

Developing a Multi-Rater Indonesian Essay Scoring Dataset for Automated Essay Scoring

Pietra Wahyu Utomo ^{1✉}, Zainal Abidin ²

(1,2) Master of Informatics, UIN Maulana Malik Ibrahim, Malang, Indonesia

Abstract: Automated Essay Scoring (AES) research requires high-quality annotated datasets that provide reliable human reference scores, particularly for languages and educational contexts that remain underrepresented in existing resources. This study develops and validates a multi-rater Indonesian essay assessment dataset designed to support future AES and educational natural language processing research. The dataset comprises 3,600 response-level records collected from 360 senior high school students answering 10 Indonesian language essay questions. Each response was independently and blindly evaluated by three qualified Indonesian language teachers using a common analytical scoring rubric. Individual rater scores were preserved alongside an aggregated reference score to maintain the multi-rater structure of the dataset. Annotation reliability was evaluated using the average-measures absolute-agreement Intraclass Correlation Coefficient (ICC(A,k)), complemented by question-level reliability analysis and maximum pairwise inter-rater score differences. The overall ICC(A,k) was 0.9905, indicating excellent inter-rater reliability, while question-level ICC values ranged from 0.9803 to 0.9934, with a mean of 0.9899. Furthermore, 96.83% of responses exhibited a maximum inter-rater score difference of no more than one point. These findings demonstrate that the dataset provides highly consistent human annotations and a reliable basis for constructing reference scores. By documenting the dataset structure, scoring rubric, annotation protocol, rater qualifications, and reliability evidence, this study provides a documented research resource for future Indonesian AES research. The preserved multi-rater annotations may support subsequent investigations of human-AI agreement, rater variability, scoring severity, and alternative reference-score construction, subject to dataset access restrictions.

Article history:

Received: 22 September 2026

Revised: 13 October 2026

Accepted: 17 October 2026

Published: 21 October 2026

Keywords:

automated essay scoring, multi-rater annotation, inter-rater reliability, educational dataset

This is an open-access article under the CC-BY-SA License.



How to cite: Utomo, P.W., & Abidin, Z. (2026). Developing a Multi-Rater Indonesian Essay Scoring Dataset for Automated Essay Scoring. RESWARA: Jurnal Riset Ilmu Teknik, 4(4), 1732–1751. <https://doi.org/10.70716/reswara.v4i4.908>

INTRODUCTION

Artificial Intelligence (AI) has become increasingly important in educational technology, including automated assessment. Automated Essay Scoring (AES) aims to evaluate written responses automatically and can help address the time and workload associated with manual essay scoring (Dikli, 2006; Page, 1966; Shermis & Burstein, 2013). Advances in Natural Language Processing (NLP), particularly transformer-based language models, have improved the ability of AES systems to represent the semantic and contextual characteristics of student writing (Ludwig et al., 2021; Li et al., 2022; Ramesh & Sanampudi, 2022; Sun et al., 2025).

A key challenge in AES research is the systematic development and documentation of assessment datasets. Information about assessment instruments, participant characteristics, scoring rubrics, annotation procedures, and data preparation is necessary to understand how reference scores are produced and to assess the reproducibility of research findings. Research on data-centric artificial intelligence has emphasized the importance of dataset quality, documentation, and responsible data management in the development of reliable AI systems (Bender & Friedman, 2018; Gebru et al., 2021;

Dodge et al., 2021). In educational assessment, these considerations are particularly important because the data represent student performance and the judgments used to evaluate it.

Another challenge concerns the reliability of human annotation. Essay scoring involves evaluative judgments that may vary according to raters' severity, experience, interpretation of scoring criteria, and decision-making processes (Leckie & Baird, 2011; Wiseman, 2012; Zhang, 2016; Huang & Chen, 2022). Consequently, a dataset intended for AES research should document its scoring procedures and provide empirical evidence of agreement among raters. Without such evidence, the consistency of the reference scores may be difficult to evaluate, potentially complicating the interpretation and comparison of subsequent AES experiments (Williamson et al., 2012; Shermis & Burstein, 2013).

These considerations motivate the development of well-documented Indonesian-language essay assessment datasets with explicit multi-rater annotation procedures and reliability evaluation. The present study addresses this need by developing and validating a multi-rater Indonesian essay assessment dataset derived from senior high school Indonesian language assessments. The dataset contains responses from 360 students to 10 essay questions, resulting in 3,600 response-level records. Each response was independently and blindly scored by three Indonesian language teachers using a common analytical scoring rubric. The individual rater scores were retained, and their arithmetic mean was used as the reference score.

This study makes three main contributions. First, it documents a structured dataset development process encompassing assessment instrument design, essay response collection, rubric-based annotation, and dataset curation. Second, it evaluates annotation reliability using the average-measures absolute-agreement Intraclass Correlation Coefficient, ICC(A,k), and examines question-level reliability and the magnitude of score differences among raters. Third, it preserves individual rater scores alongside the aggregated reference scores, allowing the annotation structure and scoring consistency to be examined in subsequent research. Together, these contributions provide an empirical foundation for investigating Indonesian AES methods and human-AI scoring agreement. The dataset is currently retained for internal research purposes and is not publicly available. Figure 1 summarizes the broad evolution of Automated Essay Scoring research, from rule-based and machine-learning approaches to deep learning, transformer-based methods, and the growing emphasis on data quality.

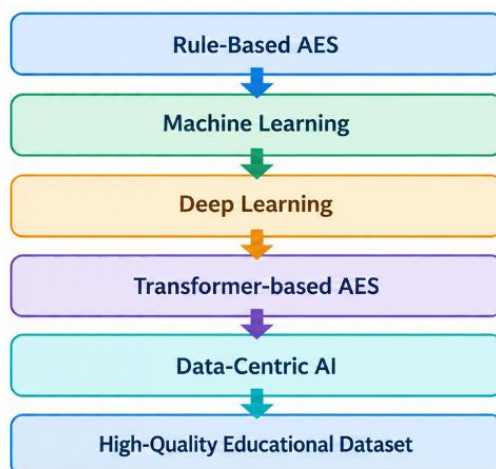


Figure 1. Evolution of Automated Essay Scoring Research

Figure 1 summarizes the broad evolution of AES research, from rule-based and machine-learning approaches to deep learning, transformer-based methods, and the growing emphasis on data quality. Despite the growing recognition of dataset quality in educational AI, the existing literature indicates that the development of educational datasets for Automated Essay Scoring remains substantially underreported compared with advances in scoring algorithms. Most published AES studies primarily emphasize model architectures, feature engineering, or predictive performance, while providing comparatively limited attention to the systematic construction and validation of the underlying assessment datasets (Ramesh & Sanampudi, 2022; Sun et al., 2025). Critical stages such as assessment instrument design, participant selection, rubric development, annotation workflow, data preprocessing, and quality assurance are frequently summarized in only a few paragraphs, making it difficult for other researchers to reproduce the dataset construction process or to evaluate the credibility of the resulting ground-truth labels (Bender & Friedman, 2018; Gebru et al., 2021; Dodge et al., 2021). Consequently, the lack of methodological transparency has become one of the major challenges limiting reproducibility and long-term usability in educational AI research.

Another important issue concerns the reliability of human annotation. Since essay assessment inherently involves subjective judgment, the quality of an AES dataset is largely determined by the consistency of the reference scores assigned by human raters. Previous research has demonstrated that rater severity, scoring tendencies, experience, and decision-making processes can introduce systematic variation into essay scores (Leckie & Baird, 2011; Wiseman, 2012; Zhang, 2016; Huang & Chen, 2022). In many cases, the final scores are presented as ground truth without sufficient explanation regarding scorer training, scoring guidelines, disagreement resolution, or inter-rater agreement analysis. Such limitations raise concerns regarding the validity of datasets used for model development and make it difficult to compare findings across different studies. Therefore, annotation quality should be empirically evaluated using appropriate reliability measures before educational datasets are used as reference resources for machine learning applications (Koo & Li, 2016; Powers et al., 2015; Cohen et al., 2018).

For educational AES research, the development of a useful dataset requires more than collecting student responses and assigning scores. The dataset development process should document the assessment instrument, scoring rubric, annotation procedure, and the consistency of human ratings. These elements are important because the resulting scores may serve as reference labels in subsequent model training and evaluation. In the Indonesian educational context, a documented essay assessment dataset with preserved independent ratings and reported inter-rater reliability can provide a basis for further AES research. The present study addresses this need by integrating dataset construction, independent multi-rater annotation, and statistical reliability evaluation within a single development framework.

Table 1. Research Gaps and Proposed Solutions

Research Gap	Evidence from Previous Studies	Proposed Solution
Dataset construction is insufficiently documented	Limited methodological reporting of dataset development	A systematic framework covering assessment design, data collection, annotation, validation, and documentation
Annotation procedures are inconsistently reported	Lack of transparent multi-rater protocols	Standardized annotation protocol involving three independent teachers
Annotation reliability is not consistently validated	Limited reporting of inter-rater reliability statistics	ICC-based validation of annotation consistency
Indonesian educational AES datasets remain limited	Few comprehensively documented benchmark datasets	Development of a multi-rater Indonesian essay assessment dataset
Limited dataset transparency and reusability	Incomplete metadata and documentation	Comprehensive dataset documentation supporting future reuse

Table 1 summarizes the main dataset-development gaps addressed in this study and the corresponding methodological responses. To address these gaps, this study develops and evaluates a multi-rater Indonesian essay assessment dataset for AES research. The dataset contains 3,600 essay responses from 360 senior high school students who answered 10 Indonesian Language questions. Each response was independently scored by three Indonesian language teachers using a common analytical scoring rubric. Annotation reliability was assessed using the average-measures absolute-agreement Intraclass Correlation Coefficient, ICC(A,k), together with question-level reliability analysis and maximum pairwise score differences. These analyses provide empirical evidence of scoring consistency across the assessment items.

The contribution of this study is methodological as well as empirical. It documents a dataset development framework that combines assessment instrument design, standardized multi-rater annotation, dataset curation, and statistical reliability evaluation. By preserving the three individual ratings alongside an aggregated reference score, the dataset structure also allows future researchers to examine rater variability and compare alternative reference-score construction strategies. However, the present study focuses on dataset development and annotation reliability; it does not directly evaluate AES model performance or establish the validity and generalizability of automated scores. Because the dataset is not publicly available, its current contribution lies in the documented development process and reliability evidence rather than in providing an openly accessible benchmark.

Figure 2 positions the present study within this data-centric perspective by linking educational assessment, assessment instrument design, multi-rater annotation, and ICC-based reliability validation to the development of an Indonesian AES dataset.

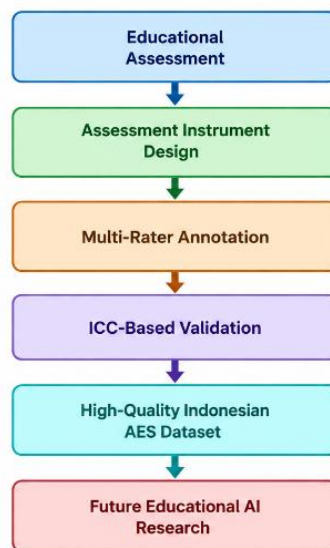


Figure 2. Positioning of the Proposed Study

RESEARCH METHOD

Dataset Development Design

This study employed a systematic dataset development design to construct and validate a multi-rater Indonesian essay assessment dataset for Automated Essay Scoring (AES). The development process was designed to integrate principles from educational assessment, natural language processing, dataset curation, and inter-rater reliability analysis. Rather than treating dataset preparation as a preliminary technical step, the study positioned dataset development as a structured research process involving assessment instrument design, data collection, scoring rubric development, multi-rater annotation, data quality validation, and dataset documentation.

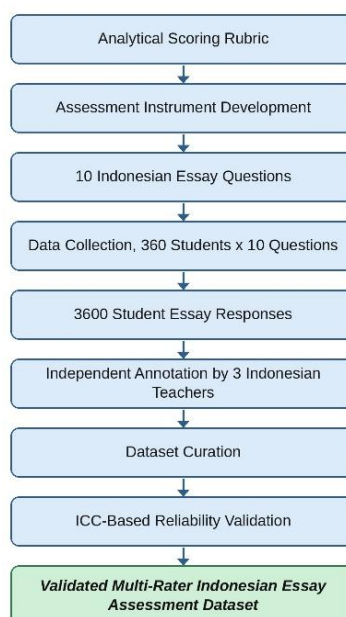


Figure 3. Dataset Development and Validation Workflow

The dataset development workflow consisted of five main stages: (1) assessment instrument development, (2) essay response collection, (3) rubric-based multi-rater annotation, (4) annotation reliability validation, and (5) dataset curation and documentation. This workflow was designed to ensure that the resulting dataset was not only sufficiently large for subsequent AES experiments but also educationally meaningful, systematically annotated, and supported by empirical evidence of scoring reliability. The overall dataset development and validation workflow is presented in Figure 3.

Assessment Instrument and Data Collection

The dataset was developed from essay-based assessments in the Indonesian language subject at the senior high school level. The assessment instrument consisted of 10 essay questions designed for students in Grades XII. The questions were selected to represent open-ended responses requiring students to construct answers using their own language rather than selecting from predetermined response options. This characteristic makes the resulting data suitable for AES research because the responses contain variations in content, wording, sentence structure, completeness, and language use.

A total of 360 students participated in the assessment. Each student responded to all 10 essay questions, resulting in 3,600 student essay responses. The resulting dataset therefore represents a response-level structure in which each observation corresponds to one student response to one essay question. This structure allows the dataset to be analyzed both globally and at the individual-question level, which is important for evaluating the consistency and generalizability of automated scoring models across different assessment prompts.

Table 2. Dataset Characteristics

Characteristic	Description
Educational level	Senior high school
Subject	Indonesian Language
Number of students	360
Number of essay questions	10
Total essay responses	3,600
Response format	Open-ended essay responses
Number of raters	3 certified Indonesian language teachers
Scoring scale	1–10 analytical scoring scale
Annotation type	Independent and blind multi-rater scoring
Reference score	Arithmetic mean of three independent rater scores
Scoring instrument	Analytical scoring rubric
Reliability validation	Intraclass Correlation Coefficient (ICC(A,k))
Primary intended use	Automated Essay Scoring research and educational assessment

The data collection process followed the structure of the educational assessment activity in which the essay responses were originally produced. The collected responses were subsequently organized into a structured dataset containing the question identifier, question text, student response, and scores assigned by the participating raters. Student responses were retained as the primary textual data for subsequent AES experiments, while the human scores served as reference labels for dataset validation and future model evaluation.

Scoring Rubric Development

A scoring rubric was developed to provide a standardized basis for evaluating the quality of student essay responses. The rubric was designed to support consistent scoring across the 10 essay questions while preserving the characteristics of analytical assessment. Analytical scoring rubrics can provide raters with explicit criteria for evaluating different aspects of written performance and can contribute to greater scoring consistency when the criteria are clearly operationalized (Barkaoui, 2007; East, 2009). The scoring criteria considered the extent to which a response addressed the expected content, demonstrated an appropriate level of completeness, and used language consistent with the requirements of the assessment.

Table 3. Analytical Scoring Rubric Used in the Multi-Rater Annotation Process

Score range	Category	Primary assessment aspect	Operational description
9–10	Excellent	Content accuracy, response completeness, and conceptual understanding	Highly accurate response covering all essential points with coherent and conceptually correct explanation.
7–8	Good	Content accuracy and substantial completeness	Generally accurate response covering most essential points with only minor omissions.
5–6	Fair	Partial understanding and limited completeness	Response covers only part of the expected content and may contain minor conceptual inaccuracies.
3–4	Poor	Limited understanding and low relevance	Very limited response with inadequate explanation or weak conceptual understanding.
1–2	Very poor	Irrelevance to the assessment task	Response is irrelevant, extremely incomplete, or provides no accessible information.

The rubric provided a common scoring framework for all raters throughout the annotation process. Its purpose was not to eliminate all differences in human judgment, as variation among raters is inherent in subjective assessment, but to establish a shared interpretation of the scoring scale. A standardized rubric is particularly important in AES dataset development because inconsistent human scoring can introduce noise into reference labels and subsequently affect model training and evaluation (Wind et al., 2018; Cohen et al., 2018).

Before the annotation process, the raters were provided with the scoring criteria and the expected interpretation of the rubric. The resulting scores were subsequently analyzed to determine whether the common scoring framework produced sufficiently consistent judgments across the three raters.

Multi-Rater Annotation Protocol

Each student essay response was independently evaluated by three Indonesian language teachers from public senior high schools. All three raters held a bachelor's degree (S1) in Indonesian Language Education and had professional teaching certification. Each rater had approximately 10–20 years of teaching experience and regularly participated in essay assessment, including the assessment of Grade XII school examinations. The use of experienced and qualified teachers was intended to ensure that the annotation process was conducted by raters with relevant subject-matter expertise and established experience in evaluating Indonesian language essays. The use of experienced raters is

also relevant because previous research has shown that rater experience and individual scoring tendencies can influence essay scoring outcomes (Leckie & Baird, 2011; Wiseman, 2012; Zhang, 2016).

The annotation protocol consisted of three main principles. First, all raters used the same analytical scoring rubric and assessment criteria. Second, each response was scored independently so that the initial score assigned by one rater would not directly influence the judgment of another rater. Third, the scoring process followed a blind-scoring procedure: student identities were concealed from the raters, and each rater did not have access to the scores assigned by the other raters during the annotation process. This combination of independent and blind scoring was implemented to minimize potential influence from other raters and from student identity during assessment (Shermis et al., 2015).

The resulting dataset therefore contains three independent human scores for each of the 3,600 student responses. The individual scores were retained separately in the dataset rather than immediately collapsed into a single label. This multi-rater structure allows the reliability of the annotation process to be empirically evaluated, enables comparison between individual rater scores and the aggregated reference score, and provides a richer basis for studying human-AI agreement in future AES experiments.

Multi-rater annotation is particularly important for essay assessment because human scoring involves interpretive judgment and may be affected by differences in scoring severity, interpretation of rubric criteria, and evaluation preferences (Artstein & Poesio, 2008; Gwet, 2021).

Dataset Structure and Data Curation

Following annotation, the collected data were organized into a structured tabular dataset. Each row represented one student response to one essay question. The primary data fields included the student identifier, question identifier, question text, student response, the three individual rater scores, and the aggregated reference score.

Table 4. Dataset Schema

Field	Description
no_id	Unique response identifier
no_siswa	Student identifier
no_soal	Essay question identifier
teks_soal	Essay question text
jawaban_murid	Student essay response
skor_g1	Score assigned by Rater 1
skor_g2	Score assigned by Rater 2
skor_g3	Score assigned by Rater 3
rerata_skor	Aggregated reference score

The aggregated reference score (rerata_skor) was calculated as the arithmetic mean of the three independent rater scores. This aggregated score was used as the primary reference label for supervised Automated Essay Scoring experiments because it represents the collective judgment of multiple qualified human evaluators while reducing the influence of individual rater variability. The use of aggregated human ratings as reference information is consistent with the broader concern in AES research that automated scores should be evaluated against carefully established human scoring standards (Powers et al., 2002; Powers et al., 2015; Cohen et al., 2018)

$$\text{Reference Score} = \frac{s_1 + s_2 + s_3}{3}$$

The dataset was curated to ensure consistency of structure and completeness of the information required for subsequent AES research. The curation process included checking the correspondence between student identifiers and question identifiers, verifying the presence of student responses and rater scores, standardizing the organization of textual and numerical fields, and preserving the relationship between each response and its corresponding assessment item.

The final dataset contains 3,600 response-level records generated from 360 students answering 10 essay questions. Because the dataset retains the original scores assigned by all three raters, it can support different research configurations. Researchers may use the individual rater scores for multi-rater reliability analysis, use an aggregated score as a reference label for supervised AES training, or compare automated predictions with individual human scores to investigate human–AI agreement.

Dataset curation was conducted with emphasis on transparency and future reusability. This principle is consistent with broader recommendations for scientific data management, which emphasize that datasets should be sufficiently documented and structured to support discovery, interpretation, and reuse (Wilkinson et al., 2016; Bender & Friedman, 2018; Gebru et al., 2021).

Annotation Reliability Validation

The reliability of the three-rater annotation process was evaluated using the Intraclass Correlation Coefficient (ICC). ICC was selected because the study involved continuous or approximately continuous essay scores assigned by multiple raters to the same set of responses. Unlike measures designed primarily for categorical agreement, ICC is appropriate for evaluating the degree to which multiple raters provide consistent and reliable quantitative measurements of the same targets (Koo & Li, 2016).

The primary reliability statistic reported in this study was the average-measures absolute-agreement Intraclass Correlation Coefficient (ICC(A,k)), calculated from the scores assigned by the three independent raters to the same student responses. ICC(A,k) was selected because the study employed a fixed set of three raters and the resulting reference score was constructed as the arithmetic mean of their scores. The absolute-agreement formulation was used because the analysis was intended to assess whether the raters assigned sufficiently similar scores to the same responses, rather than merely maintaining a consistent ranking of responses. The average-measures form was appropriate because the final reference score was derived from the combined scores of all three raters. Thus, ICC(A,k) provided an estimate of the reliability of the aggregated three-rater score used as the reference annotation for subsequent AES research (Koo & Li, 2016; Gwet, 2021).

Table 5. Reliability Analysis Framework

Analysis Level	Data Used	Purpose
Overall dataset	Scores from all 3 raters across 3,600 responses	Evaluate overall annotation reliability
Question-level	Scores from all 3 raters for each essay question	Examine reliability across assessment items
Reference score	Aggregated human scores	Provide labels for subsequent AES research

The ICC(A,k) analysis was conducted using the scores independently assigned by the three raters to the same student responses. The analysis evaluated the degree of absolute agreement among the raters and provided empirical evidence regarding the reliability of the resulting reference annotations. In addition to an overall reliability analysis, the annotation scores were examined at the question level to identify potential differences in scoring consistency across the 10 assessment items.

In addition to ICC analysis, inter-rater score differences were examined at the response level. For each student response, the maximum pairwise score difference was calculated as the difference between the highest and lowest scores assigned by the three raters. This measure was used to summarize the extent of scoring variation among raters for each response and was subsequently reported as a frequency distribution across the 3,600 annotated responses.

The ICC results were interpreted according to established reliability guidelines. Higher ICC values indicate stronger agreement and consistency among raters, whereas lower values indicate greater variability in the scoring judgments. The interpretation of ICC was conducted with consideration of the study design, the number of raters, and the intended use of the scores as reference labels for AES research. Appropriate selection and reporting of ICC specifications are important because different ICC models may reflect different assumptions regarding raters, targets, and the intended form of reliability (Koo & Li, 2016).

Dataset Quality Assurance and Documentation

Dataset quality assurance was conducted by combining structural checks with statistical validation of the annotation process. Structural quality assurance focused on the completeness and consistency of the dataset records, including the correspondence between questions, student responses, and rater scores. Statistical quality assurance focused on determining whether the three human raters produced sufficiently consistent scores across the response set.

The dataset documentation was designed to describe the origin and structure of the data, the assessment instrument, the scoring rubric, the annotation procedure, the participating rater structure, and the reliability validation procedure. This approach follows the broader movement toward transparent dataset documentation in artificial intelligence and natural language processing research, in which information about dataset development and intended use is considered essential for reproducibility and responsible reuse (Bender & Friedman, 2018; Gebru et al., 2021).

The final dataset was therefore considered ready for subsequent AES research only after the data structure had been organized, the multi-rater annotations had been retained, and the reliability of the scoring process had been empirically evaluated. The resulting resource was designed to support both model development and model evaluation, including future experiments involving contextual language models, traditional machine learning baselines, and human-AI agreement analysis.

RESULTS AND DISCUSSION

Dataset Development Outcome

The dataset development process resulted in a multi-rater Indonesian essay assessment dataset containing 3,600 response-level records. These records were generated from 360 senior high school students, each of whom responded to 10 Indonesian language essay questions. The resulting structure

provides a balanced relationship between the number of participants and assessment items, with each student contributing responses to the complete set of essay questions.

Table 6. Overall statistics of the Indonesian essay assessment dataset

Statistic	Value
Students	360
Essay questions	10
Total responses	3,600
Human raters	3
Overall ICC (A,k)	0.9905
95% CI	[0.9897, 0.9913]
Mean question-level ICC	0.9899
Highest question-level ICC	0.9934
Lowest question-level ICC	0.9803
Responses with ≤ 1 point difference	96.83%

The dataset was designed to preserve the relationship between the assessment instrument, the student response, and the human evaluation assigned to that response. Each record contains the corresponding essay question, the student answer, and three independent scores assigned by Indonesian language teachers. This structure is important because educational datasets should not only provide final labels but also preserve information about how those labels were generated. As emphasized by Bender and Friedman (2018), documentation of data provenance and annotation context is essential for interpreting NLP datasets, while the principles proposed by Gebru et al. (2021) further emphasize the importance of documenting dataset creation and intended use.

The resulting dataset can therefore be used in multiple research configurations. The student responses can serve as input data for AES models, while the individual rater scores can be used for inter-rater reliability analysis, human-AI agreement analysis, or alternative ground-truth construction strategies. The availability of three independent scores for every response also allows researchers to examine the extent to which automated predictions correspond to individual human judgments rather than relying exclusively on a single aggregated score. These potential research applications describe the analytical capabilities of the dataset structure; they do not imply that the dataset is currently publicly accessible. Access is presently restricted to internal research purposes.

Multi-Rater Annotation Results

Each of the 3,600 student responses was independently evaluated by three Indonesian language teachers. The use of multiple raters was intended to reduce the influence of individual scoring preferences and to provide a more reliable basis for establishing reference scores. This design is consistent with the broader principle that subjective annotation tasks should preserve information about annotator variation rather than treating a single human judgment as unquestionably correct.

The annotation process generated three independent score distributions for the same set of responses. This structure enables the dataset to represent not only the final assessment outcome but also the degree of agreement among human evaluators. In educational assessment, such information is particularly important because variation between raters may reflect differences in scoring severity, interpretation of rubric criteria, or evaluation preferences. Artstein and Poesio (2008) emphasize that

annotation agreement should be treated as an empirical property of the annotation process rather than assumed in advance.

The use of a shared scoring rubric provided a common basis for evaluating the responses. However, the rubric did not eliminate the possibility of differences in human judgment. This is methodologically important because a high-quality AES dataset should reflect authentic assessment conditions while simultaneously demonstrating that the resulting annotations are sufficiently consistent for use as reference labels. Consequently, the present study evaluated annotation reliability statistically rather than assuming that the use of a common rubric alone guaranteed consistent scoring.

Inter-Rater Reliability

The reliability analysis demonstrated a very high level of agreement among the three raters. The overall ICC value was 0.9905, indicating excellent inter-rater reliability. The question-level analysis also produced consistently high reliability values, with an average ICC of 0.9899, a highest value of 0.9934, and a lowest value of 0.9803.

The high overall ICC(A,k) indicates that the three raters produced highly consistent scores across the evaluated student responses. This consistency may be associated with the use of a common analytical scoring rubric, the application of the same scoring scale, and the independent evaluation of responses under a shared assessment framework. These procedures provided a common basis for interpreting student answers and assigning scores. Nevertheless, the present study did not experimentally isolate the contribution of each procedure; therefore, these factors should be interpreted as plausible explanations rather than demonstrated causes of the high reliability.

The consistently high question-level ICC values further indicate that the observed reliability was not limited to the dataset as a whole but was also evident across the ten essay questions. The accompanying analysis of maximum pairwise score differences provides a complementary perspective by showing how closely the raters' scores aligned at the response level. Whereas ICC(A,k) evaluates the reliability of the aggregated ratings relative to variation across responses, the score-difference analysis describes the magnitude of disagreement in the original scoring scale. Considering both measures therefore provides a more informative account of annotation consistency than relying on a single statistic.

These results indicate that the three raters produced highly consistent assessments across the 3,600 student responses. The narrow range between the lowest and highest question-level ICC values further suggests that annotation reliability was not restricted to a particular essay question. Instead, the high reliability was observed consistently across the assessment instrument.

The use of three independent raters allowed the dataset to preserve multiple human judgments for each essay response. This structure is valuable because essay scoring involves evaluative decisions, and individual raters may differ in their interpretation of rubric criteria or scoring severity. Retaining the individual scores makes these differences observable and allows future researchers to investigate rater variability rather than treating a single score as an unquestionable representation of response quality.

The arithmetic mean of the three scores was used as the reference score for subsequent AES research. Aggregating the ratings can reduce the influence of idiosyncratic judgments from any single rater and provides a consistent target for supervised model training and evaluation. However, the mean is an operational choice for constructing a reference score, not proof that it represents an error-

free or uniquely correct assessment of each response. Preserving the original scores alongside the mean is therefore important because it enables future studies to compare alternative aggregation methods, examine disagreement patterns, and evaluate automated predictions against both the aggregated reference score and individual human ratings.

The interpretation of these results is supported by established guidance on ICC reporting. Koo and Li (2016) emphasize that ICC is appropriate for evaluating the reliability of quantitative ratings assigned to the same targets by multiple raters, although the selected ICC specification must be interpreted in relation to the research design. In the present study, the high ICC values provide empirical evidence that the multi-rater annotation process generated sufficiently consistent reference scores for subsequent AES research. This finding is important because the quality of human reference scores is directly relevant to the validity and psychometric interpretation of automated scoring systems (Powers et al., 2015; Wind et al., 2018; Shermis, 2022)

The results also support the methodological decision to retain the scores assigned by all three raters rather than immediately replacing them with a single score. Preserving individual human ratings is particularly valuable because human scoring itself may contain meaningful variation that can be examined separately from the aggregated reference score (Leckie & Baird, 2011; Huang & Chen, 2022). Although an aggregated score may subsequently be used as a reference target for supervised learning, the original annotations remain valuable for evaluating the reliability of the assessment process. This approach is consistent with the broader perspective of inter-rater reliability research, in which agreement is not merely a descriptive statistic but an important indicator of the quality of the measurement process (Gwet, 2021).

Distribution of Rater Score Differences

The analysis measured the difference between the highest and lowest scores assigned by the three raters for each response. In addition to the ICC analysis, the distribution of score differences was examined to provide a more direct description of the magnitude of disagreement. Among the 3,600 responses, 1,233 (34.25%) had a score difference of zero, while 2,253 (62.58%) had a score difference of one point. A further 110 responses (3.06%) had a score difference of two points, and only 4 responses (0.11%) had a score difference of three points. Figure 4 illustrates the distribution of these maximum pairwise score differences.

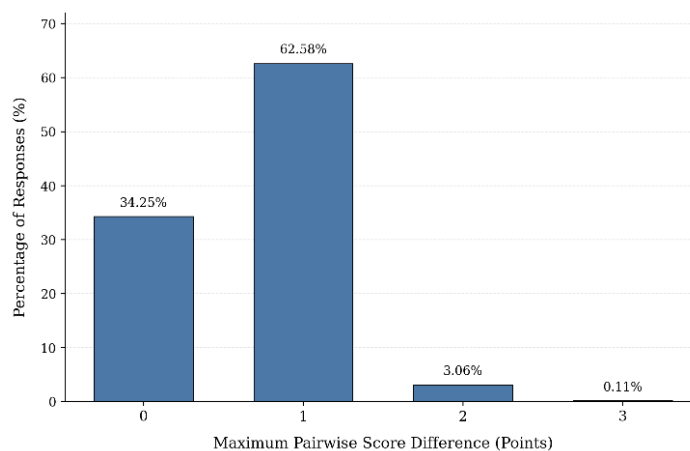


Figure 4. Distribution of Inter-Rater Score Differences

Table 7. Distribution of Inter-Rater Score Differences

Score Difference	Frequency	Percentage
0	1,233	34.25%
1	2,253	62.58%
2	110	3.06%
3	4	0.11%

This distribution provides an important complement to the ICC results. While ICC summarizes the overall reliability of the ratings, the distribution of score differences illustrated how disagreement was manifested at the response level. The fact that 96.83% of responses had a maximum score difference of zero or one point indicates that disagreement was generally limited in magnitude.

This finding is particularly relevant for AES dataset development because substantial disagreement among human raters can introduce uncertainty into the ground-truth labels used for model training and evaluation. As noted in discussions of annotation reliability by Artstein and Poesio (2008), the interpretation of agreement should consider not only whether annotators agree but also the extent and practical significance of their disagreements (Artstein & Poesio, 2008; Gwet, 2021). In the present dataset, the small magnitude of most score differences supports the conclusion that the reference annotations were highly consistent.

Question-Level Reliability and Assessment Consistency

The question-level ICC results indicate that the reliability of the annotation process remained high across the 10 essay questions. The average ICC of 0.9899 and the relatively narrow range between 0.9803 and 0.9934 suggest that the scoring rubric was applied consistently across different assessment items

Figure 5 presents the ICC(A,k) values for the ten essay questions and illustrates the consistently high annotation reliability observed across the assessment items.

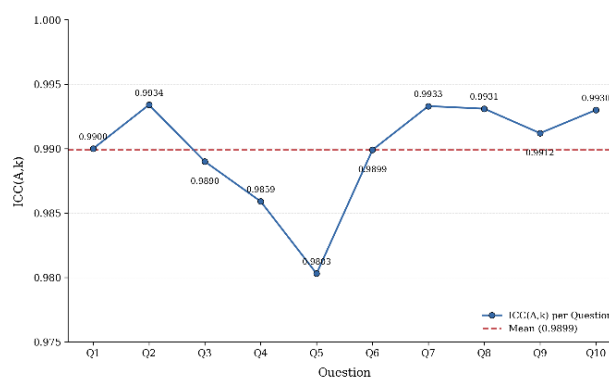


Figure 5. Question-level ICC(A,k) values

This result is important because essay questions may differ in terms of content, expected response characteristics, and potential ambiguity. A scoring framework that produces reliable results across multiple questions is therefore more valuable than one that performs well only for a single assessment item. The consistency observed in this study suggests that the combination of the assessment instrument and scoring rubric provided a sufficiently stable basis for multi-rater annotation.

Table 8. Question-level and overall ICC(A,k) values with 95% confidence intervals for the three-rater annotation process

Essay question	ICC(A,k)	95% CI	Interpretation
1	0.9900	[0.9887, 0.9921]	Excellent
2	0.9934	[0.9926, 0.9948]	Excellent
3	0.9890	[0.9874, 0.9912]	Excellent
4	0.9859	[0.9837, 0.9886]	Excellent
5	0.9803	[0.9775, 0.9843]	Excellent
6	0.9899	[0.9882, 0.9918]	Excellent
7	0.9933	[0.9922, 0.9946]	Excellent
8	0.9931	[0.9919, 0.9944]	Excellent
9	0.9912	[0.9898, 0.9929]	Excellent
10	0.9930	[0.9919, 0.9944]	Excellent
Overall	0.9905	[0.9897, 0.9913]	Excellent

From an educational measurement perspective, reliability is an essential component of the credibility of assessment data. Shermis (2022) and Ferrara and Qunbar (2022) emphasize that the interpretation of automated scores should be supported by evidence concerning the validity of the underlying assessment and scoring process. Therefore, the high question-level reliability observed in this dataset strengthens its suitability for future AES experiments because the model targets are supported by a relatively stable human assessment process.

Dataset Readiness for Automated Essay Scoring Research

The results demonstrate that the developed dataset satisfies several important requirements for subsequent AES research. First, it contains a substantial number of response-level observations generated from authentic educational assessment. Second, each response is associated with three independent human scores, allowing the reliability of the annotation process to be examined empirically. Third, the high overall and question-level ICC values provide evidence that the human annotations are sufficiently consistent to serve as reference information for machine learning research.

Table 9. Dataset Readiness for Future AES Research

Dataset Property	Evidence from This Study	Research Value
Educational authenticity	360 students and 10 Indonesian language essay questions	Represents authentic educational assessment responses
Multi-rater annotation	Three independent teacher scores per response	Enables reliability and human–AI agreement analysis
Annotation reliability	Overall ICC = 0.9905	Supports the consistency of reference annotations
Response-level granularity	3,600 individual responses	Supports flexible model development and evaluation
Dataset documentation	Structured metadata and annotation information	Supports documentation and potential reuse subject to appropriate access approval

The dataset also offers methodological flexibility. Researchers may use the student responses as input data for traditional machine learning models, neural architectures, or transformer-based language models, enabling future comparisons across different automated writing evaluation

approaches (Bridgeman & Ramineni, 2017; Ludwig et al., 2021; Voss, 2025). This flexibility is consistent with the evolution of AES research from traditional feature-based and machine-learning approaches toward neural and transformer-based architectures (Beseiso et al., 2021; Ludwig et al., 2021; Li et al., 2022; Voss, 2025). The multi-rater structure further enables alternative evaluation strategies, including comparison between model predictions and individual raters, aggregated human scores, or measures of human–AI agreement (Buzick et al., 2016; Vo et al., 2023). This flexibility is particularly valuable in AES research because the performance of an automated system should ideally be interpreted in relation to the variability of human scoring rather than solely through a single numerical benchmark.

The dataset's potential value therefore extends beyond its use in the specific AES model investigated in the accompanying research. In accordance with the FAIR principles proposed by Wilkinson et al. (2016), a well-structured dataset should be sufficiently organized and documented to support future discovery, interpretation, and reuse. Similarly, the dataset documentation principles proposed by Gebru et al. (2021) highlight that information about data origins, collection processes, annotation procedures, and intended uses is essential for responsible dataset reuse.

Accordingly, the principal outcome of this study is not merely the production of 3,600 essay responses, but the establishment of a validated multi-rater assessment resource. By combining authentic Indonesian student essays, a structured scoring rubric, independent evaluations from three teachers, and empirical reliability validation, the dataset provides a stronger foundation for future research in Indonesian AES and educational NLP than a dataset containing text responses without documented annotation quality.

Implications for Indonesian Automated Essay Scoring Research

The findings have methodological implications for the development of Indonesian Automated Essay Scoring (AES) datasets. The high inter-rater reliability observed in this study suggests that the three raters assigned scores consistently across the ten essay questions. This consistency may be associated with the use of a common analytical scoring rubric, which provided a shared basis for interpreting student responses. However, the present results do not establish which aspects of the annotation procedure contributed most to the observed agreement. Further investigation would be needed to determine the respective effects of rubric design, rater characteristics, and response characteristics on scoring consistency.

Preserving the three individual ratings alongside their arithmetic mean also has important implications for reference-score construction. The mean provides a single numerical target that can be used in subsequent supervised AES experiments, while the original ratings retain information about variation in human judgment. This distinction matters because an aggregated score can simplify model training and evaluation but may conceal disagreements between individual raters. Retaining the original scores therefore allows future studies to compare model predictions with both the aggregated reference score and individual human judgments, as well as to investigate rater severity and alternative aggregation strategies. Nevertheless, the arithmetic mean should be understood as an operational reference score rather than an error-free representation of the objectively correct score.

For Indonesian educational NLP, the study contributes a documented dataset-development framework that combines assessment instrument design, rubric-based multi-rater annotation, dataset curation, and statistical reliability evaluation. The response-level structure and preserved annotations

provide a basis for future research comparing AES approaches and examining human-AI agreement across the ten assessment questions. However, the dataset was collected from 360 senior high school students and covers a specific educational level, subject, and set of essay prompts. Its findings should therefore not be assumed to generalize automatically to other schools, student populations, educational levels, or writing tasks. In addition, the dataset is currently retained for internal research purposes and is not publicly available. Its present contribution consequently lies in the documented development process and empirical reliability evidence, rather than in offering an openly accessible benchmark for independent model comparison.

Finally, high inter-rater reliability should not be equated with overall dataset validity or demonstrated AES performance. Reliability indicates the consistency of the ratings under the conditions examined; it does not, by itself, establish that the rubric fully captures the intended writing construct, that the scores are equally appropriate across different student populations, or that an automated model trained on the data will perform well. The present study therefore establishes evidence for annotation consistency, while construct validity, generalizability, and predictive performance remain important subjects for subsequent investigation. These distinctions are essential when interpreting the dataset's readiness for further research and when planning its evaluation in future Indonesian AES studies.

Limitations and Future Work

Several limitations should be considered when interpreting these findings. First, the dataset comprises 360 students from a senior high school assessment context and ten essay questions. Although the response-level dataset contains 3,600 records, these records are not equivalent to observations from 3,600 independent students. The current data therefore do not establish that the same annotation reliability would be obtained across other schools, educational levels, student populations, or writing tasks. Broader sampling is needed to examine the generalizability of the annotation procedure.

Second, the reliability analysis evaluates the consistency of human ratings, not the validity of the scoring rubric or the predictive performance of an AES model. High inter-rater reliability does not, by itself, establish that the scores fully represent the intended writing construct, nor does it demonstrate that an automated system trained on the dataset will achieve high accuracy. Future work should therefore investigate additional sources of validity evidence and evaluate AES models using appropriate held-out data and evaluation procedures.

Third, the study used three raters and the arithmetic mean of their scores as the reference score. Although this approach provides a consistent aggregate target, alternative aggregation strategies or additional raters may yield different estimates of response quality. Future research could compare the arithmetic mean with alternative reference-score construction methods and examine the effects of rater severity, response characteristics, and question-specific scoring patterns. Further studies should also investigate the validity of automated scores across different writing populations and assessment constructs, particularly when extending the dataset beyond the present senior high school context (Hannah et al., 2023; Ferrara & Qunbar, 2022).

Future work should expand the dataset across multiple schools, educational levels, assessment topics, and student populations to examine the generalizability of the annotation framework. Further studies could evaluate different AES approaches, including multilingual and generative AI-based

scoring methods, against the aggregated reference scores and individual human ratings. The preserved multi-rater annotations may also support investigations of rater variability, scoring severity, uncertainty, and human-AI agreement. Any external data sharing should be considered only under appropriate privacy, data-governance, and access arrangements.

CONCLUSION

This study developed a systematically documented multi-rater Indonesian essay assessment dataset to support future Automated Essay Scoring (AES) research. The dataset comprises 3,600 response-level records from 360 senior high school students who answered ten Indonesian language essay questions, each independently and blindly scored by three Indonesian language teachers using a common analytical rubric. Retaining individual ratings alongside an aggregated reference score preserves information on both the consensus and variability of human assessment. The reliability analysis showed a high level of scoring consistency: the overall average-measures absolute-agreement ICC(A,k) was 0.9905, question-level ICC values ranged from 0.9803 to 0.9934 (mean = 0.9899), and 96.83% of responses had a maximum score difference of no more than one point among the three raters. These results provide empirical evidence that the annotation procedure produced consistent ratings across the ten questions under the conditions examined.

The main contribution of this work is a documented dataset development and annotation framework supported by quantitative reliability evidence, not a demonstrated improvement in AES model performance. High inter-rater reliability should not be interpreted as proof of scoring validity, generalizability to other educational contexts, or the predictive performance of an automated scoring system. Future work should extend data collection to additional schools, student populations, educational levels, and essay prompts; evaluate AES models using appropriate held-out data; examine alternative reference-score construction strategies and rater variability; and gather further validity evidence. Because the dataset is currently restricted to internal research use, any external sharing should follow appropriate privacy, data-governance, and access arrangements.

REFERENCES

- Artstein, R., & Poesio, M. (2008). Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4), 555–596. <https://doi.org/10.1162/coli.07-034-R2>
- Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing*, 12(2), 86–107. <https://doi.org/10.1016/j.asw.2007.07.001>
- Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6, 587–604. https://doi.org/10.1162/tacl_a_00041
- Beseiso, M. H., Alzubi, O. A., & Rashaideh, H. (2021). A novel automated essay scoring approach for reliable higher educational assessments. *Journal of Computing in Higher Education*, 33(3), 727–746. <https://doi.org/10.1007/s12528-021-09283-1>
- Bridgeman, B., & Ramineni, C. (2017). Design and evaluation of automated writing evaluation models: Relationships with writing in naturalistic settings. *Assessing Writing*, 34, 62–71. <https://doi.org/10.1016/j.asw.2017.10.001>
- Buzick, H., Oliveri, M. E., Attali, Y., & Flor, M. (2016). Comparing human and automated essay scoring for prospective graduate students with learning disabilities and/or ADHD. *Applied Measurement in Education*, 29(3), 161–172. <https://doi.org/10.1080/08957347.2016.1171765>

- Cohen, Y., Levi, E., & Ben-Simon, A. (2018). Validating human and automated scoring of essays against "true" scores. *Applied Measurement in Education*, 31(3), 241–250. <https://doi.org/10.1080/08957347.2018.1464450>
- Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, 5(1).
- Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., ... & Gardner, M. (2021). Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 1286–1305). <https://doi.org/10.18653/v1/2021>.
- East, M. (2009). Evaluating the reliability of a detailed analytic scoring rubric for foreign language writing. *Assessing Writing*, 14(2), 88–115. <https://doi.org/10.1016/j.asw.2009.04.001>
- Ferrara, S., & Qunbar, S. (2022). Validity arguments for AI-based automated scores: Essay scoring as an illustration. *Journal of Educational Measurement*, 59(3), 288–313. <https://doi.org/10.1111/jedm.12333>
- Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86–92. <https://doi.org/10.1145/3458723>
- Gwet, K. L. (2021). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters: Vol. 2. Analysis of quantitative ratings* (5th ed.). Advanced Analytics, LLC.
- Hannah, L., Jang, E. E., Shah, M., & Gupta, V. (2023). Validity arguments for automated essay scoring of young students' writing traits. *Language Assessment Quarterly*, 20(4–5), 399–420. <https://doi.org/10.1080/15434303.2023.2288253>
- Huang, J., & Chen, G. (2022). Individualized feedback to raters in language assessment: Impacts on rater effects. *Assessing Writing*, 52, 100623. <https://doi.org/10.1016/j.asw.2022.100623>
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155–163. <https://doi.org/10.1016/j.jcm.2016.02.012>
- Leckie, G., & Baird, J.-A. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement*, 48(4), 399–418. <https://doi.org/10.1111/j.1745-3984.2011.00152.x>
- Li, X., Yang, H., Hu, S., Geng, J., Lin, K., & Li, Y. (2022). Enhanced hybrid neural network for automated essay scoring. *Expert Systems*, 39(10), e13068. <https://doi.org/10.1111/exsy.13068>
- Ludwig, S., Mayer, C., Hansen, C., Eilers, K., & Brandt, S. (2021). Automated essay scoring using transformer models. *Psych*, 3(4), 897–915. <https://doi.org/10.3390/psych3040056>
- Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan*, 47, 238–243.
- Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2002). Stumping e-rater: Challenging the validity of automated essay scoring. *Computers in Human Behavior*, 18(2), 103–134. [https://doi.org/10.1016/S0747-5632\(01\)00052-8](https://doi.org/10.1016/S0747-5632(01)00052-8)
- Powers, D. E., Escoffery, D. S., & Duchnowski, M. P. (2015). Validating automated essay scoring: A (modest) refinement of the "gold standard." *Applied Measurement in Education*, 28(2), 130–142. <https://doi.org/10.1080/08957347.2014.1002920>
- Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring system: A systematic literature review. *Artificial Intelligence Review*, 55(3), 2495–2527. <https://doi.org/10.1007/s10462-021-10068-2>
- Shermis, M. D. (2022). Anchoring validity evidence for automated essay scoring. *Journal of Educational Measurement*, 59(3), 314–337. <https://doi.org/10.1111/jedm.12336>
- Shermis, M. D., & Burstein, J. (Eds.). (2013). *Handbook of automated essay evaluation: Current applications and new directions*. Routledge. <https://doi.org/10.4324/9780203122761>

- Shermis, M. D., Lottridge, S., & Mayfield, E. (2015). The impact of anonymization for automated essay scoring. *Journal of Educational Measurement*, 52(4), 419–436. <https://doi.org/10.1111/jedm.12093>
- Sun, J., Song, T., Peng, W., & Song, J. (2025). A survey of automated essay scoring: Challenges, advances, and future. *Neurocomputing*, 650, 130916. <https://doi.org/10.1016/j.neucom.2025.130916>
- Vo, Y., Rickels, H., Welch, C., & Dunbar, S. (2023). Human scoring versus automated scoring for English learners in a statewide evidence-based writing assessment. *Assessing Writing*, 56, 100719. <https://doi.org/10.1016/j.asw.2023.100719>
- Voss, E. (2025). Comparison of traditional machine learning and neural network approaches for automated scoring of second language English essays. *Language Testing*, 42(4), 369–396. <https://doi.org/10.1177/02655322251348959>
- Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, 31(1), 2–13. <https://doi.org/10.1111/j.1745-3992.2011.00223.x>
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018. <https://doi.org/10.1038/sdata.2016.18>
- Wind, S. A., Wolfe, E. W., Engelhard, G., Jr., Foltz, P., & Rosenstein, M. (2018). The influence of rater effects in training sets on the psychometric quality of automated scoring for writing assessments. *International Journal of Testing*, 18(1), 27–49. <https://doi.org/10.1080/15305058.2017.1361426>
- Wiseman, C. S. (2012). Rater effects: Ego engagement in rater decision-making. *Assessing Writing*, 17(3), 150–173. <https://doi.org/10.1016/j.asw.2011.12.001>
- Zhang, J. (2016). Same text different processing? Exploring how raters' cognitive and meta-cognitive strategies influence rating accuracy in essay scoring. *Assessing Writing*, 27, 37–53. <https://doi.org/10.1016/j.asw.2015.11.001>