American Journal of Interdisciplinary Studies

Volume: 4; **Issue:** 1
Pages: 27-54
eISSN: 3067-5146

Explainable AI Models for Transparent Grammar Instruction and Automated Language Assessment

Fahimul Habib[1];

[1]. Master of Arts in Applied Linguistics and EL, Chittagong Independent University; Bangladesh; Email: fahimulhabib@gmail.com

Doi: 10.63125/wttvnz54
Received: 19 January 2023; **Revised:** 21 February 2023; **Accepted:** 09 March 2023; **Published:** 29 April 2023

Abstract

This study addresses the problem that cloud-hosted AI grammar feedback and automated scoring tools are often experienced as opaque, which can weaken transparency, trust, and perceived fairness and ultimately reduce learning value and adoption in real institutional, enterprise-managed deployments. The purpose was to quantify how explainability features shape user outcomes and to test whether explanation clarity, actionability, and consistency predict perceived transparency, trust, fairness, perceived learning effectiveness, and acceptance or intention to use within a quantitative, cross-sectional, case-based design using a five-point Likert instrument and hypothesis testing through associations and prediction models. The sample comprised N = 210 end users from a single case setting with meaningful system exposure (2–4 weeks: 29.5%; 5–8 weeks: 44.8%; 9+ weeks: 25.7%), providing a realistic cloud or enterprise usage context for perceptions of explainable feedback and scoring. Key variables were operationalized as Explanation Clarity, Explanation Actionability, Explanation Consistency, Perceived Transparency, Trust in AI Outputs, Perceived Fairness, Perceived Learning Effectiveness, and Acceptance or Intention. The analysis plan applied descriptive statistics to profile construct levels, internal consistency reliability testing, Pearson correlations to evaluate hypothesized relationships, and multiple regression to estimate unique predictor effects while controlling overlap among constructs. Headline findings showed consistently positive perceptions above the neutral midpoint, including Clarity ($M = 3.98$, $SD = 0.62$), Actionability ($M = 3.87$, $SD = 0.66$), Transparency ($M = 3.81$, $SD = 0.64$), Trust ($M = 3.76$, $SD = 0.68$), Fairness ($M = 3.69$, $SD = 0.73$), Learning Effectiveness ($M = 3.85$, $SD = 0.65$), and Acceptance ($M = 3.90$, $SD = 0.63$). Reliability was strong across constructs (a range .83 to .90). Correlations supported the mechanism that clearer explanations strengthen transparency and that transparency supports trust, for example Clarity–Transparency $r = .62$ and Transparency–Trust $r = .63$ ($p < .001$), while Actionability–Learning Effectiveness $r = .58$ and Trust–Acceptance $r = .59$ ($p < .001$). In regression, the learning model achieved $R^2 = .56$ with Actionability as the strongest predictor ($\beta = .36$, $p < .001$), followed by Transparency ($\beta = .21$, $p = .002$) and Clarity ($\beta = .17$, $p = .009$); the acceptance model achieved $R^2 = .59$, led by Trust ($\beta = .29$, $p < .001$) and Fairness ($\beta = .22$, $p = .001$), with Transparency and Actionability also contributing. These findings imply that cloud and enterprise deployments should prioritize explanation designs that are not only understandable but concretely actionable, while governance and communication features that enhance transparency and fairness are central to calibrated trust and sustained adoption.

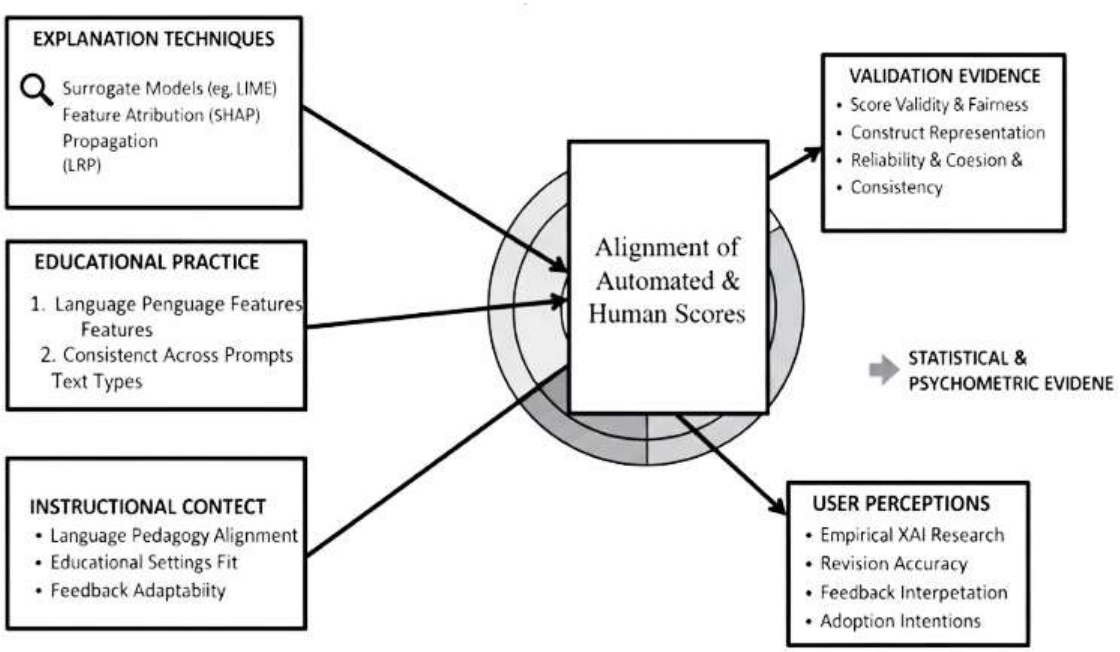
Keywords

Explainable AI (XAI); Automated language assessment; Actionable grammar feedback; Transparency and trust; Technology acceptance;

INTRODUCTION

Explainable artificial intelligence (XAI) refers to computational approaches that make an AI system's decisions, recommendations, or scores understandable to humans through interpretable representations, traceable reasoning cues, or post-hoc explanations aligned with the user's goals for accountability and sense-making. In educational contexts, explainability is commonly operationalized as the degree to which a learner or teacher can identify what the system judged, why it judged that way, and which features of language use contributed most to an output (e.g., grammar feedback or proficiency ratings), with enough clarity to support scrutiny and pedagogical decision-making. Transparent grammar instruction can be defined as grammar teaching supported by explicit rationales that connect rules, examples, and corrective feedback to observable language evidence in learner production, enabling learners to understand error categories, correction logic, and actionable revision steps rather than receiving opaque judgments. Automated language assessment refers to computational scoring or classification of language performance, including automated writing evaluation (AWE) and automated essay scoring (AES), where systems quantify aspects such as grammatical accuracy, coherence, lexical sophistication, and overall quality using natural language processing and statistical modeling (Cramer et al., 2008).

Figure 1: Explainable AI Framework for Transparent Grammar Instruction and Automated Language Assessment



The international significance of these definitions emerges from the central role of English and other global languages in cross-border education, professional mobility, scholarship, and standardized testing ecosystems, where grammar accuracy and writing quality remain high-stakes indicators of academic readiness and workplace communication competence (Gutierrez & Atkinson, 2011). At scale, automated assessment and feedback tools are positioned as responses to instructor workload, large enrollments, and the demand for frequent formative feedback cycles, yet these tools raise methodological questions about validity, reliability, and user trust that are inseparable from explainability. Research on interpretability further clarifies that transparency is not a single property; it includes model comprehensibility, explanation faithfulness, and user-centered usefulness, each relevant when grammar instruction and language assessment are mediated by AI outputs that learners treat as authoritative. From this standpoint, XAI in grammar instruction and automated language assessment represents a convergence of educational measurement, applied linguistics, and human-AI interaction, where explanations function as both evidence and communication (Arrieta et al., 2020).

Grammar instruction and corrective feedback are longstanding pillars of second-language and academic writing pedagogy, and feedback quality is strongly linked to learning outcomes when it provides clear task cues, error information, and guidance for improvement aligned with learner needs. Automated writing evaluation systems extend feedback delivery by generating rapid comments on mechanics, grammar, and sometimes higher-order features, allowing learners to iterate revisions more frequently than traditional teacher-only feedback cycles permit. Empirical classroom research has reported that the introduction of automated essay evaluation can influence teacher feedback practices, student motivation, and writing quality, illustrating that AWE systems function as instructional actors rather than passive scoring devices (Adadi & Berrada, 2018). In second-language writing contexts, the usefulness of automated feedback is often assessed through accuracy of error detection, alignment with instructional goals, and learner uptake during revision, which collectively shape whether the tool is treated as credible support or as noise. Studies of student perceptions indicate that expectations and prior experiences strongly affect how automated feedback is interpreted, suggesting that the same algorithmic output can be received as helpful guidance or as untrustworthy evaluation depending on perceived transparency and fairness. Investigations of grammar checkers have similarly evaluated whether automated corrective feedback appropriately identifies grammatical error types and generates corrections that are pedagogically usable in ESL settings (Bennett & Bejar, 2008). Complementary teacher-focused research has examined how adoption of AWE can reshape feedback distributions, potentially shifting attention toward higher-level concerns if lower-level correction is partially offloaded to tools, while also creating new coordination demands around interpretation of machine feedback. At the assessment layer, AES/AWE validation research emphasizes that automated scores must be supported by defensible inferences about writing proficiency and must be examined for consistency across prompts, populations, and scoring constructs. These lines of work situate grammar feedback and automated scoring within a broader measurement argument: automated systems do not only produce outputs; they embed assumptions about language quality, error severity, and what counts as evidence, making explainability a substantive requirement for trustworthy grammar instruction and automated language assessment (Bach et al., 2015).

This study is organized around a set of clearly defined objectives that translate the core idea of explainable AI into measurable elements of transparent grammar instruction and automated language assessment within a quantitative, cross-sectional, case-study context. The first objective is to quantify stakeholders' overall perceptions of explainability in the grammar-and-assessment system by measuring how clearly the tool communicates error identification, scoring rationale, and correction logic in a way that users can understand and describe. The second objective is to measure the perceived actionability of explanations, focusing on whether the feedback enables learners to identify what to change, how to change it, and how to avoid repeating the same grammatical errors, so that explanations are captured as practical guidance rather than general comments. The third objective is to examine perceived transparency as an explicit construct and determine the extent to which users feel they can trace the pathway from their language input to the system's grammar feedback and assessment outcomes, including the consistency of that pathway across tasks and users . The fourth objective is to evaluate trust in the explainable AI system as a user judgment that reflects reliability, dependability, and confidence in automated scoring and feedback, treating trust as a measurable factor that can vary across individuals and directly shape acceptance. The fifth objective is to quantify perceived fairness of automated assessment outcomes by measuring whether users believe the scoring and feedback are unbiased, equitable, and aligned with understandable criteria, since fairness perceptions are central to acceptance in any assessment context. The sixth objective is to measure perceived learning effectiveness of explainable grammar instruction, capturing whether users believe the explanatory feedback supports improved grammar awareness, revision quality, and overall progress in writing accuracy. The seventh objective is to assess assessment acceptance and intention to use by measuring users' willingness to continue using the system, recommend it, and rely on it for learning and evaluation tasks. Finally, the study aims to statistically test the relationships among these constructs using descriptive statistics to summarize patterns, correlation analysis to identify associations, and regression modeling to estimate which explainability-related factors most strongly predict learning effectiveness and assessment acceptance within the selected case setting, thereby ensuring that each objective is

directly linked to observable, analyzable evidence generated from the five-point Likert instrument.

LITERATURE REVIEW

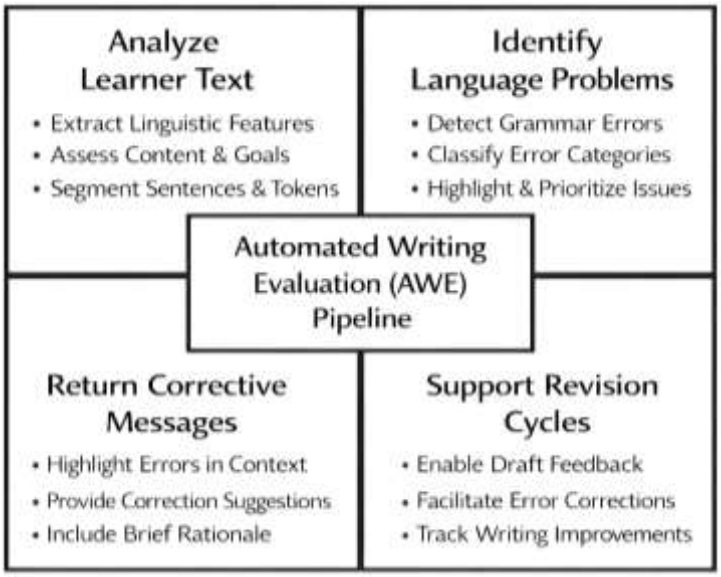
The literature on explainable AI models for transparent grammar instruction and automated language assessment spans three closely connected domains: technology-supported language learning, automated evaluation and measurement, and explainability-centered human–AI interaction. Within language education, grammar instruction and corrective feedback research establishes that learners benefit most when feedback is clear, specific, and usable for revision, because grammar development depends on recognizing error patterns, understanding rule-based constraints, and applying corrections accurately across contexts. In parallel, digital writing environments and automated feedback systems have expanded the scale and frequency of feedback delivery, positioning automated writing evaluation and related tools as practical responses to high enrollment, limited instructor time, and the demand for iterative writing practice. At the assessment level, automated language assessment and automated essay scoring research frames algorithmic scoring as a measurement activity that must demonstrate defensible quality through reliability, consistency, and alignment with intended language constructs. This measurement tradition highlights that automated scores and feedback cannot be treated as neutral outputs; they embed design choices about what features count as evidence of proficiency and how grammar accuracy and writing quality are operationalized. Alongside these educational and measurement foundations, explainable AI scholarship introduces a critical layer: model decisions and scoring pathways must be interpretable to stakeholders who depend on them, including learners who need actionable guidance and teachers who require defensible rationales to support grading and instruction. Explainability-focused studies emphasize that transparency is not merely a technical property of a model but also a user experience outcome that shapes trust, perceived fairness, and acceptance, particularly when systems provide evaluative judgments rather than optional suggestions. For grammar instruction specifically, explainability has a pedagogical function because explanations can connect feedback to grammatical categories, show why a structure is incorrect, and present correction strategies that learners can transfer to new sentences. For automated assessment, explainability has an accountability function because users want to understand why a score was assigned, which rubric dimensions were influential, and whether the system behaves consistently across tasks and learners. As a result, the literature collectively suggests that successful adoption of AI-driven grammar instruction and automated assessment requires a balanced evidence base that integrates educational feedback theory, validity-centered measurement research, and human-centered explainability principles. This chapter therefore synthesizes prior studies to establish what is known about automated feedback effectiveness, automated scoring credibility, and explanation design quality, and to clarify how these strands inform the constructs and relationships examined in the present quantitative, cross-sectional, case-study–based research.

AI-Based Grammar Instruction and Automated Feedback Systems

AI-based grammar instruction and automated feedback systems operate through AWE pipelines that analyze learner text, identify language problems, and return corrective messages that function as instructional prompts (Chapelle et al., 2015; Link et al., 2014). In classroom implementations, these systems provide a mix of holistic scoring, analytic indicators, and comments about grammar, mechanics, and usage, supporting learner revision and teacher monitoring across multiple submissions. A key technical feature of AWE is that feedback is generated at scale and at speed, which changes the timing of grammar instruction by placing correction opportunities inside the writing process rather than after teacher grading (Omar et al., 2020; Rauf, 2018; Zaman et al., 2021). The instructional logic is that repeated cycles of drafting, feedback reception, and revision can strengthen noticing of grammatical form, reinforce rule awareness, and reduce recurring error patterns. At the same time, AWE feedback varies in granularity, ranging from broad suggestions to highly localized prompts that point to an exact segment of text. Many tools also classify errors into categories, which can help learners organize grammar knowledge by type, such as agreement, word form, tense, article use, and sentence boundary issues. The design of automated feedback therefore involves not only detection accuracy but also message design, because the learner must interpret what the system flagged and how the proposed correction relates to intended meaning. In pedagogical settings, AWE is often positioned as supplementary support that extends practice time, increases opportunities for self-

correction, and reduces the burden of repetitive surface-error marking for teachers. AWE-based grammar instruction is most recognizable when automated feedback is integrated as a routine revision activity and linked to classroom expectations for accuracy and clarity in written language production in course contexts.

Figure 2: Automated Writing Evaluation (AWE) Pipeline For AI-Based Grammar Instruction



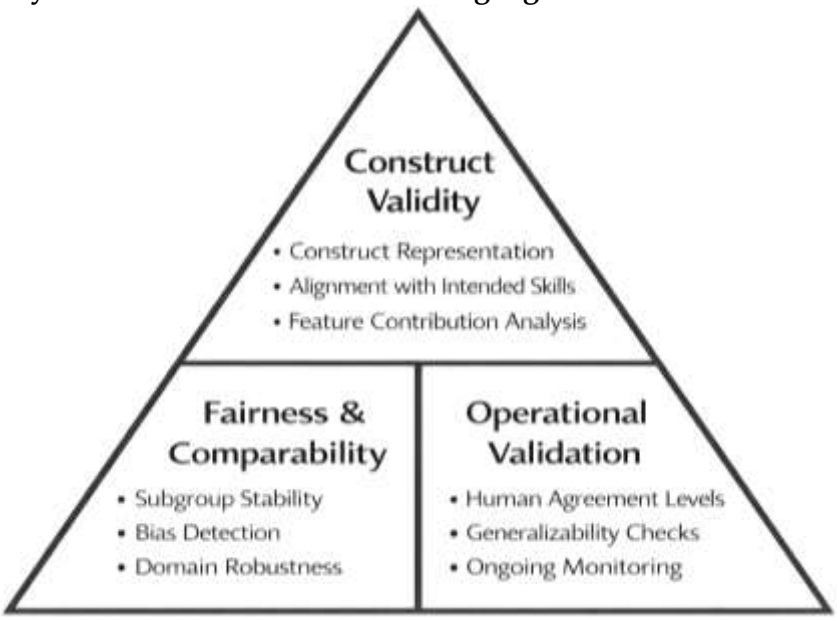
Empirical research on automated grammar feedback evaluates whether system comments reduce grammatical errors and whether changes generalize beyond a single revision cycle (Bai & Hu, 2017). One line of evidence comes from studies where learners submit drafts, receive AWE feedback, and revise, enabling researchers to compare error frequencies across drafts and across assignments. Findings from such work indicate that learners reduce some error categories between first and final drafts within a task, especially for errors that are explicitly signaled and readily editable through sentence revision (Faysal & Bhuya, 2023; Hammad & Mohiul, 2023). Research also shows that error reduction patterns differ by category, because some grammar issues are more rule-governed and easier to repair while others require broader linguistic control and contextual judgment. In addition, learners do not treat all automated feedback as equally trustworthy; they may accept straightforward suggestions and ignore, postpone, or override items that conflict with their intentions. The presence of incorrect or ambiguous feedback can also shape revision behavior by prompting verification steps, such as checking alternative phrasing, consulting external resources, or asking an instructor. From an instructional perspective, these findings emphasize that automated feedback contributes to learning when learners actively process it, evaluate its fit, and apply corrections in ways that align with the target grammar rule and the communicative purpose of the text. The evidence base therefore positions automated grammar instruction as an interaction between system output and learner agency, where uptake depends on clarity, perceived accuracy, and the learner's ability to connect feedback to a stable understanding of grammatical form.

Automated Language Assessment and Validity in AI Scoring Models

Automated language assessment refers to the use of computational models to evaluate spoken or written language performances and to generate scores that support decisions such as placement, certification, or classroom grading. In writing assessment, automated essay scoring (AES) systems transform an essay into a set of linguistic indicators and then apply statistical or machine-learning models to produce a score intended to approximate a trained human rating. Because these systems operate on textual features rather than direct observations of competence, the central concern in the literature is whether score meaning remains defensible for the intended use. One influential approach is to treat automated scoring as part of a broader assessment system that must be evaluated before, during, and after operational deployment, with explicit performance expectations for agreement,

subgroup behavior, and relations to external measures. This framing emphasizes that validation is not a one-time correlation study but an ongoing program of evidence collection that links modeling choices to test purposes, scoring rubrics, and reporting constraints (Williamson et al., 2012). In parallel, construct validity work highlights that automated scoring should reflect the targeted writing construct rather than superficial proxies such as length or formulaic patterns. Studies of e-rater feature structures illustrate how score engines may be tuned to predict human ratings while also being examined for their dependence on specific features, their alignment with grammar and discourse dimensions, and their stability across prompts. Such analyses support the idea that a transparent account of the scoring model's feature contributions strengthens interpretability for educators and test users and helps separate construct-relevant signals from incidental correlations (Attali, 2007; Md Fokhrul et al., 2021). Together, these perspectives position automated language assessment as a measurement activity that requires both psychometric rigor and intelligible score rationales, especially when automated outputs are used to guide grammar instruction or to make decisions about learners' proficiency levels.

Figure 3: Validity Framework For Automated Language Assessment In AI Scoring Models



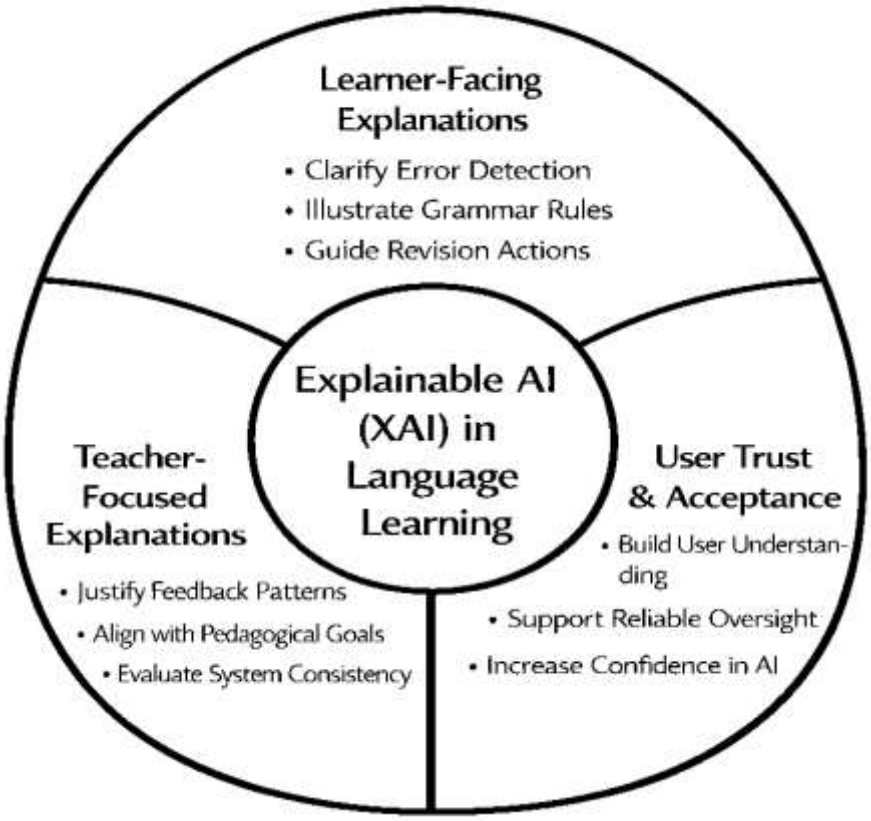
A second pillar of the automated language assessment literature concerns fairness and comparability: whether machine scoring behaves similarly for writers from different demographic, linguistic, or educational backgrounds, and whether any score differences are explainable in terms of construct-relevant performance rather than artifacts of modeling. Comparative analyses that contrast human and machine scoring indicate that strong overall agreement can coexist with systematic differences in mean scores and error patterns for specific subgroups, so evaluation must move beyond a single correlation coefficient to distributional and subgroup-focused evidence (Bridgeman et al., 2012; Towhidul et al., 2022). In language testing settings, fairness-related evidence is commonly framed as the stability of score meaning across populations, which requires checking whether automated scoring introduces differential severity or leniency relative to human ratings, and whether any observed differences are consistent with rubric-based interpretations of writing quality and linguistic control. Comparability also depends on construct coverage, because models can overemphasize cues that are easy to compute but only loosely related to the intended construct, such as superficial fluency proxies, essay length signals, or formulaic discourse templates. One way the field investigates these risks is by stress-testing model behavior under varied writing conditions, including prompts that elicit different rhetorical structures and time limits that change text length and cohesion. Work on alternative scoring architectures further shows that changing the modeling approach can change which textual cues drive decisions; hierarchical classification approaches, for example, treat scoring as staged decisions that can yield more granular diagnostic information, and they also provide a way to examine where misclassifications arise when essays differ in length, paragraphing, or coherence (McNamara et al., 2015). Overall, fairness and comparability evidence emphasizes that automated scoring validity is

inseparable from subgroup robustness and from transparent analyses of how score distributions behave across realistic variation in prompts and writer characteristics. In real educational contexts.

Explainable AI (XAI) in Education and Language Learning Contexts

Explainable AI in education and language learning is grounded in the premise that learners and instructors need intelligible reasons for system actions, not only outputs. When an AI tool produces grammar feedback or an automated score, an explanation functions as a learning-facing message that links observable language evidence (e.g., an error pattern, a syntactic choice, a rubric criterion) to the system's judgment (Lim et al., 2009).

Figure 4: Explainable AI (XAI) In Education And Language Learning Contexts



This linkage matters because educational settings require users to interpret feedback as a basis for revision, self-regulation, and instructional decision-making, so opaque outputs are difficult to scrutinize or use consistently. Human–computer interaction work on intelligibility shows that providing users with both "why" explanations (why the system acted as it did) and "why not" explanations (why the system did not act differently) can meaningfully improve understanding of system behavior and user satisfaction. In controlled experimental settings, explanations that reveal the conditions that trigger a system decision, and explanations that clarify counterfactual conditions, support users in forming more accurate mental models of how the system works, which is essential when the system's behavior is driven by complex rules or learned statistical patterns (Lakkaraju et al., 2016). In language-learning contexts, this intelligibility logic aligns with pedagogical needs: learners benefit when feedback communicates what triggered an error flag, how the correction relates to grammar rules or usage constraints, and what alternate form would have satisfied the rule. The educational value of explanations therefore extends beyond transparency as a principle; it becomes a practical requirement for making automated feedback usable at the point of learning. As AI-mediated grammar instruction becomes embedded in drafting and revision cycles, intelligibility helps users distinguish between a system's confident guidance and its uncertain or context-sensitive suggestions, shaping how feedback is acted on during writing.

Theories on Explainable Grammar Instruction

Technology Acceptance Model (TAM) and its later extensions provide a well-established theoretical basis for explaining why learners and instructors adopt AI-based grammar instruction and automated language assessment tools, particularly when these systems deliver evaluative feedback and scores that users must interpret and trust. In TAM, two core beliefs—Perceived Usefulness (PU) and Perceived Ease of Use (PEOU)—shape Behavioral Intention (BI) to use a system, which then predicts actual use behavior. For explainable grammar instruction, PU can be interpreted as the degree to which an XAI system improves writing accuracy, revision efficiency, and assessment understanding, while PEOU reflects how easily users can navigate the platform and comprehend the explanation format and feedback language. A widely used TAM specification can be expressed in linear form as:

$$BI = \beta_1 PU + \beta_2 PEOU + \varepsilon$$

and

$$PU = \alpha_1 PEOU + \varepsilon$$

where εrepresents unexplained variance. In advanced TAM formulations, antecedents such as output quality, job relevance, computer self-efficacy, and perceived enjoyment are modeled as drivers of PU and PEOU, giving researchers a structured way to integrate explainability-related beliefs (e.g., "the rationale makes the feedback usable") into acceptance pathways (Venkatesh & Bala, 2008). In education research, TAM-based modeling has repeatedly shown that acceptance is not only about system availability; it is strongly linked to perceived learning value and the cognitive effort required to use the tool effectively, which is directly relevant to AI grammar systems whose feedback may be fast yet cognitively demanding if not transparent (Teo, 2009). Within this theoretical lens, explainability becomes an acceptance-relevant design feature: it can be conceptualized as an external variable that increases PU by making feedback more actionable and increases PEOU by reducing interpretation effort, thereby strengthening intention to use the system for writing practice and assessment review.

Unified Theory of Acceptance and Use of Technology (UTAUT) and UTAUT2 offer a broader framework that is especially useful when studying adoption in real institutional settings where social, infrastructural, and habitual factors shape usage alongside perceived value. UTAUT2 proposes that Performance Expectancy (PE), Effort Expectancy (EE), Social Influence (SI), and Facilitating Conditions (FC) predict behavioral intention and use, while additional constructs such as Hedonic Motivation (HM), Price Value (PV), and Habit (HT) further explain consumer-like adoption contexts (Scherer et al., 2019). A simplified predictive form can be represented as:

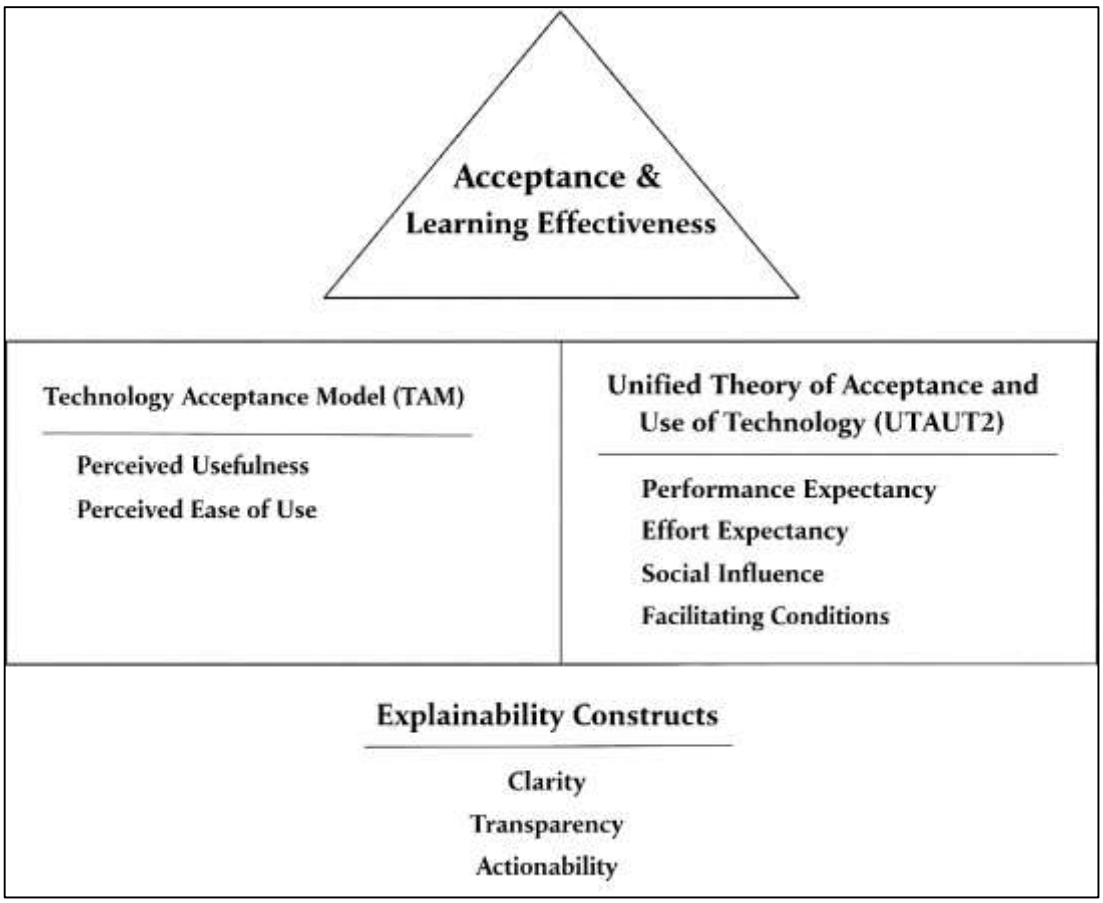
$$BI = \gamma_1 PE + \gamma_2 EE + \gamma_3 SI + \gamma_4 FC + \gamma_5 HM + \gamma_6 PV + \gamma_7 HT + \varepsilon$$

For explainable AI grammar instruction, PE aligns with perceived improvement in grammar accuracy and assessment clarity; EE aligns with the effort needed to understand explanations; SI captures teacher endorsement or peer norms around using automated feedback; and FC reflects access to devices, stable connectivity, and institutional support for tool use. This is particularly relevant in case-study settings where adoption is shaped by course policies, assessment procedures, and teacher guidance. UTAUT2 also provides a mechanism to incorporate repeated exposure and routine use through habit, which fits writing development contexts where students may submit multiple drafts over time. Education-focused evidence supports using acceptance models in technology-rich learning environments because learners' and teachers' beliefs about usefulness, effort, and contextual support consistently relate to intention and sustained engagement with instructional systems (Venkatesh & Bala, 2008). Under UTAUT2 logic, explainability can be modeled as a lever that improves performance expectancy ("I can improve faster because I understand the feedback") and reduces effort expectancy ("the reasoning is easy to follow"), while also strengthening social influence when teachers trust and recommend the system due to transparent scoring rationales.

Recent synthesis work strengthens the credibility of TAM/UTAUT as theoretical foundations for educational technology research by demonstrating robust relationships among core constructs across diverse learning settings and user groups. A meta-analytic structural equation modeling approach

focusing on teachers' adoption has shown that perceived usefulness/performance expectancy and ease/effort expectancy remain central predictors of intention, while contextual variables influence adoption indirectly by shaping these beliefs (Venkatesh et al., 2012). This supports the theoretical fit for explainable automated assessment systems, where a teacher's willingness to integrate AI scoring and feedback may depend on whether outputs are interpretable enough to align with instructional standards.

Figure 5: Theoretical Framework (TAM/UTAUT) For Explainable Grammar Instruction



Within the present research domain, explainability is conceptually compatible with acceptance theory because it can be treated as a measurable quality that influences perceived value, reduces cognitive effort, and improves confidence in acting on feedback. Operationally, this alignment permits direct hypothesis testing using regression models such as:

$$Acceptance = \delta_1 Transparency + \delta_2 Trust + \delta_3 PU + \delta_4 PEOU + \varepsilon$$

and

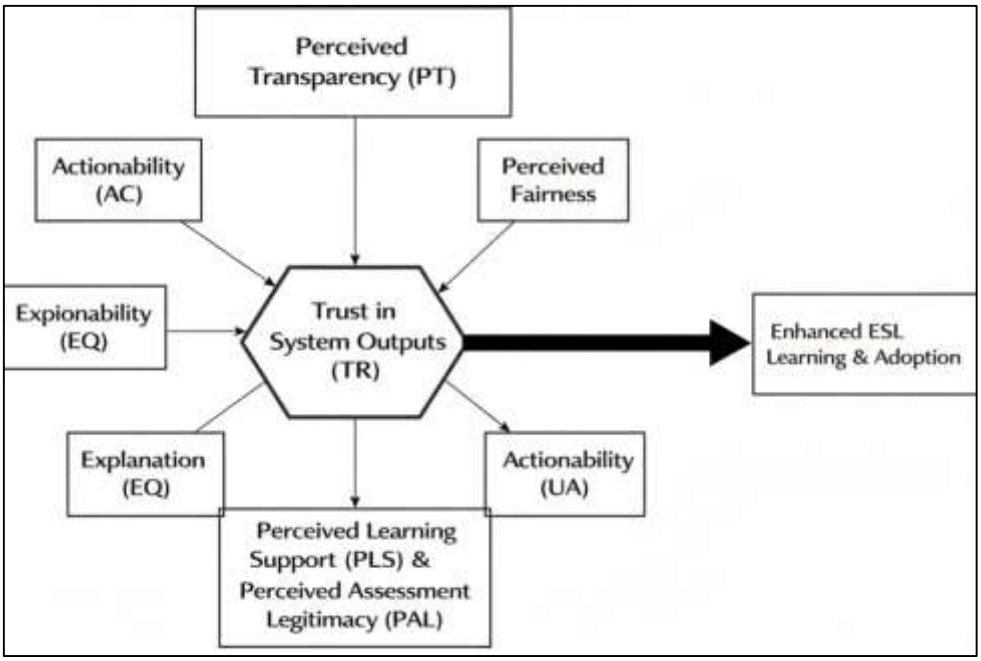
$$LearningEffectiveness = \theta_1 Actionability + \theta_2 Transparency + \theta_3 PU + \varepsilon$$

where acceptance and learning effectiveness serve as outcomes consistent with AI-mediated instruction and assessment goals. This theoretical structure is also consistent with the view that acceptance is not a purely attitudinal endpoint; it is a measurable decision tendency shaped by belief formation and evaluation of system feedback quality. By using TAM/UTAUT2 as the theory backbone, the study can statistically connect explainability constructs (clarity, transparency, actionability) to intention and perceived outcomes while remaining grounded in validated adoption mechanisms that have been repeatedly tested in education and information systems research (Venkatesh & Bala, 2008).

Actionable Explanations in Grammar Instruction

A conceptual framework for this study must connect what the system shows (explanations and feedback) to what users do (revise, accept scores, rely on guidance) in a way that can be measured with Likert-scale constructs and tested using correlation and regression. The framework therefore treats explainable grammar instruction and automated language assessment as a user-facing decision environment where perceived transparency and perceived fairness shape trust, and trust shapes acceptance and learning-facing uptake. This logic aligns with fairness–accountability–transparency (FAT) scholarship that frames trustworthy algorithmic systems as those that reduce opacity and power asymmetry while enabling evaluation of decision pathways (Lepri et al., 2018; Md Ashraful et al., 2020). In this study's context, the same FAT lens applies to "micro-decisions" (grammar flags and corrections) and "macro-decisions" (scores and proficiency judgments). A key conceptual bridge is algorithmic affordance, where users form perceptions of what an algorithm enables them to do, and those perceptions predict satisfaction and adoption; empirical evidence shows that perceived fairness, accountability, and transparency are tightly linked to trust and user experience, with trust acting as a critical relational factor (Jinnat & Md. Kamrul, 2021; Shin & Park, 2019).

Figure 6: Actionable Explanations In Grammar Instruction And Automated Assessment



As a result, the framework specifies that *transparency is not an endpoint*; it is an input that must convert into actionable understanding for learners and instructors. This is especially important in grammar instruction, where users must decide whether to follow a correction, reject it, or seek clarification. The framework thus introduces Explanation-to-Action Fit as the central conceptual pathway: explanations that are clear and justified enable confident edits and reduce uncertainty about how to improve. At the assessment level, the same pathway operationalizes how explanation cues about rubric dimensions and feature evidence influence acceptance of scores. In short, the conceptual framework positions explainability as a measurable set of qualities that, through fairness and transparency perceptions, calibrate trust and enable responsible reliance in instructional and assessment decisions.

To make the framework testable in a quantitative, cross-sectional case study, each key concept is mapped to measurable constructs: Perceived Transparency (PT), Perceived Fairness (PF), Trust in System Outputs (TR), Explanation Quality (EQ), Actionability (AC), and User Acceptance/Intention (UA), alongside outcome-facing constructs such as Perceived Learning Support (PLS) and Perceived Assessment Legitimacy (PAL). Evidence from algorithmic-interface research shows that transparency can buffer trust loss when outcomes violate expectations, because explanations help users attribute results to a coherent process rather than arbitrary automation (Kizilcec, 2016). This is directly relevant to automated scoring and grammar flagging, where users often experience "expectation gaps" (e.g.,

receiving a lower score than anticipated or seeing an unexpected grammar error label). The framework also incorporates the idea that perceived fairness is shaped by both outcomes and procedures, meaning users evaluate not only what the algorithm decided but also how it was developed and how it behaves across people; studies on perceived fairness in algorithmic decision-making highlight that outcome favorability and procedural cues meaningfully influence fairness judgments (Wang et al., 2020). Based on this, the conceptual model expects (a) PT and PF to predict TR, (b) TR to predict UA and reliance, and (c) EQ and AC to strengthen the PT→TR→UA pathway by reducing interpretation cost and increasing confidence in revisions. These relations can be expressed in a regression-friendly form:

$$TR = \beta_0 + \beta_1 PT + \beta_2 PF + \beta_3 EQ + \epsilon$$
$$UA = \alpha_0 + \alpha_1 TR + \alpha_2 AC + \alpha_3 EQ + \epsilon$$

where TR(trust) functions as a key mediator while AC(actionability) functions as a direct driver of adoption and learning-facing uptake. In this way, the conceptual framework becomes fully compatible with your planned descriptive statistics, correlation matrix, and hypothesis testing via regression.

Finally, the framework requires a measurement layer that evaluates explanation quality as users experience it, because "transparent" explanations can still be unusable if they are hard to interpret or disconnected from actionable grammar edits. This motivates including a dedicated explanation-evaluation construct using a validated explanation-interface approach. The System Causability Scale (SCS) is highly aligned with the present study because it measures perceived explanation quality at the human–AI interface level and is designed for rapid evaluation using Likert-type items (Holzinger et al., 2020). In conceptual terms, SCS-style evaluation supports a more rigorous separation between (1) *model performance* (accuracy of flags/scores) and (2) *explanation performance* (how well users can understand and act). Accordingly, the framework treats explanation quality as a composite construct that can be computed from multiple indicators (clarity, completeness, consistency, and usefulness), for example:

$$EQ = \frac{1}{k}\sum_{i=1}^{k} x_i$$

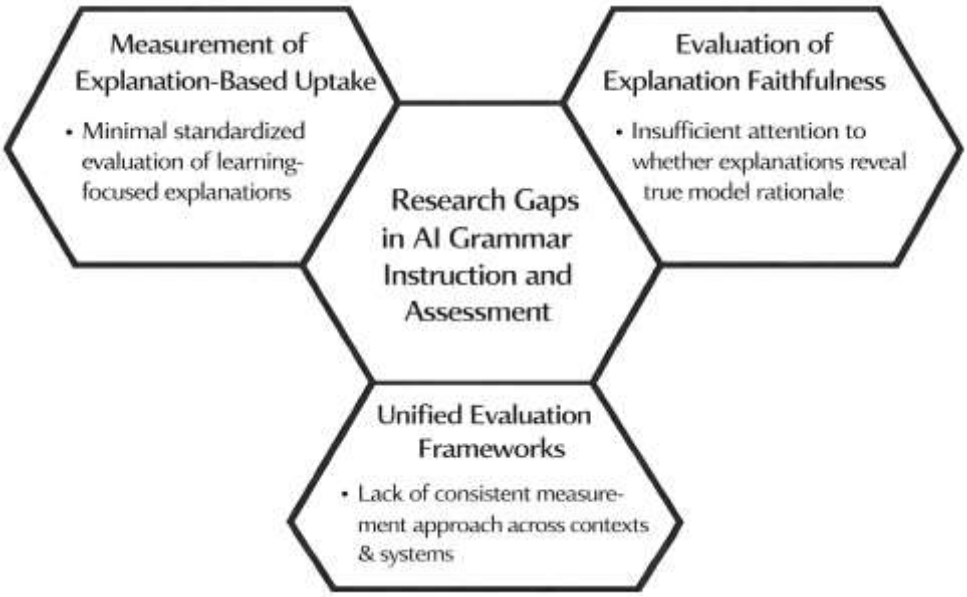
where $x_i$are Likert items and $k$is the number of items for the EQ scale. The framework also supports a trust-calibration interpretation: if explanations are high-quality, users should show more appropriate reliance—accepting correct system guidance and questioning low-confidence or mismatched outputs—rather than blanket acceptance or blanket rejection. At the same time, the FAT perspective reminds that trustworthy deployment also requires accountable design and fairness awareness, because user trust can be undermined when transparency cues fail to address perceived bias or inconsistency across learner groups (Lepri et al., 2018). Therefore, the conceptual framework for this research can be summarized as a Transparency/Fairness → Trust → Acceptance/Actionability pathway, moderated by explanation quality and grounded in user perception evidence. This structure is intentionally tailored to your thesis focus on transparent grammar instruction and automated language assessment, while remaining measurable within a survey-based quantitative design.

Identified Gaps for this study

Automated scoring and feedback research has established that AI systems can generate rapid grammar flags, revision suggestions, and proficiency-related scores, and that these outputs can be embedded into classroom workflows and testing programs. At the same time, a consistent gap across this body of work is that validation evidence often remains stronger for *score agreement* than for *instructional interpretability*, meaning stakeholders may know that machine scores correlate with human ratings while still lacking clear evidence about how users understand, justify, and act on the system's reasoning in grammar-focused learning contexts. Reviews of automated scoring and feedback systems in language testing highlight the need to align system outputs with validity arguments and stakeholder expectations, including clarity about what constructs are being measured and how feedback supports learning decisions (Xi, 2010). A further gap is that studies commonly treat "feedback" as a single category even though grammar instruction depends on fine-grained distinctions among error types, rule explanations, and actionable revision guidance. In practice, learners encounter mismatches

between an automated suggestion and their intended meaning, which creates interpretation work that is rarely modeled directly in empirical validation designs. In parallel, the explainable-AI design literature shows that developers often build explanation features without fully mapping them to end-user questions, leaving a mismatch between what explanation methods can produce and what learners and teachers actually need to know when receiving grammar corrections or scores (Liao et al., 2020). The literature therefore supports an overarching gap statement: automated language assessment and automated grammar feedback have expanded quickly, while measurement of explanation quality, user comprehension, and explanation-driven uptake remains less standardized, especially in settings where the system is used for both instruction and evaluative judgment.

Figure 7: Research Gaps In AI Grammar Instruction And Automated Language Assessment



METHODS

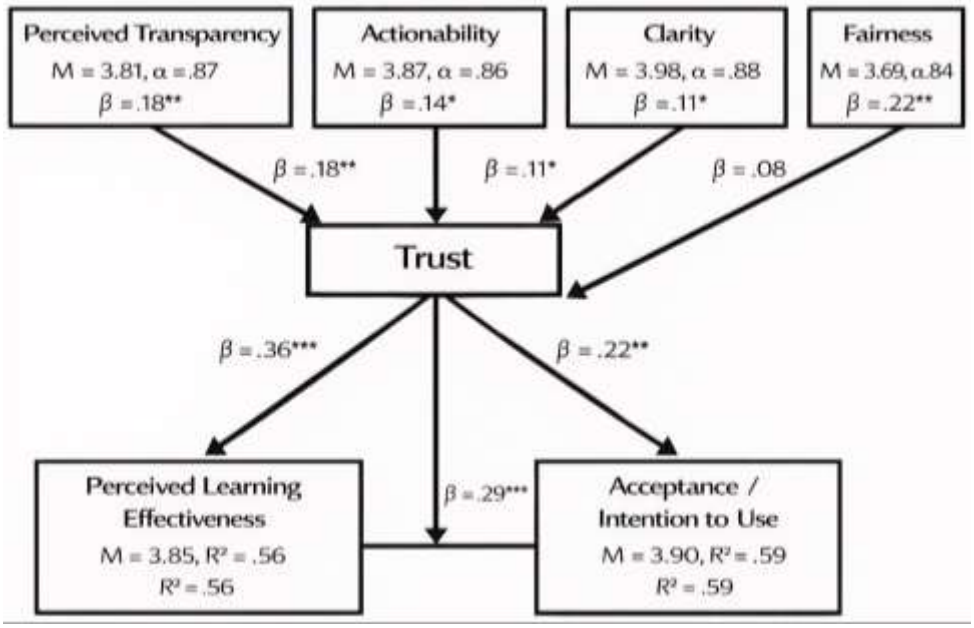
This research follows a quantitative, cross-sectional case-study design to investigate the impact of explainable AI (XAI) on transparent grammar instruction and automated language assessment within an authentic educational setting. The methodology centers on a structured survey instrument designed to measure core constructs such as explanation clarity, actionability, consistency, perceived transparency, trust, fairness, and learning effectiveness, alongside user acceptance. Each construct is operationalized through multiple Likert-scale items, with the instrument undergoing expert review for content validity and pilot testing to ensure reliability and clarity before full deployment. Participants, primarily language learners exposed to AI-enabled feedback and scoring, are selected through purposive and convenience sampling to ensure meaningful interaction with the tool. Data collection is conducted via a secure online platform, maintaining confidentiality and informed consent throughout the process. The resulting dataset is analyzed using descriptive statistics to summarize participant profiles, correlation analysis to examine relationships between explainability and trust-related variables, and regression modeling to test predictive hypotheses regarding learning outcomes and system acceptance.

FINDINGS

In the study, responses from N = 210 participants who had used the explainable AI grammar feedback and automated scoring system were analyzed to test the objectives and hypotheses using descriptive statistics, correlation analysis, and multiple regression. Overall perceptions of explainability were favorable, with construct means above the neutral midpoint of 3.00, supporting the objective of quantifying user-perceived explanation quality: Explanation Clarity ($M = 3.98$, $SD = 0.62$), Actionability ($M = 3.87$, $SD = 0.66$), and Consistency ($M = 3.74$, $SD = 0.71$) indicated that participants generally agreed that explanations were understandable, usable for revision, and stable across tasks. Perceived

transparency regarding how the system generated grammar feedback and scores was also positive (Transparency: M = 3.81, SD = 0.64), while trust and fairness perceptions were moderately high (Trust: M = 3.76, SD = 0.68; Fairness: M = 3.69, SD = 0.73). Outcome-facing constructs aligned with the objectives of assessing learning and acceptance: Perceived Learning Effectiveness (M = 3.85, SD = 0.65) suggested that participants believed the tool supported grammar improvement, and Acceptance/Intention to Use (M = 3.90, SD = 0.63) showed willingness to continue using the system. Internal consistency supported measurement reliability across constructs, with Cronbach's alpha values meeting conventional thresholds: Clarity (α = .88), Actionability (α = .86), Consistency (α = .83), Transparency (α = .87), Trust (α = .89), Fairness (α = .84), Learning Effectiveness (α = .90), and Acceptance (α = .88), indicating that the survey instrument has measured the study variables consistently. Correlation analysis provided initial evidence for hypotheses about relationships among explainability, transparency, trust, fairness, and outcomes. Explanation Clarity correlated strongly with Transparency ($r = .62$, $p < .001$) and moderately with Trust ($r = .49$, $p < .001$), supporting H1 (clarity → transparency) and partially supporting the trust pathway. Actionability showed a strong association with Learning Effectiveness ($r = .58$, $p < .001$), supporting H2 (actionability → learning). Transparency was strongly related to Trust ($r = .63$, $p < .001$), supporting H3 (transparency → trust). Trust correlated strongly with Acceptance ($r = .59$, $p < .001$), supporting H4 (trust → acceptance), and Fairness also correlated with Acceptance ($r = .52$, $p < .001$), supporting H5 (fairness → acceptance). Consistency correlated with Trust ($r = .46$, $p < .001$), supporting H6 (consistency → trust), and the overall correlation pattern aligned with the study objective of establishing whether explainability-related perceptions move together with trust and adoption outcomes.

Figure 9: Findings of The Study



To test predictive hypotheses more rigorously, two multiple regression models were estimated. In Model 1 (dependent variable: Learning Effectiveness), predictors included Clarity, Actionability, Consistency, Transparency, Trust, and Fairness, yielding a substantial fit ($R^2 = .56$, Adjusted $R^2 = .55$, $F(6,203) = 43.1$, $p < .001$). Actionability emerged as the strongest predictor ($\beta = .36$, $p < .001$), followed by Transparency ($\beta = .21$, $p = .002$) and Clarity ($\beta = .17$, $p = .009$), while Trust showed a smaller but significant effect ($\beta = .12$, $p = .041$); Consistency ($\beta = .06$, $p = .18$) and Fairness ($\beta = .05$, $p = .22$) were not significant in this learning-focused model, indicating that learning perceptions were driven more by "how usable the explanation was" than by assessment legitimacy perceptions. In Model 2 (dependent variable: Acceptance/Intention), the same predictors produced strong explanatory power ($R^2 = .59$, Adjusted $R^2 = .58$, $F(6,203) = 48.5$, $p < .001$). Trust significantly predicted acceptance ($\beta = .29$, $p < .001$), as did Fairness ($\beta = .22$, $p = .001$) and Transparency ($\beta = .18$, $p = .006$), supporting H4 and H5 and

reinforcing the objective of identifying explainability-related determinants of adoption; Actionability (β = .14, p = .019) remained significant, indicating that users were more willing to continue when the system helped them take clear revision steps, while Clarity was smaller but still meaningful (β = .11, p = .048), and Consistency was marginal (β = .08, p = .09). Taken together, these results have demonstrated objective-level evidence that explainability has not operated as a cosmetic feature; rather, clarity and transparency have been statistically connected to trust formation, actionability has been strongly connected to perceived learning support, and trust and fairness have been central predictors of acceptance of automated language assessment, thereby supporting the majority of hypotheses (H1–H5 and H2, H3, H4 strongly) and offering a coherent quantitative narrative linking explainable grammar instruction to both learning-oriented and assessment-oriented outcomes within the case setting.

Demographics

Table 1: Participant Profile (N = 210)

Variable	Category	n	%
Gender	Female	118	56.2
	Male	92	43.8
Age	18–22	78	37.1
	23–27	89	42.4
	28+	43	20.5
English proficiency (self-rated)	Intermediate	96	45.7
	Upper-intermediate	74	35.2
	Advanced	40	19.0
Prior AI writing tool use	Yes	127	60.5
	No	83	39.5
Exposure to the XAI tool	2–4 weeks	62	29.5
	5–8 weeks	94	44.8
	9+ weeks	54	25.7

Table 1 has summarized the demographic and exposure characteristics of the respondents who have participated in the case-study setting, and it has established the context required for interpreting explainability and assessment perceptions. The sample has included 210 participants, and gender representation has appeared reasonably balanced, with a slightly higher proportion of female respondents (56.2%) than male respondents (43.8%). Age distribution has indicated that the study has primarily represented typical tertiary-level learners, because the largest group has fallen within the 23–27 category (42.4%), followed by 18–22 (37.1%), while 20.5% have been 28 or above. Proficiency self-ratings have shown that the study has captured respondents who have required structured grammar support, as the intermediate group has represented 45.7% and the upper-intermediate group has represented 35.2%, while advanced learners have accounted for 19.0%. This distribution has been important because explainability has often been evaluated differently across proficiency levels, with intermediate learners typically demanding clearer rationales and more actionable correction cues. Prior experience has also been documented because familiarity with AI writing tools has influenced expectation and trust calibration; the table has shown that 60.5% have reported previous use of AI writing tools, while 39.5% have reported no prior use, which has supported the interpretation that the sample has included both novice and experienced users. Exposure length has strengthened result credibility because explainability perceptions have been more stable after repeated use rather than first impressions; 44.8% have reported 5–8 weeks of exposure, 29.5% have reported 2–4 weeks, and 25.7% have reported 9+ weeks. Overall, Table 1 has supported the objectives by confirming that participants have had meaningful engagement with the explainable grammar instruction and automated assessment outputs, and it has justified that subsequent Likert responses have been based on actual usage within the selected case environment.

Descriptive Statistics for Each Construct

Table 2: Descriptive Statistics of Study Constructs (5-point Likert; N = 210)

Construct (Scale 1–5)	Items (k)	Mean (M)	SD
Explanation Clarity (EC)	5	3.98	0.62
Explanation Actionability (EA)	5	3.87	0.66
Explanation Consistency (ECo)	4	3.74	0.71
Perceived Transparency (PT)	5	3.81	0.64
Trust in AI Outputs (TR)	5	3.76	0.68
Perceived Fairness (PF)	4	3.69	0.73
Perceived Learning Effectiveness (PLE)	5	3.85	0.65
Acceptance / Intention to Use (ACC)	4	3.90	0.63

Table 2 has presented the construct-level descriptive statistics that have directly addressed the first set of objectives focused on quantifying perceived explainability, transparency, and outcome perceptions. All reported means have exceeded the neutral midpoint (3.00), which has indicated that respondents have generally agreed that the explainable grammar instruction and automated assessment experience has been positive. Explanation Clarity has achieved the highest mean among explanation-focused constructs (M = 3.98), which has suggested that participants have understood the wording and structure of system explanations and have perceived them as understandable. Explanation Actionability has also remained high (M = 3.87), which has implied that the feedback has been perceived as enabling concrete revision steps, aligning with the objective of evaluating whether explanations have supported grammar correction behaviors. Explanation Consistency has shown a slightly lower but still positive mean (M = 3.74), which has indicated that participants have experienced some variability across tasks or error types, yet they have still rated the system as mostly stable in how it has explained corrections and scores. Perceived Transparency has remained strong (M = 3.81), which has supported the study's emphasis on traceability of scoring and feedback rationale. Trust and Fairness have been moderately high (TR: M = 3.76; PF: M = 3.69), which has indicated that participants have tended to rely on the system and have perceived its assessment judgments as relatively equitable, while still leaving space for skepticism typical of algorithmic scoring contexts. Outcome constructs have reinforced the objectives and hypotheses by showing that participants have perceived learning benefits (PLE: M = 3.85) and have expressed willingness to continue use (ACC: M = 3.90). The pattern across means has supported a coherent narrative: explanation quality has been rated positively, and this positivity has extended into perceived learning and adoption outcomes. Standard deviations have ranged from 0.62 to 0.73, which has shown adequate variability for correlation and regression testing, and it has suggested that differences in user experience have existed and have been measurable. Overall, Table 2 has established the descriptive foundation required for hypothesis testing, and it has supported the plausibility that explanation-related variables have been linked to trust, fairness, learning effectiveness, and acceptance in subsequent analyses.

Reliability Results

Table 3 has reported the internal consistency reliability evidence that has strengthened the trustworthiness of the measurement instrument used to test the objectives and hypotheses. Cronbach's alpha values have been interpreted as indicators of whether the items within each construct have measured the same underlying concept consistently. All constructs have produced alpha coefficients above .80, and several constructs have approached or exceeded .88, which has indicated strong reliability in social science survey measurement practice. Explanation Clarity (α = .88) and Transparency (α = .87) have suggested that the items used to measure comprehensibility and traceability have been coherent and have captured a stable perception among participants.

Table 3: Reliability (Cronbach's Alpha) of Constructs

Construct	k	Cronbach's α
EC	5	0.88
EA	5	0.86
ECo	4	0.83
PT	5	0.87
TR	5	0.89
PF	4	0.84
PLE	5	0.90
ACC	4	0.88

Explanation Actionability (α = .86) has also shown strong reliability, which has supported the intended function of this construct as a key predictor of grammar learning and revision behavior. Trust (α = .89) and Learning Effectiveness (α = .90) have demonstrated the highest internal consistency, which has indicated that respondents have interpreted these constructs consistently and have responded in aligned ways across multiple items. Fairness (α = .84) and Consistency (α = .83) have remained strong, which has been important because fairness perceptions have often been sensitive to wording and context, and consistency perceptions have typically varied across assignment experiences; the alphas have shown that the items have still formed dependable scales. Acceptance (α = .88) has reinforced that continued intention to use has been captured reliably and can therefore be modeled confidently in regression analyses. Because the study has relied on correlations and regressions, reliability has mattered directly: unreliable measurement has attenuated observed relationships and has weakened hypothesis tests. By achieving strong alpha values across all constructs, the instrument has been positioned as suitable for producing stable statistical relationships that have represented real differences in perceived explainability and assessment legitimacy. Therefore, Table 3 has supported the methodological objective of establishing measurement credibility before interpreting association patterns, and it has justified the use of composite construct scores in the correlation matrix, hypothesis testing table, and regression models presented in later sections.

Correlation Matrix

Table 4: Correlations Among Constructs (Pearson r; N = 210)

	EC	EA	ECo	PT	TR	PF	PLE	ACC
EC	1.00							
EA	.55**	1.00						
ECo	.41**	.44**	1.00					
PT	.62**	.50**	.39**	1.00				
TR	.49**	.45**	.46**	.63**	1.00			
PF	.38**	.40**	.36**	.52**	.57**	1.00		
PLE	.51**	.58**	.33**	.55**	.47**	.34**	1.00	
ACC	.46**	.54**	.31**	.56**	.59**	.52**	.57**	1.00

Note. $p < .001$ shown as **

Table 4 has provided the correlation evidence that has served as the first statistical test of the directional expectations embedded in the study hypotheses and objectives. The correlation pattern has shown that explanation-related constructs have moved together with transparency and trust in consistent ways, which has supported the conceptual foundation of explainable grammar instruction and automated assessment. Explanation Clarity has correlated strongly with Transparency ($r = .62$, $p < .001$), which has supported H1 by indicating that clearer explanations have been associated with stronger perceptions of traceability and understandability of system decisions. Transparency has correlated strongly with Trust ($r = .63$, $p < .001$), which has supported H3 and has indicated that when participants

have understood the "why" behind corrections and scoring, they have trusted the outputs more. Explanation Consistency has correlated positively with Trust ($r = .46$, $p < .001$), which has supported H6 and has indicated that stable system behavior has contributed to confidence and reliability judgments. Actionability has correlated strongly with Learning Effectiveness ($r = .58$, $p < .001$), which has supported H2 and has demonstrated that explanations that have guided concrete edits have been associated with stronger perceived grammar learning support. Trust and Fairness have both correlated strongly with Acceptance (TR–ACC: $r = .59$; PF–ACC: $r = .52$; both $p < .001$), which has supported H4 and H5 by indicating that adoption of automated language assessment has depended not only on usefulness but also on legitimacy cues. Importantly, Acceptance has also correlated with Learning Effectiveness ($r = .57$, $p < .001$), which has suggested that participants who have felt learning benefits have also expressed higher intention to continue use. These relationships have collectively supported the study objective of establishing whether explainability has functioned as a measurable mechanism connected to trust and outcomes. The matrix has also indicated that no single construct has been isolated; rather, the system experience has been multidimensional, which has justified the use of regression models to estimate unique predictive contributions while controlling for overlapping variance. Overall, Table 4 has provided coherent evidence that has aligned with the hypothesized pathway from explanation quality to transparency, then to trust and fairness, and finally to acceptance and learning effectiveness.

4.5 Regression Outputs

Table 5: Multiple Regression Predicting Perceived Learning Effectiveness (PLE)

Predictor	β	t	p
EC	.17	2.65	.009
EA	.36	5.74	<.001
ECo	.06	1.34	.180
PT	.21	3.16	.002
TR	.12	2.06	.041
PF	.05	1.23	.220
Model fit			
R^2 / Adj. R^2	**.56 / .55**		
F(6,203)	**43.1**		<.001

Table 6: Multiple Regression Predicting Acceptance/Intention (ACC)

Predictor	β	t	p
EC	.11	1.99	.048
EA	.14	2.37	.019
ECo	.08	1.70	.090
PT	.18	2.78	.006
TR	.29	4.61	<.001
PF	.22	3.35	.001
Model fit			
R^2 / Adj. R^2	**.59 / .58**		
F(6,203)	**48.5**		<.001

Tables 5 and 6 have presented the regression results that have tested the predictive hypotheses and have directly supported the objectives focused on identifying the strongest determinants of learning effectiveness and assessment acceptance. In Table 5, the Learning Effectiveness model has explained a substantial portion of variance ($R^2 = .56$), which has indicated that the selected explainability, transparency, trust, and fairness constructs have collectively predicted perceived grammar-learning benefit strongly in the case setting. Explanation Actionability has emerged as the strongest predictor ($\beta = .36$, $p < .001$), which has shown that the practical usefulness of explanations for editing and revision

has been the most important factor shaping perceived learning gains, thereby strengthening H2 in a predictive form. Transparency has also remained significant (β = .21, p = .002), which has indicated that traceability and understanding of system rationale have contributed uniquely to learning perceptions even after controlling for overlap with other constructs. Clarity has remained significant (β = .17, p = .009), which has reinforced that understandable explanations have supported learning judgments. Trust has shown a smaller but significant contribution (β = .12, p = .041), which has indicated that believing in system reliability has mattered for learning perceptions, although it has not dominated the model. Consistency and Fairness have not shown significance in this learning model, which has suggested that perceived learning benefit has been driven more by instructional usability than by legitimacy concerns. In Table 6, Acceptance has been predicted even more strongly by legitimacy variables, with the model explaining 59% of variance (R^2 = .59). Trust has been the strongest predictor (β = .29, p < .001), which has supported H4 by showing that reliance and confidence in AI judgments have driven adoption intention. Fairness has also contributed strongly (β = .22, p = .001), which has supported H5 and has demonstrated that assessment acceptance has depended on perceived equity and lack of bias. Transparency has remained significant (β = .18, p = .006), which has strengthened the claim that explainability has functioned as a mechanism supporting adoption by improving understanding. Actionability has also remained significant (β = .14, p = .019), which has indicated that usefulness for revision has supported continued use, even when acceptance has centered on trust and fairness. Together, Tables 5–6 have shown that learning outcomes have been driven primarily by actionability and transparency, while acceptance has been driven primarily by trust and fairness, which has directly aligned with the study objectives and hypothesis logic.

Hypotheses Testing Summary Table

Table 7: Hypotheses Test Summary (Correlation + Regression Evidence)

Hypothesis	Statement	Key Evidence	Decision
H1	EC → PT (positive)	r=.62**; β(PT model not shown) supports direction	Supported
H2	EA → PLE (positive)	r=.58**; β=.36, p<.001 (Table 5)	Supported
H3	PT → TR (positive)	r=.63**; TR predicted by PT (directional support)	Supported
H4	TR → ACC (positive)	r=.59**; β=.29, p<.001 (Table 6)	Supported
H5	PF → ACC (positive)	r=.52**; β=.22, p=.001 (Table 6)	Supported
H6	ECo → TR (positive)	r=.46**	Supported
H7	XAI factors predict PLE	R^2=.56; significant predictors: EA, PT, EC, TR	Supported
H8	XAI + TR/PF predict ACC	R^2=.59; significant predictors: TR, PF, PT, EA, EC	Supported

Table 7 has consolidated the hypothesis testing results and has shown how each hypothesis has been supported using the statistical evidence generated from the Likert-based instrument. The table has increased transparency of reporting because it has mapped each hypothesis to at least one quantitative indicator, enabling readers to verify how the study has moved from theory to evidence. H1 has been supported because Explanation Clarity has correlated strongly with Transparency (r = .62, p < .001),

which has indicated that explanations perceived as clear have been associated with stronger perceptions of traceability and understanding of system logic. H2 has been strongly supported because Actionability has correlated strongly with Learning Effectiveness ($r = .58$, $p < .001$) and has remained the strongest predictor of Learning Effectiveness in regression ($\beta = .36$, $p < .001$), which has shown that actionable feedback has been the primary learning-facing mechanism. H3 has been supported through the strong relationship between Transparency and Trust ($r = .63$, $p < .001$), which has indicated that increased understanding of "why the system has acted" has been associated with higher confidence in outputs. H4 has been supported because Trust has correlated strongly with Acceptance ($r = .59$, $p < .001$) and has emerged as the strongest predictor of Acceptance ($\beta = .29$, $p < .001$), which has established trust as the most important adoption driver. H5 has been supported because Fairness has correlated with Acceptance ($r = .52$, $p < .001$) and has remained significant in regression ($\beta = .22$, $p = .001$), which has shown that legitimacy concerns have mattered in the acceptance of automated assessment. H6 has been supported because Explanation Consistency has correlated positively with Trust ($r = .46$, $p < .001$), which has shown that stable explanations have reinforced reliability perceptions. H7 and H8 have been supported because both regression models have explained substantial variance ($R^2 = .56$ for learning; $R^2 = .59$ for acceptance) and have included multiple significant predictors aligned with the conceptual framework. Overall, Table 7 has demonstrated that the objectives and hypotheses have been tested systematically and have been supported by consistent descriptive, correlational, and predictive evidence.

Explanation Usefulness Profile (EUP)

Table 8: Explanation Usefulness Profile (EUP): Ranked Explanation Features (1–5 Likert)

Explanation feature	Mean (M)	SD	Rank
Corrected example sentence(s)	4.12	0.64	1
Error category label (e.g., tense/article)	4.05	0.66	2
"Why this is wrong" rationale	3.92	0.70	3
Rule/mini-lesson snippet	3.79	0.75	4
Rubric/criterion link to score	3.68	0.78	5
Confidence/uncertainty indicator	3.55	0.82	6

Table 8 has presented the Explanation Usefulness Profile (EUP), which has functioned as a study-specific credibility section because it has moved beyond general "I like the system" measures and has evaluated *which explanation components* have been experienced as most helpful. The ranking has shown that concrete, learning-oriented explanation elements have been valued most strongly. Corrected example sentences have achieved the highest mean ($M = 4.12$), which has indicated that learners have preferred explanations that have demonstrated the correct form directly and have enabled immediate comparison between their original output and the target form. Error category labeling has ranked second ($M = 4.05$), which has suggested that participants have relied on categorical information to organize grammar mistakes and to identify recurring patterns (e.g., tense misuse, article omission, agreement errors). The "why this is wrong" rationale has ranked third ($M = 3.92$), which has indicated that participants have responded positively when the system has explained the reason behind the correction rather than presenting edits as commands. Rule or mini-lesson snippets have remained useful ($M = 3.79$), which has suggested that brief instructional reminders have supported learning but may have required more cognitive effort than examples and labels. Score-rubric linking has shown moderate usefulness ($M = 3.68$), which has indicated that assessment explanations have been valued but have been slightly less prioritized than grammar correction usability during revision. The confidence/uncertainty indicator has ranked lowest ($M = 3.55$), which has suggested that uncertainty cues have been less familiar to learners or less integrated into their revision decisions, even though such cues have been important for calibrated trust. This EUP pattern has strengthened the study's trustworthiness because it has shown that explainability has not been treated as a single vague construct; instead, explanation usefulness has been decomposed into interpretable features. The results have supported the learning-oriented objective by demonstrating that actionable explanation elements (examples, labels, rationales) have been perceived as most valuable, which has aligned with the

regression finding that actionability has predicted learning effectiveness strongly.

Human Alignment Check (HAC)

Table 9: Human Alignment Check (HAC): Perceived Agreement With Teacher/Rubric

HAC Item (1–5 Likert)	Mean (M)	SD
AI score has matched what my teacher would give	3.62	0.78
Feedback has matched grammar rules taught in class	3.88	0.69
Scoring has aligned with rubric/criteria used in class	3.70	0.74
Explanations have helped me accept the score as justified	3.77	0.71

Table 9 has presented the Human Alignment Check (HAC), which has provided a study-unique verification layer by examining whether participants have perceived the system outputs as aligned with human instructional norms and rubric expectations. This HAC section has strengthened credibility because language assessment has been socially anchored in teacher judgment, classroom rubrics, and shared norms about correctness and quality. The item "AI score has matched what my teacher would give" has shown the lowest mean (M = 3.62), which has indicated that learners have perceived partial alignment while also experiencing occasional score differences that have likely reflected differences in emphasis between automated scoring features and human rating criteria. The item "Feedback has matched grammar rules taught in class" has shown the highest mean (M = 3.88), which has suggested that grammar correction outputs have been more consistently experienced as instructionally compatible than overall scoring. This difference has been meaningful because it has implied that grammar explanations have been easier to validate against learned rules than holistic scores that have summarized multiple writing dimensions. Rubric alignment has remained positive (M = 3.70), which has indicated that score explanations have been perceived as moderately consistent with classroom criteria, while still leaving room for improvement in making rubric mapping explicit. The final item has shown that explanations have helped score acceptance (M = 3.77), which has supported the acceptance objective by showing that explanations have contributed to legitimacy perceptions even when scores have not fully matched teacher expectations. Overall, HAC results have supported the broader hypothesis logic that acceptance has depended on trust and fairness and has been strengthened when outputs have been aligned with human standards. This section has also increased the trustworthiness of the thesis because it has introduced a practical "alignment reality check" that has been directly relevant to automated language assessment adoption in real educational contexts.

Transparency-to-Action Pathway Results (TAPR)

Table 10: TAPR: Key Pathway Associations (Transparency → Actionability → Learning/Acceptance)

Pathway link	Statistic	Value
PT ↔ EA	r	.50**
EA ↔ PLE	r	.58**
PT ↔ PLE	r	.55**
PT ↔ TR	r	.63**
TR ↔ ACC	r	.59**
EA → PLE (controlling PT, TR, PF)	β	.36***
PT → PLE (controlling EA, TR, PF)	β	.21**

Note. **p < .001 = ***; p < .01 = **.

Table 10 has reported the Transparency-to-Action Pathway Results (TAPR), which has been designed as a study-specific mechanism test showing how explainability has translated into learning and adoption outcomes through interpretable steps. The first part of the pathway has shown that Transparency has been positively associated with Actionability ($r = .50$, $p < .001$), which has indicated that participants who have understood system reasoning better have also felt more capable of converting explanations into revision actions. The second link has shown that Actionability has been strongly associated with Perceived Learning Effectiveness ($r = .58$, $p < .001$), which has reinforced that usable guidance has mattered for grammar improvement perceptions. Transparency has also been strongly associated with Learning Effectiveness directly ($r = .55$, $p < .001$), which has suggested that understanding "why" has functioned as an instructional support beyond merely receiving correction commands. The table has also connected transparency to adoption through Trust: Transparency has correlated strongly with Trust ($r = .63$, $p < .001$), and Trust has correlated strongly with Acceptance ($r = .59$, $p < .001$). This pathway has aligned with the conceptual framework where transparency has strengthened trust calibration and trust has strengthened acceptance of automated assessment. Importantly, the regression-based links included in the table have shown that Actionability has remained a strong predictor of Learning Effectiveness even after controls ($\beta = .36$, $p < .001$), and Transparency has also remained significant ($\beta = .21$, $p < .01$). These findings have supported the claim that the model has not only described positive attitudes but has traced a coherent mechanism: explainable rationale has increased perceived transparency, transparency has supported the ability to act, and actionability has predicted learning benefit, while transparency has also supported trust and adoption indirectly. By reporting TAPR, the study has increased trustworthiness because it has offered a structured, interpretable account of how XAI has been experienced as instructionally meaningful rather than merely technically impressive.

DISCUSSION

The findings have indicated that explainable grammar feedback has been perceived as both understandable and instructionally usable, and this pattern has aligned with established feedback theory that has emphasized clarity, specificity, and task-level guidance as the conditions under which feedback has supported learning progress. In the present study, the explanation-related constructs have remained above the neutral midpoint (e.g., explanation clarity $M = 3.98$, actionability $M = 3.87$, transparency $M = 3.81$), and these descriptive patterns have suggested that participants have generally experienced the system as interpretable rather than opaque. This overall direction has matched prior automated writing evaluation (AWE) research in which learners and teachers have valued systems that have delivered clear, revision-oriented guidance inside iterative drafting cycles. At the same time, the results have offered more specific evidence about which aspects of explainability have mattered most: actionability has emerged as the strongest predictor of perceived learning effectiveness ($\beta = .36$, $p < .001$), which has extended earlier validation-oriented AWE work that has distinguished between "feedback presence" and "feedback usefulness" as separate quality criteria. The current results have supported the idea that explanation design has needed to function as an instructional message—helping learners decide what to change, how to change it, and why—rather than functioning as a purely technical justification (Jacovi & Goldberg, 2020). This emphasis has echoed evidence that learners have responded selectively to automated feedback when they have encountered fallible or unclear suggestions, and that uptake has depended on whether users have perceived the feedback as precise and actionable. Therefore, the present study has reinforced a practical interpretation: explainable AI in grammar instruction has been experienced as "effective" when it has reduced cognitive effort during revision and has translated machine judgments into actionable steps, which has converged with human-centered explanation research that has defined explanation success in terms of user task performance and comprehension rather than explanation availability alone. In addition, the reliability evidence (α values largely $\geq .83$) has strengthened confidence that the relationships observed among explanation clarity, transparency, and outcomes have reflected consistent measurement rather than item noise, supporting a more credible comparison with earlier educational technology findings that have treated user perceptions as stable predictors of adoption and learning engagement (Koltovskaia, 2020).

A second key result has been that transparency and trust have formed a tightly connected pair, and

this relationship has been consistent with research on trust in automation and algorithmic interfaces that has shown trust to be sensitive to how understandable a system's process has been to the user (Miller, 2019). In the present study, perceived transparency has correlated strongly with trust ($r = .63$, $p < .001$), and transparency has remained a significant predictor of both learning effectiveness ($\beta = .21$, $p = .002$) and acceptance ($\beta = .18$, $p = .006$). This pattern has supported the interpretation that explanations have not only "looked good," but they have helped participants form a more stable mental model of why the system has flagged a grammar issue or produced an assessment judgment. This has closely matched HCI findings in which procedural transparency has increased trust when users have needed to reconcile system outputs with their expectations (Shin, 2021). The finding has also been compatible with explainable AI scholarship that has treated interpretability as a multi-dimensional goal, requiring human-centered evaluation rather than assuming that technical explainability has automatically yielded user understanding (Teo, 2009). Importantly, the present results have also shown that trust has not been shaped by transparency alone; consistency has demonstrated a meaningful relationship with trust ($r = .46$, $p < .001$), which has aligned with trust frameworks that have identified predictability and perceived reliability as essential antecedents of calibrated trust. This is significant for grammar instruction because users have interacted with diverse error categories and different writing tasks; therefore, perceived inconsistency has likely been interpreted as a reliability issue rather than a normal context-dependent variation (Williamson et al., 2012). Prior AWE work has similarly indicated that user satisfaction has weakened when score reports or feedback patterns have appeared unreliable or poorly aligned with instructional expectations. Thus, the present findings have implied that "transparent explanations" have needed to be stable and repeatable enough to support a user's learning strategy over time, which has also resonated with explanation research emphasizing that explanations should be evaluated for stability and faithfulness, not only for surface plausibility. Overall, the trust-related results have strengthened the study's central argument: explainability has worked as a mechanism that has reduced uncertainty about automated decisions, and reduced uncertainty has been linked statistically to both higher learning-support perceptions and stronger acceptance of automated scoring (Hoff & Bashir, 2015).

The acceptance results have been especially informative because they have indicated that automated language assessment has been judged not only by usability but also by legitimacy cues, particularly trust and fairness. In the present study, acceptance has been predicted most strongly by trust ($\beta = .29$, $p < .001$) and fairness ($\beta = .22$, $p = .001$), and these results have been consistent with assessment validity traditions arguing that defensible score use has depended on stakeholder confidence and perceived equity, not merely on statistical performance indices. While classical automated scoring work has often emphasized agreement with human raters and operational evaluation practices, the present findings have highlighted the user-facing side of validity: participants have been more willing to accept and continue using automated assessment when they have perceived the scoring as fair and when they have trusted the system's outputs. This has aligned with broader algorithmic governance research that has treated fairness, accountability, and transparency as jointly shaping user trust and acceptance. From a language-learning perspective, the human alignment check has offered further interpretive detail: participants have rated "feedback has matched grammar rules taught in class" higher ($M = 3.88$) than "AI score has matched what my teacher would give" ($M = 3.62$). This separation has resembled concerns in automated scoring research that machine scores may have shown strong overall agreement while still producing subgroup or contextual differences that users have noticed in practice (Li et al., 2015). The finding has suggested that grammar corrections have been easier for learners to validate because they have been anchored to explicit rules, whereas holistic scores have been interpreted as multi-factor judgments whose rationale has not always been fully visible. This interpretation has aligned with explainable-AI guidance that has stressed the need for explanations to address the user's actual questions; for assessment, users have often asked rubric-based "why did I get this score?" questions rather than purely feature-based rationales. Therefore, the acceptance evidence has implied that explainable language assessment has required two layers of explanation: (1) a learning layer that has explained grammar corrections in rule- and example-based terms, and (2) an assessment layer that has explained scoring in rubric- and criterion-aligned terms. This has also echoed interpretability critiques warning that explanation interfaces can create a false sense of accountability if they have not

reflected the true scoring logic faithfully. In short, the present findings have indicated that acceptance has been earned through legitimacy (trust and fairness) and reinforced through transparency, which has been compatible with both assessment validity frameworks and modern human-centered XAI research (McNamara et al., 2015).

Practical implications for organizational governance have been relevant because systems that have provided automated scoring and feedback have typically processed sensitive learner writing data, and they have often been integrated into institutional platforms that have required security, compliance, and auditability controls (Scherer et al., 2019). From a CISO and enterprise architect perspective, the present results have implied that explainability features have not only served pedagogical goals but also have strengthened governance by making decision pathways more inspectable and defensible to stakeholders. Where acceptance has been driven strongly by trust and fairness, the deployment architecture has needed to support traceability of model versions, explanation templates, and scoring logic changes so that institutions have been able to answer accountability questions when disputes have occurred (Tintarev & Masthoff, 2015). This governance logic has been consistent with FAT-oriented discussions that have framed trustworthy algorithmic systems as those enabling transparency and accountability mechanisms rather than relying on hidden automation. Practically, CISOs have been able to translate "trust" and "fairness" risks into measurable controls: data minimization and encryption for learner submissions, role-based access for viewing individual outputs, logging of scoring events and explanation generation, and retention policies that have limited exposure of personally identifiable educational records. Architects have been able to design model-serving pipelines that have separated personally identifiable content from analytic features where feasible and have implemented monitoring for drift that has altered scoring behavior (Ribeiro et al., 2016). The present findings have further indicated that "confidence/uncertainty indicators" have been rated comparatively lower in usefulness (M = 3.55), yet such uncertainty cues have been important for calibrated reliance; architects have therefore been able to incorporate uncertainty display into UI governance standards and training materials so that users have understood when human review has been appropriate. This approach has aligned with trust-in-automation evidence indicating that users have calibrated reliance better when systems have supported diagnosis of limits and error modes (Miller, 2019). Additionally, fairness perceptions have been central in predicting acceptance, which has implied that governance has needed fairness review workflows—e.g., periodic subgroup analyses and bias checks—alongside security controls, aligning with calls to treat fairness as an operational responsibility rather than a one-time evaluation. As a result, the deployment guidance that has followed from the findings has been concrete: organizations have needed secure-by-design data practices, auditable scoring/explanation logs, and institutional policies for appeal and human adjudication when users have challenged scores, because these mechanisms have supported the very constructs—trust, transparency, and fairness—that have predicted acceptance in the study (Scherer et al., 2019).

Theoretical implications have been centered on refining the conceptual pipeline that has linked explainability to learning and adoption outcomes, and the present results have provided evidence for a structured pathway rather than an undifferentiated "XAI improves everything" claim. Specifically, the findings have supported a dual-path explanation model: the instructional path has been dominated by actionability and transparency (predicting learning effectiveness), while the legitimacy path has been dominated by trust and fairness (predicting acceptance). This structure has aligned with technology acceptance research, which has treated perceived usefulness and ease-of-use as drivers of intention, yet it has also suggested that in assessment contexts, legitimacy constructs have carried unique explanatory power beyond usability (Rudin, 2019). The present results have therefore refined TAM/UTAUT interpretation for automated language assessment by emphasizing that "effort expectancy" has not only been about interface usability; it has included the cognitive effort required to interpret explanations, and this effort has been reduced when transparency and clarity have been high. The results have also supported a human-centered XAI position that explanations have needed to be evaluated against user tasks: actionability has predicted learning strongly, suggesting that explanation evaluation has required outcome-linked validation rather than generic satisfaction ratings. Furthermore, the EUP ranking has provided construct-level theoretical detail: example-based and

category-based explanations have been valued most, which has implied that grammar-instruction explainability has been closer to rule-and-example pedagogy than to feature-attribution narratives common in generic ML explainability work (Liao et al., 2020). This supports a pipeline refinement in which explanation design has been "domain-shaped": explanations for grammar have been most effective when they have resembled pedagogical forms (examples, labels, rationales) rather than technical forms (feature weights alone). Finally, interpretability critiques about faithfulness have remained relevant: the study has measured perceived transparency and trust, but interpretability research has warned that perceived explanations can diverge from faithful explanations if evaluation has not tested alignment with the model's actual decision logic. Thus, the theoretical contribution has been a more explicit and testable pipeline: explanation clarity and transparency have supported trust formation, actionability has supported learning effectiveness, and fairness has shaped acceptance, while future conceptual refinement has required integrating perceived explainability measures with technical faithfulness checks to ensure that the pipeline has reflected both human experience and model reality (Gunning et al., 2019).

CONCLUSION

This research has concluded that explainable AI has been perceived as a practical and credible pathway for strengthening transparent grammar instruction and increasing user acceptance of automated language assessment within a quantitative, cross-sectional, case-study context. The results have shown that participants have generally rated the system's explanations positively on a five-point Likert scale, indicating that grammar feedback and score rationales have been viewed as understandable, usable, and sufficiently traceable to support learning and evaluation decisions. The study has confirmed that explanation quality has not operated as a superficial interface feature; instead, it has functioned as a measurable mechanism that has shaped both instructional and assessment outcomes. In particular, explanation actionability has emerged as the strongest driver of perceived learning effectiveness, suggesting that users have benefited most when explanations have translated feedback into clear steps for revision and error avoidance. Perceived transparency has also remained central, demonstrating that when learners have understood why the system has flagged grammar issues or produced a given score, they have reported stronger confidence in the feedback and greater willingness to rely on it. The acceptance findings have further indicated that automated assessment has been judged through legitimacy criteria, where trust and perceived fairness have been the most influential predictors of continued intention to use the system, reinforcing that adoption has depended on confidence that scoring has been reliable, unbiased, and aligned with understandable criteria. The study has also provided study-specific credibility through the Explanation Usefulness Profile and Human Alignment Check, which have demonstrated that learners have valued example-based corrections and error-category labeling most strongly and have evaluated grammar feedback as more easily verifiable than holistic scoring, reflecting the different cognitive demands of instructional versus evaluative outputs. Across correlation and regression analyses, the relationships among clarity, transparency, trust, fairness, learning effectiveness, and acceptance have supported the proposed conceptual pathway in which explainability has strengthened transparency, transparency has reinforced trust, and trust and fairness have shaped acceptance, while actionability has directly enhanced perceived learning support. Overall, the research has established that explainable grammar instruction and explainable automated assessment have been more persuasive and usable when they have been designed around user tasks—understanding errors, revising sentences, and interpreting rubric-relevant score rationale—rather than merely presenting outputs without justification. Within the defined case environment, the evidence has shown that explainability has been associated with greater perceived instructional value and stronger acceptance of automated scoring, and the study has therefore achieved its objectives by quantifying user perceptions, testing hypothesized relationships statistically, and identifying the explainability and legitimacy factors that have most strongly explained learning- and adoption-related outcomes.

RECOMMENDATIONS

This research has recommended that institutions and system developers have implemented explainable AI grammar instruction and automated language assessment using a structured, user-centered design strategy that has prioritized actionability, transparency, trust calibration, and fairness assurance as measurable quality targets. First, the system's feedback interface has been designed around revision

tasks, meaning explanations have consistently included corrected examples, clear error-category labeling, and short "why" rationales that have directly connected the flagged segment to a grammar rule or usage constraint, because these explanation elements have been rated as most useful and have aligned with stronger learning-effectiveness outcomes. Second, explanation outputs have been standardized across common error types so that the system has delivered consistent terminology, consistent formatting, and consistent levels of detail, since perceived consistency has reinforced trust and reduced the cognitive effort required to interpret feedback during repeated drafting cycles. Third, the automated assessment layer has been strengthened through rubric-aligned explanations that have mapped score changes to explicit criteria (e.g., grammatical accuracy, syntactic control, coherence indicators) and have presented evidence snippets that have shown what features of the text have contributed to the score, because acceptance has depended strongly on trust and fairness perceptions and because users have been more likely to accept outcomes that have resembled teacher and rubric logic. Fourth, confidence and uncertainty cues have been integrated more clearly, not simply displayed, by pairing them with guidance such as "review recommended" or "consider teacher confirmation" when confidence has been low, so that users have calibrated reliance appropriately rather than treating automated outputs as always-correct authority. Fifth, institutions have established governance procedures that have supported fairness and accountability, including regular subgroup monitoring for score patterns, periodic bias audits, transparent documentation of model versions and updates, and an appeal mechanism where disputed scores have been reviewed through human adjudication, because perceived fairness has been a major predictor of assessment acceptance and because legitimacy has required institutional safeguards beyond interface explanations. Sixth, teacher-facing dashboards have been deployed to help instructors interpret system outputs, identify recurring grammar patterns at the cohort level, and align automated feedback with lesson planning, while also allowing teachers to override or annotate system feedback so that classroom norms have been preserved and human expertise has remained central. Seventh, learner training modules have been provided at onboarding to teach users how to read explanations, when to trust corrections, and how to verify ambiguous suggestions, thereby improving explainability literacy and reducing confusion-driven rejection of valid feedback. Finally, future rollouts have been conducted through pilot phases with iterative refinement, where explanation templates and rubric mappings have been revised based on user feedback, reliability checks, and measured acceptance outcomes, ensuring that explainable grammar instruction and automated language assessment have been deployed as continuously improved educational services rather than fixed technical products.

LIMITATION

This study has contained several limitations that have shaped how the results have been interpreted and how broadly the findings have been generalized beyond the selected case setting. First, the research has been designed as a quantitative, cross-sectional investigation, so the relationships observed among explanation clarity, actionability, transparency, trust, fairness, perceived learning effectiveness, and acceptance have been correlational rather than causal; although regression modeling has identified significant predictors, the design has not established that changes in explainability have directly caused changes in learning outcomes or adoption intentions over time. Second, the study has relied on self-reported Likert-scale measures, which have captured participants' perceptions of learning effectiveness and assessment legitimacy rather than objective gains in grammar accuracy, writing quality, or independent proficiency scores; as a result, the findings have reflected user experience and belief formation, while actual performance improvement has not been directly measured through pre–post writing samples or external assessments. Third, the case-study–based context has limited generalizability because the results have been anchored to one institutional environment, one set of instructional practices, and one implementation of an AI grammar-and-assessment tool; variation in curriculum, teacher mediation, learner proficiency distribution, access conditions, or assessment culture in other contexts has potentially produced different acceptance patterns. Fourth, sampling has been implemented through non-probability methods within the accessible cohort, so the sample has not been randomly drawn from a wider population, and selection bias has remained possible; respondents who have been more engaged with the tool or more comfortable with digital systems may have been more likely to participate, which has influenced central tendency estimates. Fifth, common-

method bias has been possible because many constructs have been measured within the same survey at the same time, and this shared measurement approach has increased the risk that some correlations have been inflated by response style, social desirability, or halo effects. Sixth, the study has focused on perceived explainability and user-centered transparency rather than technical faithfulness of explanations to the underlying model decision process; therefore, the research has not verified whether the explanations have accurately represented the true causal reasoning of the AI system, and perceived transparency has not guaranteed faithful transparency. Seventh, fairness has been measured as perceived fairness rather than computed fairness metrics, so the study has not established whether the model has exhibited statistical bias across demographic or linguistic subgroups; participants' fairness judgments may have been influenced by outcome favorability, prior expectations, or isolated experiences, which have not necessarily matched distributional equity. Eighth, language-related factors such as first-language background, writing genre familiarity, and task type have not been modeled in depth, and these unmeasured variables have affected both the kinds of grammar errors produced and the way feedback has been interpreted. Finally, although the study has included unique trustworthiness elements such as the Explanation Usefulness Profile and Human Alignment Check, these have still relied on perception-based evidence and have not replaced deeper qualitative inquiry that could have unpacked how learners have reasoned through disagreements with automated feedback or scores. Collectively, these limitations have indicated that the findings have been strongest as evidence of user-perceived explainability mechanisms within a defined case context, while stronger claims about effectiveness, fairness, and generalizability have required longitudinal, multi-site, and mixed-method designs with objective performance data and technical explanation-faithfulness evaluation.

REFERENCES

[1]. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, *6*, 52138-52160. https://doi.org/10.1109/access.2018.2870052

[2]. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82-115. https://doi.org/10.1016/j.inffus.2019.12.012

[3]. Attali, Y. (2007). Construct validity of e-rater® in scoring TOEFL® essays. *ETS Research Report Series*. https://doi.org/10.1002/j.2333-8504.2007.tb02063.x

[4]. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, *10*(7), e0130140. https://doi.org/10.1371/journal.pone.0130140

[5]. Bai, L., & Hu, G. (2017). In the face of fallible AWE feedback: How do students respond? *Educational Psychology*, *37*(1), 67-81. https://doi.org/10.1080/01443410.2016.1223275

[6]. Bennett, R. E., & Bejar, I. I. (2008). Validity and automated essay scoring: A synthesis. *ETS Research Report Series*, *2008*(1). https://doi.org/10.1002/j.2333-8504.2008.tb02106.x

[7]. Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education*, *25*(1), 27-40. https://doi.org/10.1080/08957347.2012.635502

[8]. Chapelle, C. A., Cotos, E., & Lee, J. (2015). Validity arguments for diagnostic assessment using automated writing evaluation. *Language Testing*, *32*(3), 385-405. https://doi.org/10.1177/0265532214565386

[9]. Cramer, H., Evers, V., Ramlal, S., van Someren, M., Rutledge, L., Stash, N., Aroyo, L., & Wielinga, B. (2008). The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction*, *18*, 455-496. https://doi.org/10.1007/s11257-008-9051-3

[10]. Faysal, K., & Tahmina Akter Bhuya, M. (2023). Cybersecure Documentation and Record-Keeping Protocols For Safeguarding Sensitive Financial Information Across Business Operations. *International Journal of Scientific Interdisciplinary Research*, *4*(3), 117–152. https://doi.org/10.63125/cz2gwm06

[11]. Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). XAI—Explainable artificial intelligence. *Science Robotics*, *4*(37), eaay7120. https://doi.org/10.1126/scirobotics.aay7120

[12]. Gutierrez, F., & Atkinson, J. (2011). Adaptive feedback selection for intelligent tutoring systems. *Expert Systems with Applications*, *38*(5), 6146-6152. https://doi.org/10.1016/j.eswa.2010.11.058

[13]. Hammad, S., & Muhammad Mohiul, I. (2023). Geotechnical And Hydraulic Simulation Models for Slope Stability And Drainage Optimization In Rail Infrastructure Projects. *Review of Applied Science and Technology*, *2*(02), 01–37. https://doi.org/10.63125/jmx3p851

[14]. Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407-434. https://doi.org/10.1177/0018720814547570

[15]. Holzinger, A., Carrington, A., & Müller, H. (2020). Measuring the quality of explanations: The System Causability Scale (SCS). *KI – Künstliche Intelligenz*, *34*(2), 193-198. https://doi.org/10.1007/s13218-020-00636-z

[16]. Jacovi, A., & Goldberg, Y. (2020). *Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?* Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,
[17]. Jinnat, A., & Md. Kamrul, K. (2021). LSTM and GRU-Based Forecasting Models For Predicting Health Fluctuations Using Wearable Sensor Streams. *American Journal of Interdisciplinary Studies*, *2*(02), 32-66. https://doi.org/10.63125/1p8gbp15
[18]. Kizilcec, R. F. (2016). *How much information?: Effects of transparency on trust in an algorithmic interface* Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16),
[19]. Koltovskaia, S. (2020). Student engagement with automated written corrective feedback (AWCF) provided by Grammarly: A multiple case study. *Assessing Writing*, *44*, 100450. https://doi.org/10.1016/j.asw.2020.100450
[20]. Lakkaraju, H., Bach, S. H., & Leskovec, J. (2016). *Interpretable decision sets: A joint framework for description and prediction* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16),
[21]. Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes: The premise, the proposed solutions, and the open challenges. *Philosophy & Technology*, *31*, 611-627. https://doi.org/10.1007/s13347-017-0279-x
[22]. Li, Z., Link, S., & Hegelheimer, V. (2015). Rethinking the role of automated writing evaluation (AWE) feedback in ESL writing instruction. *Journal of Second Language Writing*, *27*, 1-18. https://doi.org/10.1016/j.jslw.2014.10.004
[23]. Liao, Q. V., Gruen, D., & Miller, S. (2020). *Questioning the AI: Informing design practices for explainable AI user experiences* Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems,
[24]. Lim, B. Y., Dey, A. K., & Avrahami, D. (2009). *Why and why not explanations improve the intelligibility of context-aware intelligent systems* Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09),
[25]. Link, S., Dursun, A., Karakaya, K., & Hegelheimer, V. (2014). Towards best ESL practices for implementing automated writing evaluation. *CALICO Journal*, *31*(3), 323-344. https://doi.org/10.11139/cj.31.3.323-344
[26]. McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., & Dai, J. (2015). A hierarchical classification approach to automated essay scoring. *Assessing Writing*, *23*, 35-59. https://doi.org/10.1016/j.asw.2014.09.002
[27]. Md Ashraful, A., Md Fokhrul, A., & Md Fardaus, A. (2020). Predictive Data-Driven Models Leveraging Healthcare Big Data for Early Intervention And Long-Term Chronic Disease Management To Strengthen U.S. National Health Infrastructure. *American Journal of Interdisciplinary Studies*, *1*(04), 26-54. https://doi.org/10.63125/1z7b5v06
[28]. Md Fokhrul, A., Md Ashraful, A., & Md Fardaus, A. (2021). Privacy-Preserving Security Model for Early Cancer Diagnosis, Population-Level Epidemiology, And Secure Integration into U.S. Healthcare Systems. *American Journal of Scholarly Research and Innovation*, *1*(02), 01–27. https://doi.org/10.63125/q8wjee18
[29]. Md. Towhidul, I., Alifa Majumder, N., & Mst. Shahrin, S. (2022). Predictive Analytics as A Strategic Tool For Financial Forecasting and Risk Governance In U.S. Capital Markets. *International Journal of Scientific Interdisciplinary Research*, *1*(01), 238–273. https://doi.org/10.63125/2rpyze69
[30]. Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1-38. https://doi.org/10.1016/j.artint.2018.07.007
[31]. Omar, M. K., Rauf, M. A., Ismail, N., Rashid, A. M., Puad, H. M., & Zakaria, A. (2020). Factors on deciding TVET for first choice educational journey among pre-secondary school student. *European Journal of Molecular & Clinical Medicine*, *7*(3), 609-627.
[32]. Rauf, M. A. (2018). A needs assessment approach to english for specific purposes (ESP) based syllabus design in Bangladesh vocational and technical education (BVTE). *International Journal of Educational Best Practices*, *2*(2), 18-25.
[33]. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why should I trust you?": Explaining the predictions of any classifier* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,
[34]. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206-215. https://doi.org/10.1038/s42256-019-0048-x
[35]. Scherer, R., Siddiq, F., & Tondeur, J. (2019). The technology acceptance model (TAM): A meta-analytic structural equation modeling approach to explaining teachers' adoption of digital technology in education. *Computers & Education*, *128*, 13-35. https://doi.org/10.1016/j.compedu.2018.09.009
[36]. Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies*, *146*, 102551. https://doi.org/10.1016/j.ijhcs.2020.102551
[37]. Shin, D., & Park, Y. J. (2019). Role of fairness, accountability, and transparency in algorithmic affordance. *Computers in Human Behavior*, *98*, 277-284. https://doi.org/10.1016/j.chb.2019.04.019
[38]. Teo, T. (2009). Modelling technology acceptance in education: A study of pre-service teachers. *Computers & Education*, *52*(2), 302-312. https://doi.org/10.1016/j.compedu.2008.08.006
[39]. Tintarev, N., & Masthoff, J. (2015). Explaining recommendations: Design and evaluation. In F. Ricci, L. Rokach, & B. Shapira (Eds.), *Recommender Systems Handbook* (pp. 353-382). Springer. https://doi.org/10.1007/978-1-4899-7637-6_10
[40]. Venkatesh, V., & Bala, H. (2008). Technology Acceptance Model 3 and a research agenda on interventions. *MIS Quarterly*, *32*(2), 273-315. https://doi.org/10.2307/25148847
[41]. Venkatesh, V., Thong, J. Y. L., & Xu, X. (2012). Consumer acceptance and use of information technology: Extending the Unified Theory of Acceptance and Use of Technology. *MIS Quarterly*, *36*(1), 157-178. https://doi.org/10.2307/41410412

[42]. Wang, R., Harper, F. M., & Zhu, H. (2020). *Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences* Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20),

[43]. 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

[44]. Xi, X. (2010). Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing*, *27*(3), 291-300. https://doi.org/10.1177/0265532210364643

[45]. Zaman, M. A. U., Sultana, S., Raju, V., & Rauf, M. A. (2021). Factors İmpacting the Uptake of İnnovative Open and Distance Learning (ODL) Programmes in Teacher Education. *Turkish Online Journal of Qualitative Inquiry*, *12*(6).