

Research Article

Exercising Professional Judgment in AI-Assisted Assessment: Selective Mediation of AI-Generated Evaluation by English Lecturers

Nurhadi Hamka^{1*}, Akmal Jaya¹, Ikmal Muhammad¹, Nur Fajrhi¹
Nursastri Nurdin¹, Dhini Yustia Widhyah Saputri¹
Universitas Khairun, Ternate, Indonesia

*nurhadihamka@unkhair.ac.id

ABSTRACT

This qualitative case study examines how English lecturers exercise professional judgment when generative artificial intelligence (GenAI) is introduced into student writing assessment and feedback. Ten English lecturers at Universitas Khairun participated, producing 30 documented judgment episodes across three student writing samples. Each episode linked an initial lecturer assessment, AI-generated information from ChatGPT, Claude, or Gemini, and a final judgment. Data from assessment records, follow-up interviews, student texts, AI-generated outputs, and contextual documents were examined through within-episode and cross-episode thematic analysis. The findings show that AI did not move lecturers' scores in a single direction. Rather, it prompted them to retain, lower, or raise initial judgments after reconsidering the evidence. Lecturers evaluated AI output in relation to construct alignment, evidentiary weight, learner needs, task purpose, local context, and feedback usability. Professional agency was evident in how they determined which parts of the AI output could influence the final assessment. The study describes this process as selective evaluative mediation and argues that GenAI introduces a second object of professional judgment: lecturers judge not only student performance but also the relevance, usability, and authority of AI-generated interpretations of that performance.

Keywords: AI-Assisted Assessment; Assessment Literacy; Generative Artificial Intelligence; Professional Judgment; Teacher Agency

ARTICLE HISTORY

Received: 28.08.2026

Accepted: 13.10.2026

Published: 17.10.2026

ARTICLE LICENCE

Copyright ©2026 The Author(s): This is an open-access article distributed under the terms of the Creative Commons Attribution ShareAlike 4.0 International (CC BY-SA 4.0)

1. Introduction

Generative artificial intelligence (GenAI) is becoming part of assessment and feedback in higher education, particularly in language and writing instruction. Large language models can evaluate student texts, identify strengths and weaknesses, generate corrective or formative feedback, and suggest revisions. Research has therefore moved from asking whether GenAI can produce feedback to examining how AI-generated and teacher-generated feedback differ, how students respond to them, and how the two may work together. Escalante et al. (2023), Lu et al. (2024), Guo and Wang (2024), and Han and Li (2024) show that AI feedback can support writing development, although its focus, quality, consistency, and uptake do not always match teacher feedback. Similar differences have been reported for written corrective feedback, content and organization, linguistic accuracy, and student engagement (Lin & Crosthwaite, 2024; Muñoz Muñoz et al., 2025; Ding et al., 2025; Wang et al., 2026). Crosthwaite and Sun (2026) therefore place GenAI feedback within a developing area of L2 writing research rather than treating it as a direct replacement for human evaluation.

Useful AI feedback does not remove the need for human mediation. Steiss et al. (2024) found that AI can perform well against explicit criteria while differing from human evaluators in how feedback is prioritized and contextualized. Research on teacher-AI collaboration also shows that teachers often adapt or supplement AI feedback before using it with learners. Yao et al. (2025) identify complementary roles for AI-guided and teacher-initiated feedback, while Asadi et al. (2025) report benefits from combining ChatGPT with teacher feedback but also note continuing concerns about contextual understanding. The same issue is visible in studies of uptake: receiving AI feedback does not mean that students will engage with it or use it in revision (Zou et al., 2025). Zhu et al. (2025) further show that productive use of GenAI feedback requires knowledge of prompting, feedback design, and reflective evaluation, and Pretorius and Thewissen (2026) emphasize the value of teacher moderation in L2 writing revision. These studies suggest a simple but important point: AI-generated assessment information becomes useful only after someone decides how it should be interpreted, adapted, and used.

These decisions are closely tied to professional judgment and assessment literacy. Assessment is not a mechanical matching of student performance to criteria. Teachers interpret evidence, compare sources of information, and develop a defensible judgment. Phung and Michell (2022) show that experienced language teachers can move through different assessment pathways as initial judgement Gestalts are reconsidered in light of further evidence. Xu and Brown's (2016) Teacher Assessment Literacy in Practice framework similarly treats assessment literacy as an interaction among assessment knowledge, teacher conceptions, practice, learning, and sociocultural conditions rather than a fixed set of technical skills. Earlier work also shows that teachers' assessment practices are shaped by professional experience and institutional circumstances (Xu & Liu, 2009), while Xu (2019) emphasizes the need to interpret linguistic and communicative performance in relation to learners, tasks, purposes, and contexts. Subsequent work reinforces the multidimensional and context-sensitive nature of language assessment literacy (Xu & Brown, 2017; Vogt et al., 2024). In GenAI-enhanced assessment, this knowledge base has to extend to the evaluation of AI-generated information itself (Cui et al., 2025). Xu et al. (2026) likewise show that working with GenAI feedback can reshape teachers' feedback knowledge, capabilities, and dispositions when that feedback becomes an object of professional reflection.

Professional judgment, however, matters only if teachers can act on it. Mansouri et al. (2021) show that assessment literacy and professional agency are related but not identical: teachers may know what they consider appropriate while institutional conditions shape whether they can enact that judgment. Institutional arrangements can also influence teacher autonomy, agency, and assessment identity (Nazari et al., 2026; Derakhshan et al., 2026). This issue becomes more important when GenAI adds another source of evaluative information. Eedelouei's (2026) review shows that teachers' GenAI-related judgments are shaped by technological-pedagogical competence, professional beliefs, identity, ethical concerns, emotions, and contextual constraints. Jia and Pun (2025) similarly describe agency as a process of negotiating the affordances and limitations of ChatGPT rather than simply accepting or resisting it. At the institutional level, Jiang et al. (2025) point to difficulties associated with unclear guidance, limited support, and implementation. Barrot (2026) places teacher agency at the center of responsible GenAI-supported essay scoring, while Chen and Pi (2026) show that assessment decisions vary with teachers' AI familiarity, pedagogical beliefs, and interpretation of the assessment situation. In Indonesia, national guidance also emphasizes responsible GenAI use, institutional governance, competency development, and continued human responsibility in higher education (Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi, 2024).

Despite this growing literature, the evidence remains spread across several related lines of research. Some studies compare AI and human feedback, others examine teacher-AI collaboration, and more recent work addresses assessment literacy, feedback literacy, teacher agency, or institutional governance. Together, these studies explain much about what GenAI can provide and which factors shape teachers' engagement with it. Less is known, however, about the episode-level process through which a lecturer who has already formed an initial judgment responds when AI-generated assessment information is introduced. The closest studies still address different analytical objects. Lu et al. (2024) examine ChatGPT as a complement to teacher assessment, Han and Li (2024) focus on ChatGPT-supported teacher feedback, Chen and Pi (2026) examine teachers assessing GenAI-assisted student writing, and Barrot (2026) develops a conceptual account of teacher agency in AI-supported scoring. What remains less clear is how lecturers move from an initial evaluation to a final judgment while also judging the AI-generated evaluation itself.

This distinction matters because AI-assisted assessment may involve two connected judgments. One concerns the student work: what does it demonstrate in relation to the construct and criteria? The other concerns the AI-generated representation of that work: which observations are relevant, how much weight they should receive, whether the feedback is usable for a particular learner, and how much influence the output should have on the final decision. These questions move the analysis beyond broad categories of trust, acceptance, or resistance. They point instead to selective evaluative mediation, in which an AI comment may be plausible but poorly aligned with the construct, a suggested score may need to be reweighted against the rubric, or technically appropriate feedback may still need to be adapted for the learner and context. Examining these decisions provides a more precise account of professional judgment when AI enters assessment as an additional evaluative reference rather than as an autonomous assessor.

Against this background, the present study examines how English lecturers at Universitas Khairun exercise professional judgment when AI-generated evaluation and feedback are introduced into student writing assessment. It draws on 30 documented judgment episodes involving ten lecturers. Each episode links an initial lecturer assessment, exposure to AI-generated assessment information, further scrutiny of the student work and relevant evidence, and a final lecturer judgment. The study does not compare the accuracy of ChatGPT, Claude, and Gemini. Instead, it examines how lecturers decide whether AI-generated information should confirm, modify, qualify, or have limited influence on the final assessment. The analysis pays particular attention to construct alignment, evidentiary weighting, contextual and pedagogical knowledge, assessment literacy, and professional agency within the institutional setting.

Accordingly, the study addresses the following research questions:

- a. What patterns characterize lecturers' initial-to-final judgments when AI-generated assessment information is introduced into student evaluation and feedback?
- b. How do English lecturers exercise professional judgment when evaluating and selectively mediating AI-generated assessment information?
- c. How do assessment literacy, contextual and pedagogical knowledge, teacher agency, and institutional conditions shape professional judgment in AI-assisted assessment?

2. Method

2.1 Research Design and Context

This study used a qualitative case study design to examine how English lecturers exercise professional judgment when GenAI is introduced into the evaluation of student writing and feedback. A case study was appropriate because the focus was on lecturers' interpretations, reasoning, and decision-making within a bounded institutional and assessment context, rather than on comparing the accuracy or effectiveness of particular AI tools.

The case was bounded within the English Literature Department at Universitas Khairun, Ternate, Indonesia. The unit of analysis was the judgment episode. A judgment episode was defined as a documented sequence in which a lecturer first assessed a student writing sample, then encountered AI-generated assessment information related to the same sample, reconsidered the available evidence, and recorded a final professional judgment. Each episode therefore included four analytically connected stages: (1) initial lecturer assessment, (2) exposure to AI-generated assessment information, (3) comparison and professional scrutiny, and (4) final lecturer judgment. This structure allowed the analysis to examine not only whether lecturers agreed with AI-generated information, but also how they maintained, revised, or qualified their judgments after engaging with it.

The final corpus comprised 30 documented judgment episodes from ten lecturers, with each lecturer contributing three episodes. The three episodes followed predetermined writing-sample and AI-tool pairings: WS-A with ChatGPT, WS-B with Claude, and WS-C with Gemini. These pairings were used as structured assessment conditions, not as a design for comparing the three AI tools. The analysis therefore focused on lecturers' professional reasoning across episodes and made no claims about differences in the relative performance of ChatGPT, Claude, and Gemini.

2.2 Participants and Assessment Materials

Ten English lecturers participated in the study. Purposive selection was based on three criteria: involvement in English language teaching, responsibility for assessing student work, and previous experience with or engagement in GenAI for teaching, assessment, feedback, or related academic activities. The participants differed in teaching and assessment experience, familiarity with GenAI, and knowledge of the students whose writing was used in the assessment task. These differences were relevant because the study examined how assessment knowledge, pedagogical experience, learner knowledge, and contextual understanding entered professional judgment.

Three anonymized student essays on political discourse in the 2024 North Maluku gubernatorial election served as common assessment materials. The samples were selected to create contrasting assessment conditions rather than a performance hierarchy. WS-A showed relatively strong language control and clear organization but remained largely descriptive in its discourse analysis. WS-B provided more explicit analysis of pronouns, lexical choices, modality, evaluation, representation, identity, and power. WS-C gave greater attention to local context, political framing, regional identity, language choice, social media, and audience interpretation. Thus, the three samples differed in the balance between surface language quality, analytical development, and contextual interpretation.

A common assessment task and the same assessment criteria were used across the documented episodes. AI-generated assessment information was provided as an

additional evaluative reference that lecturers could compare with their own reading of the student work. The purpose was not to identify the most accurate AI system. Rather, the design created situations in which lecturers had to decide whether, and to what extent, AI-generated information was relevant to the final assessment.

2.3 Data Collection Procedure

Data collection combined structured judgment episodes, follow-up interviews, and supporting assessment materials. For each predetermined writing-sample and AI-tool pairing, lecturers first reviewed the student writing and recorded an initial judgment, including an overall score and brief assessment notes. They were then shown AI-generated assessment information for the same sample. After reading the AI output, lecturers compared it with their initial evaluation, reconsidered the student text and other relevant evidence, and recorded a final judgment. This record included the final score, the lecturer's response to the AI-generated assessment, and the principal resource informing the decision.

This procedure produced three documented episodes for each lecturer and 30 episodes in total. The assessment data collection sheet recorded the lecturer code, writing sample, AI tool, initial score and notes, final score and notes, response to the AI output, interview date, and primary judgment resource. In this way, each episode provided a traceable sequence from the initial evaluation, through engagement with AI-generated information, to the final decision.

Follow-up interviews explored the reasoning behind these decisions. The interview records covered all 30 episodes and asked lecturers to explain why they maintained or revised their initial judgments, which parts of the AI-generated assessment they considered useful or problematic, and what evidence informed their final decisions. The interviews also elicited references to assessment criteria, task purpose, discourse and language knowledge, learner history, proficiency level, local context, professional experience, and pedagogical considerations. The records therefore provided accounts of why particular AI observations were accepted, reweighted, contextualized, or excluded from the final assessment.

Institutional and assessment-related materials were examined to describe the conditions in which the judgments were made. These materials included course-level assessment criteria and rubrics, available guidance on responsible AI use, and records concerning lecturer discretion and assessment practices. They were used to contextualize professional judgment, not to establish causal links between institutional conditions and individual decisions.

2.4 Data Analysis, Trustworthiness, and Ethical Considerations

The 30 judgment episodes formed the primary analytic corpus. Analysis proceeded in two stages. First, within-episode analysis traced the movement from initial to final judgment. Each episode was examined for the initial score and rationale, the AI-generated assessment information considered, points of agreement or disagreement, evidence used in reconsidering the student work, the final score, and the explanation for the final decision. Initial and final scores were also compared descriptively to identify the direction of revision. Score changes were classified as decreased, unchanged, or increased. These patterns were used to describe the judgment process rather than to test effects of AI.

Second, cross-episode thematic analysis identified recurring patterns in how lecturers evaluated and mediated AI-generated assessment information. Initial coding

focused on observable forms of judgment, including construct alignment, assessment criteria, task purpose, evidentiary weighting, score revision, feedback adaptation, interpretation of student performance, learner knowledge, contextual knowledge, and pedagogical usability. Codes were then compared across episodes and grouped into broader patterns. Four interconnected findings were developed: (1) patterns of AI-assisted assessment and judgment revision, (2) construct-sensitive and selective mediation of AI-generated information, (3) the role of contextual and pedagogical knowledge in AI-mediated judgment, and (4) professional agency in defining the boundaries of AI authority.

Score change and response to AI were treated as analytically distinct. An unchanged final score did not automatically indicate acceptance of the AI assessment because lecturers could retain a score while revising the feedback or rejecting part of the AI interpretation. In the same way, a revised score was not taken as evidence that the lecturer had followed the AI recommendation. This distinction was maintained throughout coding so that mediation could be examined at the level of score, feedback, interpretation, or evidentiary weighting.

Trustworthiness was supported through triangulation across assessment records, student writing samples, interview accounts, AI-generated assessment information, and contextual materials. Episode-level comparison allowed interview explanations to be checked against the recorded initial and final judgments and, where relevant, against features of the student texts. Contrasting cases were also examined, including episodes in which scores decreased, remained unchanged, or increased. This helped prevent the analysis from privileging a single direction of response.

Lecturers were identified by codes L01-L10, and the writing samples as WS-A, WS-B, and WS-C. Identifying information was removed from the analysis and reporting. Participation was based on informed consent, and the assessment records were used only for research purposes. Interview quotations originally recorded in Indonesian were retained in the original language and followed by English translations in brackets so that the lecturers' wording remained visible to international readers.

3. Result

3.1 Patterns of AI-Assisted Assessment and Judgment Revision

The traceable dataset comprised 30 judgment episodes involving ten lecturers and three student writing samples. Each lecturer contributed three episodes. Every episode recorded an initial assessment, exposure to AI-generated assessment information, reconsideration of the student work, and a final lecturer judgment. WS-A was paired with ChatGPT output, WS-B with Claude, and WS-C with Gemini. These episodes are therefore treated as cases of AI-assisted professional judgment, not as evidence for comparing the relative performance of the three AI tools.

Across the 30 episodes, lecturers' initial scores averaged 4.07 out of 5 and their final scores averaged 3.70. Figure 1 shows that the movement from initial to final judgment was not uniform. Sixteen episodes (53.3%) ended with a lower score, ten (33.3%) retained the initial score, and four (13.3%) ended with a higher score.

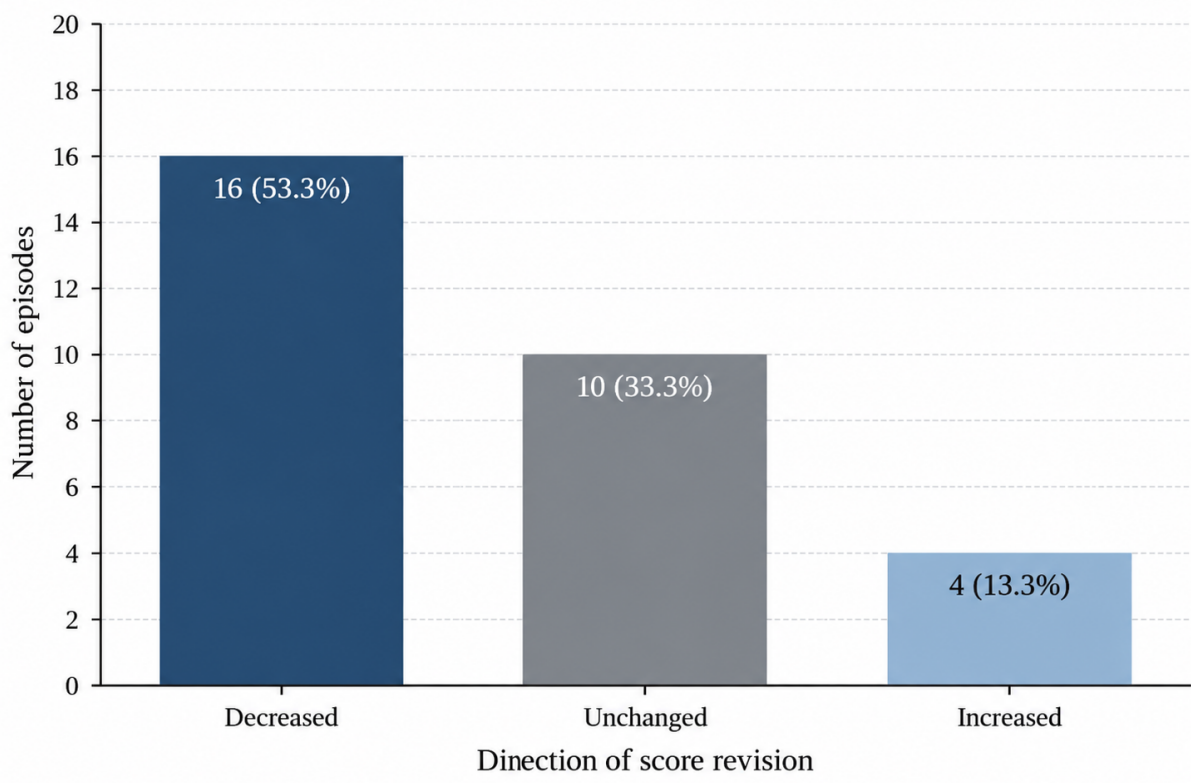


Figure 1. Patterns of score revision across the documented judgment episodes

The score pattern shows that exposure to AI-generated assessment information did not move lecturers in one numerical direction. Some lecturers lowered a score after comparing their judgment with the AI output; others retained the original score, and several raised it. AI therefore entered the assessment process as an additional point of reference rather than as a source that consistently pushed judgments upward or downward.

The interview records further show that an initial judgment was not always treated as fixed. In A001, L01 explained that the initial score of 4/5 had been assigned before the AI output was examined and that the lecturer deliberately avoided 'locking' the judgment at that point. After reading the ChatGPT assessment and returning to the essay, L01 reconsidered the quality of the argument and reduced the final score to 3.5/5. The change did not result from simply adopting the AI evaluation. Instead, the AI output prompted a closer reading, after which the lecturer judged the writing to be stronger descriptively than analytically.

The same pattern appeared in other episodes. In A015, L05 initially gave WS-C a score of 5/5 because of its strong contextual orientation. After reading the Gemini output and returning to the essay, the lecturer identified insufficient textual support for several contextual claims and revised the score to 4/5. L05 described the AI output as a reason to read the essay more carefully, not as a direct basis for changing the score. In A023, a score of 5/5 was similarly revised to 4/5 after the lecturer reconsidered whether strong vocabulary and clear organization were matched by equally strong development of the argument.

Reconsideration did not always lead to a lower score. Four episodes resulted in upward revisions. In A030, L10 initially gave WS-C 3/5 but considered the Gemini

assessment too focused on surface weaknesses and not sufficiently attentive to the student's use of local context. After returning to the essay, the lecturer increased the score to 4/5 while still noting weak links between contextual claims and textual evidence. Here, engagement with AI led the lecturer away from, rather than toward, the emphasis of the AI evaluation.

Overall, AI-generated assessment functioned as an additional evaluative reference within a lecturer-led process. Its most consistent role was not to determine the final score, but to provide another point of comparison that could confirm, unsettle, or revise an initial reading of student performance. The resulting judgments moved in different directions because lecturers returned to the student work and reconsidered the evidence in relation to the AI-generated assessment.

3.2 Professional Judgment through Construct-Sensitive and Selective Mediation

Lecturers did not evaluate AI-generated information only by asking whether it appeared reasonable or whether its score matched their own. They also asked whether the output reflected what the task was intended to assess and whether particular observations deserved the weight given to them. The assessment records show several professional resources behind these decisions. Rubric was recorded as the primary resource in eight episodes, Task purpose in four, and Learning objectives and Language knowledge in three each. Other episodes combined rubric and discourse knowledge or referred to assessment criteria, linguistic or discourse knowledge, professional experience, and learner-related knowledge. The relevance of AI-generated information was therefore judged against the task and assessment context rather than in isolation.

This construct-sensitive scrutiny was clear when lecturers separated surface language quality from the substantive performance required by the task. In A001, L01 initially scored WS-A at 4/5 and recognized that the writing was relatively accurate and well organized. After examining the ChatGPT output and rereading the essay, however, the lecturer reduced the score to 3.5/5 because the argument remained largely descriptive. L01 stated, "*Tugas ini bukan tugas gramatika, melainkan tugas analisis wacana*" ["*This is not a grammar task, but a discourse-analysis task*"]. The distinction was consistent with WS-A, which identified campaign themes, slogans, and development issues but offered more limited analysis of how linguistic choices constructed identity, power, ideology, or representation.

A similar distinction appeared in A023. L08 initially gave WS-B 5/5 but reconsidered the score after comparing the essay with the Claude assessment. The lecturer acknowledged the strength of the vocabulary and the apparent organization of the text, yet questioned whether these features were matched by equally strong argument development. L08 explained, "*Bagi saya, kosakata yang kuat tidak otomatis berarti argumen yang kuat. Ini adalah dua dimensi yang berbeda*" ["*For me, strong vocabulary does not automatically mean a strong argument. These are two different dimensions*"]. After rereading the essay and identifying claims that lacked sufficient textual support, the lecturer revised the score to 4/5. The episode shows that an AI observation could be plausible without providing sufficient evidence of overall task achievement.

Lecturers also judged the weight that AI-generated observations should receive. In A029, L10 found several parts of the Claude assessment consistent with the rubric, but considered some recommendations to give too much emphasis to criteria that were not priorities in the lecturer's rubric. The lecturer retained some observations while reducing their importance in the final judgment. The score moved from 4/5 to 3.5/5, with stronger textual support receiving greater emphasis while the student's analytical strengths were

still recognized. Professional judgment in this episode concerned not only whether an AI observation was acceptable, but how much influence it should have.

The records also show that lecturers did not always treat AI-generated scores, interpretations, and feedback as one package. Different components could be used differently in the same episode. In A014, the lecturer retained the initial score of 4/5 but used the AI output to identify recurrent grammatical problems for feedback. In A011, the lecturer considered the AI score broadly reasonable but found the accompanying feedback insufficiently connected to the classroom context and task expectations. The feedback was therefore contextualized rather than adopted directly, while the final score moved from 3/5 to 3.5/5. L04 described the AI output as a source of consideration but stated that it *"tidak bisa langsung dipakai sebagai feedback akhir"* ["could not be used directly as final feedback"] .

The distinction between score and feedback was even clearer in A027. L09 retained the initial score of 5/5 because the Gemini assessment was broadly consistent with the lecturer's evaluation of WS-C. The feedback, however, was rewritten because its terminology was considered too advanced for the student. As L09 explained, the terminology could be academically appropriate but *"tidak akan membantu mahasiswa memahami apa yang perlu dia perbaiki"* ["would not help the student understand what needed to be improved"] . Professional mediation therefore occurred even though the numerical judgment did not change.

These episodes show several forms of selective mediation within the assessment process.

Table 1. Forms of selective mediation observed in the documented judgment episodes

Form of mediation	What was reconsidered	Illustrative episodes
Construct alignment	Whether AI observations reflected what the task was intended to assess	A001, A023
Evidentiary weighting	How much importance particular AI observations should receive	A023, A029
Score mediation	Whether the numerical judgment should be revised after reconsidering the evidence	A001, A015, A023
Feedback mediation	Whether AI-generated feedback could be retained, contextualized, or rewritten	A011, A014, A027
Interpretive mediation	Whether the lecturer accepted the way AI interpreted particular features of student performance	A003, A012, A015

Across the 30 episodes, professional judgment was exercised through selective use of AI-generated information. A lecturer could accept an observation but change its weight, retain a score but rewrite the feedback, or revise an initial judgment after deciding that the AI interpretation did not adequately reflect the construct or task. This distinction matters because an unchanged score did not necessarily mean uncritical agreement with AI, and a changed score did not necessarily mean that the lecturer followed the AI recommendation. The final judgment depended on how different parts of the AI output were selected, weighted, and adapted in relation to the task and available evidence.

3.3 Contextual and Pedagogical Knowledge in AI-Mediated Judgment

The episodes also show that lecturers did not rely only on the student text, rubric, or AI-generated assessment. In several cases, they drew on knowledge of the learner, previous performance, communicative purpose, local context, and pedagogical considerations when deciding whether AI-generated information was appropriate for a particular student or task. These resources became most visible when the lecturer's interpretation differed from the emphasis of the AI output.

Knowledge of individual students was clear in A019. L07 initially gave WS-A 4/5, partly because the essay appeared stronger than the student's earlier work. The lecturer, however, interpreted this improvement in relation to the student's previous performance rather than treating the submitted text as an isolated product. After reading the ChatGPT output, L07 explained, *"AI menilai esai ini sebagai produk yang berdiri sendiri, tanpa konteks historis. Bagi saya, itu masalah"* [*"AI evaluates this essay as a stand-alone product, without historical context. For me, that is a problem"*]. The final score was reduced to 3/5 because the lecturer judged the current performance against what the student had previously demonstrated and what still required development. The episode shows that the evidentiary basis of judgment extended beyond the immediate written product.

Learner proficiency also shaped how lecturers evaluated AI-generated feedback. In A020, L07 considered the score suggested by Claude broadly reasonable but found the feedback too advanced for the student. The lecturer explained, *"AI memberi saran yang terdengar bagus, tapi kalau saya sampaikan ke mahasiswa ini, dia akan bingung. Feedback itu terlalu advanced"* [*"AI gives suggestions that sound good, but if I convey them to this student, the student will be confused. The feedback is too advanced"*]. The final score increased from 3/5 to 3.5/5, but the feedback was not adopted in its original form. The issue was therefore not only whether the AI feedback was plausible, but whether it was usable for this learner.

A related pattern appeared in A027. L09 regarded the Gemini assessment as broadly consistent with the lecturer's own evaluation of WS-C and retained the initial score of 5/5. The AI-generated feedback was nevertheless rewritten because its terminology was considered too sophisticated for the student. L09 stated that the terminology could be academically appropriate but *"tidak akan membantu mahasiswa memahami apa yang perlu dia perbaiki"* [*"would not help the student understand what needed to be improved"*]. Again, professional mediation occurred without a change in score.

Lecturers also drew on the communicative purpose of the task when deciding how much weight to give AI-generated corrections. In A021, L07 considered some of Gemini's corrections unnecessarily strict because the expressions remained acceptable for the communicative demands of the discourse-analysis task. The lecturer did not treat every linguistic issue identified by AI as equally relevant to task achievement. Instead, the emphasis was adjusted and the score was reduced only slightly, from 4/5 to 3.5/5, while the essay's contextual strengths were retained in the judgment. The episode shows that an AI-identified language issue could matter differently depending on the purpose of the assessment.

Local knowledge was particularly visible in judgments of WS-C. In A003, L01 questioned Gemini's interpretation because some linguistic choices were considered meaningful within the sociopolitical context of North Maluku rather than simply problematic at the surface level. L01 stated, *"AI kurang peka terhadap konteks sosial-politik Maluku Utara yang menjadi latar esai ini"* [*"AI is less sensitive to the sociopolitical context of North Maluku that forms the background of this essay"*]. The lecturer therefore

interpreted the text in relation to linguistic choice, local context, and task purpose. This reading was consistent with WS-C, which connected language choices with regional identity, audience, local linguistic resources, and political positioning.

The episodes therefore show several forms of contextual and pedagogical knowledge entering AI-mediated assessment:

Table 2. Contextual and pedagogical knowledge informing AI-mediated judgment

Knowledge resource	How it informed judgment	Illustrative episode
Student history	Interpreting current performance in relation to previous work	A019
Learner proficiency	Judging whether AI feedback was usable for the student	A020
Communicative purpose	Determining whether linguistic corrections were central to task achievement	A021
Pedagogical knowledge	Rewriting feedback to make it accessible to the learner	A027
Local/contextual knowledge	Interpreting linguistic choices within the sociopolitical context of the task	A003

These episodes, as depicted on table 2, show that lecturers extended the evidentiary basis of assessment beyond the information contained in the AI output. Their judgments included knowledge of who the learner was, what the learner had previously demonstrated, what the task was intended to achieve, and what kind of feedback was likely to be useful. AI-generated assessment information was therefore read against learner- and context-specific knowledge rather than applied as a context-independent evaluation.

3.4 Professional Agency and the Boundaries of AI Authority

Professional agency was visible not simply because lecturers retained responsibility for the final assessment, but because they determined how far AI-generated information could influence their decisions. The procedure itself required lecturers to make the final judgment, so final decision-making alone is weak evidence of agency. Stronger evidence appeared in the ways lecturers retained, qualified, adapted, or overrode particular parts of the AI-generated assessment after comparing them with other evidence.

This was clear when lecturers retained their original scores despite differences in how AI represented student performance. In A006, L02 maintained the initial score of 4/5 for WS-C while questioning Gemini's interpretation of several linguistic and contextual features. The lecturer explained, "*saya tidak ingin mengubah skor hanya karena AI memiliki framing yang berbeda dari saya*" ["I do not want to change the score simply because AI has a different framing from mine"]. The original score was retained on the basis of classroom context, task purpose, and knowledge of the student. Agency in this episode was expressed through the decision not to let a competing AI interpretation automatically alter the outcome.

Agency was also visible when lecturers retained the numerical judgment but changed how AI-generated information was used pedagogically. In A027, L09 kept the score at 5/5 but rewrote the feedback because Gemini's terminology was too advanced for the learner. The lecturer noted that the terminology might be academically appropriate

but "*tidak akan membantu mahasiswa memahami apa yang perlu dia perbaiki*" ["*would not help the student understand what needed to be improved*"]. Retaining the score therefore did not mean adopting the AI assessment unchanged. The lecturer accepted one component while modifying another.

In other episodes, lecturers reduced the influence of AI-generated information because its interpretation was too general or insufficiently aligned with the task. A028 illustrates this pattern. L10 considered the ChatGPT comments too general to the student text and observed, "*Komentar AI bisa saja diberikan kepada hampir semua esai mahasiswa*" ["*The AI comments could have been given to almost any student essay*"]. Rather than allowing the fluency of the AI evaluation to determine the final decision, the lecturer returned to the essay and reduced the score from 4/5 to 3.5/5 because of limited analytical depth. The AI output therefore had limited authority because its relevance to the specific assessment evidence was judged to be weak.

The opposite pattern also occurred. In A030, L10 increased the score for WS-C from 3/5 to 4/5 after deciding that the Gemini assessment gave too little weight to the student's contextual achievement. The lecturer explained that "*AI terkesan terlalu fokus pada kekurangan permukaan dan tidak cukup menghargai usaha mahasiswa dalam membawa konteks lokal ke dalam analisis*" ["*AI seemed too focused on surface weaknesses and did not sufficiently recognize the student's effort to bring local context into the analysis*"]. The AI output was not simply dismissed. Instead, the disagreement prompted another reading of the essay, after which the lecturer gave greater recognition to the student's use of local context. This upward revision shows that agency could also involve resisting an AI evaluation judged to be overly restrictive.

The practical authority given to AI therefore varied across episodes. In A024 and A025, AI information mainly confirmed an existing judgment: scores were retained while the output informed feedback. In A029, parts of the Claude assessment were retained, but their weighting was changed because some criteria were not central to the lecturer's priorities. AI authority was therefore neither fixed nor all-or-nothing. It was negotiated at the level of scores, interpretations, recommendations, and feedback.

The institutional records place this discretion within a particular setting. Lecturers worked with course-level assessment rubrics and broad national guidance on responsible AI use and human oversight, but no university-level policy specifically governing AI-assisted assessment was identified. The records also noted lecturer discretion over grading, a predominantly lecturer-led assessment culture, and time pressure during grading periods. These conditions are not treated as direct causes of individual decisions. They describe the professional setting in which lecturers decided how AI-generated information could be used.

Overall, professional agency in AI-assisted assessment was expressed through the selective allocation of authority to AI-generated information. Lecturers could use an AI observation to confirm an existing judgment, change the weight assigned to it, adapt its feedback, disregard an interpretation considered insufficiently relevant, or revise their own assessment after further scrutiny. The boundary of AI authority was therefore established through situated professional judgment rather than through wholesale acceptance or categorical rejection.

4. Discussion

The findings are better explained as professional mediation than as a transfer of evaluative authority from lecturers to AI. AI-generated assessment did not move lecturers' decisions in one direction, and lecturers did not treat scores, interpretations,

and feedback as a single package. They could retain one component, modify another, or set part of the output aside within the same episode. This pattern is consistent with work that positions GenAI as a complement to teacher assessment. Lu et al. (2024) show that ChatGPT can contribute useful assessment information while differing from teachers in important ways, and Han and Li (2024) similarly show that teachers adapt AI-generated feedback rather than simply pass it on. The present study makes this complementarity more visible at the level of the judgment episode. AI mattered because it introduced another representation of student performance that lecturers themselves had to judge.

One part of this mediation involved reconsidering judgments that had already begun to form. An initial score was not always a fixed endpoint. When AI offered a different or partly similar interpretation, lecturers sometimes returned to the student text, reconsidered the evidence, and then retained or revised the original decision. This pattern is close to Phung and Michell's (2022) account of assessment pathways, in which initial judgement Gestalts are developed and refined as assessors engage with further evidence. GenAI adds a different element to that pathway because the additional input is an interpretation of the student's performance, not simply another feature of the performance itself. Xu et al. (2026) show that engagement with GenAI feedback can support professional reflection, while Chen and Pi (2026) demonstrate that teacher reasoning in GenAI-related assessment remains shaped by professional knowledge and pedagogical beliefs. In the present study, this process can be described as reflective recalibration: AI provided another reading against which lecturers reconsidered their own, without requiring them to accept the AI conclusion.

The reconsideration was also construct-sensitive. Lecturers asked not only whether an AI comment was plausible, but whether it addressed what the task was intended to assess and how much weight it deserved. This distinction is consistent with Xu and Brown's (2016) view of assessment literacy as situated practice rather than the mechanical use of criteria. Xu (2019) makes a similar point for language assessment, where judgments depend on the construct, task purpose, communicative demands, and characteristics of learner performance. GenAI makes this work more demanding because teachers must also evaluate the assessment information generated about the student work. Cui et al. (2025) therefore argue that assessment literacy in GenAI-enhanced environments needs to include critical evaluation of AI-generated information. The present findings show what this looks like in practice: lecturers had to decide what should count as assessment evidence and how much it should count.

This is what the present study refers to as selective evaluative mediation. Lecturers could retain, reweight, adapt, contextualize, or exclude different parts of AI-generated assessment before allowing them to influence the final judgment. The same output could be treated differently at the level of score, interpretation, and feedback. Han and Li (2024) likewise show that AI-supported teacher feedback involves adaptation, while Barrot (2026) argues that responsible AI-assisted scoring depends on active teacher mediation rather than the displacement of evaluative expertise. The present findings make the location of that mediation more specific. It occurred in judgments about construct alignment, evidentiary weight, numerical score, interpretation of student performance, and the wording and pedagogical purpose of feedback. For that reason, an unchanged score cannot be read automatically as agreement with AI, just as a changed score cannot be taken as evidence that the lecturer followed the AI recommendation.

Professional mediation also drew on forms of knowledge that were not fully represented in the AI output. Lecturers referred to previous student performance, proficiency level, communicative purpose, local sociopolitical context, and the likely usefulness of feedback for a particular learner. This is consistent with work showing that

language assessment knowledge is situated in professional practice and is closely tied to knowledge of learners and instructional contexts (Xu & Liu, 2009; Xu, 2019). Crosthwaite and Sun (2025) likewise position GenAI feedback within the broader pedagogical traditions of L2 writing feedback, while Zou et al. (2025) show that providing feedback does not guarantee that learners will take it up productively. Pretorius and Thewissen (2026) further demonstrate the value of teacher moderation when ChatGPT feedback is used in L2 writing revision. These studies help explain why technical or academic plausibility alone was not enough for lecturers in the present study.

Pedagogical usability was one of the clearest examples. Some AI feedback was considered academically defensible but unsuitable for the student's proficiency level or learning needs. Lecturers therefore asked not only whether the feedback identified a problem correctly, but whether the student could understand and act on it. Human mediation, from this perspective, is not limited to correcting AI errors. A recommendation can be factually reasonable and still require revision because it is too abstract, gives too much weight to a minor issue, does not fit the task, or is expressed at a level that limits its instructional value. Professional judgment therefore operated at both evaluative and pedagogical levels. The lecturer had to decide what the work warranted and how that judgment could be communicated in a form that supported learning.

The same evidence provides a more precise account of professional agency. Agency was not demonstrated simply because lecturers made the final decision, since the research procedure required them to do so. It was more visible in how they regulated the practical influence of AI-generated information. A lecturer could preserve an interpretation despite AI disagreement, accept only part of an AI assessment, rewrite feedback while retaining the score, or revise an initial judgment after deciding that the AI had drawn attention to relevant evidence. Mansouri et al. (2021) distinguish assessment knowledge from the capacity to enact that knowledge under particular professional conditions. Eedelouei (2026) likewise shows that GenAI-related judgments emerge from professional, technological, affective, and contextual factors, while Jia and Pun (2025) describe teacher agency with ChatGPT as situated and negotiated. The present findings suggest that agency in AI-assisted assessment is best understood as the selective allocation of AI authority.

This formulation moves beyond a simple acceptance-versus-rejection model. Accepting AI-generated information may itself be an agentic decision when the lecturer has examined it and decided that it deserves influence. In the same way, disagreement does not require rejection of the entire output. Authority can be allocated differently to a score, an interpretation, a recommendation, or a feedback statement. Barrot (2026) similarly emphasizes active teacher mediation and professional autonomy in AI-supported scoring. The present study extends this argument by locating agency in the practical decisions through which lecturers decide whether AI information should confirm an existing judgment, receive less weight, be adapted for pedagogical purposes, or be excluded. Agency therefore concerns the terms under which AI-generated information is allowed to become assessment evidence.

Institutional conditions set the boundaries within which this discretion was exercised. The lecturers worked with course-level assessment structures and broad national guidance on responsible AI use and human oversight, but the contextual records did not identify a university-level policy specifically governing AI-assisted assessment. The same records indicated substantial lecturer discretion over grading, a lecturer-led assessment culture, and time pressure during grading periods. These conditions do not explain individual judgments directly. Rather, they define the professional space in which such judgments are made. Mansouri et al. (2021) show that institutional policies can mediate the relationship between assessment literacy and agency, while Nazari et al.

(2023) and Derakhshan et al. (2024) show how institutional arrangements shape autonomy, identity, and assessment practice. Jiang et al. (2025) similarly identify issues of guidance, implementation, and professional support in higher education GenAI use. Indonesian national guidance provides a broader framework for responsible use and human responsibility (Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi, 2024), but those broad principles still require interpretation in concrete assessment situations.

These findings point to a broader interpretation of AI-assisted assessment. In conventional assessment, lecturers interpret student performance in relation to the purpose of the task and the available evidence. GenAI adds a second object of professional judgment: lecturers must also evaluate an AI-generated representation of that performance. This second judgment concerns at least four matters: whether the AI interpretation is relevant to the construct, how much evidentiary weight it deserves, whether the feedback is pedagogically usable, and how much authority it should have in the final decision. This interpretation connects assessment pathways (Phung & Michell, 2022), situated assessment literacy (Xu & Brown, 2016), emerging work on assessment literacy in GenAI environments (Cui et al., 2025), and recent accounts of teacher mediation and agency in AI-assisted assessment (Barrot, 2026; Chen & Pi, 2026). The present study adds episode-level evidence showing how this second layer of judgment operates.

AI-assisted assessment is therefore better understood as selective evaluative mediation than as a choice between human and artificial judgment or a continuum between trust and distrust. Lecturers have to decide what AI-generated information should count, how much it should count, and under what conditions it should influence assessment. This also gives a more practical meaning to human oversight. Keeping a lecturer formally 'in the loop' is not enough if oversight means only final approval. It requires assessment literacy, contextual and pedagogical knowledge, and professional agency to scrutinize, reweight, adapt, or disregard AI-generated information when necessary. For professional development and institutional policy, this means that AI competence in assessment should extend beyond prompting and familiarity with particular tools. Lecturers also need to judge construct alignment, evidentiary relevance, learner-appropriate feedback, contextual adequacy, and the limits of AI authority. These capacities allow GenAI to support professional judgment without replacing the situated responsibility of the human assessor.

5. Conclusion

This study shows that AI-assisted assessment is not best described as a transfer of evaluative authority from lecturers to GenAI. Across the 30 documented judgment episodes, AI-generated information served as an additional evaluative reference that could prompt lecturers to retain, lower, or raise an initial score after returning to the student work. Lecturers did not use AI output as a single assessment package. They judged construct alignment, adjusted the weight of particular observations, revised or retained scores, contextualized interpretations, and adapted feedback in relation to learner needs, task purpose, and local context. Professional agency was therefore visible in the selective allocation of authority to different parts of the AI output rather than in final decision-making alone. Institutional structures and broad AI guidance provided the setting for this discretion without directly determining individual judgments. Taken together, the findings suggest that GenAI introduces a second object of professional judgment: lecturers assess the student performance and, at the same time, assess the relevance, evidentiary weight, pedagogical usability, and legitimate authority of AI-

generated interpretations of that performance. The study describes this process as selective evaluative mediation. From this perspective, meaningful human oversight depends on lecturers' assessment literacy, contextual and pedagogical knowledge, and capacity to decide what AI-generated information should count, how much it should count, and when it should influence the final assessment.

References

- Asadi, M., Ebadi, S., & Mohammadi, L. (2025). The impact of integrating ChatGPT with teachers' feedback on EFL writing skills. *Thinking Skills and Creativity*, 56, 101766. <https://doi.org/10.1016/j.tsc.2025.101766>
- Barrot, J. S. (2026). Generative artificial intelligence for automated essay scoring: Exploring teacher agency through an ecological perspective. *Assessing Writing*, 67, 100990. <https://doi.org/10.1016/j.asw.2025.100990>
- Chen, X., & Pi, X. (2026). Assessing EFL students' GenAI-assisted writing: Teachers' pains, perceptions and practices. *Assessing Writing*, 67, 100996. <https://doi.org/10.1016/j.asw.2025.100996>
- Crosthwaite, P., & Sun, S. (2026). Generative AI and L2 written feedback studies: A scoping review. *RELC Journal*, 57(1), 207–219. <https://doi.org/10.1177/00336882251386530>
- Cui, Y., Meng, Y., & Tang, L. (2025). Reconsidering teacher assessment literacy in GenAI-enhanced environments: A scoping review. *Teaching and Teacher Education*, 165, 105163. <https://doi.org/10.1016/j.tate.2025.105163>
- Derakhshan, A., Molana, K., & Nazari, M. (2026). "I wonder who I am when my assessment ideas are not respected": A case-study inquiry into the role of institutional policies in English language teachers' assessment identity. *Asia Pacific Journal of Education*, 46(1), 116–131. <https://doi.org/10.1080/02188791.2024.2328237>
- Ding, L., Zou, D., & Kohnke, L. (2025). ChatGPT as an automated writing evaluation tool: How students perceive it and how it affects their writing. *Education and Information Technologies*, 30, 26777–26799. <https://doi.org/10.1007/s10639-025-13775-3>
- Eedelouei, S. (2026). Factors influencing language teachers' judgements and decision-making about the use of generative AI tools: A systematic review. *Teaching and Teacher Education*, 171, 105333. <https://doi.org/10.1016/j.tate.2025.105333>
- Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education*, 20, 57. <https://doi.org/10.1186/s41239-023-00425-2>
- Guo, K., & Wang, D. (2024). To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing. *Education and Information Technologies*, 29, 8435–8463. <https://doi.org/10.1007/s10639-023-12146-0>
- Han, J., & Li, M. (2024). Exploring ChatGPT-supported teacher feedback in the EFL context. *System*, 126, 103502. <https://doi.org/10.1016/j.system.2024.103502>
- Jia, W., & Pun, J. (2025). Exploring EMI teachers' agency in addressing language-related challenges with ChatGPT: A multiple case study. *Computer Assisted*

Language Learning. Advance online publication.
<https://doi.org/10.1080/09588221.2025.2511061>

- Jiang, Y., Xie, L., & Cao, X. (2025). Exploring the effectiveness of institutional policies and regulations for generative AI usage in higher education. *Higher Education Quarterly*, 79(4), e70054. <https://doi.org/10.1111/hequ.70054>
- Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi. (2024). *Panduan penggunaan generative artificial intelligence pada pembelajaran di perguruan tinggi*. Direktorat Pembelajaran dan Kemahasiswaan, Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi.
- Lin, S., & Crosthwaite, P. (2024). The grass is not always greener: Teacher vs. GPT-assisted written corrective feedback. *System*, 127, 103529. <https://doi.org/10.1016/j.system.2024.103529>
- Lu, Q., Yao, Y., Xiao, L., Yuan, M., Wang, J., & Zhu, X. (2024). Can ChatGPT effectively complement teacher assessment of undergraduate students' academic writing? *Assessment & Evaluation in Higher Education*, 49(5), 616–633. <https://doi.org/10.1080/02602938.2024.2301722>
- Mansouri, B., Molana, K., & Nazari, M. (2021). The interconnection between second language teachers' language assessment literacy and professional agency: The mediating role of institutional policies. *System*, 103, 102674. <https://doi.org/10.1016/j.system.2021.102674>
- Muñoz Muñoz, B. C., Nassaji, H., & Bello Carrillo, F. I. (2025). ChatGPT-generated versus human direct corrective feedback on L2 writing. *System*, 134, 103805. <https://doi.org/10.1016/j.system.2025.103805>
- Nazari, M., De Costa, P. I., & Karimpour, S. (2026). The role of institutional policy in English language teacher autonomy, agency, and identity: A poststructural perspective. *Language Teaching Research*, 30(2), 885–907. <https://doi.org/10.1177/13621688221143476>
- Phung, D. V., & Michell, M. (2022). Inside teacher assessment decision-making: From judgement Gestalts to assessment pathways. *Frontiers in Education*, 7, 830311. <https://doi.org/10.3389/feduc.2022.830311>
- Pretorius, M., & Thewissen, J. (2026). Mediating L2 writing revision: The impact of teacher-moderated ChatGPT feedback on L2 accuracy and lexical diversity. *Journal of Second Language Writing*, 71, 101286. <https://doi.org/10.1016/j.jslw.2026.101286>
- Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction*, 91, 101894. <https://doi.org/10.1016/j.learninstruc.2024.101894>
- Vogt, K., Bøhn, H., & Tsagari, D. (2024). Language assessment literacy. *Language Teaching*, 57(3), 325–340. <https://doi.org/10.1017/S0261444824000090>
- Wang, Y., Hu, W., & Guo, Q. (2026). GPT versus teacher feedback on English writing: An analysis of feedback characteristics and student engagement. *System*, 138, 103967. <https://doi.org/10.1016/j.system.2025.103967>
- Xu, Y. (2019). English language teacher assessment literacy in practice. In X. Gao (Ed.), *Second handbook of English language teaching* (pp. 517–539). Springer. https://doi.org/10.1007/978-3-030-02899-2_22

- Xu, Y., & Brown, G. T. L. (2016). Teacher assessment literacy in practice: A reconceptualization. *Teaching and Teacher Education*, 58, 149–162. <https://doi.org/10.1016/j.tate.2016.05.010>
- Xu, Y., & Brown, G. T. L. (2017). University English teacher assessment literacy: A survey-test report from China. *Papers in Language Testing and Assessment*, 6(1), 133–158.
- Xu, Y., & Liu, Y. (2009). Teacher assessment knowledge and practice: A narrative inquiry of a Chinese college EFL teacher's experience. *TESOL Quarterly*, 43(3), 492–513. <https://doi.org/10.1002/j.1545-7249.2009.tb00246.x>
- Xu, Y., Zhong, J., & Han, Y. (2026). How teacher feedback literacy changed by using generative artificial intelligence (GenAI) feedback as exemplars: Case studies of English as a foreign language school teachers in China. *Teaching and Teacher Education*, 176, 105501. <https://doi.org/10.1016/j.tate.2026.105501>
- Yao, Y., Zhu, X., Xiao, L., & Lu, Q. (2025). Secondary school English teachers' application of artificial intelligence-guided chatbot in the provision of feedback on student writing: An activity theory perspective. *Journal of Second Language Writing*, 67, 101179. <https://doi.org/10.1016/j.jslw.2025.101179>
- Zhu, S., Li, J., Yao, Y., Guan, Y., & Zhu, X. (2025). What they provide and how: An intervention study on pre-service teachers' GenAI-assisted writing feedback. *The Internet and Higher Education*, 67, 101040. <https://doi.org/10.1016/j.iheduc.2025.101040>
- Zou, S., Guo, K., Wang, J., & Liu, Y. (2025). Investigating students' uptake of teacher- and ChatGPT-generated feedback in EFL writing: A comparison study. *Computer Assisted Language Learning*. Advance online publication. <https://doi.org/10.1080/09588221.2024.2447279>