TESOL Bangladesh Journal
www.tesolbangladeshjournal.com
Volume 3, Issue 1, July 2026
DOI: https://doi.org/10.69907/tbj.v3i1.197

Student Perspectives on Hybrid AI–Human Writing Assessment in Bangladeshi Tertiary EFL Education: A Sequential Explanatory Mixed-Methods Study

Md. Mohaiminul Haq Mim
Graduate Student, Institute of Education and Research, University of Dhaka
E-mail: mdmohaiminulhaqmim@gmail.com | ORCID: 0009-0006-3522-4794

Tasmia Nawrin
Assistant Professor, Institute of Education and Research, University of Dhaka
E-mail: tasmianawrin.ier@du.ac.bd | ORCID: 0000-0002-6518-4731

Abstract
Artificial intelligence (AI) tools are increasingly used by university students to revise English writing, yet their role in formal assessment remains contested. This sequential explanatory mixed-methods study examined Bangladeshi tertiary EFL students' perceptions of hybrid AI–human assessment for academic writing. Eighty survey responses were received; after two invalid test entries were removed, 78 cases were analyzed. Eleven survey respondents then completed asynchronous written interviews. The instrument was informed by the Technology Acceptance Model and TPACK, while quantitative data were analyzed descriptively and with paired-samples tests and correlations. Students rated AI significantly more effective for mechanical skills ($M = 3.756$, $SD = 0.805$) than for higher-order skills ($M = 3.128$, $SD = 0.697$), $t(77) = 5.765$, $p < .001$, $dz = .653$. Two-thirds of the sample (84.62%) preferred a hybrid workflow, whereas one respondent preferred AI-only assessment. Thematic analysis identified efficiency in mechanics, trust in human contextual judgment, a speed–depth trade-off, and concern about opaque errors and grade security. The findings support low-stakes, transparent AI assistance with teachers retaining responsibility for interpretation, feedback, appeals, and final grades.

Article History:
Received: 26/07/2026
Accepted: 18/09/2026
Published: 01/10/2026

Keywords:
AI-assisted assessment; EFL writing; hybrid assessment; student perceptions: Bangladesh

Introduction

Academic writing plays a dominant role in tertiary education. Learners need to write precise and comprehensible English while working on their academic assignments, compositions, research or reports and also while being assessed by their teacher for those works. In Bangladesh, although English plays a significant role across public and private universities, students' opportunities to receive individualized writing feedback vary prominently. Even though grammar, spelling, punctuation and organization are visible indicators of academic control, commenting on these features for large classes creates a recurring workload for instructors and can delay feedback. This practical pressure has

increased interest in digital tools that can provide immediate, repeatable support during drafting and assessment.

Automated writing evaluation and generative AI systems now offer sentence-level correction, rephrasing, vocabulary suggestions and clarifications. Students can access tools such as Grammarly, ChatGPT, Google Gemini, Google Translate and QuillBot without institutional intervention. Though effects depend on task design, learner proficiency and the quality of teacher guidance, reviews of AI-supported language assessment report potential benefits for efficiency and formative feedback (Chen et al., 2025; Fleckenstein et al., 2023). The same systems that detect surface errors may also produce credible but inaccurate comments, over-standardize students' voice or reward features. As a result, such feedback correlates only weakly comments with meaning and argument quality (Kapxhiu, 2025; Tate et al., 2024).

Nonetheless, the distinction between feedback and grading is therefore important. A learner may willingly consult AI feedback while revising an essay, but may not allow the authority of an algorithm to determine their consequential assessments. Assessment decisions affect progression, scholarships and students' sense of objectivity. For that reason, technical accuracy alone is insufficient. Furthermore, an acceptable system also requires transparent criteria, opportunities for review, data protection and an identifiable human decision maker. Research on teachers' adoption of educational AI similarly shows that trust and pedagogical beliefs shape acceptance beyond perceptions of convenience (Choi et al., 2023).

Bangladeshi research studies give some insights about the use and comprehension of digital language teaching learning and assessment. In one of them, learners remarked positive attitude towards social media in second language learning even though they assumed linguistic and technological challenges associated with (Nawrin & Moon, 2024). Authenticity, transparency and consistency in language assessment could be reinforced by unambiguous norms as suggested by studies about using rubrics in tertiary education (Nawrin & Sadek, 2022). Another study, analyzing the classroom discourse in Dhaka city, has reported teacher-controlled classrooms with very limited students' talk time where pedagogical theories, their practices and socio-cultural conditions were not adequately addressed. (Nawrin & Rahman, 2020). Composing these studies, it is predictable that teacher and learner agency, interaction among the participants of the classroom, sociocultural norms need to be incorporated along with technology as a part of assessment system and it is possible when technology will be used something beyond as a software.

Still, evidence from Bangladesh on automated writing evaluation remains limited, particularly from students who would receive the feedback or grades. However, research on Bangladeshi EFL students has shown that learners perceive AI chatbots as useful for organizing content and obtaining personalized feedback while also expressing concerns regarding overdependence, inaccurate information, and ethical issues (Nafeesa, 2025). Much research foregrounds system performance or teachers' intentions, while students' perspectives and preferences regarding the automation and human intervention receive meager consideration. When the concern is both theory and practice studies should be conducted in the concerned context. Thus, the present study examines a Bangladeshi case

through international research on technology acceptance, automated writing assessment, human-in-the-loop decision making and a mixed method inquiry.

The study intends to explore the answer of the following questions:

1. Which writing assessment model do students prefer and how do they evaluate AI and teacher trust?
2. Do students perceive AI as more effective for mechanical writing features than for creativity, originality and paragraph coherence?
3. How are prior AI experience (frequency and duration of use), self-rated writing proficiency and academic year related to trust in AI and to students' intention to accept AI-supported assessment?

A sequential explanatory design was used so that survey patterns could be interpreted alongside students' written accounts. The study is intended to inform cautious, locally responsive policies rather than to endorse full automation.

Literature Review and Theoretical Framework

AI-Supported Writing Assessment: Efficiency and Limits

Natural language processing systems can identify recurring syntactic, lexical and orthographic patterns and return feedback at a speed that is difficult to reproduce manually. In formative settings, this immediacy may increase opportunities for revision and reduce the time teachers spend marking repetitive errors. Meta-analytic evidence indicates that automated feedback can improve writing performance but effects are heterogeneous and depend on how feedback is integrated into instruction (Fleckenstein et al., 2023). A systematic review of AI-enabled language assessment similarly concludes that design, implementation and pedagogical alignment are central to effectiveness (Chen et al., 2025). These findings support viewing AI as an instructional resource rather than an autonomous evaluator.

In addition, Mechanical features are particularly compatible with automation because many can be represented as rule-governed or statistically regular patterns. Spelling, punctuation, agreement and sentence level grammar are easier to detect than rhetorical purpose, cultural implication, originality or the quality of an argument. Studies of automated scoring therefore distinguish between local textual accuracy and higher-order discourse judgments (Alshehri, 2025; Chen et al., 2025). Even when a system generates fluent explanations, fluency does not guarantee that it has interpreted the writer's intended meaning or the assignment's context correctly.

Moreover, holistic scoring introduces additional validity concerns. Systems may be sensitive to length, vocabulary or predictable organization and may assign high scores to texts that display surface sophistication without conceptual depth. The reverse error is just as real, for an algorithm may penalize culturally specific examples, an emerging writer's voice, or unconventional rhetorical choices simply because they depart from its trained patterns. Research on AI essay scoring show that useful output is possible, but performance varies across prompts, criteria and scoring conditions (Tate et al., 2024). Before a tool is used for high stakes decisions, its accuracy must be directly verified

against the target population and writing task because a general reputation for competence provides insufficient evidence on its own.

Generative AI also brings alteration to the assessed material. It can modify, enlarge or create content where the traditional automated writing evaluation generally comments on assessed resources. If students accept extensive rewriting, the final product may no longer represent their unassisted competence. Institutions consequently need clear rules distinguishing feedback, editing, translation and authorship. AI literacy should include the ability to question suggestions, preserve intended meaning, disclose support when required and understand that viable systems may recollect submitted text.

Human Judgment, Rubrics, and Hybrid Models

Human assessors contribute contextual and relational knowledge that automated systems do not possess. Teachers can interpret a response against course objectives, prior performance, classroom discussion and the developmental level expected from a particular learner. They can also ask follow up questions, recognize effort, explain a judgment and revise a score after an appeal. These functions are not incidental rather they are components of procedural fairness. Studies of teacher perceptions commonly value AI for workload relief while retaining human authority over final interpretations (Eriksen & Elstad, 2026; Karaoğlu & Doğan, 2025, Saleh, 2025).

Furthermore, rubrics can connect human and automated assessment by making criteria visible. Nawrin and Sadek (2022) argue that by clarifying standards and increasing reliability rubrics uphold objective language assessment. In a hybrid model, the same principle can constrain AI use where the tool may comment on specified mechanical criteria while teachers evaluate content, audience awareness, reasoning and final achievement. A rubric also gives students a basis for challenging both human and automated feedback. However, transparency requires more than displaying categories where users must understand how evidence leads to a judgment and where uncertainty remains.

Besides, hybrid assessment can be organized in several sequences. In an AI-first workflow, the system provides preliminary comments on mechanics, students revise and the teacher reviews the subsequent work before making the final decision. In a teacher-first workflow, the instructor assesses the work and AI is used for supplementary checking or individualized practice. Parallel review is also possible, with discrepancies prompting discussion. None of the sequence is an absolute for every purpose. The appropriate design depends on stakes, teacher capacity, student proficiency, institutional policy and whether the tool has been evaluated for the relevant genre and learner population.

Nonetheless, the strongest rationale for a hybrid model is therefore not that humans are always accurate and machines are always fast. Human raters can be inconsistent, fatigued or influenced by irrelevant features while automated systems can apply a rule consistently. At the same time, algorithmic consistency can reproduce a consistently invalid rule. Combining sources of judgment is useful only when roles are deliberately allocated, criteria are clear and the human reviewer has time and authority to interrogate the automated output.

Bangladeshi Context and International Relevance

Students, from varied educational settings in Bangladesh, vastly use mobile and social media and much digital and online language practices occur there. Nawrin and Moon's (2024) study of Facebook for second language learning found positive perceptions but also identified linguistic and technological difficulties. This pattern is relevant to AI assessment where familiarity with a platform can encourage use without ensuring critical understanding of its boundaries. Informal adoption may therefore overtake institutional policy, teacher preparation and safeguards for assessment data.

Evidence from universities in Bangladesh indicates that students and teachers alike are integrating artificial intelligence tools like ChatGPT into their academic work. This indicates the expanded use of AI in education (Rahman et al., 2025). Findings from the study also raised questions about responsible use, dependability, and the overall educational impact of the use of AI, indicating that new technology gets adopted without being fully accepted for use. Additionally, local classroom interaction also matters. Nawrin and Rahman (2020) reported teacher-dominated English classroom discourse and limited learner response across secondary schools with varied performance in Dhaka. Although the present study concerns tertiary education, the earlier findings highlight a broader risk and that is technology can either widen opportunities for learner participation or add another one-way source of authoritative feedback. A hybrid model should not simply replace teacher monologue with algorithmic monologue. It should create opportunities for students to interpret feedback, explain choices, request clarification and revise with support.

Although studies from other educational systems describe a comparable tension between efficiency and trust, these local patterns must be addressed within the broader international landscape because the specific interplay of teacher authority, language norms, data protection and local access fundamentally determines how that concern unfolds in practice. TBJ specifically expects single country research to demonstrate such connections. The present study therefore treats Bangladesh as a theoretically informative setting in which rapid student adoption of AI intersects with teacher-centered traditions, multilingual writing practices and emerging national policy discussions.

Emerging national AI frameworks necessitate local institutional governance to ensure assessment integrity. Bangladesh's draft National Artificial Intelligence Policy 2026–2030 emphasizes responsible adoption and it actively promotes the establishment of structured governance mechanisms across educational sectors (ICT Division, Government of Bangladesh, 2026). Although an unratified national draft cannot unilaterally regulate classroom instruction, it compels universities to define their own accountability standards before introducing automated systems into evaluative settings. Consequently, higher education institutions must formulate localized protocols for approved tools, data privacy, disclosure mandates, human oversight and appeal pathways rather than delegating evaluative integrity to unsupervised student experimentation. By institutionalizing these comprehensive safeguards, universities protect pedagogical validity while balancing technological innovation with rigorous human accountability.

Conceptual Framework and Hypotheses

Two frameworks structure the present study. The Technology Acceptance Model (TAM) provides the primary lens through perceived usefulness, trust and behavioral intention

(Davis, 1989), while a subordinate reading of Technological Pedagogical Content Knowledge (TPACK) explains why students endorse a differentiated division of labor between AI and human teachers (Mishra & Koehler, 2006; Koehler et al., 2013). Figure 2 maps the measured variables in this study onto the two frameworks and shows the points at which they converge on the same empirical outcome.

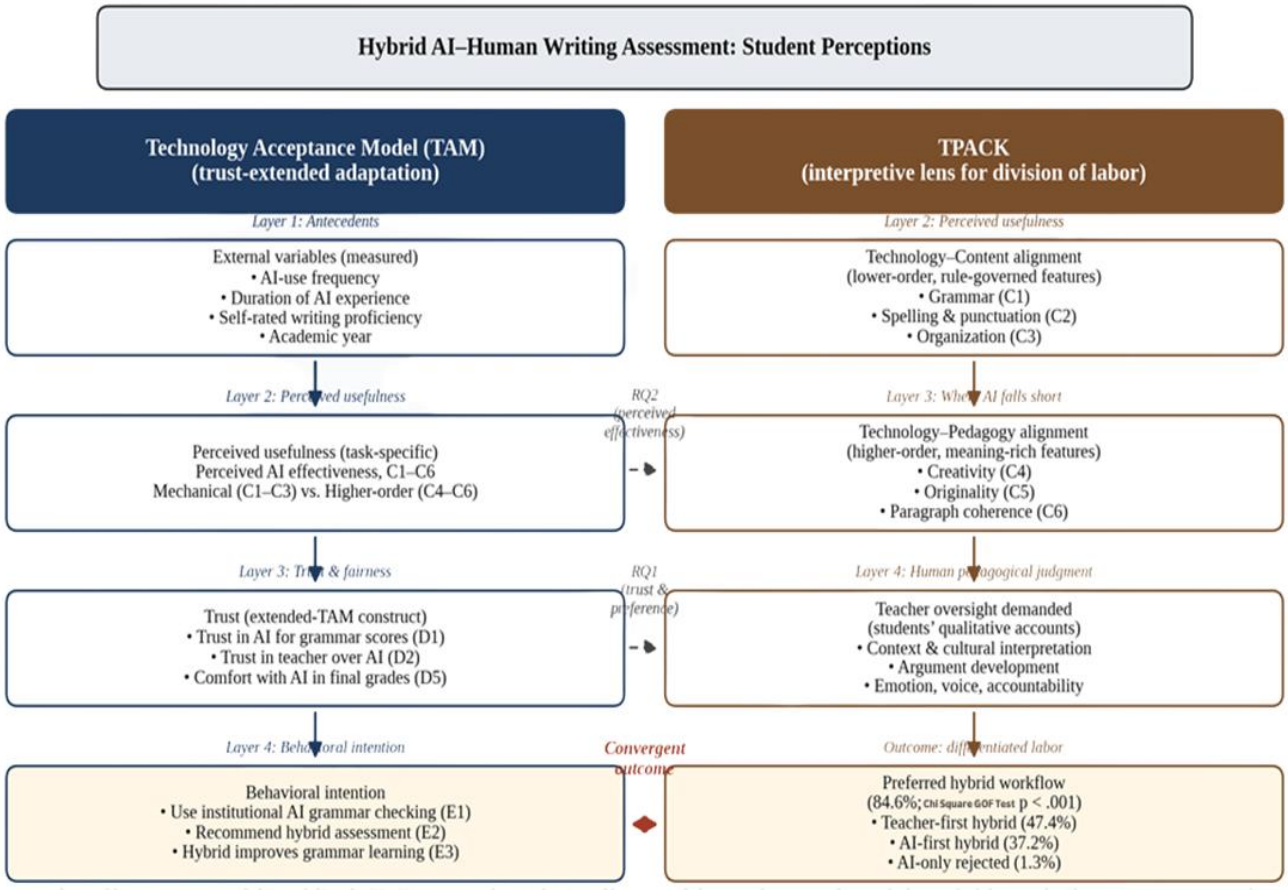

Figure 1. Conceptual Framework

TAM posits that perceived usefulness and perceived ease of use determine user attitudes and behavioral intentions (Davis, 1989). Evaluative settings complicate this dynamic, because students neither choose their grading engines nor control system adoption. Automated scores carry real academic consequences, so acceptance depends on perceived fairness and on reliable avenues for human review. The present inquiry operationalizes a trust-extended adaptation of TAM. Perceived usefulness is measured through task-specific perceived effectiveness (C1–C6). Trust, assessment preference and behavioral intention are evaluated via willingness to adopt and accept AI-assisted grading. The survey contained no validated multi-item ease-of-use scale; items on the frequency and duration of AI use therefore represent experience indicators rather than ease-of-use metrics. High-stakes evaluative adoption, in this reading, depends far more on procedural justice and human accountability than on operational usability alone.

TPACK supplies the second lens. If automated tools demonstrate superior proficiency in detecting rule-governed linguistic structures such as grammar, spelling and organization,

then student confidence in automated diagnostics represents a perceived technology and content alignment. Conversely, pedagogic evaluation requires understanding the learner within a specific instructional context, so learner insistence on teacher oversight reflects a perceived technology-pedagogy fit. The survey did not evaluate participants' intrinsic TPACK competence nor did it measure instructors' professional knowledge bases. TPACK therefore serves strictly as an interpretive lens as it organizes student effectiveness ratings and workflow preferences without advancing psychometric claims about teacher knowledge. Used in this way, the framework clarifies how students negotiate technological efficiency alongside indispensable pedagogical human judgment.

Two propositions structure the empirical tests. The first is framed as a research proposition rather than a hypothesis because the relevant survey item asked respondents to choose one of five categorical options instead of rating intensities. The proposition can nonetheless be evaluated inferentially against a null of no preference using a chi-square goodness-of-fit test on the five observed category counts. The second is a hypothesis tested with a paired comparison.

Respondents will not be uniformly distributed across the five assessment preference categories; in particular, a majority will prefer one of the two hybrid AI–human workflows over AI-only, teacher-only or no-preference options. RP1 was tested with a chi-square goodness-of-fit test against the null of equal preference across all five categories (expected frequency per category = N/5).

Students will rate AI as more effective for lower order mechanical features (grammar, spelling, punctuation and organization) than for higher-order features (creativity, originality and paragraph coherence). H2 was evaluated with a paired samples t-test comparing within person composite scores.

Method

Research Design

The study employed a sequential explanatory mixed-method design (Creswell & Plano Clark, 2018). Quantitative data were collected first to identify patterns in AI experience, perceived effectiveness, trust, assessment preference and behavioral intention. Then, qualitative written interviews followed to explain why students held those preferences and what safeguards they considered necessary. Integration occurred during participant selection, interpretation and joint display where the survey results guided the qualitative questions. Also, interview responses were compared with quantitative trends and meta-inferences were developed from convergence and divergence across the two strands.

Because preference percentages alone cannot show what students mean by trust or why they distinguish grammar correction from final grading, a mixed-method design was used. On the other hand, a small set of interviews cannot establish how common a preference is in the surveyed group. Sequential integration allowed the numerical pattern to identify issues requiring explanation and allowed students' own written accounts to qualify broad statistical summaries.

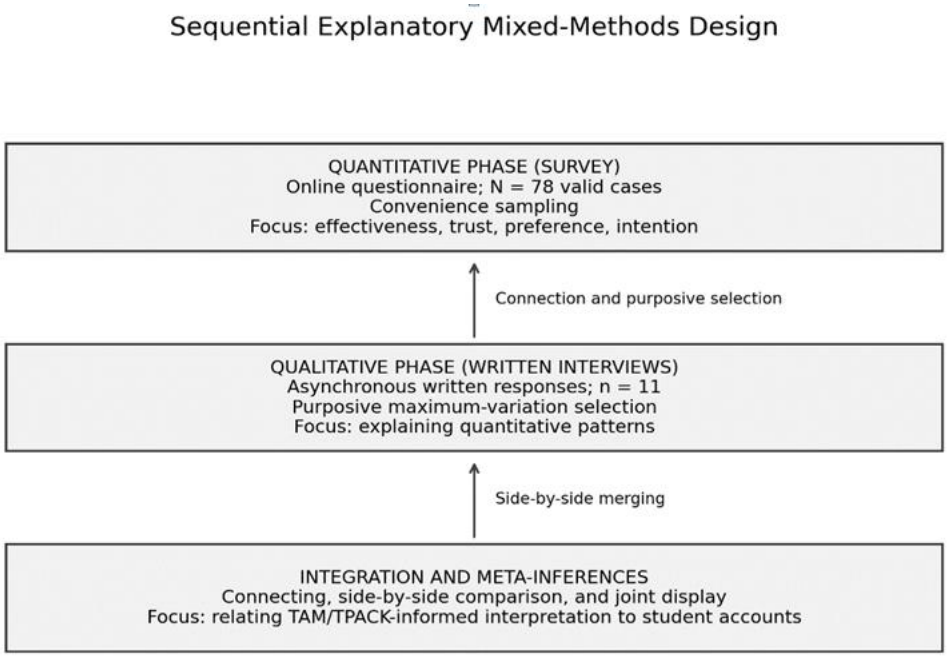

Figure 2. Sequential explanatory mixed-methods design adapted from Creswell and Plano Clark (2018)

Participants and Sampling

The final quantitative sample comprised 78 EFL students. Of these, 72 (92.31%) were drawn from public or publicly funded institutions, while six (7.69%) came from private universities. This skewness toward public institutions is a defining feature of the sample and constrains the inferences that follow. The findings should not be read as a portrait of all Bangladeshi tertiary EFL students. Rather, they describe the views of predominantly public sector tertiary EFL learners and it should be treated as suggestive rather than representative of private sector or mixed cohorts. Subgroup comparison by institution type was not undertaken because the private university cell (n = 6) was too small to support inferential comparison. All undergraduate years were represented, and the largest single cohort consisted of second year students (n = 24).

For the qualitative phase, 11 respondents from the survey pool completed asynchronous written interviews (Bakken, 2023). Participants were selected purposively with maximum variation sought across institution type, preferred assessment sequence and level of trust in AI. Purposive sampling is appropriate when participants are selected because they are information-rich cases relevant to the research questions (Patton, 2002). Recruitment was also supported by snowball sampling, a chain-referral approach that can help researchers reach participants with relevant characteristics through existing social networks (Naderifar et al., 2017). This maximum variation logic was intended to obtain contrasting explanations and not to judge qualitative prevalence. This qualitative sampling is generally directed toward relevant experiences and information-rich perspectives and not random population representation (Jalali, 2013). No names are reported and quotations are attributed only by participant number.

Instruments

The questionnaire contained five sections. Section A collected institution, academic year, department and a self-rating of English writing proficiency. Section B asked which AI

tools students had used, how often they used them, whether they had received AI-generated feedback on their writing and for how long they had been using such tools. Section C held six five-point agreement items on perceived AI effectiveness, three covering mechanical skills (grammar, spelling and punctuation, organization) and three covering higher-order skills (creativity, originality, paragraph coherence). Section D addressed trust in AI scoring, comparative confidence in teachers, the preferred assessment workflow and comfort with AI contributing to a final grade. Section E measured behavioral intention through three items: 1. willingness to use institutional AI grammar checking, 2. willingness to recommend hybrid assessment and 3. belief that hybrid assessment would help them learn grammar. For the correlation analysis, behavioral intention was computed as the mean of these three items together with the grade acceptance item from Section D. The reason was all four express willingness to let AI into formal assessment. This four item compound is the one used for RQ3.

The written interview protocol contained eight open-ended semi-structured questions. It covered the use of tool, perceived strengths and failures, differences between teacher and AI feedback, preferred sequencing, trust in combined scoring, fairness and privacy as well as recommendations to universities. As noted above, questions 2 and 3 referred explicitly to the two survey patterns. The Google Form was given to submit the answers. Though written responses are usually less spontaneous than interactive interviews (Bakken, 2023; Fritz & Vandermause, 2017; Ratislavová & Ratislav, 2014), the process reduced transcription-related error. Respondents wrote in Bangla, English or a mix of the two. All quotations reported in this article were translated into English by the authors and checked against the originals for meaning.

Instrument Review and Pilot Study

Before the main regulation, two specialists in English language testing and one educational technology specialist examined the survey for clarity and content relevance. A pilot study ivolving 25 students and served three purposes. They were identifying confusing wording, checking the range of tools students actually mentioned and examining the consistency of the proposed scales. The pilot led to tangible changes and Gemini and Claude were added to the tool list. Pilot respondents named both reported ChatGPT 96% of usage and 84% Gemini use. The trust section was expanded. Although pilot responses were suggestive, they were not confirmatory. 84% of pilot respondents preferred a hybrid modality and mean trust was 3.24 on the five-point scale. They were not included in the main dataset.

For the final sample, internal consistency of the six perceived effectiveness items was acceptable for a short exploratory scale (Cronbach's $\alpha = 0.69$) and the three-item mechanical composite performed well ($\alpha = 0.82$). The higher-order composite was weaker ($\alpha = 0.65$) which is common for short sets of heterogeneous items and is one reason the two composites were analyzed separately rather than merged. The four-item behavioral intention composite reached $\alpha = 0.66$. Because the questionnaire deliberately spans different constructs, no single overall alpha would be meaningful and none is reported (Nunnally, 1978).

Data Collection and Ethics

All responses were collected online. The first page of the survey explained the purpose of the study, the voluntary nature of participation, confidentiality and the right to stop before submitting. Interview volunteers received the same information in writing before the protocol link was sent. Participation was anonymous and no names, institutional identifiers or other identifying details are reported in the report. The quotations were screened to remove anything that could identify a participant. The consent information used in the online forms is stored by the authors.

Two further precautions were taken because the study concerns commercial AI tools. Participants were not asked to upload their assignments to any AI system for the purpose of the study. While the study drew on participants' reported perceptions and experiences of AI-supported assessment to examine how they understood the technology's capabilities and limitations, it did not independently test the accuracy, reliability or validity of any particular commercial AI tool.

Data Analysis

Frequencies, percentages, means, standard deviations and 95% confidence intervals were calculated to describe the sample. RP1 was evaluated using a Pearson chi-square goodness-of-fit test across five observed preference categories. The null hypothesis assumed equal representation across categories, so the expected frequency was E = N/k = 78/5 = 15.6 per category. The test statistic was calculated as

$$x^2 = \sum_{i=1}^{5} \frac{(O_i - E_i)}{E_i}$$

with $df = k-1 = 4$. The null hypothesis was rejected when the calculated χ^2 statistic exceeded 9.488 at $\alpha = 0.05$ or 13.277 at $\alpha = 0.01$. This is the appropriate framework for testing whether observed categorical frequencies depart from a specified theoretical distribution. (C. Ireland, 2010; J. I. Hoffman, 2019; and N. Ahad et al., 2019)

Cohen's $w = \sqrt{X^2/N}$ was used to quantify effect size. Following a significant omnibus result, standardized residuals, $Z_i = (O_i - E_i)/\sqrt{E_i}$, were inspected to identify contributing categories, using $|z| > 1.96$ as the per-category threshold. The supplied description specifies the method and decision criteria but does not provide the five observed counts or resulting $\chi 2$, p-value, or w; therefore, it does not independently demonstrate that RP1 was statistically significant. (Sharpe, 2015; Beaman & Michel, 1998)

For H2, each participant's mechanical and higher-order mean scores were compared with a paired samples t test, and Cohen's dz was reported as the standardized mean difference. One sample t tests against the scale midpoint (3) were run for the trust, grade acceptance and intention items to show where each stood relative to indifference. For RQ3, Pearson correlations linked AI-use frequency, duration of experience, self-rated writing proficiency and academic year to AI trust and behavioral intention. These variables were selected because they represent key components of the acceptance pathway in the conceptual framework. These variables were selected because they represent the acceptance pathway described in the conceptual framework, with frequency and duration

indicating exposure to the technology, self-rated proficiency reflecting learner-side capability as an external TAM variable and academic year providing a rough measure of seniority. Given the sample size, the correlations were treated as exploratory descriptions of association and they were not interpreted as tests of a structural model. Although statistical significance was evaluated at $\alpha = 0.05$, effect sizes and confidence intervals were also considered to ensure that conclusions were not based on p-values alone.

Table 1. Alignment of Research Questions, Variables, and Framework Constructs

RQ	Variables	Framework construct	Analysis
RQ1	Preferred workflow (choice item); trust in AI (D1); comparative trust in teachers (D2)	Extended TAM: trust; assessment preference	Frequencies; paired-samples t test
RQ2	Perceived effectiveness items C1–C6 (mechanical C1–C3; higher-order C4–C6)	TAM: task-specific perceived usefulness; TPACK: technology–content vs. technology–pedagogy fit	Composites; paired-samples t test; one-sample t tests
RQ3	AI-use frequency and duration (Section B); self-rated proficiency and academic year (Section A); behavioral intention (grade-acceptance item + E1–E3)	TAM: experience → trust → intention pathway	Pearson correlations

The qualitative data were analyzed using an adapted version of Creswell and Creswell's (2018) five-stage approach because the interviews were conducted asynchronously in written form. The participants submitted responses constituted the primary dataset and did not require transcription. This format allowed participants time to formulate reflective responses but did not provide access to vocal or non-verbal cues (Harris et al. 2024). The responses were first anonymized then organized by participant number and later checked against the original submissions for completeness and accuracy. The researchers then read the responses repeatedly to become familiar with the dataset and recorded preliminary observations in a private journal. Initial codes were generated inductively from recurring words, phrases and concepts relevant to the research questions. These codes were compared, refined and grouped into broader themes representing commonalities and differences across participants. The themes were reviewed against the complete dataset and presented in a qualitative narrative supported by anonymized quotations. The researcher then interpreted the findings in relation to the research questions, the study context and relevant literature. To enhance procedural rigor, a colleague independently reviewed selected anonymized responses, coding and themes (15%). Any differences were discussed and the coding framework was clarified where necessary. The analysis focused on meanings expressed in the written responses and did not infer vocal tone or other unrecorded non-verbal cues.

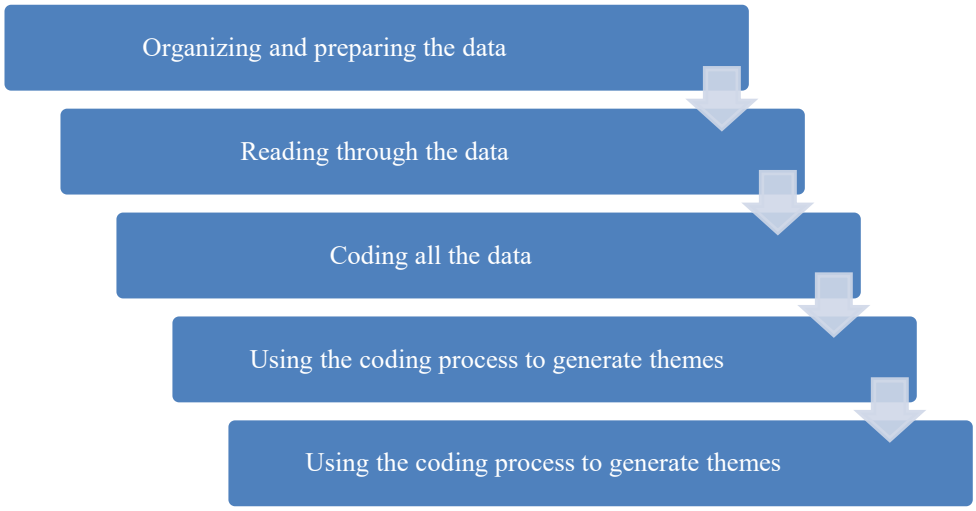

Figure 3: A diagram of the five-step data analysis process

Quantitative Results

Participant Profile and AI Use

The final sample consisted of 72 students from public or publicly funded institutions (92.31%) while six students were from private universities (7.69%). All undergraduate years were represented. Self-rated English writing skill averaged 3.30 (SD = 0.63) on the five-point scale. ChatGPT was used by 63 respondents (80.8%), Gemini was used by 48 respondents (61.5%), Grammarly was used by 33 respondents (42.3%), Google Translate was used by 31 respondents (39.7%), QuillBot was used by 22 respondents (28.2%), Claude was used by 21 respondents (26.9%) and Microsoft Word Editor was used by 10 respondents (12.8%). Formal exposure to AI feedback was much rarer. These figures describe frequency of use, not the quality or criticality of that use. Eighteen respondents (23.08%) used AI tools daily and 39 (50.00%) used them several times a week. Three respondents (3.85%) used AI tools about once a week, 17 (21.79%) used them occasionally and only one (1.28%) used them rarely or never. Forty-three students (55.13%) had used AI writing tools for one to two years and 20 (25.64%) had used them for more than two years. Twenty one respondents (26.92%) had never received AI-generated feedback from a university or teacher, while 16 (20.51%) had received it from their institution. Although all undergraduate years were represented, the largest group consisted of second-year students, with 24 respondents (30.77%). Because formal exposure to AI feedback was limited, only 16 respondents (20.51%) reported receiving such feedback from their institution. While 63 respondents (80.8%) had used ChatGPT, other AI tools were also widely used including Gemini (61.5%), Grammarly (42.3%) and Google Translate (39.7%). Although AI use was widespread, formal exposure to AI-generated feedback was considerably less common. Because the data measure reported use rather than evaluative competence, these figures do not indicate the quality or criticality of students' AI use.

Table 2. Participant Profile and AI-Use Behaviour (N = 78)

Variable	Category	n	%
Institution type	Public/other institution	72	92.31
	Private university	6	7.69
Academic year	1st year	7	8.97
	2nd year	24	30.77
	3rd year	11	14.10
	4th/final year	13	16.67
	Master's/postgraduate	23	29.49
AI-use frequency	Daily	18	23.08
	Several times a week	39	50.00
	About once a week	3	3.85
	Occasionally	17	21.79
	Rarely/never	1	1.28
Duration of AI experience	Less than 6 months	2	2.56
	6–12 months	13	16.67
	1–2 years	43	55.13
	More than 2 years	20	25.64
Received AI-generated feedback	Never	21	26.92
	Yes, from my university/teacher	16	20.51
	Both institutional and own tool use	20	25.64
	Own tool uses only	8	10.26
	Cannot tell the difference	13	16.67

Note. *Percentages may not sum to 100 due to rounding. The sample is heavily skewed toward public institutions (92.31%); findings therefore describe predominantly public sector tertiary EFL learners and should not be generalized to all Bangladeshi tertiary EFL students.*

Assessment Preference and Trust

RP1 was supported. A Pearson chi-square goodness-of-fit test was conducted on the five preference categories to test the null hypothesis that respondents were equally distributed

across the five options (no preference). The observed frequencies were: teacher only = 8, AI-only = 1, teacher-first hybrid = 37, AI-first hybrid = 29 and no preference = 3. Under the equal distribution null, the expected frequency for each category was N/5 = 78/5 = 15.6. The observed distribution differed significantly from the null, χ^2 (4, N = 78) = 68.41, $p < 0.001$, Cohen's w = 0.94. Inspection of standardized residuals indicated that each of the teacher-first hybrid category (z = 5.42), the AI-first hybrid category (z = 3.39), the AI-only category (z = −3.70) and the no preference category (z = −3.19) contributed disproportionately to the overall chi-square statistic ($|z|>1.96$). The teacher-only category approached but did not exceed the per-category threshold (z = −1.92). The two hybrid categories were chosen far more often than expected and the AI only, no preference, and teacher only options were chosen far less often than expected.

Within the hybrid group, 37 respondents (47.44%) preferred the teacher to assess first with AI providing subsequent feedback, while 29 (37.18%) preferred AI checking first followed by teacher review. Eight respondents (10.26%) preferred teacher only assessment, three (3.85%) reported no preference and a single respondent (1.28%) chose AI-only assessment. Support for combining human and machine judgment was therefore broad, although no single sequencing commanded a majority of the full sample.

Table 3. Observed and Expected Frequencies for the Five-Category Preference Item (N=78)

Preference category	Observed (O)	Expected (E=N/5)	O−E	$(O-E)^2/E$	Standardized residual (z)
Teacher-only	8	15.6	−7.6	3.70	−1.92
AI-only	1	15.6	−14.6	13.66	−3.70
Teacher-first hybrid	37	15.6	+21.4	29.36	+5.42
AI-first hybrid	29	15.6	+13.4	11.51	+3.39
No preference	3	15.6	−12.6	10.18	−3.19
Total	78	78.0	0.0	$\chi^2 = 68.41$	—

Note. *Expected frequencies assume equal preference across the five categories (null hypothesis of no preference). The test statistic was computed as* $\chi^2 = \Sigma (O_i - E_i)^2 / E_i$ *with df = 4. The critical value at a = 0.05 is 9.488; the observed* $\chi^2 = 68.41$ *exceeds this threshold,* $p < 0.001$. *Cohen's w = 0.94 indicates a large effect. Highlighted rows mark the two hybrid categories, which were endorsed significantly more often than expected under the null of no preference.*

Trust ratings were moderate rather than extreme. Mean agreement with "I trust AI to give fair scores on my grammar" was 3.154 (SD = 1.094), statistically indistinguishable from the scale midpoint, t (77) = 1.242, p = 0.218; mean agreement with "I trust my

teacher more than AI for grammar feedback" was 3.231 (SD = 1.268), t (77) = 1.607, p = 0.112. Since ratings came from the same respondents, they were compared directly with a paired samples t-test and the difference was not significant, t (77) = −0.363, p = 0.718, dz = −0.041. Students did not demonstrate a strong preference for either source. Such a balanced distribution sits far more comfortably with the hybrid preference than with any single-authority reading.

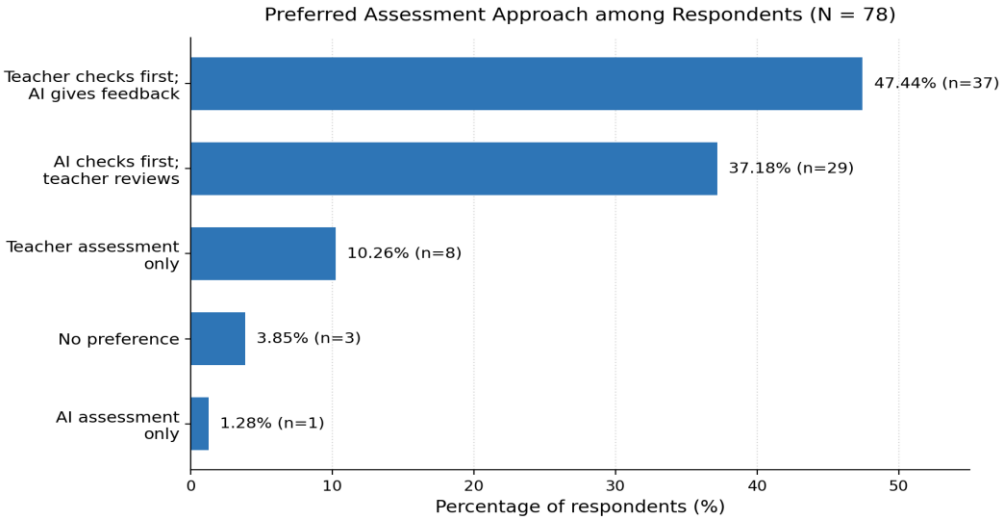

Figure 4. Preferred assessment approach among respondents (N = 78).

Perceived Effectiveness by Skill Domain

H2 was supported. Although the mechanical composite averaged 3.756 (SD = 0.805) compared with 3.128 (SD = 0.697) for the higher-order composite, the within-person difference was statistically significant, t(77) = 5.765, p < .001, with a medium-to-large effect (dz = .653). Students clearly perceived AI as stronger at rule-governed and structural features than at meaning-rich and creative judgment.

Spelling and punctuation received the highest rating (M = 3.859), followed by grammar (M = 3.769) and organization (M = 3.641); paragraph coherence (M = 3.321), creativity (M = 3.064) and originality (M = 3.000) received progressively lower ratings. The one-sample tests in Table 3 sharpen the picture: every mechanical item sat significantly above the scale midpoint, whereas creativity and originality did not differ from it and coherence only barely cleared it. These are perceptions rather than accuracy scores. They tell us what students believe AI does well, not how any particular tool would perform against expert raters.

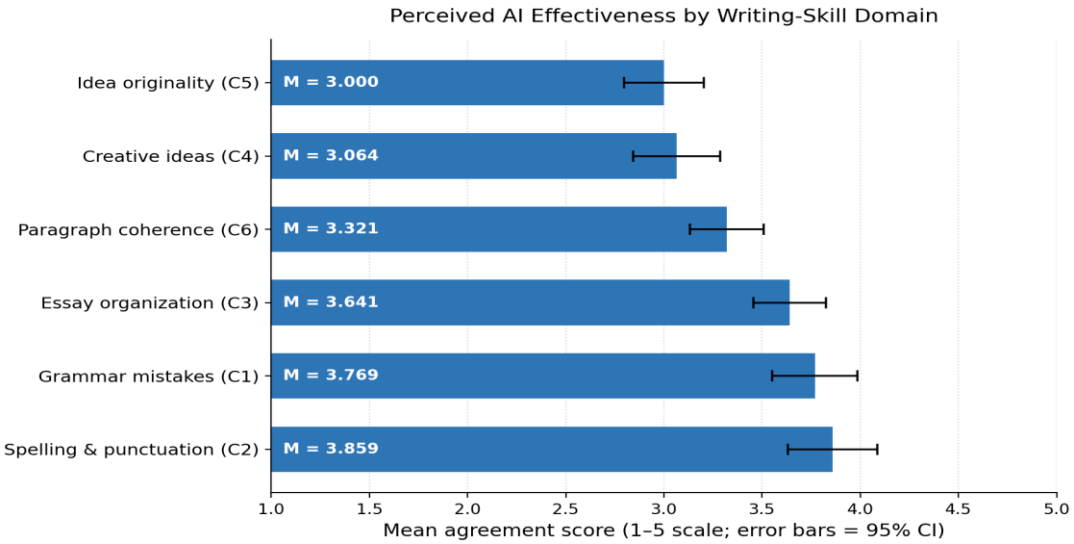

Figure 5. Perceived AI effectiveness ratings by writing-skill domain

Table 4. Perceived AI Effectiveness, Trust, and Behavioral Intention Items (N = 78)

Domain	Item	M	SD	95% CI
Lower-order	C1: AI detects grammar mistakes	3.769	0.966	[3.551, 3.987]
Lower-order	C2: AI checks spelling and punctuation	3.859	1.016	[3.630, 4.088]
Lower-order	C3: AI checks essay organization	3.641	0.821	[3.456, 3.826]
Higher-order	C4: AI judges idea creativity	3.064	0.985	[2.842, 3.286]
Higher-order	C5: AI evaluates idea originality	3.000	0.897	[2.798, 3.202]
Higher-order	C6: AI checks paragraph coherence	3.321	0.830	[3.134, 3.508]
Trust/preference	D1: Trust AI for grammar scores	3.154	1.094	[2.907, 3.401]
Trust/preference	D2: Trust teacher more than AI	3.231	1.268	[2.945, 3.517]
Trust/preference	D5: Comfortable if final grade partly AI-based	2.615	1.047	[2.379, 2.851]
Behavioral intention	E1: Use institutional AI grammar checking	3.679	0.904	[3.475, 3.883]
Behavioral intention	E2: Recommend hybrid assessment	3.603	0.998	[3.378, 3.828]

Domain	Item	M	SD	95% CI
Behavioral intention	E3: Hybrid assessment improves grammar learning	3.808	1.033	[3.575, 4.041]

> ***Note***. *Items were rated on a five-point scale (1 = strongly disagree, 5 = strongly agree). CI = confidence interval.*

Correlations with Experience, Proficiency, and Year

Table 5. Correlations with Experience, Proficiency, and Year

Variable	AI-Use Frequency	Duration of Experience	Self-Rated Writing Proficiency	Academic Year	AI Trust (D1)	Behavioral Intention (E1–E3)
1. AI-Use Frequency	—					
2. Duration of Experience	—	—				
3. Self-Rated Writing Proficiency	—	—	—			
4. Academic Year	—	—	—	—		
5. AI Trust (D1)	—	—	r= .085 (p = .462)	r= -.044 (p = .700)	—	
6. Behavioral Intention (E1–E3)	r= .211 (p = .064)	r= -.217 (p = .056)	r= .206 (p = .071)	r= .077 (p = .504)	—	—

> ***Note.*** *N = 78. All bivariate associations were evaluated using two tailed Pearson correlation coefficients at a = .05. None of the correlations reached statistical significance (p > .05). The negative coefficient for Experience Duration (r = -.217) reflects the cleaned dataset (r = -.2174) and denotes an inverse association with behavioral intention. D1 = Trust AI for grammar scores; E1–E3 = Behavioral Intention composite scale.*

Pearson correlation analysis revealed that student background antecedents were not statistically significant predictors of AI evaluative trust or behavioral intention. Although none of the correlations reached statistical significance at the conventional .05 alpha level, the overall configuration of associations remains coherent. AI use frequency correlated weakly and positively with behavioral intention ($r = .211$, $p = .064$) but duration of experience exhibited a weak negative relationship ($r = -.217$, $p = .056$). The data suggest that extended familiarity with automated tools accompanies a slight decline in willingness to accept institutional scoring. Self-rated writing proficiency correlated weakly with both AI trust ($r = .085$, $p = .462$) and behavioral intention ($r = .206$, $p = .071$). Academic year served as an exploratory seniority indicator. However, it demonstrated negligible associations with both behavioral intention ($r = .077$, $p = .504$) and evaluative trust ($r = -.044$, $p = .700$). When interpreted through a trust-extended Technology Acceptance Model (TAM), these weak relationships are theoretically illuminating because tool usage was widespread across the sample. Frequent operational use did not cultivate a willingness to be evaluated by algorithms; technical experience and evaluative acceptance were clearly decoupled. The demographic and usage variables examined here cannot explain why some frequent users accept institutional AI while others resist it. Thus, institutional acceptance appears governed by underlying trust structures and procedural fairness, which must be measured directly by future instruments.

Qualitative Results

This section highlights patterns and recurring themes identified through thematic analysis in response to the research questions. To protect participants' confidentiality and anonymity, all participants are designated by pseudonymous code numbers (e.g., Participant 1 through Participant 11).

Efficiency Appreciation in Mechanics

All 11 interview respondents described some value in AI for surface-level revision. They referred to spelling, punctuation, capitalization, agreement, tense and sentence restructuring. Speed was repeatedly emphasized as automated comments were available immediately and could be requested more than once. Students viewed this function as especially useful before submitting a draft or before asking a teacher to focus on ideas. Their accounts qualify the survey finding by showing that perceived usefulness was linked to preliminary editing rather than unrestricted scoring authority.

Participant 7 wrote: "AI is extremely effective for correcting grammar, spelling, punctuation and basic sentence structure. It identifies tiny grammar errors that humans easily miss, makes my writing sound much cleaner and saves an immense amount of time." The quotation illustrates both appreciation and a possible overgeneralization "extremely effective" that institutions would need to address through AI literacy training and independent checking.

Trust in Human Contextual and Pedagogical Judgment

A second qualitative distinction separated correcting surface text from genuinely knowing the writer. While automated tools merely scan surface syntax, human teachers interpret classroom context, trace an argument's development across drafts and provide essential

encouragement. Human evaluation also guarantees dialogic accountability; students can question a teacher's marginal comments and seek immediate clarification whenever feedback seems unconvincing.

Participant 5 highlighted this relational dimension directly: "My teacher is fully aware of the context and emotion of my writing... AI can check the text structurally, but it has no human emotion". A single student's reflection cannot measure overall teaching quality. Nevertheless, such accounts capture the personal mentorship and procedural fairness that learners demand. Because automated diagnostics cannot replicate relational care or transparent dialogue, an acceptable hybrid framework must deliberately protect the teacher-student partnership rather than reducing evaluation to mechanical scores.

Speed–Depth Trade-off and Local Context

Students reported that fast output could be generic, literal or culturally misplaced. Examples included Bangladeshi historical references, local social realities and National Curriculum and Textbook Board materials. Some respondents also felt that AI suggestions introduced unnecessarily difficult vocabulary or a formulaic style. In these accounts, the problem was not only factual error; it was a mismatch between technically polished language and the writer's intended audience, voice or local knowledge.

Participant 9 explained: "AI often understands the literal word meaning, but it fails to grasp local Bangladeshi cultural context. For instance, when I was writing about a specific NCTB primary curriculum, both ChatGPT and Gemini got confused and gave inaccurate answers. It also overuses difficult vocabulary, making my writing look formulaic and artificial." This theme connects the Bangladeshi case with international concerns about validity across contexts and genres.

Machine Error, Privacy and Grade Security

The strongest objections concerned summative grading. Respondents worried that scoring criteria would be opaque, that an incorrect result could be difficult to appeal and that student writing might be retained or reused by commercial systems. These concerns were amplified by the final grade item, which had a relatively low mean of 2.615 (SD = 1.047). Students were more willing to use AI for practice than to allow it to determine a consequential outcome.

Participant 11 wrote: "I would feel very insecure if my final grades were based purely on AI. You cannot argue with an algorithm when it makes a mistake. Plus, we don't know how safe our personal essays are on their servers. There must always be an appeal process where our teachers review the score." The requested safeguard is institutional rather than merely technical. Students requested a named human reviewer and a clear process for contesting a decision.

Table 6. Qualitative Thematic Analysis Map

Theme	Evidence	Definition and initial codes	Illustrative extract
Efficiency in mechanics	11/11	Grammar, spelling, punctuation, sentence structure; rapid feedback	"AI … identifies tiny grammar errors … and saves an immense amount of time."
Human contextual judgment	11/11	Teacher knowledge of learner, argument, development, emotion, and motivation	"Teacher checks my writing with human touch … AI has no emotion."
Speed–depth trade-off	11/11	Literal meaning; generic vocabulary; local culture and NCTB context missed	"AI cannot understand Bangladeshi social reality or NCTB textbook examples."
Machine error and grade-related security concerns	9/11	Opaque grading; appeal; privacy and data-security concerns	"What if AI gives a wrong score with no appeal?"

Discussion

The two strands of this study did more than sit side by side. This section first states where they agree, then where they were in tension and finally what they support together, before connecting the integrated findings to the international as well as Bangladeshi literature and to the conceptual framework.

Convergence between the survey and the interviews is unusually tight on three points. Firstly, the 84.62% hybrid preference has a direct qualitative counterpart. For instance, every one of the eleven interviewees described hybrid assessment as the natural agreement. Secondly, the mechanical composite (M = 3.756) matches the theme what is present in all eleven interviews. These findings suggest that AI's real strength lies in rapid mechanical correction. Thirdly, the survey's lowest item is comfort with AI in final grades (M = 2.615) which is below the midpoint. It is echoed in the nine interviews that raised opaque scoring, privacy and appeal concerns. In short, students did not evaluate AI as uniformly good or bad. They separated rapid feedback on visible mechanical features. Both datasets drew the same line in the same place.

Two tensions deserve attention precisely because a single strand reading would have missed them. The first concerns trust. While the paired comparison of the two trust items yielded a null result, the model choice item revealed an overwhelming preference for combining the two sources. Also, it suggests that the two near midpoint means may understate the decisiveness with which students rejected single authority arrangements. The second concerns experience as duration of AI use showed a marginally negative

correlation with behavioral intention. The interviews help resolve the apparent contradiction. They suggest that longer experience taught students where the tool tends to fail and thus, familiarity bred greater specificity instead of greater faith. These patterns are the most analytically useful results in the study.

The datasets support a domain specific model of acceptance. Perceived usefulness is conditional on task. Besides, trust is distributed by function rather than held or withheld wholesale. The intention to accept AI-supported assessment depends on human accountability. In TAM terms, perceived usefulness was not only real but also bounded. Trust and assessment preference did the gatekeeping between use and institutional acceptance. Table 5 summarizes these joint conclusions alongside the evidence behind them.

Table 7. Joint Display of Quantitative and Qualitative Integration

Quantitative finding	Qualitative explanation	Integrated meta-inference
Observed frequencies differed significantly from an equal-distribution null, $\chi^2(4, N = 78) = 68.41$, $p < .001$, Cohen's $w = 0.94$. 84.62% preferred hybrid assessment; 1.28% preferred AI only.	AI was valued for preliminary mechanics but not as the sole decision-maker.	Usefulness is task-specific; acceptance depends on human accountability.
Mechanical composite $M = 3.756$ exceeded higher-order $M = 3.128$, $p < .001$.	Students described rapid corrections but weak handling of local context, voice, and creativity.	Technology–content fit is stronger for rule-governed features than for meaning-rich judgments.
Comfort with AI contributing to final grades was low ($M = 2.615$), below the midpoint, $t(77) = -3.243$, $p = .002$.	Students worried about opaque criteria, commercial data use, and the absence of appeal.	Formative piloting, transparent rubrics, and human review should precede any summative use.
Trust comparison was not significant, $t(77) = -0.363$; duration of experience correlated negatively with intention, $r = -.217$.	Interviewees trusted teachers for context and accountability while accepting AI for mechanics; longer experience had revealed the tool's limits.	Trust is distributed by function, and familiarity produces discrimination rather than acceptance. (Divergence row added during this revision.)

> ***Note***. *The first row of the integration was corroborated by a Pearson chi-square goodness-of-fit test against an equal-distribution null (χ^2 (4, N = 78) = 68.41, $p < 0.001$, Cohen's $w = 0.94$).*

The hybrid preference is consistent with international research that values human oversight in educational AI (Eriksen & Elstad, 2026; Üretmen Karaoğlu & Doğan, 2025). What this study adds is the students' side of the bargain is a specific list of what they want humans to retain namely: interpretation, emotional and developmental knowledge, the ability to explain a judgment and final accountability. The list matters because it converts an abstract principle, human oversight into concrete design requirements that a university can actually implement.

The H2 result is more precise than a simple verdict on AI. The significant gap between mechanical and higher-order means, together with the interview accounts, shows students operating with a differentiated model of the technology which is fast on grammar, literal on local context and voice. Comparable concerns surface in critiques of holistic automated scoring where surface features can pass for quality (Kapxhiu, 2025; Tate et al., 2024). The study does not establish how accurate any tool actually is. It rather establishes that students already distinguish a tool's claims from its competence and that discrimination itself is an AI-literacy asset that institutions can build on.

The Bangladeshi studies read differently after these findings. The ambivalence Nawrin and Moon (2024) documented about Facebook while commenting 'useful but linguistically and technically demanding' reappears here as students welcoming access and speed without equating use with trust. The rubric argument of Nawrin and Sadek (2022) now has a concrete application where a transparent rubric is the instrument that would tell both parties which dimensions AI may address and how teacher review should proceed. Furthermore, the teacher-dominated discourse pattern Nawrin and Rahman (2020) observed warns against a specific failure mode. If a hybrid model simply substitutes algorithmic feedback for teacher feedback without creating room for student response, the one-way flow of authority remains intact under a new name.

Within the framework, the findings support a trust-extended TAM interpretation and a bounded use of the TPACK lens. Confidence in grammar checking tracked perceived technology and content fit as the demand for teacher review tracked technology and pedagogy fit. Behavioral intention depended on trust and accountability rather than exposure alone. Neither framework was validated as a measurement model as the survey measured neither ease of use nor teacher knowledge and no latent variables were tested. However, as interpretive guides they organized patterns that a purely descriptive account would have left loose.

The null trust comparison also needs a careful reading because a casual glance could mistake it for indifference. Each trust means sat near the midpoint while the model choice item revealed a strong desire to combine the two sources. Trust, in other words, appears distributed by function as the same student may trust AI to catch a comma and a teacher to judge whether the argument meets the course standard. Future instruments should therefore separate trust in feedback accuracy, score fairness, privacy, explainability and institutional governance instead of asking about trust as a single quantity.

Implications

Universities considering AI in writing assessment should begin where students already place their confidence namely: formative, low stakes, mechanical support. The

sequencing evidence argues for piloting more than one workflow rather than standardizing prematurely. Teacher first arrangements were preferred by 47.44% of respondents and AI first arrangements by 37.18%. Thus, an institution that adopts a single sequence without consulting its own students would simply be guessing. An AI first drafting pipeline could still be worth piloting as it flags mechanics, students decide on revisions and the teacher assesses the complete work. However, the choice should follow local evidence and not assumption.

Nevertheless, implementation should arrive as a package and not as a tool rollout. A published policy should list approved and prohibited uses, disclosure requirements, data protection expectations, named human review and appeal procedures. Training matters on both sides since students need practice in checking AI suggestions, recognizing hallucinations and preserving their own voice and teachers need both the time and the authority to interrogate automated output. Rubrics should state the division of labor explicitly as advisory AI comments on grammar and punctuation, human responsibility for construct interpretation, final scores and appeals.

The implications travel beyond Bangladesh. Every institution faces the same conceptual problem of separating assistance from authority. However, the operational answer is local. The NCTB example in the qualitative data shows why a system validated in one country cannot be assumed to read another country's curricula or multilingual writing correctly. The contribution of this study is therefore not a portable percentage. Rather, it is evidence for a design principle as task-limited automation combined with accountable human judgment being grounded in one specific educational context.

Recommendations

Three recommendations follow from the integrated findings. First, universities should pilot AI in formative and low stakes writing processes before considering any summative use and they should pilot more than one workflow, since this sample split its support between teacher first (47.44%) and AI first (37.18%) sequences. Second, every workflow should be governed by transparent rubrics, disclosure rules, data protection procedures and mandatory human review. Thus, the safeguards students asked for, specifically explanation, appeal and named responsibility, do exist before and not after implementation. Third, AI literacy should be taught as critical evaluation rather than prompt production where students should learn when to reject a suggestion, how to preserve authorship and voice and how to document assistance. Under these conditions, AI can reduce repetitive correction without displacing the pedagogical and ethical responsibility of teachers.

Limitations

Several constraints bound what this study can claim. Convenience sampling and the concentration of participants in accessible university networks, most visibly the dominance of one large public university, limit representativeness. Also, the private university subsample (n = 6) is far too small for comparison, thus institution type was treated as descriptive only. The sample suits the reported paired comparison but not

latent variable modeling or subgroup estimates. Self-rated proficiency and self-reported tool use carry recall and social desirability risk. The study measured perceptions and not the agreement between AI scores and expert raters. The qualitative phase used written responses which supported reflection and direct quotation but permitted no probing. The frameworks served as interpretive lenses rather than validated measurement models and validated multi-item measures of trust, ease of use, usefulness and fairness remain on the agenda for future research. Finally, the interview quotations were translated from Bangla and mixed language originals as the translations were checked for meaning, but translation inevitably involves interpretation.

Conclusion

Bangladeshi tertiary EFL students in this study treated AI as a useful but bounded participant in writing assessment. They rated it clearly higher for grammar, spelling and punctuation and organization than for creativity, originality and paragraph coherence. More than four in five participants preferred a hybrid workflow, a preference that was statistically corroborated by a chi-square goodness-of-fit test against an equal-distribution null (χ^2 (4, N = 78) = 68.41, $p < 0.001$, Cohen's w = 0.94) and exactly one respondent accepted AI-only grading. The interviews explained the numbers. Speed on mechanics was welcomed but students wanted teachers to interpret context, recognize development, explain judgments, protect grade security and hear appeals.

The next step is to put perceptions next to performance. Trained raters and several AI systems could score the same locally situated writing samples so that agreement, error patterns, bias and explanations can be examined directly. Classroom studies could follow teacher only, AI first and teacher first sequences over time, tracking revision quality and learner agency. Students themselves should also be involved in designing the appeal procedures and data governance rules they are asking for. Work of this kind would move the debate from abstract enthusiasm or fear toward empirically tested assessment designs.

References

Alshehri, A. (2025). AI's effectiveness in language testing and feedback provision. *Social Sciences & Humanities Open*, *12*. doi.org/10.1016/j.ssaho.2025.101892

Andrade, C. (2020). The inconvenient truth about convenience and purposive samples. *Indian Journal of Psychological Medicine*. doi.org/10.1177/0253717620977000

Bakken, S. A. (2023). App-based textual interviews: Interacting with younger generations in a digitalized social reality. *International Journal of Social Research Methodology, 26*(6), 633–645. doi.org/10.1080/13645579.2023.2184931

Beaman, J.J., & Michel, G. (1998). Statistical tests and measures for the presence and influence of digit preference.

Chen, A., Zhang, Y., Jia, J., Liang, M., Cha, Y., & Lim, C. P. (2025). A systematic review and meta-analysis of AI-enabled assessment in language learning: Design, implementation, and effectiveness. *Journal of Computer Assisted Learning, 41*(1), e13064. doi.org/10.1111/jcal.13064

Choi, S., Jang, Y., & Kim, H. (2023). Influence of pedagogical beliefs and perceived trust on teachers' acceptance of educational artificial intelligence tools. *International Journal of Human–Computer Interaction, 39*(4), 910–922. doi.org/10.1080/10447318.2022.2049145

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage.

Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). SAGE.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 13*(3), 319–340. doi.org/10.2307/249008

Eriksen, H., & Elstad, E. (2026). Norwegian secondary teachers' perceptions of artificial intelligence in summative assessments. *International Journal of Assessment Tools in Education, 13*(2), 481–495. doi.org/10.21449/ijate.1791004

Fleckenstein, J., Liebenow, L. W., & Meyer, J. (2023). Automated feedback and writing: A multi-level meta-analysis of effects on students' performance. *Frontiers in Artificial Intelligence, 6*. doi.org/10.3389/frai.2023.1162454

Fritz, R. L., & Vandermause, R. (2017). Data collection via in-depth email interviewing: Lessons from the field. *Qualitative Health Research, 28*(11), 1640–1649. doi.org/10.1177/1049732316689067

Harris, S. L., Shaw, M., & Green, J. F. (2024). Utilizing asynchronous email interviewing for qualitative research among multiple participant groups: Perspectives on met and unmet needs from chaplain staffing. International *Journal of Qualitative Methods*. doi.org/10.1177/16094069241257943

ICT Division, Government of Bangladesh. (2026). *National artificial intelligence policy 2026–2030* (Draft V2.0). https://ictd.gov.bd/

Ireland, C. (2010). *CABI, Analysis of frequency data*. doi:10.1079/9781845935375.0264

Jalali, R. (2013). Qualitative research sampling. *Qualitative Research in Health Sciences, 1*(4), 310–320.

Hoffman, J. I. E. (2019). Categorical and cross-classified data: Goodness of fit and association. In *Basic biostatistics for medical and biomedical practitioners,* pp. 197-231. doi.org/10.1016/B978-0-12-817084-7.00014-0.

Kapxhiu, L. (2025). AI in language learning assessment: Examining the limitations of online testing and the reliability of teacher judgement. *Academic Journal of Business, Administration, Law and Social Sciences, 11*(3), 59–72. doi.org/10.2478/ajbals-2025-0028

Koehler, M. J., Mishra, P., & Cain, W. (2013). What is technological pedagogical content knowledge (TPACK)? *Journal of Education, 193*(3), 13–19. doi.org/10.1177/002205741319300303

Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record, 108*(6), 1017–1054. doi.org/10.1111/j.1467-9620.2006.00684.x

Naderifar, M., Goli, H., & Ghaljaie, F. (2017). Snowball sampling: A purposeful method of sampling in qualitative research. *Strides in Development of Medical Education*.

Nafeesa, S. (2025). Embracing AI in Language Learning: Perspectives of ESOL Students on Using Chatbots for English Writing Practices. *TESOL Bangladesh Journal (TBJ), 2*(2).

Ahad, N. A., Alipiah, F. M., & Azhari, F. (2019). Applicability of G-test in analyzing categorical variables. *AIP Conf. Proc.* 21 August 2019; 2138 (1): 050003. doi.org/10.1063/1.5121108

Nawrin, & Moon, N. N. (2024). Understanding students' perception on the use of Facebook for second language learning. *Teacher's World: Journal of Education and Research, 49*(1), 117–136. doi.org/10.3329/twjer.v49i1.70265

Nawrin, & Rahman, M. F. (2020). Comparative classroom discourse analysis among excellent, good and average secondary schools of Dhaka city in Bangladesh. *Teacher's World: Journal of Education and Research, 46*(Mujib Borsho), 112-129.

Nawrin, & Sadek, A. (2022). Role of rubric in assessment of language learning in higher education. *Teacher's World: Journal of Education and Research, 48*(2), 112–129. doi.org/10.3329/twjer.v48i2.67555

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill.

Patton, M. Q. (2002). *Qualitative research and evaluation methods* (3rd ed.). Sage.

Rahman, A., Islam, M. K., Al-Mamun, A., & Islam, M. S. (2025). Teachers' and students' use of ChatGPT at Social science faculty in the public and private Universities of Bangladesh. *F1000Research*, *14*(269). doi.org/10.12688/f1000research.161351.1

Ratislavová, K., & Ratislav, J. (2014). Asynchronous email interview as a qualitative research method in the humanities. *Human Affairs, 24*(4), 452–460. doi.org/10.2478/s13374-014-0240-y

Saleh, A., Al-Mamun, A., & Hasan, A. (2025). Educators' perspectives on integrating AI tools in educational practices: Opportunities, challenges, and ethical considerations. *The International Journal of Technologies in Learning*, *33*(1), 25. doi.org/10.18848/2327-0144/CGP/v33i01/25-50

Sharpe, D. (2015). Your Chi-Square Test Is Statistically Significant: Now What? *Practical Assessment, Research and Evaluation, 20*, 1-10.

Tate, T. P., Steiss, J., Bailey, D., Graham, S., Moon, Y., Ritchie, D., Tseng, W., & Warschauer, M. (2024). Can AI provide useful holistic essay scoring? *Computers and Education: Artificial Intelligence, 7*. doi.org/10.1016/j.caeai.2024.100255

Üretmen Karaoğlu, S., & Doğan, C. (2025). EFL teachers' insights on incorporating AI in language education. *Kuramsal Eğitimbilim Dergisi, 18*(3), 632–659. doi.org/10.55733/keg.1658985