Educalingua
International Conference
FPBS UPGRIS

The AI Assessment Literacy Gap: Preparedness of Indonesian Pre-Service EFL Teachers for AI-Driven Language Assessment

I Putu Andre Suhardiana
Universitas Hindu Negeri I Gusti Bagus Sugriwa Denpasar, Indonesia, putuandresuhardiana@gmail.com

Luh De Liska
Universitas PGRI Mahadewa, Indonesia, luhdeliska86@gmail.com

Ni Putu Yuniarika Parwati
Universitas PGRI Mahadewa, Indonesia, parwatiyuniarika@gmail.com

Ni Putu Pande Mirah Surya Dewi
Universitas Lambung Mangkurat, Indonesia, mirah@ulm.ac.id

ABSTRACT

Integrating AI into language assessment requires EFL teachers to acquire/develop new competencies that go beyond the traditional Language Assessment Literacy (LAL). Nevertheless, studies on pre-service teachers' readiness in AI-driven assessment are still rare, especially in the Indonesian context. The current research examines the readiness of pre-service EFL teachers at Universitas Hindu Negeri I Gusti Bagus Sugriwa Denpasar to competently and ethically design, carry out, and assess language tests using AI. The research design was a mixed-methods sequential explanatory, with data collected through classroom questionnaire from 26 pre-service teachers and semi-structured interviews from 12 participants and 3 EFL lecturers. Quantitative data were analyzed with descriptive statistics, and qualitative data were subjected to thematic analysis. Results found participants held positive attitudes to AI-based assessment (M = 3.41), moderate ethical awareness (M = 3.34), low knowledge (M = 2.89) and very weak skills (M = 2.67). Qualitative findings revealed a total absence of experiential knowledge, with none of the participants interviewed having designed/implemented any AI-driven assessment. Academics confirmed that nothing systematic in the curriculum ensures students develop AI assessment literacy. Qualitative data revealed that there are three major gaps observed: practical skills, technical knowledge and ethical application. The study finds that all the constructs of self-efficacy in the

pre-service teachers provide a limited to moderate level of preparedness, and while positive attitudes and ethical awareness are associated with practice skills, they are insufficiently supported by these competencies. Recommendations of the Task Force include: integrating ethical training into the curriculum; investing in faculty development; and institutional investment.
Keywords: AI-driven assessment, assessment literacy, language assessment, pre-service EFL teachers, teacher education

Introduction

Artificial intelligence (AI) is radically transforming the nature of language teaching and learning through its use in educational assessment, presenting never-before-seen opportunities as well as daunting challenges for teachers. Language Assessment Literacy (LAL) is typically understood as the combination of knowledge, skills and principles required for designing, administrating and interpreting language assessments (Puspawati et al., 2024), making it a long-standing pillar for effective EFL teacher preparation. Coombe et al. (2020) noted that mainstream LAL frameworks have concentrated on familiar assessment issues such as validity, reliability and practicality, and the pedagogical capacity to design and evaluate conventional testing tools (Puspawati et al., 2024). The arrival of generative AI tools (ChatGPT, automated writing evaluation systems, and speech recognition systems) transform pedagogical practices and necessitates an urgency for much broader definitions of teacher proficiency (Létourneau et al., 2025). However, these technologies go beyond just providing ancillary services, they are heralding a new paradigm as they bring features like automated grading, immediate feedback generation and adaptivity in evaluation that offers greater efficiency and demonstrate novel assessment capabilities (Suhardiana et al., 2026). This technological evolution requires that modern-day language teachers develop a new skill set which goes way beyond traditional LAL, encompassing the practical how to use AI tools and being critically aware of their complexities and limitations (Esmaeilzadeh et al., 2025).

However, there is a research gap in this area especially for pre-service EFL teachers despite the increased role of AI in education. Although research related to the use of AI in language assessment has become more common, much of it focused on the efficacy (technical) and reactions from in-service teachers, raising further inquiry into the current understanding (and lack thereof) of future educators (Sidqi et al., 2025). A recent systematic review of AI-driven language tutoring systems shows that they can offer targeted interventions and measure the proficiency level of learners, but the problems of algorithmic bias, data privacy, and transparency remain largely unresolved issues in teacher preparation programs (Esmaeilzadeh et al., 2025). Preliminary evidence from the Indonesian context suggests that despite many pre-service teachers believing AI brings efficiency and objectivity, substantial uncertainty remains concerning reliability and self-efficacy for integrating AI tools into classroom assessment (Sidqi et al., 2025). An additional nationwide study on 200 future teachers from 60 Indonesian universities revealed that even though the participants had very high levels of traditional LAL, they were only moderately ready for AI-based assessment, which indicates a gap in this foundational assessment knowledge and what is needed for AI-driven practices (Ayuningtryas & Emaliana, 2025). Particularly with ethical considerations, prospective

teachers lack frameworks to analyze AI tools for fairness / cultural fit (Esmaeilzadeh et al., 2025), creating a gap between positive educator perceptions of AI tools and negative academic results. Until now no research investigates the habitat of Universitas Hindu Negeri I Gusti Bagus Sugriwa Denpasar, thus it is needed to explore the extent pre-service EFL teachers will be able to design, implement and batch AI-powered language assessments in a well-informed, substantive and ethical way.

Hence, this study explores the assessment literacy of pre-service EFL teachers for AI-driven language assessment by answering the question 'To what extent do pre-service EFL teachers have the assessment literacy to develop, conduct and evaluate AI-powered language evaluations successfully and ethically?' This study uses a sequential explanatory mixed-methods design, collecting quantitative data on self-perceived knowledge, skills and attitudes through an online survey then qualitatively exploring experiences and challenges through semi-structured interviews. Such study has a theoretical and practical importance. In theory, it would help build an AI-Supported LAL (AS-LAL) framework by specifying competencies anchored for assessment integrated with AI, which expands the reach of existing LAL approaches that so far concentrate on conventional testing. In practical terms, the results will give evidence-based data on pre-service teacher preparedness (in Indonesian) as well as recommendations for programs and policies that develop curriculum to fill gaps in readiness. This research will also provide policy recommendations on how higher education institutions can integrate AI assessment literacy into the curriculum in order to prepare future EFL teachers who are competent and responsible in using IA-Assisted language Assessment.

Methods

Design

This research was guided under the randomized quantitative phase followed by the qualitative exploratory interpretive sequential explanatory mixed-methods design. Phase one consisted of a quantitative study using a structured questionnaire to assess AI language assessment-related self-reported knowledge, skills and attitudes among the participants. Phase two consisted of qualitative data collection through semi-structured interviews was conducted with a purposively selected subset of the survey sample. The qualitative phase was intended to illuminate, interpret and contextualize the statistics derived from the survey, specifically for why pre-service teachers scored low (unsatisfactory) / high (satisfactory) preparedness scores, ethical issues that were raised by students in relation to their experiences of perceived/actual 'challenges' faced as a teacher.

Participants

The participants of this study consisted of pre-service EFL teachers who took a part in English Language Education Study Program at Universitas Hindu Negeri I Gusti Bagus Sugriwa Denpasar. In particular, the research was focused on third and fourth-year undergraduates as they were likely to have completed introductory classes in language testing, teaching approach and learning technology which made them more suited to reflect on their preparedness for AI-assisted assessment practices. However, in order to fill-up the context and gain some institutional perspectives on how AI assessment literacy was currently being addressed within or neglected from the teacher training program the

study also involved EFL lecturers who taught assessment related courses. Voluntary recruitment of all participants and informed consent prior to data collection. In the quantitative phase, 26 pre-service teachers were contacted to complete a survey and in the qualitative stage, a smaller sample of 10–15 participants was purposively chosen for detailed interviews to gather further qualitative data into their experiences, perceptions and perceived readiness towards AI-based language assessment.

Instruments

This study developed and used two main instruments. The first instrument was a questionnaire on pre-service teachers assessment literacy in AI. The questionnaire was based on LAL frameworks and modified to include attitudes, knowledge of available AI tools for assessment design, skills in using such tools with assessments for evaluation and ethical considerations about their use. Questionnaire was used in the study as five-point Likert scale ranging from 'Strongly Disagree' to 'Strongly Agree' and it was pilot-tested with few pre-service teachers who were not part of the main sample, for clarity, reliability and validity purpose. The second data-gathering tool was a semi-structured interview protocol, which included some open-ended questions related to participants' experiences regarding AI-driven assessment, obstacles faced, ethical challenges and recommendations for improving their training. The survey results informed the development of the interview protocol, providing opportunities to focus on themes identified in the survey and also explore divergent findings.

Data Collection

The data collection was conducted in two sequential phases. Thus, in the quantitative phase, my personal plan was to hand out the structured questionnaire to all of the 26 pre-service teachers during their regular class sessions at (with permission from the course instructors). The questionnaire took no longer than 20 to 30 minutes, and the researcher clarified questions on the items in the questionnaire as needed. In the qualitative phase, after completion of the quantitative analysis, 10 to 15 survey respondents were invited for individual semi-structured interviews by the researcher. We purposively sampled these participants based on their survey scores to obtain a variety of preparedness levels (low, moderate, and high) to represent multiple perspectives. Each participant is interviewed individually by researchers in a comfortable and private setting, lasting for around 30-45 minutes. After receiving informed consent, all interviews were audio-recorded and transcribed verbatim for further analysis. In addition, a triangulated perspective was incorporated via interviews with EFL lecturers to complement the student data and provide an overview of AI assessment literacy training conducted in the present time.

Data Analysis

The researchers analyzed the quantitative and qualitative data separately, then integrated both datasets. Quantitative data was analyzed through descriptive statistics (frequencies, means, standard deviations and percentages) to depict a general overview of self-reported AI-driven assessment literacy among the pre-service teachers for all items in the questionnaire. The data was then analysed using SPSS. Qualitative data was analysed using thematic analysis, according to the six-step framework recommended by Braun and Clarke (2006) on transcribed interview recordings. The analysis consisted of several

stages, becoming familiar with the data, generating initial codes, searching for themes, reviewing themes, defining and naming themes, and finalizing the report. The researchers conducted the qualitative analysis manually, with qualitative data analysis software supported coding and theme organization (NVivo). The last step was the integration of qualitative and quantitative findings during interpretation, with the qualitative themes used to explain, elaborate on, and challenge quantitative results. This integration helped the researcher make broader and deeper inferences regarding pre-service EFL teachers' preparedness for AI-based language assessment and provided a basis for valid recommendations on curriculum enhancement.

Findings and Discussion

Findings

Quantitative Findings

Profile of Participants

Twenty-six pre-service EFL teachers participated in the quantitative phase. Most were fourth years (n = 16, 61.5%) and others third-year students (n = 10, 38.5%). While 76.9% (n = 20) were female, there were 23.1% (n = 6) male participants. In terms of course completion, 22 participants (84.6%) had completed a language assessment course and 19 participants (73.1%) had completed an educational technology course. Self-reported familiarity of participants with AI tools, rated on a scale from 1 (Very Unfamiliar) to 5 (Very Familiar), had a mean score of 2.73 (SD = 0.83), which suggests moderate familiarity. The most indicated tool was ChatGPT (n = 21, 80.8%) followed by Grammarly (n = 18, 69.2%), and QuillBot (n = 12, 46.2%). Five participants (19.2%) reported some familiarity with automated scoring systems like E-rater and three participants (11.5%) reported no experience using AI tools at all.

Awareness of AI-Driven Language Assessment (Finding 1)

Table 1. Pre-service Teachers' Awareness of AI-Driven Language Assessment (Items K1-K7)

Item Code	Survey Item Description	Mean (M)	Standard Deviation (SD)	Participants Agreeing/Strongly Agreeing (n, %)
K1	I know what AI-driven language assessment is.	3.42	0.81	18 (69.2%)
K2	I know many AI tools for assessing language skills.	3.08	0.93	14 (53.8%)
K3	I know how AI tools assign scores or feedback.	2.73	0.96	9 (34.6%)
K4	I know the strengths of human vs AI assessment.	2.81	0.94	~13 (50%)*
K5	I understand the basic algorithms/mechanisms behind AI tools.	2.38	0.94	5 (19.2%)
K6	I am aware of recent trends in AI related to language assessment.	2.54	0.95	6 (23.1%)
K7	I know how to determine whether an AI assessment tool is suitable for a given purpose.	2.92	0.89	12 (46.2%)

Item Code	Survey Item Description	Mean (M)	Standard Deviation (SD)	Participants Agreeing/Strongly Agreeing (n, %)
Mean (K1-K7)	(Average of all items)	2.89	0.78	N/A

In terms of self-reported knowledge, the mean score for knowledge of AI potential language assessment (Section B items K1-K7) was 2.89 (SD =.78), which is below the midpoint on a five-point Likert scale designating moderate exposure to this technology. Thus, it indicates that participants neither strongly agreed nor disagreed with statements about knowing how to implement AI-driven assessment items, but skewed toward the lower end of this scale. Item-by-item analysis revealed considerable variation. The item with the highest mean score was K1 'I know what AI-driven language assessment is' (M = 3.42, SD = 0.81), where 18 participants (69.2%) either agreed or strongly agreed to this statement. That means most participants had a very basic conceptual understanding of the term. The item K2, *'I know many AI tools for assessing language skills'* (M = 3.08, SD = 0.93), scored relatively high, too with other fourteen participants (53.8%) agreed to this statement. These results match the demographic evidence that participants had some knowledge about a few AI tools, but such knowledge was not extensive. But participants had much less knowledge in some important areas. Item K3, 'I know how AI tools assign scores or feedback' (M = 2.73, SD = 0.96), with only 9 participants (34.6%) agreeing and item K4, 'I know the strengths of human vs AI assessment' (M = 2.81, SD = 0.94), with only about half responding positively to agree; in summary: Scores were lowest for item K5, 'I understand the basic algorithms / mechanisms behind AI tools' (M = 2.38, SD = 0.94), in which only 5 participants (19.2%) expressed agreement, followed by item K6, 'I am aware of recent trends in AI related to language assessment' (M = 2.54, SD = 0.95) with only a total of 6 participants agreeing (23.1%). Item K7 (as appropriate to the task) — 'I know how to determine whether an AI assessment tool is suitable for a given purpose', M = 2.92, SD = 0.89, with 12 participants (46.2%) agreeing. These results suggest that while pre-service teachers had a basic knowledge of what is involved in AI-based language assessment, and could name some tools relevant to the topic, their understanding of the foundational mechanisms underscoring these tools (and more) including algorithms, strengths and limitations, and state-of-the-art developments was still quite limited. The difference between K1 (conceptual understanding) versus K3, K5 (technical knowledge), indicates shallow rather than deep learning.

Skills in AI-based Assessment Design and Implementation (Finding 2)

Table 2. Pre-service Teachers' Skills in AI-Based Assessment Design and Implementation (Items S1-S8)

Item Code	Survey Item Description	Mean (M)	Standard Deviation (SD)	Participants Agreeing/Strongly Agreeing (n, %)
S1	I have a language assessment I can combine with AI tools.	2.50	0.91	8 (30.8%)
S2	I can choose the appropriate AI tools to evaluate particular language skills.	2.62	0.90	9 (34.6%)
S3	I am able to use AI tools to	2.54	0.90	7 (26.9%)

Item Code	Survey Item Description	Mean (M)	Standard Deviation (SD)	Participants Agreeing/Strongly Agreeing (n, %)
	develop appropriate activities for assessment tasks.			
S4	I have the necessary skills to successfully implement AI-based assessments in a classroom.	2.46	0.91	6 (23.1%)
S5	I can interpret and use results generated through AI to inform my teaching decisions.	2.85	0.88	12 (46.2%)
S6	I can give constructive feedback to students based on AI-generated assessment data.	2.96	0.91	13 (50.0%)
S7	I can merge AI-run assessments with traditional human assessments.	2.73	0.87	10 (38.5%)
S8	I can troubleshoot basic technical issues.	2.31	0.93	5 (19.2%)
Mean (S1-S8)	(Average of all items)	2.67	0.85	N/A

The mean score for skills in designing, implementing and evaluating AI-driven language assessments (Section C, items S1-S8) was 2.67 (SD = 0.85), which was lower than the knowledge section and indicates a moderate to low level of self-reported skills. This indicates that participants felt they had limited real capacities to implement AI tools for assessments. For the item-level analysis, respondents rated their skill level for areas more related to pedagogical application rather than technical implementation with the highest rankings. The highest mean score (M = 2.96, SD = 0.91) within Item S6, *'I can give constructive feedback to students based on AI-generated assessment data,'* was agreed with by 13 participants (50.0%). The statement in S5: *'I can interpret and use results generated through AI to inform my teaching decisions'* (M = 2.85, SD = 0.88); only agreed by 12 participants (46.2%). These results indicate, perhaps not surprisingly, that participants felt relatively more confident about the pedagogical usage of data generated by AI about than they did in designing / conducting an assessment utilizing AI. In contrast, respondents felt significantly less confident with skills relating to design and implementation. Item S1, *'I have a language assessment I can combine with AI tools'* (M = 2.50, SD = 0.91), was endorsed by just eight participants (30.8%). Item S2: *'I can choose the appropriate AI tools to evaluate particular language skills,'* M = 2.62, SD = 0.90 and 9 participants (34.6%) agree. The item S3, *'I am able to use AI tools to develop appropriate and appropriate activities for assessment tasks'* (M=2.54, SD= 0.90) had the least agreement of only 7 participants (26.9%). They scored lowest in practical implementation and technical troubleshooting. Item S4, *'I have the necessary skills to successfully implement AI-based assessments in a classroom'* (M = 2.46, SD = 0.91), only endorsed by 6 participants (23.1%). The lowest mean was for Item S8 (M = 2.31, SD = 0.93), wherein only 5 participants (19.2%) agreed with the statement: *'I can troubleshoot basic technical issues'*. Item S7 *'I can merge AI-run assessments with traditional human*

assessments' (M = 2.73, SD = 0.87) had agreement from ten participants (38.5%). This trend showed that participants were more confident in interpreting and applying AI-generated data than they were at designing, implementing and troubleshooting AI-based assessments. This implies that while participants have been somewhat comfortable in working with the output from AI assessments, they did not possess foundational skills to design and implement their own AI-enabled assessments independently.

Ethical Awareness Level in AI-Driven Assessment (Finding 3)

Table 3. Pre-service Teachers' Ethical Awareness in AI-Driven Assessment (Items E1-E7)

Item Code	Survey Item Description	Mean (M)	Standard Deviation (SD)	Participants Agreeing/Strongly Agreeing (n, %)
E1	I know about ethical issues related to using AI in assessment.	3.54	0.86	20 (76.9%)
E2	I know that it is essential to make sure my work does not favour one group over another.	3.62	0.85	21 (80.8%)
E3	I know how to safeguard student data privacy.	3.38	0.94	17 (65.4%)
E4	I can assess whether a given AI tool is culturally and linguistically appropriate.	3.08	0.93	13 (50.0%)
E5	I appreciate that it is needed to share information with students openly.	3.42	0.86	18 (69.2%)
E6	I am aware of the implications of becoming over-reliant on AI to make high-stakes decisions.	3.27	0.96	15 (57.7%)
E7	I know how to guard against academic misconduct when students use AI tools.	3.15	0.92	14 (53.8%)
Mean (E1-E7)	(Average of all items)	3.34	0.82	N/A

The mean score for ethical awareness (Section D, items E1-E7) was 3.34 (SD: 0.82), indicating moderate to high ethical awareness - higher than knowledge and skills sections overall. This indicates that the participants were more aware of ethical implications than actual expertise or ability to apply in real life AI-powered assessment. Item-level analysis showed that participants were most aware of ethics and data privacy items. Item E2, *'I know that it is essential to make sure my work does not favour one group over another',* (M = 3.62, SD = 0.85) had a mean score higher than all but one other questionnaire items, with 21 participants (80.8%) agreeing of disagreeing with this statement positively. MIS E1. *'I know about ethical issues related to using AI in assessment'* (M = 3.54, SD = 0.86) and be agree by 20 participants (76.9%). Moving on to the last measure (Item E3, *'I know how to safeguard student data privacy'*), 17 out of 26 participants (65.4%) agreed with this item sufficiently (M = 3.38, SD = 0.94). Participants also displayed moderate levels of awareness with respect to transparency and over-reliance concerns. Item E5*, 'I appreciate that it is needed to share information with students openly'* (M = 3.42, SD = 0.86), agreement, n=18 (69.2%). 15 (57.7%) agreed with the statement in Item E6, *'I am aware of the implications of becoming over-reliant on AI to make high-stakes decisions'*

(M = 3.27, SD = 0.96). In contrast, awareness was slightly lower in areas that required critical reflection and application. A total of 13 participants (50.0%) agreed on item E4: *'I can assess whether a given AI tool is culturally and linguistically appropriate'* (M = 3.08, SD = 0.93). For item E7, *'I know how to guard against academic misconduct when students use AI tools'* (M = 3.15, SD = 0.92), there were 14 participants (53.8%) who agreed. These results afford an insight, participants were generally aware of ethical notions and valued fairness, bias, privacy and transparency. However, the participants reported not having quite as much confidence in applying those principles in practice, as far as assessing tools for cultural appropriateness and assessing academic integrity.

Attitude Towards AI-Driven Language Assessment (Finding 4)

Table 4. Pre-service Teachers' Attitudes Towards AI-Driven Language Assessment (Items A1-A8)

Item Code	Survey Item Description	Mean (M)	Standard Deviation (SD)	Participants Agreeing/Strongly Agreeing (n, %)
A1	I have confidence in my ability to use AI tools for assessment.	2.77	0.91	9 (34.6%)
A2	I would like to learn more about AI assessment.	4.12	0.77	23 (88.5%)
A3	I think AI tools can improve the quality of assessment.	3.65	0.85	19 (73.1%)
A4	I am open to incorporating AI tools in my future assessment practices.	3.85	0.83	20 (76.9%)
A5	I feel that my teacher training program has prepared me well.	2.50	0.95	7 (26.9%)
A6	I think AI will help reduce the workload for teachers without impacting quality.	3.54	0.86	18 (69.2%)*
A7	AI could afford more objective and consistent scoring.	3.38	0.90	16 (61.5%)
A8	I have concerns about AI-driven assessment.	3.62	0.85	20 (76.9%)
Mean (Positive Attitude Items A1-A7)	(Average of items A1-A7)	3.41	0.79	N/A

The mean score of the positive attitude items (Section E, A1-A7) was 3.41 (SD = 0.79), and thus indicates that respondents harboured generally positive attitudes (highest mean over all sections). This shows that participants had a more positive attitude than knowledge / skills. At the item level, participants showed highest willingness to learn and interest. The item received the highest score across the entire questionnaire (M = 4.12, SD

= 0.77), 23 out of the participants (88.5%) indicated agreement with Item A2 *'I would like to learn more about AI assessment'*. Item A4, *'I am open to incorporating AI tools in my future assessment practices'* (M = 3.85 SD = 0.83), received agreement from 20 respondents (76.9%). That these findings show high levels of motivation on the part of pre-service teacher to develop competence in this area. There were also positive (overlapping) beliefs about the benefits they expect benevolent AI to confer. Item A3, I think AI tools can improve the quality of assessment (M = 3.65, SD = 0.85) saw agreement from 19 participants (73.1%). Nineteen agreed with item A6, *'I think AI will help reduce the workload for teachers without impacting quality',* (M = 3.54, SD = 0.86) corresponding to 69% of participants agreeing. Item A7, *'AI could afford more objective and consistent scoring'* (M = 3.38, SD = 0.90) finds support with 16 respondents (61.5%). But the participants were much less sure of their own capabilities or those of their training program. For item A1, *'I have confidence in my ability to use AI tools for assessment'* (M = 2.77 SD = 0.91), only 9/26 participants (34.6%) agreed. The positive attitude item A5, *'I feel that my teacher training program has prepared me well'* (M = 2.50, SD = 0.95) had the lowest mean score of all with only 7 participants (26·9%) agreeing. 982 Item A8, which measures concerns around AI received a mean score of 3.62 (SD = 0.85), with 20 participants (76.9%) indicating agreement that they had concerns. The combination of positive attitudes with a high level of attention and low confidence implies ambivalence to AI driven assessment amongst pre-service teachers.

Qualitative Findings

The qualitative part consists of semi-structured interviews with 12 pre-service teachers (8 fourth year, 4 third year) and 3 EFL instructors. Participants were intentionally selected to provide representation of low (n = 4), moderate (n = 4), and high (n = 4) scores on the survey. Ten interviews were completed, which allowed five themes to emerge from the thematic analysis based on the structure of the interview protocol.

Experiences of and Attitudes Towards AI Assessment (Finding 5)

Participants showed some ability to convey, at least an understanding of the material but perfect for depth. Participants who scored well were able to provide a much richer account, while those who did poorly clearly had no idea what was going on. Participant P07 (high score) articulated: *'AI-driven language assessment is seeking to artificially intelligence and automating measuring students how the nature of providing scores. ChatGPT can be used to assess writing, and there are apps for assessing speaking'*. Alternatively, for the low score, Participant P03 said:

'I'm not quite sure what it means to be honest. I have heard of educational AI, but do not see how it is used to assess a person. Or maybe it is as if you are using laptops to analyze your sentences and rate them?' Participants were only familiar with ChatGPT, Grammarly and QuillBot (not introduced to language assessment tools via comprehension-expressiveness effects). Participant P11 (high score) explained: *'I use ChatGPT a quite bit, especially as a writing assistant I've also used Grammarly. However, I am not sure if these are technically 'assessment' tools since they function as writing assistant applications. I don't know of a lot of tools that are designed specifically for assessment.'* Several quotes suggested that their contact only came through using it as students (as opposed to the formal course). Participant P06 (moderate score) shared: *'I discovered using ChatGPT and QuillBot through the introductions of friends and my*

own use' At least we were not taught specifically about AI tools in any of our classes, especially related to assessment'.

Respondents noted that the strengths here were one of efficiency and consistency. Participant P09 (high score) elaborated: *'I believe AI can assist in evaluating fast and large numbers of students. AI can provide instant feedback, which human teachers are unable to do. Besides, AI may be less biased because it uses uniform criteria for every student'.* But participants also identified limitations, especially when it comes to how well AI can measure creativity and things that require cultural context and meaning. Participant P04 (moderate score) expressed: *'AI likely does not get the more nuanced nature of what students are trying to say, especially in writing. Sure, it can be a site used to check grammar and vocabulary but could it really identify when a student is being creative and conveying something more intricate?'* Also Participant P12 (high score) reported: *'AI might have biases. Depending on what the data is trained on, it could be biased to students of different English varieties and backgrounds'.* The lack of any practical experience using AI-driven language assessment was one of the most imposing points raised. None of the 12 participants had designed / implemented an assessment using AI tools Participant P02 (low score) stated: *'No, never. But I have never used AI to design an assessment. To be honest, I would not even know where to start. We have never trained for using AI to assess, we learnt about traditional assessment methods such as written tests and rubrics'.* Participant P05 (moderate score) explained: *'Grammarly works by giving my writing feedback, but as a student, not as a teacher. As a teacher, I had no idea how Grammarly would assess my students and what to make of the output'.* Participant P08 (high score) noted: *'In my mind, I could use AI for feedback on writing. However, I have no idea how to configure it in practice. What platform would I use? How would I know that the scores are true? I have no idea.'*

None of the three lecturers reported that AI-driven assessment was systematic in their curriculum. Lecturer L02 explained: *'So I know that this is an important situation, but unfortunately we do not have a particular course / module at the moment for AI-based language assessment. Or, fine, we write about technology in education and this is mostly general ICT skills. With AI only appearing in the briefest of mentions'.* Lecturer L03 added: *'We talk about technology and assessment, just not from the angle of students using technology to cheat. We have not developed content on how teachers may use AI to design or conduct assessments'.* Lecturer L01 noted: *'I see that our students are well aware of ChatGPT in general but apply it to their learning and homework, rather than writing. They don't really consider using these tools for the purpose of assessment. Mostly, they seem interested but confused when I have asked them about AI in assessment'.* Lecturer (L02) discussed institutional barriers as well: *'We have to ramp up infrastructure and the training. And most of our lecturers are not experts in AI assessment tools themselves so we can't deliver what we don't know. Faculty first. We need professional development soon'.*

Ethical Concerns and Challenges (Finding 6)

The ethical issue that stood out the most was data privacy. Participant P01 (low score) expressed: *'I'm really worried about the privacy of our students. If we have AI tools testing students, everything they do / who they are may be somewhere in a database. But what if they utilize the data for non-salvageable purposes? What if there*

is a breach in the data?' Participant P09 (high score) elaborated: *'When we utilize online services we are making our students' data available often without understanding the terms of service. As educators, we need to be aware of where our trust lies. But really, I suck at estimating these privacy policies'*. The stakes here are high, fairness and bias were hot topics of discussion. Participant P10 (moderate score) stated: *'An AI may be biased if it is trained on non-representative data. The dataset might not include the entity of all students. For instance if it is trained on American English, it will mistake Aspiring Indonesian Students' English as wrong when in their context its perfectly fine. We are not trained enough to check for this'*. Participant P07 (high score) added: *'I am concerned about students who may not have access to technology / who are not tech-savvy. Will AI Assessment labor more Inequality? Those are thoughts I ponder but do not hold answers to'*. There was concern about individuals relied on AI too heavily and whether AI dehumanized assessment. Participant P12 (high score) noted: *'It is not only assessment that means giving a score. It relates to knowing students and relationships, therefore helping them grow. If we are reliant with AI, we risk to lose that human contact which is the clear value of being a teacher'*. Participant P08 (high score), who had similar sentiments stated: *'AI has so much potential, but can it really empathise with a student's struggle / be genuinely happy for their progress? I believe that teachers have a role in which no AI can replace them. However, I fear schools will try to use AI instead of teachers in order to cut costs'*. Participants with limited infrastructure within their schools struggled to access AI tool and lacked high-speed internet availability. Participant P06 (moderate score) explained: *'Even if I wanted to use AI tools, there would be no opportunity in my future school. Most Indonesian schools do not even have a decent internet connection to access AI platforms, let alone use them. If the technology is not available, then your question may be 'How can I use AI for assessment?'* Lecturer L03 further elaborated: *'We face infrastructure challenges. The students need technology and reliable internet. Initially, they will serve as the basis for training on AI-led assessment; without them we cannot do that'*.

Attitudes, Readiness and Recommendations (Finding 7)

Even with a confidence rating of between only 5 to 6, they had great interest in knowing more about AI-driven assessment. Participant P04 (moderate score) stated: *'I am really looking forward to the possibilities it could bring. I could make a better teacher with AI because it could offer me more feedback for students on specific behaviors. I want to know how to utilize these tools correctly'*. Participant P02 (low score) added: *'I know not a great deal about AI assessment, but that I am very eager to learn. I believe this is the future. We need more training on this'*. In accordance with the quantitative findings, interviews revealed that participants had little confidence in their preparedness. Participant P03 (low score) expressed: *'I have not trained in assessment with AI and I have no idea where to begin. I have no clue what all these tools are going to look like, and when I become a teacher, how am I supposed to use these?'* Participant P06 (moderate score) stated: *'I guess I am somewhat prepared as I am quite tech-savvy but when it comes to AI assessment, I lack any experience'*. Respondents provided suggestions for enhancing their training. Participant P09 (high score) suggested: *'We need workshops where we can actually practice using AI tools to assess knowledge. We need experience. So we would be able to practice designing assessments and*

interpreting results'. Participant P11 (high score) added: *'We should be learning to question AI tools. There are a multitude of tools out there and we need to recognize what tools are appropriate for what context and for which students'*. Participant P05 (moderate score) proposed: *'There should be a dedicated technology and assessment course that really emphasizes AI. Currently, it is spread out in multiple courses. It would be better if it was a dedicated course'*. Participant P07 (high score) emphasized: *'The training on ethics is required in the following manner. How to be fair, how not to invade privacy, how to be transparent with students? These are essential, and we have not written about them at all'*. Lecturers recommended how improvements could be made to the curriculum. Lecturer L01 stated: *'We need not one course, but a curriculum with assessment literacy in AI. Students are exposed to these ideas in assessment courses, educational technology courses, and teaching methodology courses'*. Lecturer L02 suggested: *'We have to generate guidelines for the ethical use of AI-based assessment. These standards should inform our institutional policies as well as be shared amongst all pre-service teachers'*. Lecturer L03 emphasized: *'Before we can train anyone else, we have to prepare the trainers. Without faculty who understand and know AI-assessment, we cannot prepare our students effectively'*.

Combining quantitative and qualitative findings

Preparedness (Finding 8)

Quantitative and qualitative findings mostly corroborated a picture of generally poor preparedness. Interview participant descriptions of having a basic conceptual understanding of AI-driven assessment, with little knowledge about the underlying mechanisms or current developments, confirmed the moderate knowledge scores (M = 2.89). By contrast, the low skills scores (M = 2.67) were robustly validated by the complete absence of any practical experience reported across all 12 interviews, all of whom had neither designed nor implemented an AI-driven assessment. Interview participants elaborated on concerns regarding privacy, bias, fairness, and transparency supporting higher ethical awareness scores (M = 3.34). Quantitative and qualitative data showed a disconnect between knowledge and use, with participants reporting the limited capacity to assess tools for their cultural appropriateness or utilize ethical safeguards. The interview participants expressed a positive attitude (M = 3.41) and strong interest in learning (M = 4.12) with their enthusiasm to get involved in using AI assessment, despite perceptions of low confidence. The qualitative data shed light on the finding that participants had higher scores in attitudes, but lower scores in skills. Although interview responses made plain that subjects did not have a detailed understanding of the assessment capabilities of AI, learning instead just from the general technological optimism and an abstract acknowledgment about some efficiencies AI could bring, nevertheless the directions registered were positive. On the contrary, low performance skill scores were attributed to a lack of any experience and overall awareness about AI assessment tools built into their training program. The low K5 and K6 scores in relation to algorithms and technical mechanisms were attributed to their comments claiming that they had not been given education as to how AI tools work, or how they might be evaluated. Participants also discussed the gap between their awareness of principles versus practice, noting that while they recognised ethical considerations there seemed to have been no training in actual implementation; for example deciding if and how to

examine tools for possible bias or approaches to communicating with students around AI use. By their work, three gaps in preparation emerged. Most obviously, participants fared worst when it came to practical skills, it became apparent that they had virtually no experience designing or implementing AI-based assessments. Second, the knowledge gap of technical background was large by the fact that participants did not know how scores were produced by AI tools and quality assessment. Third, the application gap demonstrated that while participants identified ethical principles, they required practical knowledge of how to implement these principles, particularly in relation to evaluating tools (feasibility) and cultural responsiveness. However, the perspectives of lecturers reinforced gaps established by pre-service teachers. The absence of practical experience is clarified by Lecturer L02 who observes that *'we don't have a dedicated course and module on AI language assessment'*. Lecturer L01 described the superficial understanding as: *'students know ChatGPT in general, but they do not think about applying these tools to our assessment'*.

Discussion

A finding of this study is a disparity between positive attitudes about AI-driven assessment by pre-service teachers and their readiness to practice with it. Participants proved to be highly interested in AI-based assessment (M = 4.12), willing to use AI tools (M = 3.85) and perceived that using AI-powered tools could provide benefits but had very low knowledge of AI (M = 2.89) and skills (M = 2.67). This pattern is consistent with the research of Sidqi et al (2025), who found that pre-service teachers in Indonesian contexts recognized the effectiveness of AI, they remained uncertain about their own readiness. Likewise, Ayuningtyas and Emaliana (2025) found out that participants across 60 universities in Indonesia had high traditional LAL, but their AI readiness only sufficiently medium-high-based on the fact how essential it is to analyze have a strong understanding of foundational assessment competency compared with the nature of practice involving AI. Ironically, such separation between attitudes and ability is characterized as a *'training paradox'* throughout the literature that manifests in experiences with actual support (Burriss & Leander, 2024). Positive attitudes are a good start, but they alone do not ensure effective integration of AI; what you really need is a systematic process of developing competencies through experiential learning.

The qualitative results really highlight why the skills scores for participants were so low across all dimensions, as none of the 12 pre-service teachers interviewed had ever created / done an AI-based assessment, so much creative potential left untapped. This lack of embodied experience exemplifies what Abrahams (2026) refers to as a call for *'AI Pedagogical Judgement'*, the ability to trade off generative efficacy with curriculum alignment, ethical reflexivity and professional agency when judging whether or not to endorse, adapt or reject AI outputs. Judgement of this nature is something which requires practical experience, as opposed to theoretical exposure. This gap explains the systemic nature of this course that is buried deep down in all meetings from these lectures confirming that *'we are not running a dedicated module (even a module) on AI powered language assessment'*. This result is in agreement with the study of Puspawati et al (2024) have also observed that although LAL is increasingly needed, teacher education programmes cover alternative assessment practices poorly. In the Iranian context, Arefian (2025) also found that teachers' underdeveloped LAL was compounded by

insufficient professional development resources reporting solutions as needing AI-human hybrid feedback models whereby AI adds to human judgment rather than replaces it. The results of the present study indicate that Indonesian teacher education programs need to move beyond traditional LAL frameworks and more systematically incorporate AI-specific competencies especially via workshops where pre-service teachers can explore various aspects of AI through authentic assessment design tasks.

Another finding is the disparity between ethical awareness and ethical action. In awareness, participants were relatively aware of ethical principles (M = 3.34), with an acknowledgement of fairness (M = 3.62) and privacy (M = 3.38). Despite this, their confidence in applying these principles to practice was significantly lower especially with respect to assessing AI tools for cultural and linguistic appropriateness (M = 3.08). Thus, it is worrisome that these ethical principles are sidelined in AI assessment literature since they are not peripheral to the validity of education assessment but rather core. Research conducted on the Indian higher education context indicate that blindly implementing AI is fraught with serious risks to valid assessment, including bias towards different English dialects; further entrenching the digital divide and de-skilling of educators (Biju et al., 2024). The teacher education literature consistently points out a similar gap, where the teachers can identify these issues, but are unable to do anything about it. Esmaeilzadeh et al. (2025) found similar results where while AI tools are promising, issues like bias, data privacy, and algorithm transparency were insufficiently addressed in their teacher preparation programs. The results of these studies indicate that ethical training needs to go beyond recognition of principles and include practical methods (like appraising the AI tools, talking with students about AI use, building safeguards against algorithmic bias).

What is even more striking, however, is the participants low feelings of personal competence (M = 2.77) and desire to be qualified through their training program that had not sufficiently trained them (M = 2.50). Those low confidence levels, paired with relatively high concern regarding AI use (M = 3.62), match what (Burriss & Leander, 2024) refer to as *'conscious resistance,'* a stage characterized by both a high degree of technical comfort and low levels of pedagogical trust in AI. This resistance stems from a realization that one does not have the skills to harness AI responsibly and well. As Burriss & Leander (2024) state, acceptance is determined by *'familiarity'* with tools and items of technology: fear and anxiety stemming from inadequate exposure often fuel resistance to novelty in thinking, whereas further habituation enables understanding or demystification resulting in conscious integration. These findings from the current study provide strong evidence to support this assertion, as first I found that participants reporting more familiarity with AI (scoring ≥3) consistently outperformed those reporting less familiarity for each outcome measure. It was hypothesized that repeated, structured, supervised experience with the AI tools will improve both competence and confidence. The demographic analysis suggested some trends. Fourth-year students consistently had higher scores than third-year students on the taken tests, and those who performed / completed-all language assessment / educational technology courses scored higher than those who did not. That indicates that a little bit of existing coursework contributes to AI evaluation literacy, although the differences were small. However, much larger differences (M = 2.82 vs. M = 2.31) were found in the skills section between those who completed technology courses compared to controls, suggesting that coursework focused on technology is especially important with respect to preparedness for practical

competencies. Nevertheless, even the top scores were well below moderate; curricula do not seem to fully prepare students with the depth of competency needed for appropriate assessment and moral action with AI. A challenge for an institution is the acknowledgement by the lecturers themselves that even faculties at colleges are not familiar with AI assessment tools, and as such, teacher educators cannot teach what they do not know. This determination is consistent with work from Mohammadi (2024), who argues for professional development programs focused on enhancing the capacities of faculty in AI assessment literacy before expectations are placed upon them to prepare pre-service teachers.

The results have implications for how teacher AI literacy is conceptualized. Traditional LAL frameworks are made available with great references to Puspawati et al. (2024), which emphasize knowledge of assessment principles, test design and the interpretation of results. Specifically, this study argues that AI language assessment needs a broader definition of literacy than (Burriss & Leander, 2024) refer to as *'Critical AI Literacy,'* a multidimensional construct with three variates, the conceptual understanding of algorithmic logic, critical agency (to evaluate ethical aspects including bias and data privacy), and pedagogical governance (to differentiate when a task should be absorbed by AI). The wider literacy framework is echoed in the work of (Stockwell (2026), who argue for social, ethical and AI literacy dimensions that go beyond digital literacy surrounding computer-assisted language learning. The findings from this study call for the assumption of such a multi-perspective framework in teacher education where training programs should be more than technical training, but that they also teach critical evaluation of data sources, ethical reasoning and pedagogy. The persistently short list of institutional constraints in the lecturers' identification points to several overarching issues such as limited infrastructure, access to AI tools, and insufficient internet connectivity in Indonesian higher education. This finding echoes a study conducted by Ayuningtyas and Emaliana (2025) that despite the ongoing Digital Transformation Initiative, asymmetric digital competency and access among pre-service cohorts remains prevalent. For instance, in the context of Indonesia, it has been found that perceived usefulness and ease of use significantly influenced creating positive attitudes and intentions to adopt AI but were negatively associated with perceived risk (Helmiatin et al., 2024). Participants in this present study had similar concerns, especially about the issues around data privacy, fairness and more latterly trust or using AI instead of rather than as well as human judgment. To reform the curriculum, these institutional barriers need to be addressed because pre-service teachers are unable to develop competence with tools they have never heard of. The faculty development must come first according to the lecturers in this study, without knowledgeable and confident faculty, preservice teacher preparation cannot happen meaningfully. The study also highlights a tension between what participants see as the promise of AI in terms of increasing efficiency and their desire for dehumanization and, thus, loss of teacher agency. This tension captures well what Burriss & Leander (2024) which has described as *'technocratic drift'* in which formative judgment becomes computational automatism. The worries expressed by participants have been supported in prior research investigating hybrid models based on teacher-AI feedback (Arefian, 2025), stressing that AI should support rather than replace the human assessments made by teachers, furthermore, such models to enable sustainable LAL development must not undermine teachers' control over their decision-making. The

suggestions from participants for more experiential workshops, ethical training, and dedicated AI in assessment courses are consistent with early best practice models emerging in the domain that require a practical, experience-based component to develop both competence and judgement (Abrahams, 2026; Stockwell, 2026).

Conclusion

The purpose of this study was to find out how well-prepared the pre-service EFL teachers at Universitas Hindu Negeri I Gusti Bagus Sugriwa Denpasar are in terms of assessment literacy for AI-driven language assessments. Using explanatory sequential mixed-methods research consisting of 26 survey respondents and qualitative interview data through 12 interviews with pre-service teachers and 3 EFL lecturers, this study answers the research question succinctly, readiness to teach as future EFL trainees is limited-moderate amongst these preservice EFL teachers, having a gap between positive attitudes and perceived competency. The results show that the participants possessed relatively most positive attitudes towards AI-based assessment (M = 3.41), high learning interest (M = 4.12) and intention to use AI tools (AI tool integration M = 3.85). Ethical awareness was also fairly high (M = 3.34) with the principles of fairness (M = 3.62) and privacy recognized (low M=3.38). But knowledge scores were only moderate (M = 2.89) and skills scores were lowest (M = 2.67). Gaps were largest regarding technical knowledge of how AI tools produce scores (M = 2.73), skills for practical implementation (M = 2.46), and troubleshooting skills (M = 2.31). Importantly the qualitative findings showed that none of the 12 people interviewed had designed or used an AI based assessment at any point providing a plausible reason for their lower skills scores and supporting low levels of practical experience. The lecturers stated that AI assessment is not systematically embedded into the curriculum, and they also find even educators do not use them for assessment. There were three gaps in preparedness. Most significant was the gap in practical skills, as participants had limited experience designing, implementing, and debugging AI-supported assessments. Huge gaps in technical knowledge existed, as many participants were unable to explain the process by which AI tools score respondents or assess their appropriateness for a given task. The gap of knowledge application around ethics showed that participants recognized ethical principles, however there was practical under-educatedness in how to implement them especially with respect tool evaluation for cultural appropriateness and also bias. Institutional barriers, such as basic infrastructure, internet access, and access to tools for creating and using AI, further widen these gaps and are not addressed without integration of AI assessment material into the curriculum. This study has theoretical implications for the reconceptualization of Language Assessment Literacy. Traditional LAL frameworks are focused on knowledge of assessment principles and test design. The present study shows that AI-driven assessment needs a broader conception, which includes technical knowledge about the logic of algorithms, practical skills for the design and implementation of AI-integrated assessments, critical evaluation of ethical implications, and pedagogical judgment to strategically decide when to delegate tasks to AI. Practically, the findings show the need for curriculum change to incorporate workshops, ethical training, and courses on AI in assessment. This kind of transformation requires faculty professional development and institutional investment in infrastructure.

References

Abrahams, W. (2026). Exploring Pre-Service Teachers' AI Pedagogical Judgement in Assessment Design: A Document Analysis of Technology-Integrated Tasks. *International Journal of Learning, Teaching and Educational Research*, *25*(6), 616-635.

Arefian, Mohammad. (2025). Supporting Iranian Teachers' Language Assessment Literacy: A Hybrid AI–Human Feedback Approach within Collaborative Action Research. 10.21203/rs.3.rs-8281675/v1.

Ayuningtryas, V., & Emaliana, I. (2025). Are Pre-Service EFL Teachers Ready for AI-Assisted Assessment? The Role of Assessment Literacy in the Digital Era. *PANYONARA: Journal of English Education*, *7*(2), 267-294.

Biju, N., Abdelrasheed, N. S. G., Bakiyeva, K., Prasad, K. D. V., & Jember, B. (2024). Which one? AI-assisted language assessment or paper format: An exploration of the impacts on foreign language anxiety, learning attitudes, motivation, and writing performance. Language Testing in Asia, 14(1), 45.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, *3*(2), 77-101.

Burriss, S. K., & Leander, K. (2024). Critical posthumanist literacy: Building theory for reading, writing, and living ethically with everyday artificial intelligence. *Reading Research Quarterly*, *59*(4), 560-569.

Esmaeilzadeh, F., Ghahari, S., & Rohani, G. (2025). Artificial intelligence in language teaching and assessment: A systematic review of research (2020-2024). *The Asian Journal of Applied Linguistics*, *9*(1), 1222-1222.

Helmiatin, Hidayat, A., & Kahar, M. R. (2024). Investigating the adoption of AI in higher education: a study of public universities in Indonesia. *Cogent Education*, *11*(1), 2380175.

Létourneau, A., Deslandes Martineau, M., Charland, P., Karran, J. A., Boasen, J., & Léger, P. M. (2025). A systematic review of AI-driven intelligent tutoring systems (ITS) in K-12 education. *npj Science of Learning*, *10*(1), 29.

Mohammadi, M. (2024). Language teachers' assessment literacy in AI-aided adaptive learning environments. *Journal of Research in Applied Linguistics*, *15*(2), 73-88.

Puspawati, I., Khansa, M., & Widiati, U. (2024). Developing EFL Teachers' Language Assessment Literacy: A Systematic Literature Review on Teacher Training Programs. *TESL-EJ*, *28*(2), n2.

Sidqi, M. F., Matra, S., & Ulya, I. (2025). Pre-Service Language Teachers' Perceptions of AI-Driven Language Assessment: A Preliminary Investigation. *IDEAS: Journal on English Language Teaching and Learning, Linguistics and Literature*, *13*(1), 1892-1903.

Stockwell, G. (2026). Legal, Affective, and Policy Dimensions of the AI Ecosystem in Language Education. *International Journal of Computer-Assisted Language Learning and Teaching (IJCALLT)*, *16*(1), 1-14.

Suhardiana, I. P. A., Mahendra, M. W., Juliari, I. G. A. I. T., Phalaguna, I. B. G., & Wijayanti, A. (2026). Attitudes and readiness for Global Englishes in language assessment among Indonesian university English lecturers. *Metathesis: Journal of English Language, Literature, and Teaching*, *10*(1), 29-48.

Coombe, C., Vafadar, H., & Mohebbi, H. (2020). Language assessment literacy: What do we need to learn, unlearn, and relearn?. Language Testing in Asia, 10(1), 3.