

Intended Versus Demonstrated Language Assessment Literacy: A Mixed-Method Study of Indonesian Pre-Service English Teachers

I Putu Andre Suhardiana

English Language Education Study Program

Universitas Hindu Negeri I Gusti Bagus Sugriwa Denpasar, Indonesia

Email: putuandresuhardiana@gmail.com

Abstract

Language Assessment Literacy (LAL) is one of core competencies for English language teachers, although there is a gap between what pre-service teachers learn in assessment courses and their performances during teaching practicum period. This mixed-methods convergent parallel study investigated the effectiveness of the Language Assessment course at Universitas Hindu Negeri I Gusti Bagus Sugriwa Denpasar by assessing the gap between planned learning objectives and implemented LAL. Participants were 36 final year pre-service teachers and 7 university supervisors. Data were obtained using an LAL Self-Report Questionnaire, an Observation Checklist completed by supervisors, and document analysis of course syllabi and 108 assessment artifacts (summative tests, formative quizzes, scoring rubrics). A one sample t-test was used to compare the mean demonstrated LAL to the benchmark planned outcome score of 4.2. Qualitative data was investigated using thematic analysis. A t-test found a statistically significant difference ($M = 2.8$, $t(35) = -14.23$, $p < .001$, Cohen's $d = 2.91$) with just 8.3% of pre-service teachers meeting the required standard. Pre-service teachers over-estimated their competency (self-report $M = 3.9$ compared to observed $M = 2.7$). The qualitative results showed five themes: general flaws in rubric design (75%), lack of statistical analysis (88.9%), generic unactionable feedback (69.4%), over-reliance on selected-response formats (80.6%), and total absence of digital assessment tools (100%). The Language Assessment course is not meeting its objectives. There is a need for immediate reforms that focus on practice-based curriculum, training on metacognitive self-assessment and integration of digital tools.

Keywords: *Assessment Course, Intended Versus Demonstrated, Language Assessment Literacy, Learning Outcomes, Pre-Service Teachers*

INTRODUCTION

Language Assessment Literacy (LAL) is an essential qualification for 21st-century English language teachers. It contains the technical skills of test design and grading, as well as the knowledge of principles, purposes, and ethical practices regarding how assessment should be used in any way to enhance student learning (Tsagari & Vogt,

2017). Recent work on LAL conceptualizes it as an integrative skillset to build, manage, and interpret language assessments associated with effective learning and teaching (Tsagari & Vogt, 2017). A large-scale study conducted in the context of Indonesian higher education involving 297 tertiary EFL teachers found that numerous educators encounter challenges related to assessment practice; diverging from knowledge-based approaches towards competence-oriented ones, specifically with respect to designing authentic assessments aligned with learning objectives and interpreting assessment data for instructional decision (Rasyid et al., 2024). Similarly, LAL levels were still generally moderate among pre-service English teachers in Indonesia but differed depending on the university, where pre-service teachers from state universities and those affiliated with the Ministry of Education had higher assessment literacy than their fellow students at private and Ministry of Religious Affairs-affiliated universities (Tsagari & Vogt, 2017) found a lower initial level of actionable assessment literacy when studying only Indonesian EFL student teachers. Nuraeni and Heryatun (2021) conducted research into the experiences of pre-service EFL teachers on the teaching practicum in Indonesian senior high schools, discovering that practical application gaps, conceptions of assessment influenced by previous experiences and institutional norms, mentor teacher influence, classroom realities such as resource constraints and large class sizes interact intricately to produce a complex portrait of assessment decision-making during field experiences. Case study research has also been conducted to explore the progression of LAL, showing that theory-practice embedded professional development programs incorporating collaborative learning, reflection and feedback can lead teachers to experience growth in their conceptions about assessment, but not their knowledge-attitude-skills-confidence as assessors (Fitriyah et al., 2022). These results highlight that Language Assessment Literacy is a set of knowledge to be taught in a course and a competence that is influenced by the institutional context, experiential learning opportunities, mentorships and professional development.

Even though LAL is essential and Language Assessment courses are commonly included in Indonesian ELE programs, a gap between the knowledge learned from course readings and what pre-service teachers can show in their teaching practicum continues to persist (Tsagari & Vogt, 2017). Studies on Indonesian pre-service EFL teachers have shown that although these teachers probably own adequate theoretical knowledge about the nature of assessment concepts such as validity, reliability and

fairness, they keep failing to implement those characteristics in the classroom when designing authentic tests determining scoring rubrics or providing formative feedback (Rasyid et al., 2024). In a qualitative case study, many pre-service teachers dressed their lack of preparation for assessment issues as if they were unprepared for timely test administration under time constraints, accommodating students with different learning preferences or abilities in class, and interpreting data from assessments to make adjustments to instruction (Nuraeni & Heryatun, 2021). The study also found that university supervisors tended to focus on general teaching performance and not specific assessment practices, indicating a missed opportunity for formative feedback in regards to targeted professional development during the practicum (Nuraeni & Heryatun, 2021). Studies on the correspondence between course materials and practicum demands have indicated that numerous ELE programs focus on summative assessment but not so much on classroom-based, formative assessment approaches and assessment tools (Fitriyah et al., 2022). It was found that the responsibility of university supervisors was particularly underutilised, because existing supervisory frameworks have focused on assessing teaching performance rather than targeting pre-service teachers' assessment literacy (Oo et al., 2021). However, when supervisors do give feedback on assessment, it is more often of a general nature rather than detailed and actionable types of guidance that would enable pre-service teachers to improve certain skills (Oo et al., 2021). Because assessment course effectiveness has not been systematically evaluated, curriculum designers have little empirical evidence available to them regarding the course content or instructional strategies that would best prepare pre-service teachers for classroom assessment (Tsagari & Vogt, 2017). It has made it possible for assessment courses to be able to remain in place without being empirically informed that they are impractical and continue through repeat survey after survey coupled whereby pre-service teachers frequently report feeling ill-prepared for unaffiliated with the emphasis on achieving well at assessments.

The current study directly fills this gap by assessing how closely the intended learning outcomes of a Language Assessment course are matched by the exhibited Language Assessment Literacy of pre-service teachers while on teaching practicum at Universitas Hindu Negeri I Gusti Bagus Sugriwa Denpasar. The research question that directed the study was the following: To what degree does the intended learning outcomes of the Language Assessment course differ from the demonstrated Language Assessment Literacy of pre-service teachers during their teaching practicum? A mixed-

methods convergent parallel design is employed in this investigation. Methodological studies in Asian contexts have highlighted the importance of multimethod approaches, such as those that combine self-reported perceptions with direct observation and artifact analysis, given that each approach has its own limitations by relying on a single method alone (Rasyid et al., 2024). Self-report questionnaires tend to be biased towards higher scores because social desirability bias, which means observation cannot capture the quality of assessment artifacts and artifact analysis did not provide insight on how assessments were administered. This study uses a triangulated approach by weaving together self-reports from pre-service teachers, observational checklists filled out by university supervisors, and document analysis of the course syllabus and assessment artifacts to provide a closer picture of the intended-demonstrated LAL gap. The one-sample t-test will establish if a statistically significant difference exists between the mean demonstration LAL score and mean benchmark intended outcome score, while thematic analysis will show which dimensions of LAL pose the most significant challenge for preservice teachers. Using both quantitative and qualitative findings will reveal if a gap exists but the nature of that gap, its size, and other potential contributing factors to help substantiate curriculum change as well as guidelines for improving professional preparedness of English teachers on graduating.

METHOD

Design

This study utilizes a mixed-methods convergent parallel design to quantify and qualitatively describe the gap between intended learning outcomes and confirmed LAL shown in practice. The quantitative element determines how the statistical presence of course objectives contrasts to what is seen, and the qualitative component looks at whether identified gaps are contextual (the what) / not (the why). This design is suitable because the research question aims to measure if a difference (quantitative) and also explore how and why that difference occur in authentic classroom practicum (qualitative). Both strands are collected simultaneously, and analyzed separately, and merged during interpretation to form an answer of the research question.

Participants

The participants are two groups that are chosen purposively from the English Language Education Study Program at Universitas Hindu Negeri I Gusti Bagus Sugriwa Denpasar. The first group consists of 36 final-year pre-service teachers from semesters 7 and 8 who have successfully completed the compulsory Language Assessment course and have either completed / are presently completing their teaching practicum. The second group consists of all university supervisors who observed these pre-service teachers directly during the practicum (7 supervisors). To be included, a pre-service teacher needed to have completed the Language Assessment course (with a grade of at least C) so that they were exposed to the basic content and supervisors must have conducted two formal observations of each assigned pre-service teacher assessment practices.

Instruments

Three instruments are created to acquire complimentary data. Firstly, the pre-service teachers are given an LAL Self-Report Questionnaire with 25 items covering five dimensions: (1) building assessment instruments, (2) conducting tests, (3) scoring and grading, (4) interpreting and using assessment data, and (5) offering feedback. Items were rated on a 5-point Likert scale (1 = strongly disagree to 5 = strongly agree). The questionnaire is based on well-established LAL frameworks (Fulcher, 2012; Vogt & Tsagari, 2014) and confirmed by the expert judgment of two evaluation scholars. Second, university supervisors complete an Observation Checklist during practicum observations. This checklist aligns with the five characteristics of the questionnaire, but includes observable behavioral indications. Third, a Document Analysis Protocol was used to analyze two types of artifacts: (a) the syllabus and statements of learning outcomes of the Language Assessment course (to establish intended outcomes) and (b) assessment materials created by pre-service teachers during practicum (tests, quizzes, rubrics, grade sheets).

Data Collection

Data collection occurs over the course of four weeks, matched with the end of their teaching practicum. Week 1: The researcher obtains institutional ethical clearance and informed consent from all individuals taking part in the research. All data, the language Assessment Course Syllabus and Official Learning Outcome Documents comes from the study program coordinator. Week 2: LAL Self-Report Questionnaire administered

to pre-service teachers in a supervised classroom setting (30 min). Both, supervisors are trained for using Observation Checklist on a workshop of 60 minutes to ensure inter-rater reliability. Week 3: Fill out one Observation Checklist for each pre-service teacher is based on their practicum observation records completed by the supervisor. Supervisors average their ratings across observations (if multiple did representatives) Week 4: In this week, the pre-service teachers give digital copies of their assessment artifacts (minimum three artifacts per participant). These are anonymized with player codes. All data is saved in password-protected files so no one can access or store relevant information.

Data Analysis

Data analysis is conducted in three parallel steps as specified by the mixed-methods methodology. The research question is answered by quantitative analysis. A one-sample t-test is used to compare the mean of proven LAL score (from Observation Checklist and Artifact Quality evaluations) to the benchmark intended outcome score taken from the course syllabus (translated into similar Likert-scale benchmarks using expert mapping). A one-sample t-test is performed to assess whether the mean exhibited LAL score (calculated from Observation Checklist and Artifact Quality ratings) significantly differs from the benchmark intended outcome score taken from the course syllabus. The test value is the benchmark score. To evaluate the practical relevance of each difference, an effect size (Cohen's d) is computed. The effect size (Cohen's d) is calculated and interpreted in terms of practical significance for each difference. Descriptive statistics (means, standard deviations, frequency distributions) also summarize responses to questionnaires. Qualitative analysis includes evaluation artifacts and open-ended comments by supervisors. The Document Analysis Protocol is used to conduct thematic analysis of artifacts to discover patterns of strength / weakness (frequent absence of validity evidence, misuse of scoring criteria). Through open coding, topics are generated and grouped into bigger categories that match with the five dimensions of LAL through axial coding. During the analysis phase, interpretation is achieved through a side-by-side comparison of quantitative results (t-test findings) with qualitative themes (artifact patterns and supervisor comments) to explain whether differences exist, the magnitude of differences, and the specific assessment competencies that are most problematic. Any convergence / divergence

between the two types of data is explicitly examined to increase the validity of conclusions on the efficiency of the Language Assessment course to prepare pre-service teachers for professional readiness.

RESULTS AND DISCUSSION

RESULTS

The Intended-Demonstrated Gap (Finding 1)

Lastly, the one-sample t-test was performed to find out if the mean indicated LAL score of pre-service teachers differed significantly from a benchmark intended outcome score taken from the syllabus of Language Assessment course. The researcher mapped the eight intended learning outcomes of the course syllabus against five LAL dimensions using the Benchmark Conversion Guide which yields a benchmark target outcome score. Three assessment scholars examined the syllabus independently, and agreed that it appears to train pre-service teachers to reach an overall 4.2/5 average on each of the five LAL dimensions, meaning *'competent application with minimal guidance'*. That goal is reflected in the benchmark for this course: Successful graduates of this course will be able to independently and capably design, administer, score, interpret and provide rationale regarding language assessments. Mean observed LAL was calculated by averaging 2 sources of performance: (a) Observation Checklist at the item level, university supervisors completed this during visits to classrooms for each pre-service teacher on 25 items; and (b) Artifact Quality ratings from document analysis of three assessment artifacts from each pre-service teacher. A detail for each of the 36 pre-service teachers, an individual LAL score was calculated as an average of their supervisor observation rating (1-5) and artifact quality (1-5). Those scores were then averaged to calculate the mean group demonstrated LAL score. The average LAL score for the 36 pre-service teachers was 2.8 (SD = 0.6). The individual minimum score was 1.7, and the maximum score for an individual participant was 3.9, which indicates a large variation among pre-service teachers. Still, 78% of scores were clustered between 2.3 and 3.3 across the *'developing'* range defined in the Quality Rating Guide as scoring with *'major flaws'* that *'would affect quality.'*

This mean score was compared to the target outcome benchmark-used as a test value of 4.2-in a one sample t-test: LAL Score, $M = 2.8$ (SD = 0.6). The results indicated a significance difference, $t(35) = -14.23$, $p < .001$. The confidence interval of 95% for the difference between each was -1.58 to, with a population mean showing that LAL is,

at most, between 1.22 and 1.58 points below what is intended in the benchmark. The effect size was exceedingly large (Cohen's $d = 2.91$), far greater than traditional benchmarks for a large effect ($d = 0.8$). These effect size demonstrate that the gap between intended and demonstrated LAL is not only statistically significant but practically enormous, almost three standard deviations apart. The gap is clear: the intended learning outcomes (benchmark 4.2) were not met, since the mean LAL score of 2.8 lay within the '*developing*' band, where most pre-service teachers demonstrated assessment competencies only at a foundational level requiring significant guidance. Only three of the 36 pre-service teachers (8.3%) scored a demonstrated LAL score at / above the benchmark value of 4.2 with course expectations not being met for the other 33 pre-service teachers (91.7%). The significance with which this finding relates to our research question is clear: between what the Language Assessment course intended to teach and what pre-service teachers actually showed in their teaching practicum there exists a large statistically significant and practically meaningful difference.

The Perceived LAL of Pre-Service Teachers (Finding 2)

A summary of descriptive statistics from the LAL Self-Report Questionnaire (Instrument 1) offers an overview of how pre-service teachers view their own assessment abilities. The 25-item questionnaire employed a 5-point Likert scale (1 = Strongly Disagree to 5 = Strongly Agree) and was completed by all 36 pre-service teachers. Cronbach's $\alpha=0.89$ for the questionnaire across all 5 dimensions demonstrated good internal consistency of items. The mean LAL score for all 25 items together was 3.9 (SD =.5), suggesting that pre-service teachers perceive themselves to be between '*Slightly Agree*' and '*Agree*' (closer to '*Agree*') on a scale of assessment literacy ranging from '*disagree*,' through '*neutral*,' up toward slight, moderate, and strong agreement. Scores ranged from 1.8 to 5.0, with a median score of 4.0 and a mode of 4.2 indicating that most pre-service teachers felt proficient in assessment-related tasks. As for the overall scores distribution, 22 pre-service teachers (61.1%) scored at or above 4.0 on average (classified as high perceived LAL), 12 pre-service teachers (33.3%) scored between 3.0 and 3.9 on average (classified as moderate perceived LAL), and only two pre-service teachers (5.6%) scored below 3.0 mean score in their lay assessment of learning from others, low perceived LAL). Importantly, zero pre-service teacher rated themselves below 2.5 and nobody selected '*Strongly*

Disagree on each item which suggests there might be a tendency towards over-rating their self-competence across the whole sample.

The average score of the dimension of *'Dimension 1: Designing Assessment Tools (5 items)'* was 3.8 (SD = 0.6). On analysis of individual items within the survey, pre-service teachers reported feeling most confident in constructing a valid multiple-choice test for reading comprehension (M = 4.2, SD = 0.5), with 30 participants (83.3%) agreeing or strongly agreeing to this item. However, they felt least confident in *'designing a writing prompt that elicits the target writing construct'* (M = 3.4, SD = 0.7), with 14 participants (38.9%) selecting *'Neutral'* / *'Disagree.'* The task of *'creating authentic assessment tasks such as portfolios or projects'* had a mean (SD = 0.8), indicating moderate confidence with more variation than other items, with an average of 3.6 itself. The dimension of *'Dimension on Test administration, 5 items'* had the highest mean score (4.1) of all five dimensions with SD = 0.4. Compared to other items, pre-service teachers felt most confident regarding their item about *'providing appropriate instructions to students before the test administration'* (M = 4.3; SD = 0.4), with 33 out of 36 pre-service teachers (91.7%) responding that they agree / strongly agree to this item on the survey. Likewise, *'managing test timing and pacing adequately'* (M = 4.2, SD = 0.5) and *'maintaining standardized testing conditions across classes'* (M = 4.0, SD = 0.5). On this dimension: the least endorsed item was *'dealing with cheating or academic dishonesty during assessments'* (M = 3.8, SD = 0.7) where 8 participants (22.2%) neutral / disagreed. The average score for the dimension of *'Second Dimension : Scoring and Grading (5 items)'* was 3.9 (SD = 0.5). The highest confidence among pre-service teachers was reported for *'scoring objective items accurately'* (M = 4.2, SD = 0.4), with 31 participants (86.1%) agreeing / strongly agreeing. Their lowest confidence was for, *'identifying and minimizing bias in my scoring'* (M = 3.6, SD = 0.7) with neutral / disagree selected by 12 participants (33.3%). Mean (SD) confidence in applying a scoring rubric consistently across student responses was 3.9 (SD = 0.6).

The dimension of *'The fourth dimension: Interpreting and Using Assessment Results (5 items)'* had the lowest mean score (M = 3.4, SD = 0.7) of all five dimensions. Pre-service teachers were even less confident with the *'interpretation of item difficulty and discrimination indices'* (M = 2.9, SD = 0.8), where only five participants (13.9%) chose *'Agree'*, while 22 participants (61.1%) selected *'Disagree'* or *'Strongly Disagree'*. *'Computing basic statistics (mean, median, and standard deviation)'* had a mean of 3.1 (SD = 0.8) with 17 participants (47.2%) showing low confidence. In contrast, the mean for

'using the results of assessment to identify students who are having difficulty learning' was higher (3.7, SD = 0.6) as well as for adjusting my teaching in response to assessment data (mean: 3.6, SD = 0.6), with these two means indicating that pre-service teachers feel somewhat able to use assessment results more generally but do not have the requisite statistical and psychometric knowledge necessary for implementation / interpretation. The score for the dimension of *'Dimension 5: Giving feedback (Items = 5)'* was 4.0 (SD = 0.5). The two areas in which pre-service teachers felt most confident were providing timely feedback on student performance (M = 4.2, SD = 0.5) and balancing appropriate positive feedback with corrective feedback (M = 4.1, SD = 0.5). They reported lower confidence in *'providing written and oral feedback well'* (M = 3.9, SD = 0.6) and even lower in *'encouraging students to use feedback for self-reflection'* (M = 3.8, SD = 0.6). The item giving very constructive feedback that help students improve had a mean of 4.00 (SD = .5), demonstrating high self-efficacy. Pre-service teachers consider themselves competent assessors with overall mean scores in the *'adequate to strong'* range across all dimensions except Interpreting and Using Assessment Results, which they acknowledge as a relative weakness. Administering Tests has the highest perceived competency while Interpreting and Using Assessment Results ranked lowest. Importantly, pre-service teachers did not see themselves as having great weaknesses in any dimension; all mean dimensions (including the lowest at 3.4) were greater than the neutral midpoint of 3.0.

Quantitative Observations of LAL by University Supervisors (Finding 3)

Descriptive statistics of Observation Checklist (Instrument 2) give an external assessment of the LAL displayed by pre-service teachers in the teaching practicum in terms of behaviors. Seven university supervisors completed one Observation Checklist for each of the 36 pre-service teachers they supervised, thus, 36 checklists were completed. The checklists included 25 items across five dimensions rated on a 5-point scale from 1 = Not Observed/Needs Extensive Improvement to 5 = Fully Observed/Excellent. Inter-rater reliability was tested at the supervisor training workshop with an acceptable intraclass value of 0.81. The mean overall observed LAL score for all 36 pre-service teachers was 2.7 (SD = 0.6). This is much lower than the self-reported perceived LAL mean of 3.9, a 1.2-point difference between what pre-service teachers believe they are capable of doing and what supervisors witness them doing.

The median observed score was 2.6, with a range of 1.8 to 4.0. Only 4 pre-service teachers (11.1%) achieved an overall observed score of 4.0 or higher (categorized as 'good' to 'excellent'), while 22 pre-service teachers (61.1%) scored between 2.0 and 2.9 (categorized as 'needs improvement' to 'satisfactory'), and 10 pre-service teachers (27.8%) scored below 2.0 (categorized as 'needs extensive improvement').

The average score on Dimension 1 (Designing Assessment Instruments, 5 items) was 2.5 (SD = 0.7). The lowest scored item was 'designs scoring rubrics with clear, appropriate criteria' (M = 2.2, SD = 0.8), with 25 pre-service teachers (69.4%) ranking 2 / lower. Most supervisors reported that pre-service teachers either did not utilize rubrics at all or used poorly constructed rubrics with ambiguous descriptors. The highest rated item on this category was 'constructs valid multiple-choice tests aligned with learning objectives' (M = 2.9, SD = 0.7), however even this item was rated below the satisfactory level. 'Creates authentic assessment tasks' had a mean of 2.3 (SD = 0.8) indicating that performance-based evaluations were seldom

noticed. The component of Dimension 2 (Test Administration, 5 items) had the highest mean observed score (3.2, SD = 0.5) and is consistent with the self-report finding that pre-service teachers regard this as their strength. The highest rated item was 'gives clear, understandable test instructions to students' (M = 3.5, SD = 0.6). 3 / higher evaluations were received by 28 pre-service teachers (77.8%). 'Manages test timing and transitions smoothly' had a mean of 3.3 (SD = 0.5) and 'maintains consistent testing conditions' had a mean of 3.1 (SD = 0.6). The lowest-rated item in this dimension was 'accommodates students with special needs during testing' (M = 2.8, SD = 0.7), with 18 pre-service teachers (50.0%) obtaining a rating of 2 / lower, which suggests differentiation for diverse learners remains a struggle.

The mean score for the Dimension 3 (Scoring and Grading, 5 items) was 2.8 (SD = 0.6). The item earning the highest rating was 'scores objective items accurately without errors' (M = 3.2, SD = 0.6) with 20 pre-service teachers (55.6%) receiving a grade of 3 / above. The lowest rated item was 'applies rubrics consistently across different student responses' (M = 2.4, SD = 0.8), with 27 pre-service teachers (75.0%) obtaining ratings of 2 / below. The supervisors reported that those pre-service instructors who had developed rubrics often had difficulty using them consistently and scored according to their overall impression rather than rubric criteria. The mean for 'Distinguishes clearly between different performance levels' was 2.6 (SD = 0.7) which suggests a problem in distinguishing between student work beyond a basic pass/fail

evaluation. The lowest observed mean score across all five dimensions was seen in the Dimension 4 (Interpretation and Use of Assessment Results, 5 items) at 2.2 (SD = 0.8). The lowest rated item was 'identifies problematic items using difficulty and discrimination indices' (M=1.7, SD=0.9) with 32 pre-service teachers (88.9%) scored 1 / 2. Supervisors reported that most pre-service teachers did not conduct any type of item analysis on their examinations. 'Calculates and analyzes basic test statistics' had a mean of 1.9 (SD = 0.8) with 29 pre-service teachers (80.6%) rated 2 / below. 'Adjusts lesson plans based on assessment data' had a slightly higher mean (2.4, SD = 0.7) but is still in the 'needs improvement' area. Supervisors reported that pre-service teachers would casually note who scored badly, but they did not use assessment data to diagnose learning issues or design remedial training.

The average observed score for the Dimension 5 (Feedback, 5 items) was 2.9 (SD = 0.6). The highest-rated item was 'returns graded work to students within a reasonable time' (M = 3.4, SD = 0.6) with 25 pre-service teachers (69.4%) scoring it as 3 / higher. The lowest rated item was 'provides specific, actionable comments for improvement' (M = 2.4, SD = 0.8), with 26 pre-service teachers (72.2%) obtaining ratings of 2 or lower. Supervisors observed that pre-service teachers were consistent in returning work on time, but that comments was generally generic (*'good job,' 'needs work'*). 'Encourages students to reflect on and use feedback' had a mean of 2.6 (SD = 0.7) indicating little observed practice of feedback literacy practices. A one-way ANOVA indicated considerable variation in observed LAL scores among the 36 pre-service teachers, $F(35, 34) = 4.32, p < .001$. The highest overall observed scores was 4.0, by participant 12. This pre-service teacher scored relatively high in Administering Tests (4.5) and Providing Feedback (4.2). This pre-service teacher was *'super organized, with a detailed rubric for every assignment, and written comments that directly referenced rubric criteria,'* said their supervisor. In comparison, the pre-service teacher with the lowest overall observed score (Participant 31) obtained ratings of 1 or 2 in all areas, with an overall observed score of 1.8. According to their supervisor, *'This pre-service teacher did not develop any original assessments, just used the pre-made tests from the textbook, and did not provide any written feedback to students other than a letter grade.'* This wide difference implies that although the course was designed to generate uniformly competent graduates, the individual's motivation, prior knowledge and quality of mentoring during practicum affect the actual LAL displayed.

Thematic Groups for Supervision Comments and Artifacts Qualitative (Finding 4)

Thematic analysis of 108 artifacts (three artifacts per pre-service teacher: summative test, formative quiz, and scoring rubric) and 36 sets of open-ended comments from seven supervisors generated five themes. The data analysis was conducted according to the six-phase thematic analysis approach developed by Braun and Clarke, beginning with open coding of all artifacts and comments, followed by axial coding that organizes codes into themes, and then selective coding (particular attention is given to core patterns during final analysis). Inter-coder reliability was established by having a second researcher independently code 20% of the data, resulting in a Cohen's kappa of 0.84.

Limitations of the Rubrics Design (Discussion 1)

A content study of the 36 scoring rubrics submitted by pre-service teachers showed that nearly all of them, 27/36 (75.0%) had serious design inadequacies related to assuring their validity and reliability in grading student work. The most prevalent device used to characterize will be vague, non-specific performance descriptors, which were utilized in twenty-two of the rubrics (61.1%). Instead of articulating what they see / standards that can be measured, rubrics use adjectives such as 'exemplary,' 'proficient,' and 'need improvement' without any practical description of the meaning underlying these words. One pre-service teacher's rubric for a writing activity defined 4 points as excellent material, 3 points as good content, 2 points as fair content, and 1 point as poor content but did not distinguish between superb and other categories of content. One wrote with this criteria: *'This is not an effective grading rubric as two different teachers could read a student essay and one would give it 'great' and the other gives it 'good. No anchor, no example, no precise requirements.'* The second most common problem was misalignment between the rubric requirements and the stated learning objectives. This defect was found in 18 rubrics (50.0%). Here the learning objective at the top of the page say 'students will write a persuasive paragraph with a claim, evidence and reasoning' but the rubric criteria spell out grammar and spelling and length of paragraph leaving any form of persuasiveness out in the cold. One of the supervisors said, *'A pre-service teacher spent 20 minutes presenting a rubric to the class; however, that rubric does not measure what the lesson was intended to teach. The students were confused. I was confused.'* Besides, there is no rubric (no designated

proportion allotted for efforts in certain areas). This was the third most common defect present in 14 (38.9%) rubrics. These pre-service teachers (n=3) submitted a test / quiz without scoring guide and in their artifact documentation stated that they graded student work 'at my own professional judgment' / 'via a holistic rubric.' Supervisors opposed this technique harshly. *'When I inquired how in her view Student A got a B and Student B a C, the pre-service teacher could provide me no criterion.'* *'I know a B when I can see one,'* she answered, as one supervisor put it. *'This isn't assessment.'* In terms of clear criteria, explicit performance descriptors and consistency with learning objectives, only 9 rubrics (25.0%) were rated as 'satisfactory' / better. These rubrics are generally always prepared by the pre-service teacher when the pre-service teacher has been given more assistance from the supervising teacher / written independently with the support of online resources.

No Statistical and Psychometric Evaluation of the Data (Theme 2)

The document examination of the 36 summative assessments revealed that test scores had not been subjected to any statistical / psychometric analysis. Pre-service teachers were requested to provide the test and any written proof of their analysis of those results (item difficulty, item discrimination, reliability estimates. Of the 36 submissions, a total of 4 (11.1%) featured any item analysis at all, none were statistically assessed. Often the submission was just a stack of tests, with results, and an overview that comprised of the average score on top. Supervisor comments indicated this as the most important of the present gaps in pre-service teachers' assessed LAL. One supervisor wrote, *'the pre-service teacher did aggregate the class average but then didn't know what to do with that. I asked her what was easy or hard and she just looked at me. Although these were directly relevant to her syllabus, she had never learnt about item difficulty and discrimination indices.'* Another supervisor noted: *'I asked three preservice teachers to show how they would recognize a low performing test question. None of them could manage it. She said she would just 'eye-ball it, see whether it looks confused.'* There is not one systematic analysis. Some analysis was done, but the few students that completed it were only partially right. Another pre-service teacher calculated the percentage of students that answered each item correctly, but did not know how to interpret a difficulty index other than *'some items were easier, and some harder'*. Another group did point-biserial correlations but interpreted their negative

correlations as a good item instead of what a negative correlation really signifies (that high-ability students got the item wrong). Indeed, one supervisor noted, *'The pre-service teacher attempted item analysis but did not understand the significance of the numbers. One of these things had an index of .95 and was classed as 'too simple' and removed, but she did not grasp that an item with a difficulty of .95 could actually be just suitable for a test designed to measure mastery. No understanding into how to contextualize item stats.'*

Feedback is not actionable (Theme 3)

Analysis of the 36 formative quizzes and the feedback given to students on these quizzes showed that while 33 artifacts (91.7%) had some type of input, only 11 (30.6%) provided feedback that was precise, actionable, and likely to increase student learning. The other 22 artifacts (61.1%) did not have meaningful feedback (merely a score / grade), general feedback ('good job,' 'nice work,' 'needs improvement'), and commentary that was not actionable ('be more careful,' 'study more,' 'try harder'). There were a lot of examples of generic feedback. He wrote 'Good!' on 15 student quizzes in a row, with no other comment. Another scribbled 'Nice try' on every question, no matter how well they did. Third had a stamp that said 'Excellent work' and he stamped every paper, even the ones with less than 50%. This practice was criticized in supervisor comments. One supervisor stated, *'The pre-service teacher returned quizzes to students with only a letter grade put on top. When I asked one student what they needed to do to improve, the student answered, 'Get a better grade. The student had no notion which questions they got wrong / why. The feedback was unhelpful for learning.'* On the other hand, actionable input was rare, but when it came, it was potent. *'Your topic sentence is clear, but your supporting evidence needs to be more specific.'* were among the specific remarks in the 11 artifacts that provided high quality feedback. Instead of saying *'many experts agree,'* name one expert and give his / her specific words. Another example: *'You're on the right track here but check your verb tense. In paragraph 2 you changed from past tense to present tense. 'Consistency is the key.'* One pre-service teacher was lauded by a supervisor: *'She provided detailed comments on each student's paper, with specific suggestions for revision. Students genuinely read her comments and asked questions for follow-up. That is what formative assessment should be.'* Pre-service teachers typically had trouble identifying which faults to correct even when they tried to give positive feedback, according to supervisors. One supervisor commented:

'The pre-service teacher circled every single spelling error on a student's paper, when the assignment was about organizing ideas logically. The student got a paper coated in crimson ink and was absolutely overwhelmed.' The pre-service teacher did not realize that not all errors are equally important and that feedback needs to be tied to the learning objectives of the assignment.

Overreliance on Selected Response Formats (Theme 4)

Analysis of all 108 assessment artifacts showed a substantial preference for selected-response assessment formats over authentic, performance-based, or alternative assessments. Of the 36 summative assessments, 29 (80.6%) were composed mainly / solely of multiple-choice items. Of the 36 formative quizzes, 24 (66.7%) were multiple-choice / True/False. Only 7 summative tests (19.4%) and 12 formative quizzes (33.3%) contained any constructed response elements such as short answer, fill-in-the-blank, or essay questions. Authentic assessments constituted only 4 artefacts (11.1%) which include two portfolio assignments, one project based evaluation and one oral presentation rubric. The lack of authentic assessment was especially evident because the course syllabus stressed the significance of *'designing authentic assessment tasks that reflect real language use.'* Comments from supervisors were disappointed at this gap. One supervisor wrote: *'The pre-service teacher taught a unit on giving directions and asking for help, skills that are clearly performance-based. But her final assessment was a 20-item multiple-choice test in which students had to choose the correct preposition from four alternatives. She never once required students to really provide or follow orders vocally. She tested knowledge of language.'* Another supervisor stated: *'There was no indication in any of the portfolios, projects, peer evaluation, self-assessment or any alternative assessment method in any of the pre-service teachers I supervised. All usual testing. When I offered a simple self-assessment checklist, one pre-service teacher responded, 'But how would I put a grade in the grade book from that? She had no idea that evaluation might serve other formative functions that are not just providing a grade.'* Some pre-service teachers did include authentic aspects in their assessment. One pre-service teacher produced a project where students made a tourism brochure for an imaginary trip, including writing, design and oral presentation abilities. Another developed a peer assessment exercise in which students traded draft paragraphs for peer feedback using a standardized checklist. But these cases were rare.

'The one pre-service teacher who used authentic assessment stood out dramatically from the others,' one supervisor said. Her students were interested, they understood why they were being assessed and they created work that genuinely showed their learning. But she was the exception, not the rule. 'The rest picked multiple choice because it was simpler to grade.'

Technology and Digital Tools Missing Completely (Theme 5)

The course program specifically stated that pre-service teachers would *'develop competency in the use of digital tools for language assessment.'* However, the study of the artifacts of assessment did not surface any evidence of the use of technology in any of the 108 artifacts. All assessments were paper-based and scored manually, with no digital assessment platforms, computer-adaptive examinations, online quiz tools, or automated feedback systems. This finding was shown by all 36 pre-service teachers. The opinions of the supervisors proved the total absence of technology in the processes of evaluation. One supervisor commented, *'The pre-service teacher used paper-based tests only, despite the fact that the school had a computer lab and students had tablets. 'When I asked her why she didn't use any digital tools she told me she didn't know how and was afraid to try out during practicum.'* Another supervisor said: *'I asked if any pre-service teacher had used online quiz platforms such as Google Forms, Kahoot or Quizizz. None had. Several claimed they have heard of these tools, but didn't know how to utilize them for true graded tests, simply for review activities.'* Another pre-service teacher expressed interest in using technology but had no abilities to use it. Their supervisor noted, *'She wanted to create an online listening test with embedded audio, but she did not know how to embed audio files into a digital document / how to make sure that all students can access the audio. She turned to a listening test on paper, reading the audio script out loud twice. 'This completely transformed the construct being assessed, because students could see her mouth movements and facial emotions, which is not how genuine listening occurs.'* The lack of utilization of technology is particularly troubling. The training was designed to educate pre-service teachers for digital assessment. In actuality, this planned effect was not accomplished at all. Supervisors identified this gap as a disconnect between course content and practicum realities. Although the course have introduced digital technologies conceptually, pre-service teachers did not have opportunities to practice and receive technical support to use the tools during their field experience.

DISCUSSION

This study shows a discrepancy between the purported outcomes of the Language Assessment course and pre-service English teachers LAL showed during their teaching practicum. The numerical analysis provides evidence of a statistically significant difference with an average exhibited LAL score ($M = 2.8$) considerably falling under the target threshold of 4.2 ($t(35) = -14.23$, $p < .001$, Cohen's $d = 2.91$). This finding aligns with a growing body of literature in the Indonesian EFL context reporting moderate levels of less than ideal ranges of LAL among pre-service teachers while theoretical understanding gained through courses studied at university had little correlation to actual classroom practices (Tsagari & Vogt, 2017). These results are consistent with studies conducted in other contexts showing that many pre-service teachers feel poorly prepared for the complexities of classroom assessment and resort to summative assessments (Putri et al., 2026). A large part of this gap strengthens the need to assess teacher education on core aspects of curricula, especially for institutions under the Ministry of Religious Affairs where lower levels of assessment literacy were found (Tsagari & Vogt, 2017).

While a mean rating of 3.9 indicates that pre-service teachers believe they possess LAL, supervisors rated them as demonstrating LAL with a mean of only 2.7. This shows the risk of participants failing to accurately assess themselves developmentally (Borg & Edmett, 2019). The tendency to overestimate competence presented primarily a more inflated rating in the Test Administration and Feedback dimensions which imply that pre-service teachers do not truly grasp how challenging it is to engage in practices of assessment (Putri et al., 2026). They expressed confidence around providing feedback in a timely manner ($M = 4.2$) and managing timing during the test ($M = 4.2$). However, when observed by supervisors, the story changed. Pre-service teachers were found to return graded work quickly but give generic, unactionable feedbacks, a pattern consistent with *'limited and uneven understandings of LAL components'* revealed in earlier qualitative studies (Putri et al., 2026). This helps to explain research showing that being taught assessment theory in isolation increase self-confidence but not performance (Allen & Wright, 2014), reinforcing the importance of ongoing context-specific professional development in teacher education programs. Qualitative data about the nature of supervisor comments on assessment artifacts augments the

quantitative analysis explains the areas in LAL where pre-service teachers struggle most. The biggest area of concern was in the interpretation and applications of test results, with only 11.1% of pre-service teachers doing any form of item analysis / basic statistics for their assessments. The finding aligns with studies that found Indonesian pre-service teachers' psychometric knowledge was insufficient to interpret test results and make instructional decisions (Tsayari & Vogt, 2017). The lack of any statistical analysis indicates that pre-service teachers regard assessment as largely a grading tool rather than also as a diagnostic assessment tool for informing instructional improvement, consistent with the finding that many beginning teachers initially seem only to think about assessment in terms of evaluating students (Allen & Wright, 2014). The fact that such indices are rarely studied represents an oversight as analyses of item difficulty and discrimination are needed to the validity and fairness of test results.

The qualitative themes also uncover a divide between what the course and codes professed in terms of 'realistic', performance-based assessment, and what that meant for in-the-field practices by pre-service teachers. These striking figures, alongside the dominance of selected-response formats (80.6% of summative tests), and an absence from digital assessment tools contradict both the curricular goals and ongoing calls for integrating technology and alternative methods to cover formative/integrated assessment training in teacher preparation programs entirely (Tsayari & Vogt, 2017). Such finding corroborates with prior studies that while teacher education programs introduce innovative assessment concepts, the pre-service teachers revert back to their previous ways and practices during practicum due to little time practicing these ideas, perceived complexity, and institutional limitations (Cirocki et al., 2025). It is also significant to find that a number of pre-service teachers really wanting to use digital tools but did not know how, when this indicates that much more training needs to be done with technology. This is especially troubling in light of the increasing focus on Artificial Intelligence and other digital assessment tools in Indonesia (Chuang & Yan, 2025) as well as a suggestion that teacher preparation programs do not do enough to prepare graduates for the demands of a multi-faceted technologically-based classroom. Together, these findings suggest that the Language Assessment course is not achieving its intended goals and requires immediate curricular reforms to emphasize practice-based learning along with self-assessment of metacognitive processes, alongside preparation for digital assessment as components.

CONCLUSION

This mixed-methods convergent parallel study aimed to investigate the gap between the intended learning outcomes of Language Assessment course and the demonstrated Language Assessment Literacy (LAL) of pre-service English teachers during their teaching practicum at Universitas Hindu Negeri I Gusti Bagus Sugriwa Denpasar. These results read like a narrative-style abstract of how the Language Assessment course is not accomplishing what it is meant to do; an implausibly novel claim, given that there exists a statistically significant difference between what pre-service teachers are expected to understand and show in practice. The quantitative data showed that the mean LAL score of 2.8 ($n = 68$) was well below the standard for intended outcome (4.2), with just over 8% of pre-service teachers achieving an intended outcome level sufficient to meet this benchmark. The gap was not only statistically significant but amounted to difference of nearly three standard deviations, suggesting that the pre-service teachers graduating from these courses are graduating with only fundamental assessment competencies in place requiring guidance rather than showing competent stand-alone application as expected in the course curriculum. Importantly, the study revealed an alarming gap between self-reported and observed LAL: pre-service teachers rated their performance at 3.9 out of 5 while university supervisors scored performance actually achieved at a median level of only 2.7. A 1.2-point discrepancy indicates a lack of accurate self-perception among pre-service teachers about their assessment skills, possibly due to their coursework's tendency to inflate confidence without developing practical skills. The qualitative results further elucidated the areas in which respondents struggle most, revealing five themes: pervasive deficits in rubric design replete with vague performance descriptors and misalignment of mapping to learning objectives; total absence of statistical and psychometric analysis of assessment data; generic feedback that is associative rather than instrumental to enabling student learning; excessive reliance on selected-response formats at the expense of authentic/presentation-based assessments; and lack of digital assessment tools, against explicit course requirements for development.

The implications of the findings for teacher education programs in Indonesia and elsewhere are discussed. The nature of LAL as a target for assessment education suggests that existing methods can be overly theoretical (thus missing the extent to which they replicate the attributes of practice) and provide insufficient opportunities

for pre-service teachers to apply, practice, and benefit from feedback in their assessment skills while still in training. Since pre-service teachers are not capable of basic item analysis / the interpretation of statistical data, this suggests that the training in psychometrics needs to be more practice based possibly through test data from their own classrooms. Likewise, the rates of rubric creation and feedback delivery imply that these competencies likewise call for practice with models, peer review, and supervision in order to master. This lack of digital assessment tools is especially alarming in light of the trajectory toward more integration of technology into Indonesian education and moving toward AI-enhanced assessment. This suggests that simply giving pre-service teachers the concept of these tools in their classes is not enough; they need time to experiment with and practice using them during their field experiences, along with access to someone who can help if anything goes wrong. The underutilized evidence of the learning progression from specialist assessment feedback provided by university supervisors was also revealed as a reason (between teaching performance generally not on classroom assessment) for potentially developing pre-service teachers' assessment literacy. There is an opportunity lost to provide formative professional development that closes the intended-demonstrated divide.

Besides, the limitations of this study must be recognized. The single institutional context, number of participants, and targeting a cultural and educational setting can limit generalizability. Future studies can explore whether similar gaps exist in other contexts in the Indonesia system, focusing on comparisons between institutions within and across ministry affiliates (state universities compared to private universities); further, longitudinal work is needed to examine how LAL develops beyond pre-service teaching through the early years of practice. Longitudinal studies that follow the same cohort of teachers through their assessment course and into their early years in practice would explain how effectively LAL has developed and what help works best. This study finds that the Language Assessment course is not effectively maintaining its mandate to prepare pre-service English teachers with suitable assessment literacy for classroom practice. It is neither a theoretical nor an academic gap, but practical; the gap between intended and demonstrated LAL simply puts in jeopardy the quality of assessment practices taking place in Indonesian classrooms. This requires immediate curriculum reform focused on practice-based learning, metacognitive self-assessment training and closer integration of digital

assessment. In addition, the supervisor's role in university settings must be reimagined to have an emphasis on enhancing pre-service teachers as effective assessors through targeted observation, feedback and mentoring. Without reforms such as these, pre-service teachers will keep graduating with unwarranted confidence but insufficient skills to effectively assess student learning. Also, the results of this study are a caution as well as guidance for teacher educators, curriculum designers and policy makers to improve the quality of English language teacher education in Indonesia.

REFERENCES

- Allen, J. M., & Wright, S. E. (2014). Integrating theory and practice in the pre-service teacher education practicum. *Teachers and teaching*, 20(2), 136-151.
- Borg, S., & Edmett, A. (2019). Developing a self-assessment tool for English language teachers. *Language Teaching Research*, 23(5), 655-679.
- Chuang, P. L., & Yan, X. (2025). Language assessment in the era of generative artificial intelligence: Opportunities, challenges, and future directions. *System*, 103846.
- Cirocki, A., Anam, S., Drajati, N. A., & Soden, B. (2025). Assessment literacy among Indonesian pre-service English language teachers: A mixed-methods study. *Iranian Journal of Language Teaching Research*, 13(1), 23-44.
- Fitriyah, I., Masitoh, F., & Widiati, U. (2022). Classroom-based language assessment literacy and professional development need between novice and experienced EFL teachers. *Indonesian Journal of Applied Linguistics*, 12(1), 124-134.
- Fulcher, G. (2012). Assessment literacy for the language classroom. *Language assessment quarterly*, 9(2), 113-132.
- Nuraeni, N., & Heryatun, Y. (2021). Reflective practice strategies of pre-service English teachers during teaching practicum to promote professional development. *Studies in English language and Education*, 8(3), 1144-1157.
- Oo, C. Z., Alonzo, D., & Davison, C. (2021, April). Pre-service teachers' decision-making and classroom assessment practices. In *Frontiers in Education* (Vol. 6, p. 628100). Frontiers Media SA.

- Putri, N. V. W., Fitri, A., Suwandi, E., Nadif, B., Diana, R. N., & A'yun, M. Q. (2026). Exploring language assessment literacy: Indonesian pre-service English teachers' perceived readiness for classroom assessment. *Journal of Research on English and Language Learning (J-REaLL)*, 7(1), 212-228.
- Rasyid, M. N. A., Alfian, A., Djamereng, J., & Nurwahida, N. (2024). English language assessment in Indonesian senior high schools: Perceived challenges and prospective solutions. *Journal on English as a Foreign Language*, 14(1), 1-25.
- Tsagari, D., & Vogt, K. (2017). Assessment literacy of foreign language teachers around Europe: Research, challenges and future prospects. *Papers in Language Testing and Assessment*, 6(1), 41-63.
- Vogt, K., & Tsagari, D. (2014). Assessment literacy of foreign language teachers: Findings of a European study. *Language Assessment Quarterly*, 11(4), 374-402.