ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL

EXPLAINABLE AI IN LANGUAGE ASSESSMENT: INTERPRETING MACHINE LEARNING MODELS FOR ESL WRITING FEEDBACK

Qaisra Honey
qaisrahoney@gmail.com

Abstract

This paper explored the use of explainable artificial intelligence (XAI) in ESL writing assessment, by creating an interpretable machine learning model which can produce scores and explanatory feedback. The accuracy of the models and perceptions of the users were determined using a mixed-methods design. The research was done at The University of Lahore and the University of Management and Technology, Lahore with the help of purposive sampling where 100 ESL students and 10 teachers were taken. Human and AI scoring of essays was used to get quantitative data, and questionnaires and semi-structured interviews were used to get qualitative data. The results demonstrated that AI-generated scores and human ratings were significantly positive ($r = .86$, $p < .01$), which means that the model is highly reliable. The feedback was also seen by the participants as easy to understand, interpret, and pedagogically valuable, but some feelings of trust were not eliminated. The research proposes that explainable AI has the potential to improve the instructional quality and the transparency of automated writing assessment ESL.

Keywords

Explainable artificial intelligence; automated writing assessment; ESL writing; interpretable machine learning; feedback; mixed-methods research

Introduction

Background of the Study

The introduction of artificial intelligence (AI) in education has dramatically changed teaching, learning, and assessment in education. Automated Writing Assessment (AWA) systems have become a potent instrument of measuring written performance of learners in the field of second language (L2) writing in an efficient and consistent manner. These systems can process linguistic features like grammar, vocabulary, and coherence, which is fueled by machine learning and natural language processing techniques, and can give fast feedback to a learner (Shermis and Burstein, 2013). Due to the growing use of digital technologies in educational institutions, AI-based assessments tools are now a part of language classes.

Although they do have benefits, traditional AWA systems are usually criticized on account of lack of transparency. The majority of machine learning systems are black-box systems, i.e., users cannot easily reason about how the input features are converted into decisions on the output side (Adabi and Berrada, 2018). This interpretability is a critical problem in the sphere of education where assessment results have significant consequences on the learning process and evaluation. Students can get scores without knowing the reasons behind the score and educators may struggle to rely on such systems or incorporate them in their teaching.

Problem Statement

Although prior studies have shown that automated scoring systems can be as reliable as human scorers (Attali and Burstein, 2006; Rudner and Liang, 2002), these systems do not necessarily give meaningful explanations of the ratings. This weakness inhibits their pedagogical potentials, since feedback is a core of language learning. In the absence of clear and interpretable feedback, learners might be unable to determine their weaknesses and strengths, and teachers might not be able to rely on AI-generated insights to aid instruction. Moreover, transparency is not present, which raises the questions of fairness, responsibility, and trust in the user. Use of opaque AI systems in high-stakes educational environments can result in doubts and opposition by educators and learners. Therefore, the immediate priority is to change the focus of models that rely on accuracy to more reliable and understandable systems.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL

Research Gap

Although the explainable artificial intelligence (XAI) area of interest has experienced an increasing trend over the last few years, its application to the ESL writing assessment has not been studied in detail. Research done on AWA has mostly focused the research on the performance of the models and the accuracy of the models in scoring but little has been done on the interpretability and user experience (Shermis and Burstein, 2013). Moreover, not many studies have investigated the perception of AI-generated feedback by teachers and learners in classrooms. This gap indicates that the research is necessary that combines the views of both technical and pedagogical approaches through the assessment of the performance of explainable models and their effects on learning. In particular, there is not much empirical research on how interpretable AI systems may increase the clarity, usefulness, and didactic value of writing assessment.

Purpose of the Study

This study was aimed to examine the effectiveness of explainable artificial intelligence in ESL writing assessment. The objective of the study was to design and implement an interpretable machine learning model that could produce both scores and explanatory feedback as well as to test its performance against the performance of humans. The research, further, aimed at investigating the perception of clarity, usefulness, and pedagogical value of AI-generated feedback by teachers and learners.

Significance of the Study

This research work has theoretical and practical significance. Theoretically, it is relevant to the new area of XAI in education because it combines notions of interpretability and the field of language assessment. It builds on the current studies of automated writing assessment by highlighting the significance of transparency besides accuracy. In practice, the research is beneficial to ESL educators, students, and schools. The introduction of an explainable AI model provides the research with a pattern of producing feedback, which is not only accurate but can be understood and acted upon. This has the potential to advance the writing abilities of the learners and assist teachers to provide effective and meaningful evaluation. Moreover, the results can guide the development of AI-based learning systems that can meet pedagogical objectives and ethical requirements. This paper makes contributions to the literature in three aspects. First, it builds upon the research on automated writing assessment by introducing interpretability to machine learning-based scoring. Second, it integrates technical assessment and teacher and learner perception in a mixed-method design. Third, it provides a real-life situation of ESL higher education in Pakistan, where the study of explainable AI in language evaluation is scarce.

Literature Review

The implementation of automated writing assessment (AWA) in second language writing teaching has changed the practice of writing evaluation by facilitating timely and scalable and more advanced assessment procedures. Initial AWA studies determined that automated scoring was able to estimate human ratings with moderate consistency especially in large-scale assessment contexts (Attali & Burstein, 2006; Rudner & Liang, 2002; Shermis & Burstein, 2013). The main focus of these early systems was efficiency and scoring reliability. More recent studies have however broadened the scope of scoring to how AWA can contribute to formative feedback and pedagogical support (Ding & Zou, 2024; Shi & Aryadoust, 2024) as well.

Modern research indicates that AWA systems can facilitate grammatical accuracy improvements, revision behavior, and writing fluency with the effective integration into instruction (Wei et al., 2023; Li, 2021). Research syntheses have also indicated that writing platforms like Grammarly, Criterion, and Pigai can play a beneficial role in the development

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

of learner writing, but their use is highly contingent on the context, engagement between the learner and the teacher, and mediation (Ding & Zou, 2024; Ebn-Abbasi et al., 2024). These results imply that AWA is becoming more and more regarded not only as an assessment tool, but also as a constituent of writing pedagogy.

Meanwhile, researchers warn that the efficiency of AWA is not just influenced by the technical quality of the system but also the ways in which learners perceive and respond to automated feedback (Ranalli, 2021; Liu, 2024). Therefore, recent studies have ceased focusing on whether automated systems can score writing correctly to how it can be used to promote meaningful learning.

One of the most criticism of AI-based assessment systems is its opaqueness despite improvements in automated scoring. Most machine learning and deep learning models are black-box systems, the inner workings of which are hard to interpret as decisions and can be made by users (Adabi & Berrada, 2018; Kumar & Boulanger, 2021). Despite the possible accuracy of the prediction generated by such systems, users are not always able to figure out how specific linguistic characteristics are used to produce scores or feedback.

Such a lack of transparency poses substantial problems in the educational setting, where the results of the assessment are supposed to be understandable, impartial, and educational. Uto (2021) asserted that the growth in complexity in model-based automated essay scoring has raised the issue of explainability and consistency with assessment objectives. In a similar vein, Kumar and Boulanger (2021) have shown that the use of deep learning-based scoring models that are highly accurate can be challenging to interpret without a clearly articulated rubric level.

These limitations are especially problematic in the field of language assessment since feedback is supposed to help one improve and learn, rather than simply to score. In cases where a learner is provided with automated evaluations without the reasons behind it, the pedagogical usefulness of feedback can be lost (Hyland & Hyland, 2006; Shi & Aryadoust, 2024). Such considerations have led to the growing popularity of explainable methods of AI-based assessment.

As a reaction to opaque artificial intelligence systems, explainable artificial intelligence (XAI) has been introduced. XAI aims to render algorithmic decisions explainable through offering explanations of how prediction is made and what features are used to drive results (Adabi & Berrada, 2018). This interpretability has gained greater significance in the field of education since learners and teachers need technically valid and pedagogically significant explanations (Khosravi et al., 2022).

New research indicates explainability is emerging as the key to legitimacy and uptake of educational AI systems. The authors of the study by Khosravi et al. (2022) have claimed that XAI in the educational field should address the concepts of trust, fairness, usability, and instructional relevance. Turkmen (2025) also discovered that explainability techniques are being utilized more often to promote transparency in educational AI studies and that transparency is closely linked to user trust.

Whilst XAI has been used more widely in the educational field of AI more generally, its use in language testing is not as well-developed. Ribeiro-Flucht et al. (2024) have also shown that explainable AI can be associated with interpretable linguistic evidence in language learning, providing a model of how one can make judgments using automated methods more transparent. This holds especially true to ESL writing assessment, in which it is advantageous to make visible and comprehensible criteria like lexical diversity, syntactic complexity, cohesion, and grammatical accuracy.

In recent studies, the importance of outcomes of AI-generated feedback has increasingly been highlighted as not merely depending on the quality of the feedback itself but also on the experience and perception of the feedback by the users. Research in student

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

engagement with automated feedback indicates that trust and digital literacy and learning orientation are critically important factors in determining whether students use automated feedback in a productive manner (Ranalli, 2021; Zhang & Hyland, 2025).

Ranalli (2021) discovered that automated feedback can be appreciated by learners but at the same time they can be mindful of depending on it completely, especially when the teacher recommendations are not aligned with their expectations or when they are vague. On the same note, Zhang and Hyland (2025) demonstrated that the digital literacy of students has an impact on their success in understanding and interacting with automated feedback systems. These results indicate that trust and digital competence can be considered as a mediating variable between the availability of feedback and learning outcomes.

The teacher mediation is also important. Li (2021) showed that varying modes of incorporating AWA in classroom usage can provide significantly different revision behaviors among students. This implies that automated feedback does not operate in a vacuum of instructional context; instead, the effectiveness of automated feedback is determined by the way teachers contextualize and position the use of automated feedback.

Second language writing still revolves around feedback. Hyland and Hyland (2006) suggest that effective feedback must be clear, particular and actionable in order that learners may know what they can do to improve on their writing. Although traditional automated systems have tended to focus on superficial feedback, more recent studies propose that the quality of feedback should also be considered with regards to interpretability and uptake by the learner (Shi and Aryadoust, 2024; Liu, 2024).

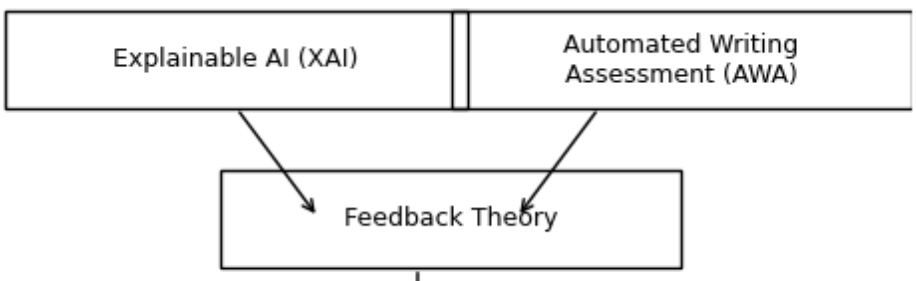
Systematic reviews revealed that AI-driven automated feedback research has grown at a rapid pace, yet issues exist about ecological validity, engagement of learners, and the instructional utility of feedback (Shi and Aryadoust, 2024). In a similar manner, Liu (2024) has found that the validity and educational effectiveness of automated feedback should be taken into account in addition to the interaction and implementation of students with the feedback.

These conclusions indicate that explainable AI could improve the quality of feedback by correlating automated judgments with interpretable linguistic evidence, making AI-based assessment more reflective of pedagogical ideals.

Theoretical Underpinning

The current research is based on the combination of three theoretical approaches: explainable artificial intelligence, automated writing assessment, and feedback theory. XAI provides the theoretical basis of transparency and interpretability (Adabi and Berrada, 2018), AWA provides the theoretical basis of computational evaluation of writing (Shermis and Burstein, 2013), and the feedback theory indicates the importance of comprehensible and practical feedback on language learning (Hyland and Hyland, 2006). All these opinions point to the conclusion that the effective AI-based writing assessment should be a mixture of technical and pedagogical relevance. Assessment systems must not only come up with the reliable scores but also give explanations to support learning. XAI informed the interpretability part of the model in this study, AWA informed the automated scoring process and the feedback theory informed the assessment of clarity, usefulness and pedagogical value. These two views combined to influence the AI system design and perception of user reactions.

Figure 1. Theoretical Framework



ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

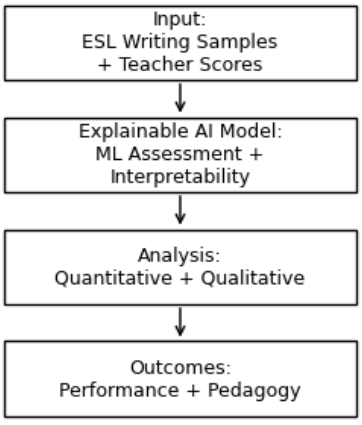
Research Gap

Although AWA has advanced and there is increased concern about XAI, the literature has a number of gaps. To begin with, automated writing assessment research has yet to be dominated by studies that focus on scoring accuracy rather than interpretability and the quality of feedback (Uto, 2021; Kumar and Boulanger, 2021). Second, not many studies have investigated the perceived explainability of AI-generated feedback by teachers and learners in real classroom settings (Khosravi et al., 2022; Türkmen, 2025). Third, mixed-methods studies that involve both technical analysis of model performance and perceptions of users are scarce, especially in the context of ESL higher education that is not a focus of research (Li, 2021; Wei et al., 2023). Such a gap is a reason to continue empirical research on how explainable AI can aid in both accurate assessment and pedagogically valuable feedback.

Conceptual Framework

According to the literature reviewed, the current study will take an integrated conceptual approach that is an integrated input, process, and output components. ESL writing samples, teacher assessment can be used as inputs, explainable AI model is the core process and their output is the model performance and user perceptions of feedback. This structure presents the two-fold interests of the study in technical accuracy and pedagogical usefulness.

Figure 2. Conceptual Framework



Methodology

Research Design

The study design w ch; it intertwined quantitative and qualitative methods to gain insight into the effectiveness and interpretability of explainable artificial intelligence (XAI) in ESL writing assessment. The quantitative element assessed the validity and quality of the AI model, comparing its scores with the ratings of the human, and

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL

the qualitative element examined the attitudes of teachers and learners toward the clarity, usefulness, and pedagogical usefulness of AI-generated feedback. This combination of both methods facilitated the triangulation of the methods and the findings became more credible.

Research Setting and Population

The research was carried out in The University of Lahore and University of Management and Technology. The target groups included undergraduate ESL students taking a course in English language or general education and ESL instructors who teach and graded academic writing.

Sample and Sampling Technique

In the study, 100 undergraduate ESL students and 10 ESL teachers took part in the study. Students were intermediate and advanced English speakers and teachers had previous experience in ESL writing instruction and assessment.

The participants were selected using a purposive sampling method to ensure that the selected participants had the relevant experience and were suitable to the research objectives. The sample size was deemed suitable in correlation analysis and exploratory mixed-methods research because it would provide a sufficient amount of quantitative data to test the agreement between AI and human scores and also offer the depth of qualitative research to gain insight into perceptions of the participants. Besides this, benchmark scoring was further enhanced by the use of several teacher raters which enhanced the reliability of the scoring.

Data Collection Instruments

ESL Writing Task

The participants were given a standardized argumentative essay work that was meant to bring out lengthy written answers. The task also provided consistency in data collection and enabled consistent assessment of all the participants.

Analytic Scoring Rubric

Teachers used an analytic scoring rubric to score the writing samples in five dimensions: grammar accuracy, vocabulary use, coherence and cohesion, organization, and task achievement. The rating of each criterion was done on a five-point scale (poor = 1, excellent = 5). The rubric was used to compare the scores generated by AI.

Explainable AI Model

An automated writing evaluation system was created and applied based on machine learning and used to assess essay by students. A supervised learning algorithm (Random Forest regression) was selected to train the model since it is well suited to work with numerous linguistic features and give results that are easy to understand. Models were developed and analysed using Python and scikit-learn library was used to model machine learning and SHAP (SHapley Additive Explanations) was selected as an interpretability tool.

The model consisted of four stages, i.e. preprocessing, feature extraction, scoring, and the generation of explanations.

During the preprocessing, the essays were processed by cleaning using tokenization, normalization and sentence segmentation. During the feature extraction step, interpretable linguistic features were derived, such as lexical diversity, syntactic complexity, markers of cohesion, and grammatical error patterns. These attributes were in line with the requirements of the human scoring rubric.

The features obtained were fed into the supervised model where it was trained on human rated essays using 80/20 training/validation split, where 80 percent of the essays were used in training and 20 percent in validation. The performance of the model was confirmed by comparing AI-generated scores with teacher ratings in correlation analysis.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL

The interpretability module with SHAP generated feature-based explanations of each score, which explained how a particular linguistic variable was positively or negatively related to the predictions of the model.

Questionnaire

A structured questionnaire was administered to both students and teachers to collect quantitative data on perceptions of AI-generated feedback. The instrument consisted of 25 items distributed across five constructs:

- Clarity (5 items)
- Usefulness (5 items)
- Interpretability (5 items)
- Trust (5 items)
- Pedagogical Value (5 items)

The answers were noted using a five-point Likert scale (strongly disagree to strongly agree).

The questionnaire is based on the earlier research on automated feedback perception and acceptance of educational technology (Ranalli, 2021; Khosrow et al., 2022). The instrument was pilot tested on 15 students outside of the main sample before administration to ensure clarity and reliability. The content validity was identified through checking it out with two experienced ESL teachers who rated the topicality, the conciseness and conformity of the instrument with the study objectives.

Semi-Structured Interviews

The semi-structured interviews were held with 10 students and 5 teachers chosen purposely as a part of the main sample to give in-depth qualitative information.

Each interview lasted 20-30 minutes and focused on the experience of participants who were provided with AI-generated feedback namely in the context of clarity, usability, trust, and instructional relevance.

The interviews were tape-recorded with the body of the participants and transcribed verbatim and analyzed using a theme analysis.

Data Collection Procedure

The data were collected in four stages.

The writing task was done under controlled classroom conditions in the first phase with the students.

The second phase involved assessing essays separately on the basis of the analytic scoring rubric by the teachers in order to establish benchmark scores.

During the third stage, the identical set of essays were fed into the explainable AI model and both scores were generated and explanatory feedback thereupon.

The last stage involved the participants going through the AI generated feedback, filling the questionnaire, and those who were selected participated in semi-structured interviews.

Data Analysis

Quantitative Analysis

The analysis of quantitative data was done using statistical methods. The correlation coefficients (r) of Pearson were determined to determine the level of agreement between AI-generated scores and human ratings. Questionnaire responses were calculated using descriptive statistics (means and standard deviations).

The reliability was analyzed in terms of the Cronbach alpha and inter-rater reliability was estimated with the help of the Intraclass Correlation Coefficient (ICC).

Qualitative Analysis

Thematic analysis was used to analyze interview data. Coding, categorizing, and organizing the transcripts into recurring themes based on clarity, interpretability, usefulness and trust were performed.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL

Ethical Considerations

Prior to data collection, ethical approval was given. Informed consent was given by participants, confidentiality of the study was upheld, and the participants were made aware of their right to withdraw at any point.

Reliability and Validity

To achieve reliability, several human raters and inter-rater agreement were used. The Intraclass Correlation Coefficient was used to determine inter-rater reliability and it was found to be strongly correlated (ICC = .88, 95% CI [.82, .92], $p < .001$).

Cronbach alpha was used to examine the internal consistency of the questionnaire with a range of between .79 and.88 which is higher than the acceptable value of.70.

Another strong positive correlation, $r = .86$, $p < .01$, between AI-generated scores and human ratings also contributed to the reliability of the AI model.

The construct validity was provided by correspondence between the linguistic features extracted and the existing ESL writing assessment criteria.

The content validity was determined by examining and judging by two qualified ESL teachers the appropriateness, relevance and suitability of the instruments to the study objectives.

Methodological triangulation was used to strengthen validity and combined both quantitative and qualitative data and improved ecological and convergent validity.

Results

Agreement Between AI-Generated Scores and Human Ratings

Pearson correlation analysis was performed to assess the correctness of the explainable AI model through the comparison of AI-generated scores and human ratings.

Table 1. Correlation Between AI Scores and Human Ratings

Variable	AI Score	Human Rating
AI Score	1.00	.86**
Human Rating	.86**	1.00

Note. p < .01

The results indicated that there is a positive correlation between AI generated scores and human rating ($r = .86$, $p < .01$) which reveals that there is a high rate of consensus between the two scoring methods. This finding implies that the elucidable AI model could have been reliable in estimating human judgment when it came to the assessment of ESL writing performance.

Inter-Rater Reliability

An Intra-class Correlation Coefficient (ICC) analysis was used to test the consistency of human raters.

Table 2. Inter-Rater Reliability

Measure	Value	95% Confidence Interval	p-value
ICC	.88	[.82, .92]	< .001

The ICC of .88 shows that there is a high degree of consensus between human raters implying that the scoring is consistent and reliable. This confirms the appropriateness of human ratings as a standard of measuring the AI model.

Descriptive Statistics of Questionnaire Responses

The descriptive statistics were determined to analyze the perceptions of the participants to AI-generated feedback.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

Table 3. Descriptive Statistics for Perception Constructs

Construct	Mean (M)	Standard Deviation (SD)
Clarity	4.12	0.64
Usefulness	4.05	0.70
Interpretability	4.18	0.58
Trust	3.89	0.75
Pedagogical Value	4.22	0.60

The respondents also indicated fairly positive attitudes towards AI-generated feedback. Pedagogical value had the highest mean score (M = 4.22, SD = 0.60), which means that the feedback was seen as useful to learning and teaching. On the same note, interpretability (M = 4.18, SD = 0.58) and clarity (M = 4.12, SD = 0.64) were rated high indicating that the explainable AI model was able to communicate how it arrived at its evaluations. Nevertheless, trust was rated with a relatively lower mean (M = 3.89, SD = 0.75), which implies that there was some skepticism amid the respondents when it comes to the dependability of AI-based feedback.

Reliability Analysis of Questionnaire

To assess internal consistency, Cronbach's alpha was calculated for each construct.

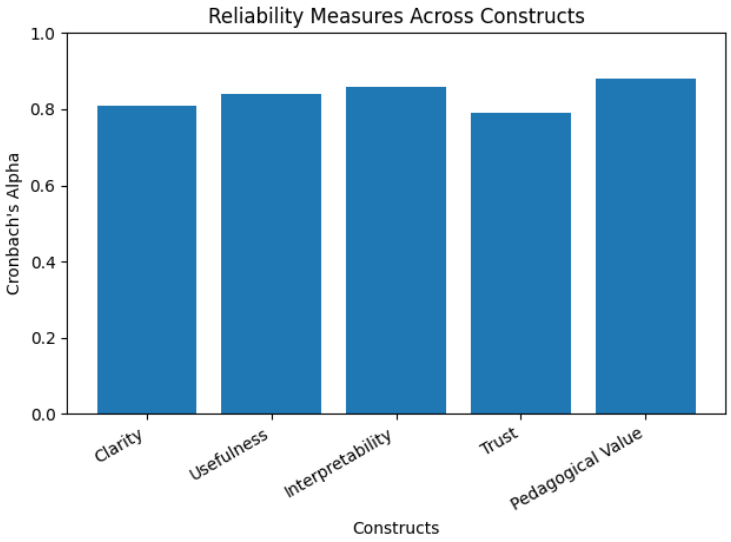
Table 4. Reliability Analysis

Construct	Cronbach's Alpha
Clarity	.81
Usefulness	.84
Interpretability	.86
Trust	.79
Pedagogical Value	.88

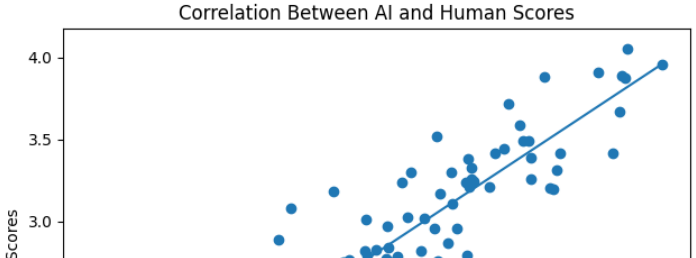
Constructs all exhibited acceptable to high reliability with Cronbach alpha values of between.79 and.88 which are higher than the recommended value of.70. This means that the questionnaire was a valid tool of assessing the perceptions of the participants.

Visual Representation of Results

Figure 4. Reliability Measures Across Constructs



The relial constructs, which is a good indicati



ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL

A scatterplot shows that there is a definite linear correlation between the AI generated scores and human ratings. The high correlation (r =.86) can be confirmed by the high density of data points in a diagonal trend, which means that the AI model was very close to human scoring.

Qualitative Findings

Qualitative data from semi-structured interviews were analyzed using thematic analysis, resulting in four major themes.

Table 5. Summary of Qualitative Themes

Theme	Description
Clarity	Feedback was easy to understand and well-structured
Interpretability	Participants appreciated explanations behind scores
Pedagogical Usefulness	Feedback supported writing improvement and teaching practices
Trust Concerns	Some skepticism regarding AI reliability

The respondents also repeatedly emphasized the understandability and lucidity of the AI-generated feedback and that the explanations made them realize their writing performance better. The theme of clarity was evident as one participant, a student, stated that the feedback was easy to comprehend since it indicated the reason behind the score given. Likewise, one of the teacher participants mentioned that the explanations contributed to finding out which writing characteristics influenced the performance, which supports the interpretability of the feedback. The pedagogical usefulness of the system was also highlighted by participants, as one teacher noted that the feedback proved useful in helping students to revise their work and assisting in classroom instruction. Meanwhile, there were still some issues pertaining to trust. One student said that the feedback was useful, but I would still like to see the teacher verify the score, which implies that, though explainability helped to improve understanding, it might take further exposure and validation to trust AI-generated assessment.

Summary of Results

Participants also mentioned the pedagogical usefulness of the system, with one teacher pointing out that the feedback was helpful in assisting the students to revise their work and aid in classroom instruction. Meanwhile, there were still some issues pertaining to trust. According to one student, the feedback was helpful, but I would still appreciate the teacher to verify the score, which means that, though explainability was helpful in enhancing the understanding, additional exposure and validation may be necessary to trust AI-generated assessment.

Discussion

The existing study examined the practicality and explicability of explainable artificial intelligence (XAI) in ESL writing scoring through a mixture of machine learning-based scoring and feature-based explanations. The results offer valuable perspectives on the technical

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

functionality of the model and its educational significance, as well as present significant questions concerning trust, adoption, and the circumstances under which AI-based evaluation can be incorporated successfully into the educational process. To begin with, the positive correlation between the scores generated by AI and human ratings ($r = .86$, $p < .01$) is strong, which indicates that the explainable AI model was highly accurate and reliable. This observation is in line with other studies that have established that automated writing evaluation systems have the ability to estimate the human judgment (Attali and Burstein, 2006; Rudner and Liang, 2002). The current research, however, builds upon the previous research by demonstrating that high predictive accuracy can be accompanied by interpretability, thus overcoming a significant weakness of traditional black-box models (Adabi et al., 2018). This is an important contribution since it implies that transparency does not always need to be associated with the costs of accuracy, which is a major concern in the context of interpretable AI. Meanwhile, the results indicate that accuracy is not enough to make the education effective. The descriptive and qualitative findings indicated that participants appreciated the clarity and interpretability of the explanations, which indicated that feedback was not received as an output but as information that can be used to facilitate learning. The observation is consistent with the feedback theory, which states that an effective feedback should be clear, specific and actionable (Hyland and Hyland, 2006). The explainable AI model, by mapping scores to particular linguistic characteristics, increased the pedagogical value of evaluation and established automated feedback as a possible learning aid as opposed to solely a scoring system. The results also show that explainable AI could be of value in other areas besides assessment by serving as a formative instructional tool. The participants found the feedback to be pedagogical, especially in aiding revision and classroom practice. It confirms the recent claims according to which AI in education is not to be regarded as an efficiency tool, but as a technology that can help to create personalized and feedback-rich learning spaces. In this regard, the research confirms the view that explainability is not just a technical enhancement but a pedagogical factor that influences the perception and use of AI-generated feedback. However, the relatively lower scores of trust cause a major tensions in the outcomes. Although the participants rated the extent of transparency and utility of the feedback as important, some of them were still not willing to trust AI-generated assessment fully. This would mean that the conditions of adoption may be technical reliability and interpretability (but not necessary). The notion of trust appears to be related to more general issues of familiarity, perceived authority, and trust in the algorithmic decision-making. This result is especially critical since it shows that the enhancement of model transparency might not necessarily result in the complete acceptance by the user. Instead, trust may be established throughout time through numerous exposures, institutional acceptance and giving of a user an opportunity to contrast AI responses with those of humans. This is a quite considerable contribution of the study such a conflict of technical confidence and user trust. Unlike a lot of the prior studies, which mainly revolved around whether AI systems can be accurate, the current results demonstrate that a more important factor in adoption in education settings is whether users feel comfortable using said systems. This brings to play a crucial human component of the existing research on automated writing assessment and shows that any future AI assessment tool should not only stop at model performance but should also include the elements of trust-building. Theoretically, the results are in line with the united theory that integrates XAI, automated writing assessment, and feedback theory. This paper illustrates that interpretability is a very important factor in relating algorithmic performance and pedagogical significance. Explainable AI allows filling the divide between assessment and instruction by making model decisions transparent and accessible, thus addressing technical and educational aspects of assessment. The results have more general implications in the local environment of this study. With the world of educational institutions

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL

increasingly considering the use of AI in instruction and evaluation, the issue of transparency, trust, and alignment of pedagogy have become even more critical. The current research adds to these wider discussions by offering empirical data that can justify the use of explainable AI to provide dependable scoring and useful feedback and also by showing that the issue of trust is one of the main obstacles to adoption. However, the results are to be taken with a grain of salt. Since the research involved short-term perceptions in a given context of higher education, it is not able to establish whether positive perceptions of explainable feedback would lead to long-term changes in the development of writing. Equally, the identified trust issues in this case can be modified in different ways in situations where learners and teachers are exposed to AI-based assessment systems over a longer period of time. In general, the research can be added to the increasing number of AI studies in the educational field as it demonstrates that successful AI-based evaluation requires not only predictive power, but interpretability, pedagogical utility, and trust in the user. It indicates that future studies are required to further explore the long-term learning results, building of trust, and the use of explainable AI in learning in various educational contexts.

Conclusion

This paper explored the use of explainable artificial intelligence in ESL writing evaluation by designing and testing a explainable machine learning model that can produce scores and explanatory feedback. The results indicated that the model was highly agreed with the human ratings as well as it gave transparent and pedagogically meaningful feedback. These findings indicate that explainable AI has the potential to improve automated writing assessment by not only ensuring technical reliability but also improving the clarity, interpretability, and instructional utility of feedback. Connecting the results of the assessment with particular linguistic characteristics, the model helped learners to analyze their performance better and helped teachers in making decisions in the teaching process. Meanwhile, the research also uncovered a significant challenge connected with trust. Participants were aware of the positive aspects of explainable feedback, yet, some of them were still skeptical about the idea of using AI-generated assessments exclusively. This implies that broader implementation of AI-based assessment can rely not just on enhancing the performance and transparency of the models, but also on building the confidence of the users through further validation and in-class expectations.

Implications

Theoretical Implications

The work has added to the literature by combining the explainable AI, automated writing assessment, and the feedback theory in that interpretability is a key element of successful AI-based assessment systems.

Practical Implications

To teachers, the results indicate that explainable AI can facilitate the assessment and teaching process by ensuring that feedback is transparent and actionable. To learners, these systems can facilitate the process of writing, but to institutions, the implementation of AI tools can be based on the pedagogical and ethical priorities.

Limitations and Future Research

The research was only done on a small sample of students and teachers in two universities, which can compromise the generalizability of the results. Also, the research concentrated on the perception in the short term as opposed to the long-term impacts on the development of writing.

Future studies must consider the long-term effects of explainable AI on writing acquisition, explore what factors affect trust in AI-aided assessment, and understand how

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL

explainable models apply to a broader range of educational settings, including more complex AI architectures with explainability methods.

Generally, the research indicates that explainable artificial intelligence is promising a great deal in improving ESL writing assessment since it integrates accuracy, transparency, and pedagogical value.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE access, 6, 52138-52160.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V. 2. The Journal of Technology, Learning and Assessment, 4(3).

Dikli, S. (2006). An overview of automated scoring of essays. The Journal of Technology, Learning and Assessment, 5(1).

Ding, L., & Zou, D. (2024). Automated writing evaluation systems: A systematic review of Grammarly, Pigai, and Criterion with a perspective on future directions in the age of generative artificial intelligence. Education and Information Technologies, 29(11), 14151-14203.

Hyland, K., & Hyland, F. (2006). Feedback on second language students' writing. Language teaching, 39(2), 83-101.

Ebn-Abbasi, F., Fattahi, N., Noughabi, M. A., & Botes, E. (2024). The strength of self and L2 willingness to communicate: The role of L2 grit, ideal L2 self and language mindset. *System*, *123*, 103334.

Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y. S., Kay, J., ... & Gašević, D. (2022). Explainable artificial intelligence in education. Computers and education: artificial intelligence, 3, 100074.Kumar, V. S., & Boulanger, D. (2021). Automated essay scoring and the deep learning black box: How are rubric scores determined? *International Journal of Artificial Intelligence in Education, 31*(3), 538–584. https://doi.org/10.1007/s40593-020-00211-5

Li, Z. (2021). Teachers in automated writing evaluation (AWE) system-supported ESL writing classes: Perception, implementation, and influence. System, 99, 102505.

Liu, W. (2024). A Systematic Review of Automated Writing Evaluation Feedback: Validity, Effects and Students' Engagement. Language Teaching Research Quarterly, 45, 86-105.Ranalli, J. (2021). L2 student engagement with automated feedback on writing: Potential for learning and issues of trust. *Journal of Second Language Writing, 52*, 100816. https://doi.org/10.1016/j.jslw.2021.100816

Ribeiro-Flucht, L., Chen, X., & Meurers, D. (2024, June). Explainable ai in language learning: Linking empirical evidence and theoretical concepts in proficiency and readability modeling of portuguese. In Proceedings of the 19th workshop on innovative use of NLP for building educational applications (BEA 2024) (pp. 199-209).

Rudner, L. M., & Liang, T. (2002). Automated essay scoring using Bayes' theorem. The Journal of Technology, Learning and Assessment, 1(2).

Shermis, M. D., & Burstein, J. (2013). Handbook of automated essay evaluation. NY: Routledge.

Shi, H., & Aryadoust, V. (2024). A systematic review of AI-based automated written feedback research. ReCALL, 36(2), 187-209.

Türkmen, G. (2025). The review of studies on explainable artificial intelligence in educational research. Journal of Educational Computing Research, 63(2), 277-310.

Uto, M. (2021). A review of deep-neural automated essay scoring models. Behaviormetrika, 48(2), 459-484.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL

Wei, P., Wang, X., & Dong, H. (2023). The impact of automated writing evaluation on second language writing skills of Chinese EFL learners: A randomized controlled trial. Frontiers in Psychology, 14, 1249991.

Zhang, Z., & Hyland, K. (2025). The role of digital literacy in student engagement with automated writing evaluation (AWE) feedback on second language writing. Computer Assisted Language Learning, 38(5-6), 1060-1085.