–100, 2021.

[44] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *International Conference on World Wide Web*, pages 1171–1180, 2017.

[45] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333, 2013.

[46] M. Zhang, S. Baral, N. Heffernan, and A. Lan. Automatic short math answer grading via in-context meta-learning. *arXiv preprint arXiv:2205.15219*, 2022.

[47] Y. Zhang, R. Shah, and M. Chi. Deep learning+ student modeling+ clustering: A recipe for effective automatic short answer grading. In *International Conference on Educational Data Mining*, page 562, 2016.

APPENDIX

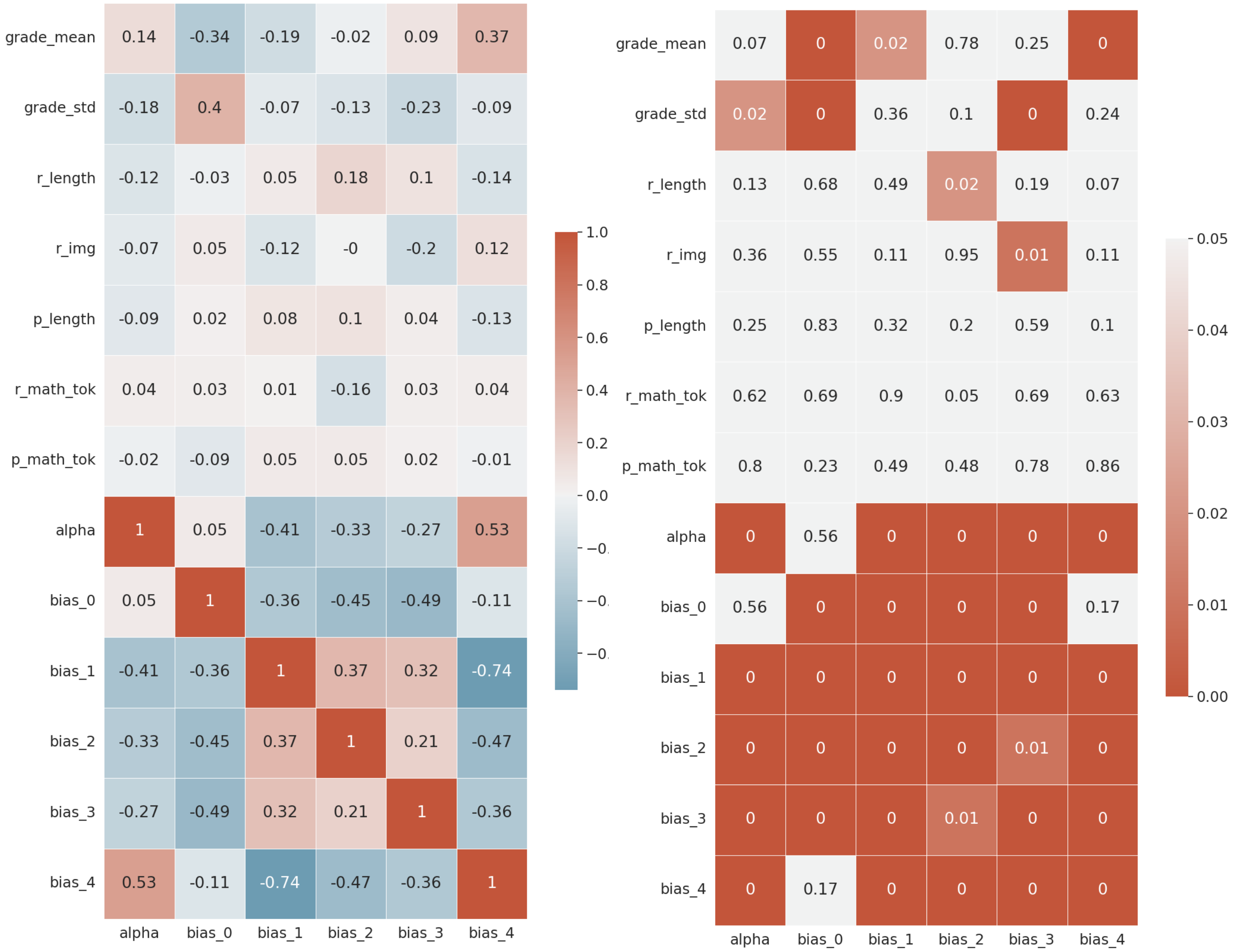
A. CORRELATION ANALYSIS

In Figure 2, we see that the learned scorer biases for each cluster are highly correlated with the observed score distribution across score categories. However, it is not obvious how the variance, i.e., the inverse of the temperature parameter (α), correlates with other model parameters and response features. Therefore, we calculate the correlation coefficient (left) and the corresponding p-value (right) between each pair of model parameters and response features and show them in Figure 3. In the left part of the figure, we see that α positively correlates with the mean of scores and negatively correlates with the standard deviation of scores. In the right part of the figure, we see that α is significantly correlated with the standard deviation of scores, which is expected since this temperature parameter is designed to capture the variation in score category assignments. We also see that α is also significantly correlated with the bias terms of each score category, with a positive correlation with the bias for score category 4 and a negative correlation with the bias for other categories.

For the bias terms, we see that most of the biases are significantly correlated with the mean and standard deviation of scores, but less correlated with question-response pair features. This observation suggests that the bias terms mainly depend on scorer behavior rather than the question-response pair, which is what the model intended to do; the question-response pair is captured by the $\mathbf{w}_c^T \mathbf{r}_i$ term of Eq in 2.2. The bias for score category 2, however, does not significantly correlate with the mean and standard deviation of scores but significantly correlates with other question-response pair features. One possible explanation is that since this score category is in the middle of all scores, scorers do not show any bias towards or against this score category and can solely rely on the actual content of the question and response. , for example, the length of the response which might show that bias 2 does not accurately represent scorer grading behavior.

B. CASE STUDY: SAME SCORER, DIFFERENT RESPONSES

Table 3 shows several examples of different questions and responses and corresponding scores for a single scorer, with the actual score, biases calculated from the content-driven scorer bias and variance model, and predicted scores for different models. The overall bias for this scorer is $[-0.043, -0.36, -0.212, 0.061, 0.439]$ across all score categories, which indicates that this scorer prefers to assign high scores (especially the full score 4) but often assigns low scores except the lowest score (0). Overall, we see that if we do not include biases in the AS model (the sixth column), the AS model tends to predict middling scores, while the human scorer tends to give students full credit (4). For Question 2, this example shows that the content-driven scorer bias model captures nuanced scorer preference: for the meaningless response “idk”, which should have a score of 0, the scorer has a strong preference towards giving it a high score (3). This bias only appears for seemingly meaningless responses but not overall (overall bias towards score category 3 is minimal at 0.061). Therefore, we see that the scorer-specific model cannot capture this information since its biases and variance are global across all question-response pairs for this scorer. As a result, content-driven scorer models are more flexible in handling these cases compared to other models, which is also evident in the quantitative results in Table 2 that this model achieves the highest overall AS accuracy.



(a) Correlation Coefficient Matrix

(b) P-values Matrix

Figure 3: Correlation coefficients and corresponding p-values across the bias, variance terms, and response features for the scorer-specific bias and variance models.

Table 3: Examples of student response and scores for a single scorer with biases $-0.043, -0.36, 0.061, -0.212, 0.439$ for all score categories. Notice that the no-bias prediction is the prediction of the content-driven model that does not scale with bias.

Question id	Response	Actual score	Content-driven prediction	Scorer-specific prediction	No bias prediction	Content-driven scorer bias
1	The graph was touching the origin, but it didn't have a straight line	4	4	4	3	[-0.61, -1.29, 0.04, -0.33, 1.33]
2	It meets the origin and it goes perfectly diagonal.	3	4	4	3	[-0.26, -1.80, -0.57, -0.39, 2.09]
	Because it's a straight line that goes through the origin	4	4	4	3	[0.13, -1.53, -0.76, -0.47, 1.71]
	its proportional because it has a straight line and starts at the bottom.	3	3	4	2	[1.19, -0.20, -3.18, 1.24, 1.21]
	idk	3	3	0	0	[-6.26, -0.09, 2.76, 4.56, 1.50]