Leveraging AI Graders for Missing Score Imputation to Achieve Accurate Ability Estimation in Constructed-Response Tests

Masaki Uto[1,*], Yuma Ito[1]

[1]The University of Electro-Communications, 1-5-1 Chofugaoka, Chofu, Tokyo, Japan

Abstract

Evaluating the abilities of learners is a fundamental objective in the field of education. In particular, there is an increasing need to assess higher-order abilities such as expressive skills and logical thinking. Constructed-response tests such as short-answer and essay-based questions have become widely used as a method to meet this demand. Although these tests are effective, they require substantial manual grading, making them both labor-intensive and costly. Item response theory (IRT) provides a promising solution by enabling the estimation of ability from incomplete score data, where human raters grade only a subset of answers provided by learners across multiple test items. However, the accuracy of ability estimation declines as the proportion of missing scores increases. Although data augmentation techniques for imputing missing scores have been explored in order to address this limitation, they often struggle with inaccuracy for sparse or heterogeneous data. To overcome these challenges, this study proposes a novel method for imputing missing scores by leveraging automated scoring technologies for accurate IRT-based ability estimation. The proposed method achieves high accuracy in ability estimation while markedly reducing manual grading workload.

Keywords

Constructed-response tests, educational measurement, item response theory, automated scoring, data augmentation, large language models

1. Introduction

Evaluating the abilities of learners is a critical component of various educational assessments, including entrance and qualification exams, as well as in-class assessments. Ability estimation is also essential for educational applications such as personalized learning support systems, including intelligent tutoring, and adaptive learning platforms because they generally require ability estimation to provide optimal recommendations for learning strategies, content, and other interventions tailored to the ability of each learner [1, 2, 3, 4].

Objective tests, typically consisting of multiple-choice questions, have been widely adopted as a popular approach for ability estimation in educational settings owing to their scalability and ease of implementation. However, modern education increasingly emphasizes the importance of 21st-century skills such as expressive abilities and critical thinking [5, 6, 7, 8, 9]. To effectively assess such abilities, constructed-response tests, including short-answer and essay-type questions, have gained increasing attention. These tests, however, necessitate substantial manual grading, which makes them both labor-intensive and costly [10, 11, 12, 13].

Item response theory (IRT) [14], a statistical method well-established in the fields of educational and psychological measurement, offers a promising solution for estimating ability using incomplete score data, where human raters grade only a subset of learner answers across multiple test items, as exemplified in Table 1. IRT typically estimates learner ability by maximizing the likelihood of observed scores based on IRT models, which define the probability of score observations as a function of learner ability and item characteristic parameters. This allows IRT to be easily applied to incomplete score data

EvalLAC'25: 2nd Workshop on Automatic Evaluation of Learning and Assessment Content, July 26, 2025, Palermo, Italy
*Corresponding author.
uto@ai.lab.uec.ac.jp (M. Uto)
https://sites.google.com/site/utomasakieng/ (M. Uto)
0000-0002-9330-5158 (M. Uto)

Table 1
Examples of incomplete data, where cells marked with a ✓indicate scored answers and blank cells indicate missing scores.

	Learner								
	1	2	3	4	5	6	7	8	⋯
Test item 1	✓		✓	✓		✓	✓		⋯
Test item 2	✓	✓		✓	✓		✓	✓	⋯
Test item 3		✓	✓		✓	✓		✓	⋯

	Learner								
	1	2	3	4	5	6	7	8	⋯
Test item 1	✓		✓	✓			✓		⋯
Test item 2	✓	✓			✓		✓	✓	⋯
Test item 3		✓	✓			✓		✓	⋯

by calculating the likelihood while excluding missing scores [15, 16, 17]. This feature is particularly advantageous for achieving ability estimation while reducing manual grading workload. However, the accuracy of ability estimation decreases as the proportion of missing scores increases.

A possible strategy for addressing this problem is data augmentation through imputation of missing scores [18, 19]. Simple methods such as mean or mode imputation [20] are widely recognized but fail to effectively capture the underlying complex patterns within data. More advanced approaches such as statistical model-based and machine learning-based methods [21, 22, 23, 24, 25, 26] aim to model such underlying patterns to predict missing scores for more accurate imputation. However, when the assumed model does not fit the target data or when the missing rate is very high, the resulting imputations often lack accuracy. Moreover, conventional imputation methods generally rely on the assumption that all data, including both observed and missing data, follow a single underlying pattern modeled as a specific data-generation process, while real-world constructed-response tests frequently violate this assumption. For instance, proficient learners could achieve lower scores due to inattention, while less proficient learners could obtain higher scores due to compatibility between the learner and the test item or other chance factors. These issues suggest that conventional imputation methods may lack the robustness for sparsity and heterogeneous data, making them unsuitable for achieving accurate ability estimation while substantially reducing manual scoring workload.

To overcome these limitations, this study proposes a novel method for imputing missing scores by leveraging automated scoring technologies [11, 27, 28, 29, 30, 31] for accurate IRT-based ability estimation in constructed-response tests. Specifically, the approach begins by developing neural automated scoring models trained on a subset of manually scored responses for each test item, or by employing zero-shot scoring models using large language models (LLMs). These models are then used to predict missing scores, generating a complete dataset. The augmented dataset is subsequently used to estimate learner ability by using IRT models. The proposed method offers several key advantages:

1. More robust imputation is achieved, even for heterogeneous data, by using learner answer text directly to predict missing scores without the need to model underlying patterns of score data.
2. Recent scoring models based on pre-trained neural models are expected to enable accurate imputations from a relatively small subset of score data, as demonstrated in recent automated scoring studies [28, 32, 33, 34, 35]. This facilitates accurate ability estimation while markedly reducing reliance on human grading.

Through empirical evaluation using real-world datasets, this study demonstrates that the proposed method achieves markedly higher accuracy in ability estimation than do conventional approaches, even with high missing ratios. While the proposed method is based on a relatively simple idea, its effectiveness in greatly improving the accuracy of ability estimation directly contributes to enhancing various educational applications, as outlined above.

2. Task Settings and Objective

The objective of this study is to estimate learner ability based on a collection of scores assigned to their answers for multiple constructed-response test items. This collection of scores is defined as follows:

$$\mathbf{U} = \{u_{ij} \in \mathcal{K} \cup \{-1\} \mid j \in \mathcal{J}, i \in \mathcal{I}\}, \tag{1}$$

where u_{ij} indicates the score assigned to the answer of learner $j \in \mathcal{J}$ for item $i \in \mathcal{I}$, and $\mathcal{I}$ and $\mathcal{J}$ represent the sets of items and learners, respectively. Furthermore, $\mathcal{K} = \{1, 2, \ldots, K\}$ represents the set of score categories, where K indicates the number of categories, and $u_{ij} = -1$ indicates missing data.

A common approach to estimating learner ability is to calculate the average or total score for each learner. However, such simple methods are not suitable for datasets with missing data, such as those exemplified in Table 1. This limitation arises because item characteristics such as difficulty and discrimination vary among items, causing average or total scores to depend heavily on which items each learner is graded on [36, 37]. This property is not suitable for our objective, which is to accurately estimate learner ability from score data with substantial missing values to reduce assessment workload. To address this limitation, we utilize IRT, a robust framework for estimating learner ability from incomplete data.

3. Item Response Theory

IRT [14] is a test theory based on statistical models that has been widely employed for ability estimation and item analysis in various educational tests. IRT estimates learner ability by considering the characteristics of test items, such as item difficulty and discrimination. This is done using probabilistic models known as IRT models, which define the probability of score observations as a function of learner ability and item characteristics.

Among the various IRT models, the generalized partial credit model (GPCM) [38] is a representative model particularly suited for Likert-scale polytomous score data, as assumed in this study. The GPCM defines the probability that learner j receives score k for constructed-response test item i as

$$P(u_{ij} = k) = \frac{\exp \sum_{m=1}^{k} [\alpha_i(\theta_j - \beta_i - d_{im})]}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} [\alpha_i(\theta_j - \beta_i - d_{im})]}, \tag{2}$$

where θ_j represents the latent ability of learner j, α_i is the discrimination parameter for item i, β_i is the difficulty parameter for item i, and d_{im} is the step difficulty parameter representing the difficulty of transitioning between scores $m-1$ and m for the item. For model identification, $d_{i1} = 0$ and $\sum_{m=2}^{K} d_{im} = 0$ are assumed.

The parameters of learner ability and item characteristics are estimated from a score collection $\mathbf{U}$, typically by maximizing the following log-likelihood function:

$$\log \mathcal{L} = \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} z_{ij} \log P(u_{ij}), \tag{3}$$

where z_{ij} is a dummy variable that equals 0 if $u_{ij} = -1$ and 1 otherwise. As evident from this equation, IRT can estimate parameters, including ability, from datasets with missing scores by calculating the likelihood while excluding missing scores [15, 16, 17]. Furthermore, IRT generally provides more accurate ability estimates compared with methods based on simple averages or total scores because it accounts for the characteristics of test items during the estimation process [36, 37]. However, even with the IRT approach, the accuracy of ability estimation diminishes as the proportion of missing scores increases. A common strategy to address this limitation is the application of data augmentation techniques to impute missing scores.

4. Data Augmentation

There are various methods for imputation-based data augmentation. Simple approaches include mean or mode imputation, in which missing scores are replaced with averages or the most frequent scores [20]. Although computationally efficient, these methods often produce biased estimates because they fail to effectively capture the underlying patterns within the data.

More advanced approaches such as statistical model-based and machine learning-based methods aim to predict missing scores by modeling such underlying patterns for more accurate imputation [21, 22, 23, 24, 25, 26]. For example, one representative approach involves constructing supervised machine learning models, such as linear regression, support vector machine, and random forests, that predict each variable as the objective variable using the remaining variables as explanatory variables [21, 25, 26]. Another approach is based on unsupervised learning and directly utilizes the similarity of observed data patterns among samples [22, 23, 24]. A typical example is k-nearest neighbors (k-NN), which estimates missing values by identifying similar samples based on their observed data [22, 24]. Other examples include matrix factorization techniques, such as singular value decomposition, which approximate the data as a low-rank matrix, estimating missing values by uncovering latent structures and dependencies among variables [23]. Multiple imputation is another approach that integrates these individual imputation methods by generating multiple plausible datasets and combining the results through statistical pooling [39]. This approach accounts for the uncertainty of missing data.

However, as discussed in Section 1, these traditional imputation methods often struggle to achieve accurate imputation in real-world constructed-response tests, particularly in situations with high data sparsity and potential violations of the assumption of an underlying consistent data-generation process. This difficulty arises because they primarily infer missing values based on the relationships observed among the available scores (e.g., correlations between items or similarities between learners). When data sparsity is high, or when the assumption of a single underlying data-generation process is violated, these observed relationships become unreliable predictors for the missing scores, leading to biased or inaccurate imputations. To address these limitations, the main idea of this study is to leverage automated scoring technologies to impute missing scores.

5. Automated Scoring for Constructed-Responses

Recently, automated essay scoring and automated short-answer grading using artificial intelligence technologies have become prominent topics of artificial intelligence in the education community [11, 27, 28, 29, 30, 31]. While various methods have been proposed, conventional automated scoring approaches typically fall into one of two categories: feature-engineering-based approaches or neural-based approaches [11, 31].

Feature-engineering-based approaches rely on manually designed features, such as text length or the number of spelling errors, to predict scores using regression or classification models [10, 40, 41, 42, 43, 44]. While this approach offers interpretability and explainability, achieving high accuracy generally requires extensive effort in feature design and selection, which often needs to be tailored for each specific test item.

To address this limitation, neural-based approaches, which automatically extract features from data using deep neural networks, have gained increasing popularity. Early neural models have primarily employed convolutional neural networks or recurrent neural networks [45, 46, 47, 48, 49, 50]. More recent advancements have focused on using pretrained transformer networks [51], such as bidirectional encoder representations from transformers (BERT) [52], which have demonstrated superior performance and accuracy in automated scoring tasks [53, 54, 55, 56, 57, 58]. BERT and its variants use extensive pretraining on large-scale text corpora, with high accuracy obtained by fine-tuning them for a target scoring task using relatively small datasets of scored responses.

Most recently, LLMs have emerged as the next frontier in automated scoring. LLMs build upon the transformer architecture, similar to BERT and its variants, but are pretrained on more massive and diverse datasets using various training techniques, such as reinforcement learning from human feedback and instruction tuning [59, 60]. A major advantage of LLMs is their capability to address various natural language tasks, including automated scoring of constructed responses, by providing a concrete task explanation as a prompt in a zero-shot setting or by including a small number of examples alongside the prompt in a few-shot setting [61]. This reduces the reliance on extensive labeled datasets, making LLMs highly adaptable to a wide range of natural language processing tasks. Recent studies exploring

the application of LLMs for essay and short-answer scoring have shown that LLMs often achieve reasonable scoring performance using zero-shot or few-shot approaches [28, 30, 62, 63, 64, 65, 66, 67], although they tend to perform worse than conventional fine-tuned scoring models based on pretrained transformers [62, 64, 65, 67, 68].

6. Proposed Method

This study proposes a novel imputation-based data augmentation approach using automated scoring technologies for accurate IRT-based ability estimation in constructed-response tests. The proposed approach consists of the following steps:

1. **Developing a scoring model**: Neural automated scoring models are prepared either by fine-tuning BERT or its variants on a subset of manually scored learner answers for each test item, or by employing a zero-shot scoring model using LLMs. The choice between these methods depends on various conditions as listed below:
    - **Fine-tuned models**: These methods are recommended when a relatively large number of scored answers, such as more than a hundred, can be prepared for each item. This is often feasible in scenarios with a large number of examinees and when manual grading costs are acceptable.
    - **Zero-shot scoring models**: These are more suitable when only a very limited number of scored answers are available for each item. For instance, this approach is preferable when the number of examinees is small, or when scoring individual answers is time-consuming. Zero-shot models are also suitable for situations in which clear scoring criteria are available or in which the evaluation task is relatively easy, because zero-shot evaluation is expected to be effective in such cases.
2. **Predicting missing scores**: Once the scoring model is prepared, it is used to predict missing scores and construct a complete score dataset.
3. **Estimating ability using IRT models**: The augmented dataset is then used to estimate learner ability by applying IRT models.

The proposed method is expected to achieve more robust imputation, even for heterogeneous data lacking a consistent data-generation process, because it directly leverages learner answer text to predict missing scores without modeling underlying patterns of score data. Additionally, it is expected to provide accurate imputations even under high missing ratios because constructing scoring models based on pre-trained neural language models or zero-shot models often achieves reasonable scoring performance with relatively few or no samples, as demonstrated in recent automated scoring studies [28, 32, 33, 34, 35]. These features make the proposed method suitable for achieving accurate ability estimation while reducing the need for human grading.

7. Experiments

We conducted empirical evaluation experiments using real-world datasets to demonstrate the effectiveness of the proposed method.

7.1. Data

For our experiments, we required datasets comprising scored constructed responses for multiple items in which the same set of learners answered all items. Therefore, popular benchmark datasets for automated essay or short-answer grading tasks, such as ASAP (automated student assessment prize)[1]

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

and ASAP–SAS (short answer scoring)[2], could not be used owing to the lack of information identifying respondents for each answer. Consequently, we utilized the following three datasets:

1. **Short-Answer Grading (SAG) Dataset**: This dataset, developed by the Benesse Educational Research and Development Institute in Japan, consists of responses from 511 Japanese university students to three short-answer items in a Japanese reading comprehension test. Scores for the responses were provided by expert raters using five rating categories for each item.
2. **Essay Scoring (ES) Dataset [69]**: This dataset consists of essays written by 327 Japanese university students in response to three essay tasks offered in a natural science lecture. Each response was scored on a five-point scale by expert raters.
3. **ELYZA-tasks-100 (ELYZA) Dataset**[3]: This dataset is designed for evaluating Japanese LLMs. It includes responses generated by 33 LLMs for 100 writing tasks, with each response scored by expert raters using a five-point scale. We used this dataset by treating individual LLMs as learners and writing tasks as constructed-response test items. Although this dataset is not specifically intended for automated essay or short-answer scoring, it appears to be suitable for applying the proposed method with a zero-shot scoring model. This is because the number of examinees is small, and LLM-generated responses are relatively straightforward to score using LLMs in a zero-shot manner.

7.2. Experimental Procedures

We conducted the following experiments for each dataset:

1. We estimated the IRT parameters from the complete score data based on the GPCM introduced earlier. The obtained ability estimates for learners were treated as gold-standard values in this experiment.
2. We created incomplete datasets from the complete score data by converting some scores into missing values. The missing ratios were varied as follows:
   - For the SAG and ES datasets, we examined three missing ratios: 33%, 50%, and 62%
   - For the ELYZA dataset, we examined five missing ratios: 10%, 20%, 50%, 65%, and 80%

   The incomplete datasets were created following a systematic design [70, 71, 72] to generate missing patterns while ensuring the applicability of IRT and the conventional missing imputation methods. The algorithm for creating missing patterns, which supports the rationale behind the selection of missing ratios for each dataset, is detailed in the Appendix.
3. Using each incomplete score dataset, we applied the GPCM to estimate learner ability through the following methods:
   - **Estimation without imputation**: This method directly applies the GPCM to the incomplete score dataset, ignoring missing scores during likelihood calculation as detailed in Section 3.
   - **Estimation with imputation by k-NN or random forest**: Ability is estimated by the GPCM after missing scores are imputed using k-NN or random forest (RF)[4], as introduced in Section 4.
   - **Estimation with imputation by the proposed method**: Ability is estimated using the GPCM based on the proposed method. In the SAG and ES datasets, BERT-based models are used for the automated scoring models fine-tuned on scored answers for each item[5]. For the ELYZA dataset, we employed a zero-shot scoring approach with GPT-4o, where the

[2] https://github.com/benhamner/ASAP-SAS

[3] https://huggingface.co/datasets/elyza/ELYZA-tasks-100/tree/main

[4] We used the `VIM` and `missForest` libraries in R for missing score imputations using k-NN and RF, respectively.

[5] We used BERT models with a linear output layer on top of the [CLS] token, which is appended to the beginning of the input text. The pre-trained BERT model was `tohoku-nlp/bert-base-japanese-v3`. The optimizer was AdamW with a learning rate of 1e-5. The mini-batch size was 64 for the SAG dataset and 16 for the ES dataset.

Table 2
Translated prompt for zero-shot scoring on the ELYZA dataset. (†) indicates that remaining details have been abbreviated. The full Japanese grading criteria are available at https://huggingface.co/datasets/elyza/ELYZA-tasks-100/blob/main/baseline/humaneval/guideline.md.

You are a grader. You will be given a test item, a reference answer, a grading rubric, and a response. Referencing the grading rubric and the reference answer, grade the response on a scale of 1 to 5, and output the number only. **Test Item:** {text of the test item} **Reference Answer:** {text of the reference answer} **Basic Grading Criteria** *1. Incorrect:* Does not follow instructions. Chooses an incorrect option in a multiple-choice question. (†) *2. Incorrect but heading in the right direction:* Usually a 3-point answer with a 1-point deduction. (†) *3. Partially correct:* Addresses the majority of a complex instruction correctly. (†) *4. Correct:* Correctly answers the question. (†) *5. Helpful:* Correctly answers the question and further anticipates user needs. (†) **Basic Deductions**: Scores may be adjusted based on the following factors. *- Unnatural Japanese (-1 point):* Syntactically awkward or unclear Japanese, repetition of the same sentence, or abrupt insertion of English words. *- Partial Hallucination (-1 point):* A response partially inconsistent with facts. (†) *- Excessive Safety Concerns (Score as 2 points):* e.g., Responds with "I cannot answer for ethical reasons." **Item-Specific Grading Criteria:** {Grading Criteria defined for each item} **Response:** {Response to be evaluated}

prompt was designed to include detailed task instructions with scoring criteria defined in the original dataset, as shown in Table 2.

4. The root-mean-squared error (RMSE) and Pearson's correlation coefficient were calculated between the estimated and gold-standard abilities. To address the scale indeterminacy inherent in IRT estimation, the ability values were normalized to have a mean of zero and a variance of one, ensuring that the estimated and gold-standard abilities were directly comparable.
5. Steps 2 to 4 were repeated 10 times, each time using a different missing pattern. To create different missing patterns using the same algorithm, the order of learners was randomly shuffled with each repetition.

Furthermore, we investigated the performance of the proposed method in situations in which the conventional approach was not applicable owing to a substantially high missing ratio. Specifically, we examined three additional missing ratios, namely, 70%, 80%, and 90%, for the SAG and ES datasets, and a 100% missing ratio for the ELYZA dataset in experimental procedure 2 for the proposed method. It should be noted that under these conditions, the conventional method cannot be applied due to the presence of learners with no assigned scores.

7.3. Experimental Results

Fig. 1 shows the experimental results. The upper plot shows the averaged RMSEs, while the lower plot shows the averaged correlation coefficients, with error bars indicating the standard deviation over 10 repetitions. An exception is the condition with a 100% missing ratio in the ELYZA dataset, in which results are based on a single trial due to the lack of variation in missing patterns. Note that the conventional methods have no results for missing ratios above 62% in the SAG and ES datasets and 100% in the ELYZA dataset because they are not applicable under these conditions, as described above.

According to the experimental results, the proposed method demonstrated lower RMSEs and higher correlation coefficients compared to conventional methods across all conditions. While all methods showed a reduction in the accuracy of ability estimation as the missing ratio increased, conventional imputation methods exhibited a more rapid decline in accuracy as the missing ratio increased than the proposed method, as highlighted in the results for the ELYZA dataset. This reflects the difficulty of conventional methods in modeling patterns underlying observed score data under high missing ratios.

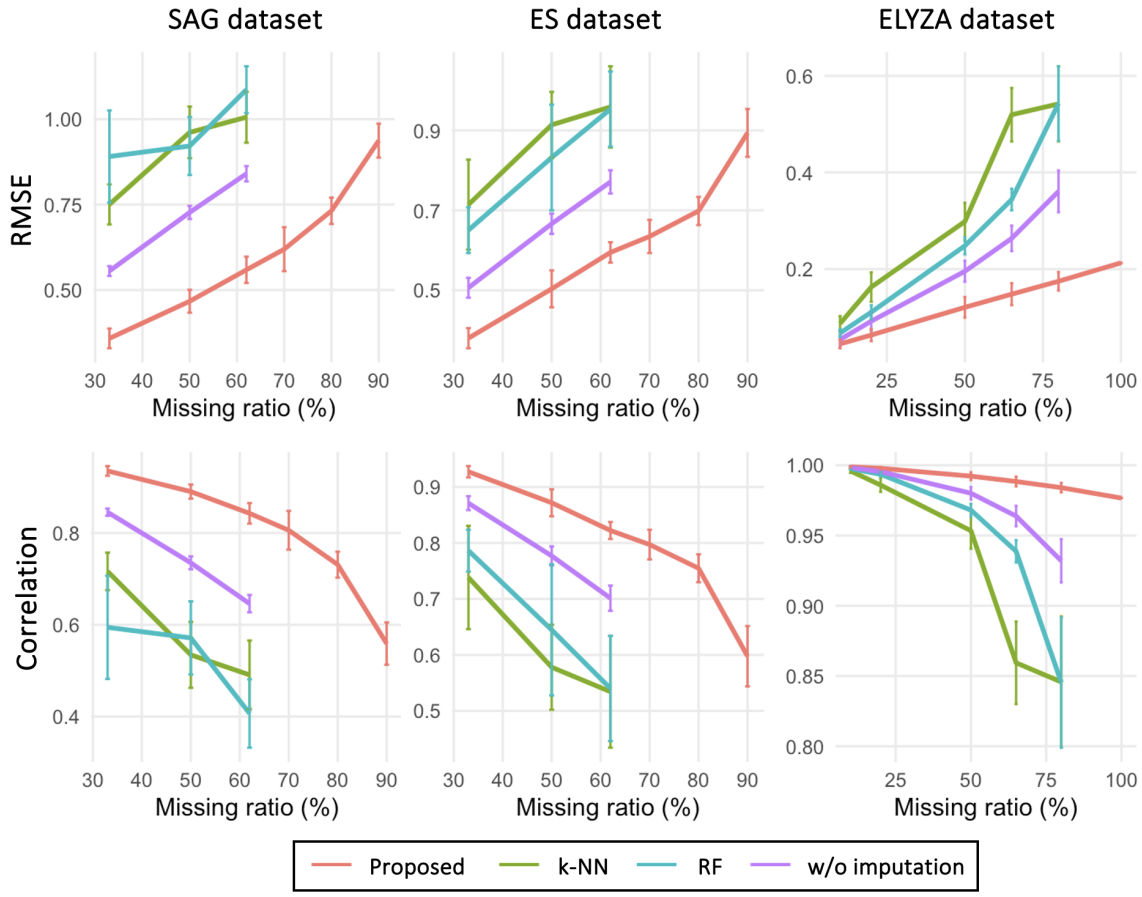

Figure 1: Comparison of ability estimation accuracy across different missing ratios. Upper plots show RMSE and lower plots show Pearson's correlation coefficient between estimated and gold-standard abilities. Since the ability values are normalized to have zero mean and unit variance, an RMSE around 0.3 corresponds to about 5% of the distribution range (± 3 standard deviations), indicating high accuracy, while an RMSE around 0.6, corresponding to about 10%, may still be acceptable.

Additionally, conventional imputation methods underperformed compared to ability estimation without imputation, suggesting that inaccurate imputation introduces heavy bias into ability estimation.

A paired t-test was conducted to compare the average accuracy differences between the proposed method and each of the other methods at each missing ratio. The results indicated significant differences at the 1% significance level for all comparisons, except for the comparison between the proposed method and the method without imputation at a missing ratio of 10% in the ELYZA dataset. These findings indicate that the proposed method achieves remarkably high accuracy in ability estimation from incomplete data, except when the missing ratio is extremely low.

Furthermore, the proposed method achieved reasonable accuracy in ability estimation in situations in which conventional methods are not applicable. Specifically, it maintained high accuracy even with a 100% missing ratio in the ELYZA dataset, corresponding to the case employing zero-shot scoring. The use of fine-tuned scoring models in the SAG and ES datasets also demonstrated relatively good accuracy with a missing ratio of up to 80%. More specifically, the accuracies under the 80% missing ratio were higher than those of conventional methods at a 62% missing ratio. These results highlight the unique and important advantage of the proposed method for accurately estimating abilities, except in cases of extremely large missing ratios, such as at a missing ratio of 90%, when fine-tuned scoring models are used.

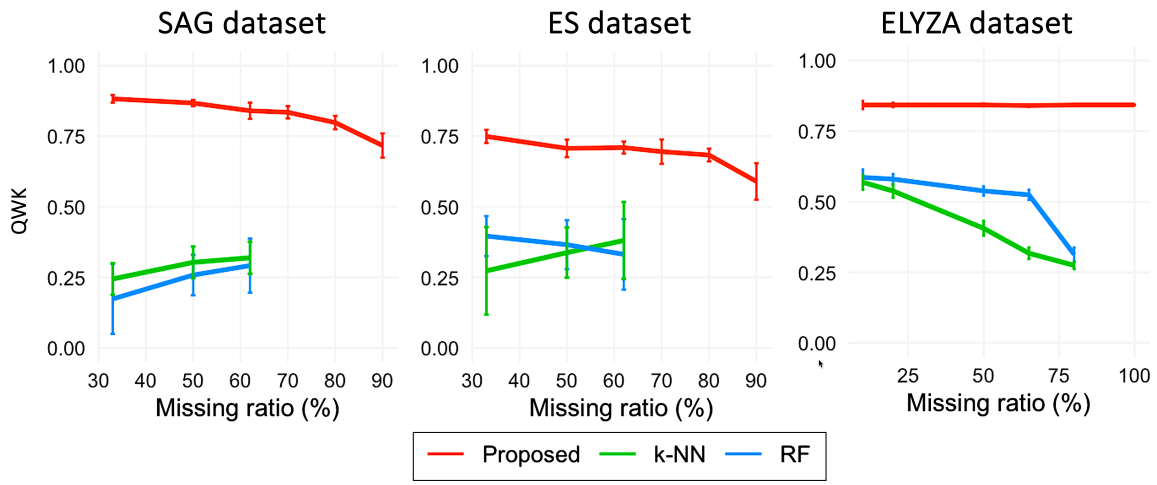

Figure 2: Accuracy of predicting missing scores for each imputation method, measured by QWK.

8. Analysis

8.1. Accuracy of Missing Score Prediction

We can infer that the high accuracy in ability estimation of the proposed method comes from its high accuracy in missing score imputation. To confirm this, we analyzed the accuracy of missing score imputation for each method. Specifically, using the results from experimental procedure 3, we calculated the agreement between the predicted scores for missing values generated by each imputation method, including k-NN, RF, and the proposed method, and their corresponding true scores. As an evaluation metric, we used quadratic weighted kappa (QWK), which is commonly employed in research on automated scoring.

Fig. 2 shows the results, with the average QWK values with error bars representing the standard deviations obtained from 10 repeated experiments. Given that higher QWK values indicate greater imputation accuracy, the results show that the proposed method achieves high accuracy in imputing missing scores, whereas conventional methods exhibit markedly lower accuracy.

Furthermore, in the SAG and ES datasets, the proposed method exhibits a decline in accuracy as the missing ratio increases due to the reduction in training data for fine-tuning, while the degraded accuracies remain higher than those of conventional methods. Moreover, in the ELYZA dataset, while conventional methods suffer a drastic decline in imputation accuracy, the proposed method maintains high accuracy regardless of the missing ratio. These results suggest that the proposed method achieves high imputation accuracy, which is likely to contribute to its high ability estimation accuracy.

8.2. Robustness to Heterogeneity

As described in Section 1, the proposed method is expected to be effective for imputing missing scores that are difficult to predict from the patterns of observed data owing to their heterogeneity. To demonstrate this, Table 3 provides examples of score imputations by the proposed method and the k-NN method for two learners sampled from the results of the SAG dataset.

In the table, the *True scores* row represents the original complete data and the *w/o imput.* row shows the incomplete data (with missing values indicated by "NA"). The *k-NN* and *Proposed* rows show the complete data created by imputing the missing values using each respective method. Additionally, the *True avg.* row indicates the averaged scores for each item calculated from the complete data, while $\hat{\theta}$ represents the estimated abilities based on each score dataset.

Considering that the average item scores follow the order of item 3 > item 2 > item 1, predicting the missing data in the left table as a score of 5, as done by the k-NN method, might seem reasonable. Similarly, in the right table, given that the score for item 3 was 3, predictions by the k-NN method for

Table 3
Examples of score imputation by the k-NN and proposed methods.

	Item 1	Item 2	Item 3	$\hat{\theta}$
True scores	4	5	3	0.35
w/o imput.	4	5	NA	1.77
k-NN	4	5	5	2.35
Proposed	4	5	2	0.09
True avg.	2.31	3.22	3.84	

	Item 1	Item 2	Item 3	$\hat{\theta}$
True scores	2	5	3	-0.25
w/o imput.	NA	NA	3	-0.84
k-NN	2	3	3	-0.63
Proposed	2	5	3	-0.12
True avg.	2.31	3.22	3.84	

the two missing data items (item 1, 2; item 2, 3) appear reasonable based on the overall trend. However, these predictions by the k-NN method are quite inaccurate. This discrepancy arises because the true scores in these cases do not follow the surrounding patterns, making them difficult to predict.

As shown in the examples above, conventional methods struggle to handle such data. In contrast, the proposed method does not rely on modeling surrounding score patterns but instead evaluates the content of individual answer texts. Therefore, the proposed method can make accurate predictions for heterogeneous score data.

9. Conclusion

This study proposed a novel method for imputing missing scores to enhance IRT-based ability estimation by leveraging automated scoring technologies. Experimental results demonstrated that the proposed method achieves higher accuracy in ability estimation compared with conventional approaches, even under conditions of high missing ratios or heterogeneous data. This indicates that the proposed method achieves high accuracy in ability estimation while markedly reducing the manual grading workload.

However, we acknowledge several limitations. Firstly, the effectiveness of our method depends on the accuracy of the employed automated scoring model. Inaccurate scoring models, whether fine-tuned on insufficient data or based on zero-shot LLMs with suboptimal prompts, could lead to biased imputations and consequently affect the validity of the final ability estimates. The observed accuracy drop at the 90% missing ratio when using fine-tuned models might partially reflect this dependency. Secondly, developing these scoring models can require substantial computational resources for fine-tuning or careful prompt engineering for zero-shot approaches, potentially offsetting the intended reduction in manual grading workload. Thirdly, while we demonstrated effectiveness on specific datasets, all datasets were limited to the Japanese language, which restricts the generalizability of our findings to broader linguistic and assessment contexts. Fourthly, the proposed method treats imputed scores as deterministic inputs to IRT estimation and does not propagate the uncertainty associated with automated scoring, which could result in overconfident or biased ability estimates. Lastly, the use of black-box models such as BERT and LLMs limits the interpretability of individual imputed scores and hinders systematic fairness evaluation, an important consideration for high-stakes testing scenarios.

Future research should address these limitations and explore further extensions. Although this study focused on unidimensional IRT, the proposed method may be applicable in various ability measurement contexts. For instance, the method is applicable to complex student models, such as multidimensional IRT [73], cognitive diagnostic models [74], and knowledge tracing [3]. These complex models are more sensitive to data sparsity, which may further highlight the advantages of the proposed approach. Furthermore, exploring methods to incorporate the uncertainty associated with automated scores could lead to more reliable ability estimates and associated error measures. In addition, methods to enhance the interpretability of imputed scores should be explored to increase transparency and trust in the system. Evaluating the method on more diverse datasets across multiple languages, domains, and assessment formats is also a key direction for improving external validity. In addition, future studies should consider comparing the proposed method with more advanced deep-learning-based imputation methods, such as deep matrix factorization [75] and variational autoencoders, which were not included in the current evaluation. Another promising direction is to leverage IRT models with rater parameters [37, 76, 77, 78],

such as many-facet Rasch models, to achieve more valid ability estimation by treating AI graders as distinct raters within a unified measurement framework. Finally, studying the integration of this method into practical applications like adaptive testing systems for constructed-response items, where real-time scoring and imputation could enhance test efficiency and personalization, would be valuable. Given the critical role of accurate ability estimation in learning support systems, we believe these future investigations hold significant promise.

Appendix: Algorithms for Generating Missing Patterns

In this appendix, we explain how incomplete data were created from complete data in the experiments. For the missing ratios of 33%, 50%, and 62% in the SAG and ES datasets, the missing patterns, which enable IRT parameter linking [70, 71, 72] and the application of other imputation methods, were generated as follows:

- The 33% missing pattern was created by repeating the patterns of learners 1–3 on the left side of Table 1.
- The 50% missing pattern was created by repeating the patterns of learners 1–6 on the right side of Table 1.
- The 62% missing pattern was created by combining one repetition of learners 1–3 with six repetitions of learners 4–6, based on the pattern on the right side of Table 1.

For the missing ratios of 10%, 20%, 50%, 65%, and 80% in the ELYZA dataset, these missing patterns were generated using Algorithm 1, which also ensures parameter linking in IRT estimation and the application of other imputation methods

Algorithm 1 Algorithm for creating missing patterns for the ELYZA dataset.

Input: $\mathcal{I}$, $\mathcal{J}$, N_m: Number of missing ratio
Initialize missing indicator variable $\{z_{ij} = 1 \mid i \in \mathcal{I}, j \in \mathcal{J}\}$ defined in Equation (3)
for $j \in \mathcal{J}$ **do**
  Set $i_{st} = (j * 10)\%100$ and $i_{ed} = (j * 10 + N_m)\%100$
  if $i_{ed} > i_{st}$ **then**
    Set $z_{ij} = 0$ for i in $[i_{st} : i_{ed}]$
  else
    Set $z_{ij} = 0$ for i in $[i_{ed} : 100]$ and $[1 : i_{st}]$
  end if
end for

For the other missing ratios in the SAG and ES datasets, where IRT and conventional imputation methods are not applicable, missing patterns were generated by randomly selecting learners for each item until the specified percentage of scores was converted to missing values.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers 23K20727 and 24H00739. We thank Yuki Doka and Yoshihiro Kato from the Benesse Educational Research and Development Institute for permission to use the SAG dataset.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-4o in order to: Grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] Y. Hajjioui, O. Zine, M. Benslimane, A. Ibriz, Intelligent tutoring systems: A review, in: Big Data and Internet of Things, 2024, pp. 663–676.

[2] M. S. Khine, Using AI for Adaptive Learning and Adaptive Assessment, Springer Nature, 2024, pp. 341–466.

[3] S. Shen, Q. Liu, Z. Huang, Y. Zheng, M. Yin, M. Wang, E. Chen, A survey of knowledge tracing: Models, variants, and applications, IEEE Transactions on Learning Technologies 17 (2024) 1898–1919.

[4] Y. Tomikawa, A. Suzuki, M. Uto, Adaptive question–answer generation with difficulty control using item response theory and pre-trained transformer models, IEEE Transactions on Learning Technologies 17 (2024) 2186–2198.

[5] Y. Abosalem, Assessment techniques and students' higher-order thinking skills, International Journal of Secondary Education 4 (2016) 1–11.

[6] H. J. Bernardin, S. Thomason, M. R. Buckley, J. S. Kane, Rater rating-level bias and accuracy in performance appraisals: The impact of rater personality, performance management competence, and rater accountability, Human Resource Management 55 (2016) 321–340.

[7] O. L. Liu, L. Frankel, K. C. Roohr, Assessing critical thinking in higher education: Current state and directions for next-generation assessment, ETS Research Report Series (2014) 1–23.

[8] R. J. Mislevy, Sociocognitive Foundations of Educational Measurement, Routledge, 2018.

[9] M. Murtonen, K. Balloo, Redefining Scientific Thinking for Higher Education: Higher-Order Thinking, Evidence-Based Reasoning and Research Skills, Palgrave Macmillan, 2019.

[10] E. Amorim, M. Cançado, A. Veloso, 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, 2018, pp. 229–237.

[11] Z. Ke, V. Ng, Automated essay scoring: A survey of the state of the art, in: Proceedings of the International Joint Conference on Artificial Intelligence, 2019, pp. 6300–6308.

[12] G. Leckie, J. A. Baird, Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience, Journal of Educational Measurement 48 (2011) 399–418.

[13] M. Uto, M. Okano, Learning automated essay scoring models using item-response-theory-based scores to decrease effects of rater biases, IEEE Transactions on Learning Technologies 14 (2021) 763–776.

[14] F. Lord, Applications of item response theory to practical testing problems, Routledge, 1980.

[15] F. Baker, S. H. Kim, Item Response Theory: Parameter Estimation Techniques, CRC Press, 2004.

[16] W. J. van der Linden, Handbook of Item Response Theory, Volume Two: Statistical Tools, CRC Press, 2016.

[17] M. L. Nering, R. Ostini, Handbook of Polytomous Item Response Theory Models, Routledge, 2010.

[18] M. Barrabés, M. Perera, V. N. Moriano, X. Giró-I-Nieto, D. M. Montserrat, A. G. Ioannidis, Advances in biomedical missing data imputation: A survey, IEEE Access 13 (2025) 16918–16932.

[19] H. Zhao, K. Sun, A. Dezfouli, E. V. Bonilla, Transformed distribution matching for missing value imputation, in: Proceedings of the 40 th International Conference on Machine Learning, 2023, pp. 42159–42186.

[20] R. J. A. Little, D. B. Rubin, Statistical Analysis with Missing Data, Wiley & Sons, 2019.

[21] D. Bertsimas, C. Pawlowski, Y. D. Zhuo, From predictive methods to missing data imputation: An optimization approach, Journal of Machine Learning Research 18 (2018) 1–39.

[22] L. P. Brás, J. C. Menezes, Improving cluster-based missing value estimation of dna microarray data, Biomolecular Engineering 24 (2007) 273–282.

[23] F. Husson, J. Josse, B. Narasimhan, G. Robin, Imputation of mixed data with multilevel singular value decomposition, Journal of Computational and Graphical Statistics 28 (2019) 552–566.

[24] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, R. B. Altman, Missing value estimation methods for DNA microarrays, Bioinformatics 17 (2001) 520–525.

[25] R. Wang, What is the impact of nonrandomness on random choice models?, Manufacturing &

Service Operations Management 24 (2022) 485–503.
[26] X. Zhang, X. Song, H. Wang, H. Zhang, Sequential local least squares imputation estimating missing value of microarray data, Computers in Biology and Medicine 38 (2008) 1112–1120.
[27] R. Chaudhari, M. Patel, Deep learning in automated short answer grading: A comprehensive review, in: ITM Web of Conferences, volume 65, 2024.
[28] G. Kortemeyer, Performance of the pre-trained large language model GPT-4 on automated short answer grading, Discover Artificial Intelligence 4 (2024) 47.
[29] S. Li, V. Ng, Automated essay scoring: A reflection on the state of the art, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024, pp. 17876–17888.
[30] H. Misgna, B.-W. On, I. Lee, G. S. Choi, A survey on deep learning-based automated essay scoring and feedback generation, Artificial Intelligence Review 58 (2025) 1–40.
[31] M. Uto, A review of deep-neural automated essay scoring models, Behaviormetrika 48 (2021) 459–484.
[32] H. Do, Y. Kim, G. G. Lee, Prompt- and trait relation-aware cross-prompt essay trait scoring, in: Findings of the Association for Computational Linguistics, 2023, pp. 1538–1551.
[33] L. Jiang, N. Bosch, Short answer scoring with GPT-4, in: Proceedings of the Eleventh ACM Conference on Learning@ Scale, 2024, pp. 438–442.
[34] R. Ridley, L. He, X. Y. Dai, S. Huang, J. Chen, Automated cross-prompt scoring of essay traits, in: Proceedings of the Association for the Advancement of Artificial Intelligence, volume 35, 2021, pp. 13745–13753.
[35] T. Shibata, M. Uto, Enhancing cross-prompt automated essay scoring by selecting training data based on reinforcement learning, in: Workshop on Automated Evaluation of Learning and Assessment Content, 2024.
[36] M. Uto, M. Ueno, Empirical comparison of item response theory models with rater's parameters, Heliyon 4 (2018).
[37] M. Uto, M. Ueno, A generalized many-facet Rasch model and its Bayesian estimation using Hamiltonian Monte Carlo, Behaviormetrika 47 (2020) 469–496.
[38] E. Muraki, A generalized partial credit model, in: W. J. van der Linden, R. K. Hambleton (Eds.), Handbook of Modern Item Response Theory, Springer, 1997, pp. 153–164.
[39] D. B. Rubin, Multiple imputation after 18+ years, Journal of the American Statistical Association 91 (1996) 473–489.
[40] S. Burrows, I. Gurevych, B. Stein, The eras and trends of automatic short answer grading, Journal of Artificial Intelligence in Education 25 (2015) 60–117.
[41] M. Dascalu, W. Westera, S. Ruseti, S. Trausan-Matu, H. Kurvers, ReaderBench learns Dutch: Building a comprehensive automated essay scoring system for Dutch language, in: Proceedings of the International Conference on Artificial Intelligence in Education, 2017, pp. 52–63.
[42] C. Leacock, M. Chodorow, C-rater: Automated scoring of short-answer questions, Computers and the Humanities 37 (2003) 389–405.
[43] H. V. Nguyen, D. J. Litman, Argument mining for improving the automated scoring of persuasive essays, in: Proceedings of the Association for the Advancement of Artificial Intelligence, volume 32, 2018.
[44] M. D. Shermis, J. C. Burstein, Automated Essay Scoring: A Cross-disciplinary Perspective, Routledge, 2003.
[45] D. Alikaniotis, H. Yannakoudakis, M. Rei, Automatic text scoring using neural networks, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2016, pp. 715–725.
[46] Y. Farag, H. Yannakoudakis, T. Briscoe, 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, 2018, pp. 263–271.
[47] C. Jin, B. He, K. Hui, L. Sun, 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, 2018, pp. 1088–1097.

[48] M. Mesgar, M. Strube, A neural local coherence model for text quality assessment, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2018, pp. 4328–4339.
[49] K. Taghipour, H. T. Ng, A neural approach to automated essay scoring, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2016, pp. 1882–1891.
[50] Y. Wang, Z. Wei, Y. Zhou, X. Huang, Automatic essay scoring incorporating rating schema via reinforcement learning, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2018, pp. 791–797.
[51] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.
[52] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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, 2019, pp. 4171–4186.
[53] T. Liu, W. Ding, Z. Wang, J. Tang, G. Y. Huang, Z. Liu, Automatic short answer grading via multiway attention networks, in: Proceedings of the International Conference on Artificial Intelligence in Education, 2019, pp. 169–173.
[54] J. Lun, J. Zhu, Y. Tang, M. Yang, Multiple data augmentation strategies for improving performance on automatic short answer scoring, in: Proceedings of the Association for the Advancement of Artificial Intelligence, volume 34, 2020, pp. 13389–13396.
[55] E. Mayfield, A. W. Black, Should you fine-tune BERT for automated essay scoring?, in: Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications, 2020, pp. 151–162.
[56] C. Sung, T. I. Dhamecha, N. Mukhi, Improving short answer grading using transformer-based pre-training, in: Proceedings of the International Conference on Artificial Intelligence in Education, 2019, pp. 469–481.
[57] J. Xue, X. Tang, L. Zheng, A hierarchical BERT-based transfer learning approach for multi-dimensional essay scoring, IEEE Access 9 (2021) 125403–125415.
[58] M. Yamaura, I. Fukuda, M. Uto, Neural automated essay scoring considering logical structure, in: Proceedings of the International Conference on Artificial Intelligence in Education, 2023, pp. 267–278.
[59] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, et al., Training a helpful and harmless assistant with reinforcement learning from human feedback, arXiv preprint (2022).
[60] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, et al., Training language models to follow instructions with human feedback, Advances in Neural Information Processing Systems 35 (2022) 27730–27744.
[61] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners, Advances in Neural Information Processing Systems 35 (2022) 22199–22213.
[62] I. Chamieh, T. Zesch, K. Giebermann, LLMs in short answer scoring: Limitations and promise of zero-shot and few-shot approaches, in: Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications, 2024, pp. 309–315.
[63] L.-H. Chang, F. Ginter, Automatic short answer grading for finnish with ChatGPT, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, 2024, pp. 23173–23181.
[64] S. Lee, Y. Cai, D. Meng, Z. Wang, Y. Wu, Unleashing large language models' proficiency in zero-shot essay scoring, in: Findings of the Association for Computational Linguistics, 2024, pp. 181–198.
[65] W. A. Mansour, S. Albatarni, S. Eltanbouly, T. Elsayed, Can large language models automatically score proficiency of written essays?, in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, 2024, pp. 2777–2786.
[66] M. Stahl, L. Biermann, A. Nehring, H. Wachsmuth, Exploring LLM prompting strategies for joint essay scoring and feedback generation, in: Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications, 2024, pp. 283–298.
[67] Y. Wang, R. Hu, Z. Zhao, Beyond agreement: Diagnosing the rationale alignment of automated

essay scoring methods based on linguistically-informed counterfactuals, in: Findings of the Association for Computational Linguistics, 2024, pp. 8906–8925.
[68] K. P. Yancey, G. Laflair, A. Verardi, J. Burstein, Rating short L2 essays on the CEFR scale with GPT-4, in: Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications, 2023, pp. 576–584.
[69] K. Takeuchi, M. Ohno, K. Motojin, M. Taguchi, Y. Inada, M. Iizuka, T. Abo, H. Ueda, Development of essay scoring methods based on reference texts with construction of research-available japanese essay data, Journal of Information Processing [In Japanese] 62 (2021) 1586–1604.
[70] M. Ilhan, A comparison of the results of many-facet Rasch analyses based on crossed and judge pair designs, Educational Sciences: Theory and Practice 16 (2016) 579–601.
[71] M. J. Kolen, R. L. Brennan, Test Equating, Scaling, and Linking, Springer, 2014.
[72] M. Uto, Accuracy of performance-test linking based on a many-facet Rasch model, Behavior Research Methods 53 (2021) 1440–1454.
[73] M. D. Reckase, Multidimensional Item Response Theory, Statistics for Social and Behavioral Sciences, Springer-Verlag New York, NY, 2009.
[74] J. L. Templin, R. A. Henson, Measurement of psychological disorders using cognitive diagnosis models, Psychological methods 11 (2006) 287–305.
[75] H.-J. Xue, X. Dai, J. Zhang, S. Huang, J. Chen, Deep matrix factorization models for recommender systems, in: Proceedings of the 26th International Joint Conference on Artificial Intelligence, 2017, pp. 3203–3209.
[76] T. Eckes, Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluating Rater-Mediated Assessments, Peter Lag Pub. Inc., 2023.
[77] M. Uto, A multidimensional generalized many-facet Rasch model for rubric-based performance assessment, Behaviormetrika 48 (2021) 425–457.
[78] M. Uto, J. Tsuruta, K. Araki, M. Ueno, Item response theory model highlighting rating scale of a rubric and rater–rubric interaction in objective structured clinical examination, PLOS ONE 19 (2024) 1–23.