

Integrated Assessment Exists but Feels Fragmented: Pre-Service English Teachers' Quality Perceptions in Indonesian EFL Teacher Education

I Putu Andre Suhardiana^{1*}, I Dewa Gede Rat Dwiyanu Putra², I Gusti Ngurah Agung Wijaya Mahardika³, Ni Putu Ratni⁴

^{1,2,3,4} *Universitas Hindu Negeri I Gusti Bagus Sugriwa Denpasar, Indonesia*
putuandresuhardiana@gmail.com¹, ratdwiyanaputra@uhnsugriwa.ac.id²,
wijayamahardika@gmail.com³, puturatni@uhnsugriwa.ac.id⁴

Received 25 May 2026 | Received in revised form 16 June 2026 | Accepted 07 July 2026

APA Citation:

Suhardiana, I. P. A., Putra, I. D. G. R. D., Mahardika, I. G. N. A. W., & Ratni, N. P. (2026). Integrated Assessment Exists but Feels Fragmented: Pre-Service English Teachers' Quality Perceptions in Indonesian EFL Teacher Education. *Journal of English Language Teaching and Linguistics*, 11(2), 2026, 215-236.
doi: <http://dx.doi.org/10.21462/jeltl.v11i2.2004>

Abstract

The traditional way of assessing pre-service English teachers has separated language competency from pedagogical competence, which experts refer to as an artificial duality. However, preliminary findings from an Indonesian teacher education program indicated that integrated assessment is already in place, but felt inadequate by students. The study focused on the gap between the existence and effectiveness of integrated assessment that was experienced by pre-service English teachers, particularly rubric transparency, feedback implementation, and genuineness. This study used a convergent parallel mixed-methods design with 64 pre-service English teachers of Universitas Hindu Negeri I Gusti Bagus Sugriwa Denpasar. Data were acquired by a 15-item questionnaire, open-ended responses and semi-structured interviews with five selected students. Quantitative data were analyzed using descriptive statistics, and qualitative data using thematic analysis. The results showed a paradox. 81.3% of students reported experiencing integrated assessment, and 76.6% rejected the idea of fragmentation; however, 87.5% still wanted assessment to be more integrated. Three quality gaps emerged: clarity deficit of rubric (79.7% strongly acknowledged the requirement of integrated rubrics), feedback integration gap (96.9% requested integrated feedback), and authenticity problem (73.4% regarded assessments as too theoretical). Students cited five features: dual-skill rubrics, simulation-based exercises, portfolio assessment, manageable workload and reflective practice. The study finds that the problem is not the lack of integrated assessment but the quality of integration as viewed by the students.

Keywords: assessment quality, feedback integration, integrated assessment, pre-service English teachers, rubric clarity

1. Introduction

Teacher competence has been re-conceptualized over the last ten years. In contrast to traditional frameworks that treat language proficiency and pedagogical ability as independent constructs, with grammar, vocabulary, and pronunciation assessed separately from lesson planning, classroom management, and student facilitation (Giraldo, 2021), contemporary scholarship acknowledges the connected nature of these domains in classroom practice (Pastore & Andrade, 2019; Giraldo, 2021). This shift signals increasing recognition that teachers use linguistic knowledge and pedagogical approaches as complementary, rather than separate resources; both are deployed based on the changing and sometimes ill-defined requirements of real-time instruction (Scarino, 2013; Pastore & Andrade, 2019). The impact of this artificial divide is tangible. It has been shown that good performance at isolated language proficiency tests is not a reliable indicator of good teaching skills any more than strong pedagogical knowledge guarantees the appropriate use of the target language in the classroom (Lindahl, 2019). One difference between mediocre and great language teachers is the ability to utilize linguistic knowledge for teaching purposes (simplifying language, scaffolding understanding, asking questions strategically to get students to engage as much as possible (Faez et al., 2021). Assessment frameworks that treat language and pedagogy as separate domains send a message implicitly to pre-service teachers that these dimensions are discrete, even though in actual practice the two academic languages are intertwined, where they cannot easily be separated (Xu & Brown, 2016).

The need for integrated assessment, defined here as evaluative tasks in which pre-service teachers simultaneously show language proficiency and pedagogical competence through authentic teaching activities (Pastore & Andrade, 2019), is an academic endeavor and a need heralded by national policy frameworks. The Indonesian Qualification Framework (KKNI) requires higher education graduates to exhibit competencies aligned with national qualification levels, and for teacher education programs to document the measurable outcomes of learning which graduates have achieved during their studies (Mufanti et al., 2024). Likewise, the approach of Outcome-Based Education (OBE), which is taken in considering KKNI requirements, stresses alignment of curriculum-instruction-assessment around explicitly stated outcomes, but does not offer much guidance as to how integrated competencies, such as cogenerative forms of language-pedagogy, can be assessed (Aminah et al., 2025). Even more pressing is the need to develop assessment models which map how pre-service teachers can actually use English for pedagogical purposes in a real classroom as teachers gain new autonomy under the Merdeka Curriculum, giving agencies wider discretion over how to construct instruction and assessment (Hunaepi & Suharta, 2024). Even with these policy demands, the English Language Education Department in Universitas Hindu Negeri I Gusti Bagus Sugriwa Denpasar, having ‘*Unggul*’ (Excellent) accreditation from LAMDIK, does not have a validated instrument that describes how graduates integrate the two aspects of English proficiency and teaching skills (Asbari & Nurhayati, 2024). Preliminary evidence from the English Language Education Department at Universitas Hindu Negeri I Gusti Bagus Sugriwa Denpasar suggests that students feel integrated assessment is present; however, they describe it as lacking transparency and meaningfulness. Students recognize that they are evaluated on both language and teaching competencies, but cannot explain how the two domains are related / what proficient integration looks like. This paradox of integrated yet not integrated, as understood by students, suggests that the issue is about transparency, clarity and intentional design focused on communicating how learning is assessed with the students.

The focus of this study shifts from whether integrated assessment is used to the meaning and experience of it for pre-service English teachers so as to address this paradox. Existing literature has focused on the lack of integrated assessment models and needs to create new frameworks (Giraldo,

2021), while little attention has been given to how students experience the quality of integration when there is one. However, three aspects of assessment quality from the student perspective are less explored: the clarity of rubrics, feedback integration, and authenticity. Rubric clarity is the extent to which students can understand how language and teaching criteria are expressed and weighted together; feedback integration refers to whether students receive feedback that considers the confluence of language and pedagogy, not separate feedback for each domain; and authenticity is the degree to which assessment tasks mirror the demands of teaching in a classroom, including managing student behaviour, responding to unpredictable questions, and simplifying language on-the-go (Darling-Hammond, 2017). Without an understanding of how pre-service teachers understand these dimensions, tailored integrated assessment models can gamble on targeting what students do not see as needed and failing to provide the features they most urgently seek. Thus, the current study examines how pre-service English teachers perceive the quality of integrated assessment in an Indonesian EFL teacher education program. This study is framed by the following research questions:

1. To what extent do pre-service English teachers perceive a gap between the presence of integrated assessment and the quality of its implementation regarding (a) rubric clarity, (b) feedback integration, and (c) authenticity?
2. What features do pre-service English teachers identify as needed for integrated assessment to be educationally meaningful, as defined by their perceptions of clarity, feedback usefulness, authenticity, and manageable workload?

2. Literature Review

2.1 The Artificial Duality of Language and Pedagogy in Teacher Assessment

In applied linguistics, the separation of language proficiency and pedagogical competence when assessing teachers has long been criticized (Giraldo, 2021; Scarino, 2013). The critique is not without foundation, both theoretically and empirically. The separation falsely depicts classroom practice, where teachers constantly navigate between linguistic and instructional decisions (Faez et al., 2021) as discrete, toggle-like modes of practice. Empirical studies supported that how well teacher candidates do on isolated language tests has weak predictive power for teaching quality, and that strong pedagogical knowledge alone does not ensure the appropriate use of language in the classroom (Macaro et al., 2018, as cited in Lindahl, 2019). Prior literature has documented this disconnect in other contexts, including English medium instruction settings (Macaro et al., 2018), and language teacher education programs (Lindahl, 2019; Xu & Brown, 2016). The continued existence of this artificial division in assessment frameworks is especially concerning since it sends the message to pre-service teachers that language and pedagogy exist as separate concerns. The classroom reality resists such a piece of information, as the teacher has to balance linguistic correctness, pedagogical clarity and student engagement (Scarino, 2013). As a result, the literature has increasingly highlighted the need for assessing language as not something separate from teaching that can be evaluated (together with outcomes), but rather an essential functioning of pedagogical practice (Giraldo, 2021; Pastore & Andrade, 2019).

2.2 Definitional and Framework Issues of Integrated Assessment in Teacher Education

Integrated assessment is the design type of evaluative tasks in which pre-service teachers perform language and pedagogical tasks concurrently, using authentic teaching and learning (Pastore & Andrade, 2019). Integrated assessment differs from traditional approaches, which measure each domain in isolation, and view classroom instruction primarily as a linguistically mediated activity. Integrated assessment is characterized by (a) tasks that reflect authentic classroom responsibilities (for writing a lesson plan, demonstrations of teaching and interaction with students)

(Darling-Hammond, 2017), (b) rubrics combining ideas about language and the teaching task to be assessed according to one scoring system (Pastore & Andrade, 2019), (c) feedback describing how particular linguistic decisions serve or hinder enactment of model-defined effectiveness in instruction (Xu & Brown, 2016). In language teacher education, integrated assessment is based on the concept of teacher language awareness in that it is necessary for good teachers to understand how languages function in teaching (Lindahl, 2019). Although there is theoretical support for integrated assessment, the practice has been inconsistent. A large number of programs still use analytic rubrics that treat language criteria (accuracy, fluency, range) separately from teaching criteria (structure, questioning, feedback), thus perpetuating the very divide that integrated tasks are intended to bridge (Scarino, 2013). Further, feedback practices separate comments on language from pedagogy, but these shortcomings of engagement are not addressed, and students continue to make inferences about the interplay between the two (Xu & Brown, 2016). For this reason, the literature has moved on from simply promoting integrated assessment to analyzing the particulars of integration as it is lived by learners.

2.3 Assessment Quality from the Student Perspective: Rubric Clarity, Feedback Integration, and Authenticity

Rubric clarity, feedback integration and authenticity are three aspects of assessment quality that emerge in particular from the learners. Rubric clarity refers to the degree to which students know how the criteria (expectations) for assessment are defined, weighted and used across domains (Pastore & Andrade, 2019). When the rubric maps language and teaching onto distinct columns with independent scores, students have no clue how their performance on one dimension affects performance on another. When the evaluation criteria are not transparent, it fails to serve the formative function of assessment, compelling students to play guessing games around how to improve and potentially directing their attention towards different aspects of a performance (Yan & Fan, 2021). Feedback integration focuses on whether students receive feedback stating how language use connects to teaching outcomes. Fragmented feedback (separate notes on grammar errors and lesson structure from different instructors) implies that the two domains are independent, while integrated feedback explains why complex sentence structures confused students / why standard vocabulary increases engagement (Giraldo, 2021). The authenticity relates to the extent to which assessment tasks mimic the often unpredictable realities of classroom teaching. Authentic assessments do not assess rote memory of pedagogical theories through written tests/peer teaching demonstrations with compliant classmates (Darling-Hammond, 2017). In terms of task, authentic tasks are real-world tasks similar to the challenges that exist in a classroom, interrupting students, probing questions, technology failures and behaviours of different learners (Munandar & Newton, 2021). Pre-service teachers who experienced assessments often engage in questioning practices, leading to higher learning gains and greater understanding of how their beliefs relate to complex academic environments (Darling-Hammond, 2017).

2.4 Policy Imperatives and Implementation Gaps

Integrated assessment has been propelled at both national policy frameworks and local educational reforms in Indonesia. The Indonesian Qualification Framework (KKNI) prescribes the competencies that higher education graduates should possess at different qualification levels, therefore necessitating teacher educator programs to keep measurable learning outcomes documentation (Mufanti et al., 2024). Outcome-Based Education (OBE), which is used to implement KKNI requirements, focuses on alignment of curriculum, instruction, and assessment around clearly stated outcomes (Aminah et al., 2025). By contrast, OBE is not particularly prescriptive about how integrated competencies, such as the simultaneous use of language for

pedagogical purposes, can be gauged. In the context of more recent developments, Merdeka Curriculum has shifted a significant amount of instructional and assessment design decision-making territory to schools/teachers, thus amplifying just-in-time assessments that signal pre-service teachers' flexible and responsive use of English as a medium of instruction (Hunaepi & Suharta, 2024). Yet even with such policy demands, implementation gaps are still there, as highlighted by the empirical studies of Darling-Hammond (2017). However, many Indonesian teacher education programs do not have validated instruments to assess the extent and how well their graduates combine English proficiency with the requisite teaching skills (Asbari & Nurhayati, 2024). This paradox indicates that the issue is integration experience as pre-service teachers. This study addresses a gap in the literature, wiring up the concepts of clarity, feedback and authenticity from an integrated assessment perspective around student perceptions as they experience integrated assessment being nominally present.

2.5 Workload and Structural Constraints on Assessment Quality

One emerging and less-explored theme in the assessment literature is how student workload influences perceptions of assessment quality. If lecturers provide students with too many assignments, deadlines, and course conflicts, any assessment, even the best-designed assessment, is beyond the student's cognitive-emotional capacity for engaging meaningfully at any point (Abdillah & Fithriani, 2023). In the Indonesian higher education context, structural constraints such as large class sizes, limited instructional time and institutional pressure to cover extensive curricula lead to heavy workloads (Multamasatika et al., 2023). This tension between workload and integration has been under-explored in integrated assessment literature, which has often conceptualized its task and rubric design without recourse to the larger structural conditions that mediate student learning. The present study aims to fill this gap by looking at how workload influences perceptions of integrated assessment quality, noting that even the best-designed model for pedagogically appropriate assessments will fail if students do not have the time and mental energy to participate with them.

2.6 Theoretical Framework

The theoretical lenses of this study are drawn from the Australian concept of assessment literacy theory by Pastore & Andrade (2019), complemented by the previous development by Xu & Brown (2016). Assessment literacy is an effort made to the field, which requires knowledge and skill sets associated with creating, administering and analyzing assessments as well as communicating assessment information and feedback (Pastore & Andrade, 2019). In teacher education, assessment literacy is about technical competence and a capacity to render assessment practices visible, transparent and purposeful for students (Xu & Brown, 2016). The study also draws on sociocultural theory (Vygotsky, 1978) that learning takes place through engagement in participation-based, authentic, socially-mediated activities. In this light, the development of integrated competence takes place through engagement in authentic teaching tasks that require the deployment of both types of knowledge together. Thus, assessment should reflect the integrated nature of practice. These theoretical lenses, assessment literacy and sociocultural theory, help orient the study around three dimensions of assessment quality: rubric clarity (describing how well the assessment criteria are communicated); feedback integration (how the feedback attends to issues related to language as it intersects with pedagogical practice); and authenticity (defining assessments in terms of whether / not they resemble real classroom tasks). These spheres describe how literacy intersects with sociocultural context to shape meaningful learning.

2.7 Gap Synthesis and Study Positioning

The literature shows that an integrated assessment approach is an element of language teacher education; however, its actual implementation proves to be challenging. At the same time, although

scholars have criticized the separation of language and pedagogy (Giraldo, 2021; Scarino, 2013) and advocated for more genuine integrated assessment designs (Pastore & Andrade, 2019; Darling-Hammond, 2017), empirical studies have mostly directed their focus to exploring absence-based scenarios without integration. What is less understood is under what STS conditions pre-service teachers experience the quality of integration when it exists as a nominal construct, whether pre-service teachers see integrated tasks as transparent, feedback as meaningful, and assessments as authentic. In particular, this gap is evident in the context of Indonesia, where national policy frameworks (KKNI, OBE) require integrated competencies but provide little guidance on assessing them (Mufanti et al., 2024; Aminah et al., 2025). This study fills the gap by investigating pre-service English teachers' beliefs as to the quality of assessment regarding rubric clarity, feedback integration, and authenticity, while identifying characteristics they see as important for integrated assessment to be educationally useful.

3. Research Methods

3.1 Research Design

The research used a convergent parallel mixed-methods design, which comprises collecting and analyzing quantitative and qualitative data concurrently in order to develop an understanding of the quality of integrated assessment by pre-service English teachers. The quantitative portion measured students' perception of a gap between integrated assessment and its quality across three dimensions: clarity of rubric, feedback integration, and authenticity. The qualitative component examined the characteristics students identified as needed for assessment integration to be educationally meaningful. This design was selected based on Creswell & Plano Clark's (2018) convergent parallel mixed-methods framework for three reasons. The research questions of study 1 require both breadth (quantifying how widespread and in what patterns perceptions are expressed across the student population) and depth (the rationale for those perceptions as illustrated by rich qualitative accounts). Second, a convergent design involves the collection and analysis of quantitative and qualitative data independently before combining results to find the convergence/divergence/complementarity between them. Furthermore, when research questions do not privilege one / the other types of data this design is appropriate (QUAN + QUAL). A sequential design was not determined because the research questions do not necessitate one strand to inform the other; rather, both strands present different ways of addressing same research questions.

3.2 Participants

The participants were 64 pre-service English teachers of the English Language Education Department, Faculty of Dharma Acarya, Universitas Hindu Negeri I Gusti Bagus Sugriwa Denpasar-Indonesia. The sample included that all participants were in semesters 6, 7, or 8 (final year) and had completed / currently taking Micro Teaching and/or Teaching Practicum courses to ensure they have been well equipped with assessment experiences to make sound judgement regarding quality integrated assessment practices. This year, the department itself has also achieved 'Unggul' (Excellent) accreditation with LAMDIK, the National Accreditation Agency for Higher Education in Indonesia; thus this context is an avenue through which to assess quality of assessment in a high-quality program. Criterion-based sampling was employed to recruit participants who met two inclusion criteria: (1) enrollment in semesters 6, 7, or 8 (final year), and (2) completion of / current enrollment in Micro Teaching and/or Teaching Practicum courses. All students meeting these criteria (N=64) were invited to participate, and all agreed, resulting in complete enumeration of the eligible population. The sample has 51 female (79.7%) and 13 male (20.3%) students. Participants ranged in age from 20 to 23 years ($M = 21.4$, $SD = 0.9$). In terms of semester

distribution, 21 students (32.8%) were in semester 6, 19 (29.7%) in semester 7, and 24 (37.5%) in semester 8. All participants had completed at least two language proficiency courses and three methodology courses at the time of data collection. Thirty-seven participants (57.8%) had completed their teaching practicum, while 27 (42.2%) were currently enrolled.

3.3 Instruments

Data collection was carried out using three instruments. First, 15 Likert-scale items (1 = Strongly Disagree to 5 = Strongly Agree) about the following three constructs: current assessment practices (measured by five items), perceptions of current assessment quality with respect to those characteristics as compared to standard-based goals (five items), and students expectations for integrated assessment (five items). This questionnaire was adapted based on the concept of integrated teaching competence to the context of Indonesian EFL.

Table 1 presents the alignment between the questionnaire items and the study's three analytic constructs, rubric clarity, feedback integration, and authenticity, organized by subscale.

Table 1: Item-to-Construct Alignment for Questionnaire

Construct	Subscale	Item Numbers	Item
Rubric Clarity	Current Practices	1, 2, 3	<i>'My course rubrics show how language and teaching skills are assessed together'</i>
Rubric Clarity	Quality Perceptions	6, 7, 8	<i>'The rubrics I receive clearly explain how language choices affect teaching effectiveness'</i>
Rubric Clarity	Expectations	11, 12	<i>'I want rubrics that combine language and teaching criteria in one scoring system'</i>
Feedback Integration	Current Practices	4, 5	<i>'I receive separate feedback on my language and my teaching' (reverse-coded)</i>
Feedback Integration	Quality Perceptions	9, 10	<i>'Feedback explains how my language use affects student understanding'</i>
Feedback Integration	Expectations	13, 14	<i>'I want feedback that addresses language and teaching together'</i>
Authenticity	Current Practices	3 (also)	<i>'My assessments reflect real classroom challenges'</i>
Authenticity	Quality Perceptions	8 (also)	<i>'Assessments prepare me for unpredictable classroom situations'</i>
Authenticity	Expectations	15	<i>'I want assessments that simulate real teaching challenges'</i>

Second, there were four open-ended questions in the questionnaire asking what students liked about current assessments, what they found hard to do, what an ideal integrated assessment looks like and other comments. Third, a semi-structured interview guide was created which served to conduct follow-up interviews with five selected students based on questionnaire responses related to rubric transparency, feedback integration, and the perceived presence versus quality of assessment. The semi-structured interview guide covered the following thematic areas: (1) How participants understand assessment criteria; what they find clear or unclear about current rubrics; (2) The nature of feedback they receive; whether feedback addresses the relationship between language and teaching; (3) How well assessments prepare them for actual teaching; what kinds of tasks they find most meaningful; (4) How the number and timing of assessments affects their ability to engage with learning; (5) What changes would make assessments more educationally meaningful. The interview protocol included 8 questions and 4-6 follow-up prompts per thematic area. Interviews were conducted in a private office on campus, lasted 30-45 minutes, and were audio-recorded with participant consent. Probes such as *'Can you give me an example?'* and *'What do you mean by that?'* were used to elicit detailed responses. Before being administered, all instruments underwent

content validation by three experts: two faculty members with PhDs in Language Assessment and one with a PhD in English Education, all with at least five years of experience in teacher education. Experts evaluated items on a 4-point scale (1 = not relevant, 4 = highly relevant) for clarity, relevance to the constructs, and appropriateness for the Indonesian context. Based on expert feedback, three items were revised for clarity (items 7, 9, and 12), and one item was removed (item 10) due to low relevance ratings. The remaining 15 items achieved a Content Validity Index (CVI) of .87, exceeding the .80 threshold recommended for content validation (Lynn, 1986). Internal consistency reliability was assessed using Cronbach's alpha. The three subscales showed acceptable reliability: Current Practices (5 items, $\alpha = .78$), Quality Perceptions (5 items, $\alpha = .81$), and Expectations (5 items, $\alpha = .84$). These values exceed the .70 threshold recommended by Field (2024) for social science research.

3.4 Data Collection Technique

Data were collected for a period of nine days. To maximize the response rate, the questionnaire was administered in paper format as part of the regular class sessions. Before conducting the administration, the researcher explained to what was the purpose of this study, describing that respondents would be treated confidentially and voluntarily by providing written informed consent from all 64 students. The questionnaire comprised a combination of Likert-scale and open-ended sections, which took each participant about 20-25 minutes to complete. On the basis of a preliminary analysis which highlighted the richness and representativeness of students' open-ended responses to the questionnaire, five students were purposefully selected for follow-up semi-structured interviews. Interviews lasted for 30-45 minutes, were recorded with consent from participants and the audio was transcribed verbatim. The Research Ethics Committee of Universitas Hindu Negeri I Gusti Bagus Sugriwa Denpasar has given ethical permission. All participants gave their written informed consent after being told the goal of the study, that participation was voluntary and that they had the right to withdraw at any time. Names of participants were coded (S01-S64) in order to ensure confidentiality and all identifying information was omitted from transcripts and reported statistics.

3.5 Data Analysis Technique

Responses from the three measured constructs for each item were analyzed with descriptive statistics (frequencies, percentages, modes, and means) for quantitative data drawn from the Likert-scale items to determine instances of central tendency and patterns in student perceptions. Due to the nature of Likert-scale data, the mode should be considered as a main measure of central tendency. Qualitative open-ended questionnaire responses and interview transcripts were analysed thematically using Braun & Clarke's (2006) six-phase framework, following the steps of on familiarizing the researchers with the data, generating initial codes, searching for themes, reviewing themes, defining and naming themes, and producing the final analysis. The analysis was conducted using an inductive (data-driven) approach at the semantic level, meaning that themes were derived from explicit patterns in participants' responses rather than imposed from pre-existing theory (Braun & Clarke, 2019). Coding was performed by hand, employing a color-coded system to pinpoint recurrent patterns around rubric clarity, feedback quality, authenticity of work, workload concerns and suggestions for improvement. To increase its trustworthiness, the following strategies were used based on Braun & Clarke's (2019) guide for quality of reflexive thematic analysis. Two researchers first independently coded 20% of the qualitative data, held reflexive discussion sessions to compare interpretations, and then determined a consensus for developing themes. This is an interactive process of analysis known as dialogic consensus (Braun & Clarke, 2019), aims at deepening interpretive rigor through dialogue between researchers. In the second step, member

checking was conducted by sharing preliminary findings with three participants, who confirmed that the themes reflected their experiences. Third, negative case analysis was used: responses that contradicted the emerging themes were analysed and reported (in Findings 4 and 6). Fourth, an audit trail documenting coding decisions, theme development and memos created during analysis was maintained. During analysis and interpretation, quantitative findings were examined with qualitative findings to determine whether convergence, divergence, and complementarity existed based on what students indicated quantitatively (presence and quality of integrated assessment) compared to how they described their perceptions of gaps qualitatively.

4. Results

4.1 The Presence-Quality Paradox

Table 1: Pre-Service English Teachers' Perceptions of Integrated Assessment

Indicator	Number of Students (N=64)	Percentage
Reported taking assessments that combine language with teaching skills.	52	81.3%
Understanding the link between language proficiency and instructional efficiency	58	90.6%
Agree they are frequently tested on both English proficiency and teaching skills	61	95.3%
Routine evaluation reported on English skills	58	90.6%
Regular assessment reported on teaching skills	61	95.3%
Disagreed that assessments feel separate and disconnected (modal response)	49	76.6%
Strongly agree / agree that assessments feel detached from teaching goals	8	12.5%
Was unconcerned about dissociation	7	10.9%
Would like assessments to be more integrated and realistic (knowing it is)	56	87.5%

The quantitative results showed an underlying contradiction in the way that pre-service English teachers perceive integrated assessment in their teacher education program. Against the pervasive assumption in the language assessment literature that teacher education programs either assess separate facets of language skills and pedagogical competence, often characterized as a problem of artificial dualism, this study presents a different finding. In particular, among all students (N = 64), 52 claimed they had tests integrating language with teaching skills (81.3%). This alone contradicts the discussion that integrated assessment does not play a role within teacher education settings. Moreover, 58, (90.6%) were aware of the connection between their language ability and teaching efficiency. Thus, students are experiencing integration and recognizing the links between the two domains in which they are required to operate. Also, 61 students (95.3%) agreed that they are regularly assessed in English proficiency and teaching skills, with 58 students (90.6) from the two groups saying that they are routinely evaluated on the English proficiency and 61 students (95.3%) a regular evaluation on their teaching skills. What these numbers suggest is that the program is already evaluating both domains. Specifically, when students are asked if the assessments they undertake are separate and disconnected, a modal Disagree was obtained (n=49; 76.6%) with roughly three-quarters of students rejecting any idea of fragmentation. No student strongly agreed / agreed that assessments feel dissociated in terms of teaching goals (8 students, 12.5%); 7 students (10.9%) remained indifferent. This result is therefore directly at odds with the artificial dualism problem, so often referred to in the literature, which predominantly assumed that language and pedagogy are evaluated independently of one another. The data from the present study reveals that teacher educators at UHN I Gusti Bagus Sugriwa Denpasar have already begun addressing a call to integration in their assessment design (set integrated assessment tasks such that students will need to perform language and teaching on simultaneously).

Though aware of the integrated nature, 56 students (87.5%) responded that they wished assessments were more integrated and realistic. This paradox of integration (but feeling excluded) reinforces the argument in this study. Underlying data show that students also differentiate between the presence of integrated tasks and their perceived adequacy. Integration is not integration by the presence of integration. Students stated that their programme included integrated assessment, but the quality of the integration was considered insufficient in terms of transparency, clarity, and meaningfulness. This tension was aptly expressed in an interview (P02): *'We do use the English language and teaching in the same task, yet it feels like two tasks because there is a separate rubric for each, and each is marked individually.'* As another interviewee (P04) stated, *'I have faced integrated tasks such as teaching demonstrations and lesson plan presentations at least five times. Even so, the way they are tested makes me feel like language and teaching remain separate entities in my mind. I receive only separate marks from the lecturer at the end.'* So while the integration of task itself was assimilated, that assessment component associated with this task was not. A questionnaire respondent (Q18) wrote: *'My program does a good job with having us teach in English. You are asked to talk and write in English about teaching for almost every assignment. For grading and feedback, though, everything is broken out into different sections. I feel disintegrated because the assessment feels disintegrated.'* Another interviewee (P01) responds, *'One of my lecturers told me that I did well in preparing and performing a teaching demonstration while also saying I have to improve on language. I asked what that meant. Were my grammar mistakes hurting my teaching? Or was my teaching still successful, even though I started speaking like a beginner?'* Even the assessors are aspiring to a distinction between categories. These comments reflect that students differentiate between how assessment tasks are designed and how assessments are implemented. Integrated tasks must be present, but they are not sufficient for students to find meaning in integration. Quantitative data revealed the presence-quality paradox (81.3% experiencing integration vs. 87.5% wanting more). Qualitative data explained this paradox by revealing that although task-level integration exists, assessment implementation (rubrics, feedback) remains fragmented. The two datasets converge on the finding that integration is present but perceived as inadequate. They diverge slightly in that quantitative data show rejection of fragmentation (76.6% disagreeing that assessments feel separate), while qualitative data reveal that students experience disconnection through fragmented rubrics and feedback. This divergence highlights that students distinguish between task design (integrated) and assessment implementation (fragmented).

4.2 Deficiency in Rubric Clarity

Table 2: Summary of Findings on Rubric Clarity Deficit

Aspect	Finding
Quantitative – Student agreement	79.7% (51/64) selected <i>Strongly Agree</i> that rubrics (integrating language & teaching skills) are needed
Neutral responses	4.7% (3/64) chose <i>Neutral</i>
Disagreement	0% selected <i>Disagree</i> / <i>Strongly Disagree</i>
Comparison with other survey items	No other item showed such extreme/concentrated responses
Qualitative - Core student complaint	Rubrics divide language and teaching (different scales, distinct scores) and do not provide integrated feedback on how they combine to form teaching proficiency
Qualitative - Student examples	• Separate scores for grammar, vocabulary, pronunciation, lesson planning → unclear how they combine • Language score (85) and pedagogy score (80) treated as isolated numbers • Rubrics across multiple courses have same problem: language criteria (accuracy, fluency, vocabulary, pronunciation) separate from teaching criteria (lesson structure, questioning, feedback, engagement)

Implication	Students do not feel the rubrics are clear enough. The existing rubrics do not reflect the interaction between language and teaching skills in actual classroom performance.
-------------	--

The quantitative result from the entire study came from student ratings of rubric clarity. The rubric clarity subscale (5 items) showed a mean score of $M = 4.21$ ($SD = 0.62$), with 79.7% of participants (51/64) selecting 'Strongly Agree' to the item *'I want clear rubrics that indicate how both language and teaching skills are assessed together.'* The remaining items in the subscale also showed high endorsement: Item 6 (*'Current rubrics explain how language affects teaching'*) had a mean of 3.85 ($SD = 0.73$), Item 7 (*'I understand how language and teaching criteria are weighted'*) had a mean of 3.76 ($SD = 0.81$), Item 8 (*'Rubrics help me understand what integrated competence looks like'*) had a mean of 4.08 ($SD = 0.69$), and Item 12 (*'Rubrics should show integration at different performance levels'*) had a mean of 4.34 ($SD = 0.58$). The subscale mean of 4.21 indicates strong agreement across the construct. 3 students (4.7%) chose Neutral and nobody selected Disagree / Strongly Disagree. None of the other items on the entire 15-item questionnaire prompted such highly concentrated and extreme responses. This demand also speaks to the nature of what students actually want; rubric clarity is their one and only highest priority in this area. The almost unanimous call for clear integrated rubrics suggests that students do not believe current rubrics are transparent enough to know how they will be assessed and how they can enhance their integrated competence. That none of these students disagreed with this statement speaks volumes and illustrates how deep the rubric clarity deficit runs. The qualitative data offered inferences to provide a deeper explanation for the demand being so high and so widespread within the student body. Micro-teaching is a sample classroom teaching which the students have to do on real students before facing their own classes, and one interviewee (P03) said that: *'The rubric for micro-teaching includes columns such as grammar, vocabulary, pronunciation, and lesson planning. At the end, I get six different scores, and they (1) are on different scales and (2) mean nothing about how it all combines into teaching proficiency. Do I have fine grammar but lousy classroom management if my score is high on grammar and low on management? Or is my grammar affecting my classroom management?'* A different interviewee (P05) recalled: *'I knew that my language had 85 and pedagogy had 80 from last time, that example was from one of my teaching demonstrations. When I remarked to my lecturer that my language score outstripped my pedagogy score and asked how the latter might have been affected by the former, she told me the numbers were for separate portions of the rubric. The numbers were isolated, and so was my perception of how I did. I need this rubric to tell me things like, 'when you wrote complex sentences, the students lost your instructions' and also 'you asked simple questions, so students paid more attention when the vocabulary was simpler.' A rubric like that would teach me how to teach something.'* A questionnaire respondent (Q22) described: *'Unless I have a rubric that links language with something related to teaching, I am only guessing every time. I concentrate on making improvements to my grammar and my lesson planning separately, since that is what is measured separately by the rubric. In the real classroom, however, I have to do both simultaneously at the same time.'* Another interviewee (P02) added: *'I have seen rubrics from three different courses in this program, and the same issue exists across all of them. Each rubric has a language section with four different criteria: accuracy, fluency, vocabulary range, and pronunciation. Then each rubric also has a teaching component that includes items like lesson structure, questioning strategies, feedback provision, and student engagement.'* In effect, these comments show that the issue is not with rubrics themselves, but the criteria embedded in them. Quantitative data revealed strong demand for integrated rubrics (subscale $M = 4.21$, $SD = 0.62$; 79.7% Strongly Agree on the item).

Qualitative data explained this demand by describing how current rubrics separate language and teaching criteria, leaving students unable to understand how the two domains combine. The

datasets converge on the finding that rubric clarity is a gap. Complementarity is evident: the quantitative data show the magnitude of the demand, while the qualitative data explain the mechanisms driving this demand (separate columns, distinct scores, no indication of interaction).

4.3 Feedback Integration Gap

Table 3: Student Feedback on Integrated Feedback (Language + Teaching Together)

Response Category	Number of Students (n = 64)	Percentage (%)
Strongly Agree	53	82.8%
Agree	9	14.1%
Neutral	2	3.1%
Disagree	0	0.0%
Strongly Disagree	0	0.0%
Total Positive (Agree + Strongly Agree)	62	96.9%

The feedback integration subscale (5 items) showed a mean score of $M = 4.38$ ($SD = 0.55$), with 96.9% of participants (62/64) agreeing / strongly agreeing with the item *'I want feedback on how to improve my English and teaching together at the same time.'* The remaining items in the subscale showed: Item 4 (*'I receive separate feedback on language and teaching'*) had a mean of 3.52 ($SD = 0.88$), Item 5 (*'Feedback from different lecturers is coordinated'*) had a mean of 3.28 ($SD = 0.92$), Item 9 (*'Feedback explains how language choices affect student understanding'*) had a mean of 3.94 ($SD = 0.76$), and Item 13 (*'I want feedback that explains the relationship between language and teaching'*) had a mean of 4.42 ($SD = 0.54$). The high subscale mean reflects strong demand for integrated feedback. 53 students (82.8%) agreed strongly with wanting feedback on how to improve their English and teaching together at the same time; A further 9 students (14.1%) agreed with this statement, giving a total of 62 out of 64 students (96.9%) indicating a positive response to integrated feedback. Only 2 (3.1%) students chose Neutral and none selected Disagree / Strongly Disagree. This finding (almost as strongly in demand as the very clear rubrics) suggests that students see a large systemic failure in how feedback is currently structured in their program. The widespread call for integrated feedback shows that there is a disconnect between the feedback students receive and their awareness of what they must do to master this junction between language and pedagogy. The qualitative data provided information into the character of this gap. One interviewee (P01) explained, *'I receive different feedback from different lecturers. My language lecturer leaves comments on my grammar, vocabulary, and pronunciation in different documents. In separate papers, my methodology lecturer comments on both my lesson planning and questioning techniques as well as classroom management. No one tells me whether my grammar errors confuse students with my instructions. No one tells me whether my probing of students failed because the language I used in my questions was too advanced for the students.'* A questionnaire respondent (Q31) says, *'On my feedback form from a teaching practicum at the school, my supervising teacher commented that my explanations confused students. That was the entire comment. Three words: 'explanations were not clear.' She did not tell me in any way what about my language made it impossible to follow.'* Another interviewee (P04) said: *'The feedback forms in this program always contain two sections. The first part is marked as Language Performance, while the latter part is marked as Teaching Performance. The feedback in the language section from my last demo lesson read, 'good range of vocabulary and accurate grammar,' while the teaching section said, 'needs to be more clear with instructions given to students.' If I had a good vocabulary and my grammar was correct, then why were my instructions unclear?'* This question is left unanswered because the feedback treats language and teaching as though they were wholly independent of each other. A questionnaire respondent (Q45) said: *'After a teaching*

demonstration, I asked my lecturer how to develop as both an English speaker and as a teacher together. She said, 'You have to work on your grammar separately and your lesson planning separately. Work on one thing at a time.' Such comments reveal that current practices do not address integrated competence. Quantitative data showed near-unanimous demand for integrated feedback (subscale $M = 4.38$, $SD = 0.55$; 96.9% positive response).

Qualitative data explained this demand by illustrating how students receive separate feedback on language and teaching from different sources, with no explanation of how the two domains interact. The datasets converge on the finding that feedback is experienced as fragmented. Complementarity is evident, the quantitative data establish the pervasiveness of the demand, while the qualitative data reveal the forms of fragmentation (separate forms, different lecturers, disjointed comments).

4.4 Authenticity in Contrast to Theoretical Overload

Table 4: Student Perceptions of Assessment Authenticity vs. Theoretical Overload

Aspect	Quotes
Prevalence of Concern	Theme emerged in 73.4% of open-ended responses (47/64) and all five follow-up interviews.
Core Problem	Assessments are too theoretical and rote-based and do not prepare students for the unpredictability of real classrooms.
Issue 1: Theory vs. Practice	The student aced the written test on lesson preparation but could not handle an actual class of 30 students with behavior problems.
Issue 2: Memorized Principles	A student passed an exam on the concepts of authentic evaluation, but failed to apply them in a practical speaking examination.
Issue 3: Fake vs. Real Classroom	Role-playing classmates as 'ideal students' did not prepare for genuine student behaviors (interruptions, unexpected inquiries, quick English).
Issue 4: Unforeseen Circumstances	Student 'went blank' after surprise question; examinations never asked for them to respond to unpredictability.
Student-Desired Change	Assessments should be real challenges: student perplexity, tech failures, unexpected replies, etc.
Conclusion	Assessments of teacher education need to be rethought to help close the gap between theory and practice and get ready for the challenges of real classrooms.

Qualitative data suggested a held view among students that present assessments are insufficiently grounded in the reality of what is required for authentic classroom teaching, tending towards being overly theoretical / dependent on more rote approaches to learning. This theme was pervasive across the qualitative data, appearing frequently in open-ended responses and all five follow-up interviews. While frequencies are not the primary measure of importance in qualitative analysis, for context, this theme appeared in 47 of 64 open-ended responses, suggesting that theory-practice gap was a pervasive concern felt throughout the student body. Quantitative data supported this concern. When asked about assessment authenticity, 47 students (73.4%) agreed / strongly agreed that *'current assessments are too theoretical and do not prepare me for real classroom challenges'* ($M = 4.02$, $SD = 0.87$). Besides, 42 students (65.6%) disagreed / strongly disagreed that *'peer teaching demonstrations accurately reflect real classroom conditions'* ($M = 3.82$, $SD = 0.92$). One interviewee (P02) described clearly, *'Most of our assessments come in the form of written exams about pedagogies. I can explain the steps for lesson planning and pass it off on an examination. I can do all of that. That alone does not mean that I can walk up to a class of thirty different students, each with their own set of behavioral challenges, and teach them something useful. The exam is a memory test. It tests what you can remember, not what you can do.'* A questionnaire respondent (Q12) expanded, *'In one of my language assessments courses last*

semester, the final exam focused on expectations like what are some principles of authentic assessment and why each principle is needed. I wrote down everything to textbook perfection and scored well. On my teaching practicum at a junior high school, I was going to have to test my students' speaking skills, and I had no idea how to put any of those memorized ideas to use. All I knew was the theory, with no knowledge of how to actually apply it in practice. The test was designed to get me through a test and not to teach.' Another interviewee (P05) elaborated, 'The teaching practice assessments done in this program usually entail my classmates role-playing as students. They sit up straight, they stretch their hands into the air without making a sound, they never cut each other off mid-sentence nor act out, and they play into your idea of an ideal student who puts you on a pedestal. In my real teaching practicum in a school, students had conversations with each other, asked me questions I hadn't thought of, and one student even told me that my English was too fast. None of my assessments prepare me at all for any of these moments.' A questionnaire respondent (Q38) said, 'In my practicum when I was working with students, one of them raised his hand and asked me a question I never saw coming. I knew the answer but went blank. I would not be able to explain in English in a way the students would understand. None of the exams I had ever taken forced me to respond to unforeseen circumstances, yet now I was expected to handle those moments. Everything about my assessments was predictable, planned, and rehearsed. Teaching is not like that.' Another interviewee (P03) said: 'I want assessments that challenge me. For example, a student who says 'I do not get it.' A projector that stops working. A student who gives an answer I did not expect. These are all things that happen every day in real schools, yet none of my assessments have asked me to grapple with any of them.' These students' stories highlight the need to re-evaluate teacher education evaluations, that actually prepares them for the challenges of the live classroom. Quantitative data supported the authenticity concern (73.4% agreeing that assessments are too theoretical; 65.6% disagreeing that peer teaching reflects real classrooms).

Qualitative data explained this concern by describing how written exams test rote memory and peer teaching demonstrations lack the unpredictability of real classrooms. The datasets converge on the finding that assessments lack authenticity. Complementarity is evident: the quantitative data show the extent of the concern, while the qualitative data provide concrete examples of what authenticity would look like (real challenges, unpredictable moments, genuine student behaviors).

4.5 Design Features Identified by Students

Table 5: Five Student-Identified Design Features for Integrated Assessment

No.	Feature	Request
1.	Dual-skill rubrics	Combine language & instructional elements in one rubric (90.6% agreed).
2.	Simulation tasks	Assess response to actual classroom uncertainty (interruptions).
3.	Portfolio-based	Replace high stakes tests with portfolios that indicate improvement over time.
4.	Feasible workload	More tasks that are shorter and more intensive for feedback and revision.
5.	Reflective practice	Include self-assessment and written reflection in the grading process.

Qualitative analyzes showed five features that identified evaluation as necessary for meaningful integrated assessment on students' educational. Finally, the most common request across 58 out of 64 responses (90.6%) was for dual-skill rubrics that incorporate demonstrations of both language and teaching criteria. The 90.6% agreement for dual-skill rubrics was derived from Likert-scale item

responses (Item 11: *'I want rubrics that combine language and teaching criteria in one scoring system'*), with 58 students selecting 'Strongly Agree' / 'Agree.' The other four features were identified through qualitative analysis: simulation tasks appeared in 41 open-ended responses (64.1%), portfolio-based assessment in 36 responses (56.3%), manageable workload in 38 responses (59.4%), and reflective practice in 29 responses (45.3%). One interviewee (P04) said, *'I want a rubric that has combined criteria, for example, 'uses language to clarify concepts' or 'modifies vocabulary for the student population' instead of adding separate columns. I also want to find out, within the same criterion, how my language choices impact my effectiveness as a teacher.'* Another interviewee (P01) expressed, *'The rubric should show me what integration looks like at various levels. That would show me what I am aiming for.'* Second, students asked for simulation-based assessment tasks that embrace the unpredictable nature of actual classrooms. A questionnaire respondent (Q27) said *'Do a mock lesson in which halfway through, the assessor indicates that a student is inattentive or that the projector has broken down.'* *'I want to see how I respond. That is exactly what happens in my classroom.'* One interviewee (P05) explained, *'An ideal assessment would be for me to teach for fifteen minutes with three unexpected things occurring during that time. The assessor would ask a really difficult question. Another student would provide an answer I never saw coming. The assessor would then ask me to speed up because we had very little time left. I want to see if I could cope with those moments using my English alongside my teaching ability.'* Third, portfolios should be the basis of assessment rather than high-stakes testing. A questionnaire respondent (Q44) said: *'An evaluation based on one class should not be capable of forecasting what the future holds for me. What if I am sick? What if I am nervous?'* Another questionnaire respondent (Q51) also suggested that portfolios could both showcase the best of their work and demonstrate progress made, adding: *'Portfolios would allow me to show my best work and also demonstrate how I have improved. One example could be showing a video from the beginning of the semester and then another video from the end to compare my progress. That would be fairer and more motivating.'* Fourth, students called for a workload that is feasible and enables proper interaction with assessment tasks. One interviewee (P02) said: *'All these requests for integrated assessment will not work if we have six courses each with three assignments. I cannot produce integrated work when I am running on fumes and trying to cram too many things into my schedule. You have to reduce the workload, or else none of this will matter.'* Another interviewee (P03) added: *'I prefer fewer assignments, with each one dealt with more deeply. I want time to get feedback, think about it, revise, and resubmit. Right now there is no time for that because the next assignment is already due.'* Fifth, students wanted reflective practice incorporated into the way they are assessed. A questionnaire respondent (Q19) said, *'After I teach, I want to write a reflection on what went well and what did not go as planned, and detail how I would modify my language and teaching the next time around. Because that reflection shows whether I can learn from my own experience, it should be part of my grade.'* Another questionnaire respondent (Q33) added, *'Self-assessment would compel me to honestly evaluate my strengths and weaknesses. Rather than just having lecturers lecture me, if I can pinpoint my own issues, that means I am capable of critiquing my own practice.'*

Those five features identified depict what pre-service English teachers need in integrated assessment to be truly educationally meaningful, and not just in name only. Quantitative data showed strong agreement for dual-skill rubrics (90.6% from Likert-scale items). Qualitative data identified four additional features (simulation tasks, portfolio-based assessment, manageable workload, reflective practice) and provided descriptions of what each feature would look like in practice. The datasets complement each other. The quantitative data establish the priority of dual-skill rubrics, while the qualitative data expand the vision of integrated assessment to include other features that students consider essential.

4.6 The Tension Between Workload and Integration

Table 6: Workload as a Barrier to Integrated Assessment

Issue	Prevalence	Student Quote (Abridged)
Overwhelming volume	59.4% replies + 5 interviews.	'18 to 30 assignments per semester... no time to think.'
Grouped deadlines	Common unsolicited theme	'7 large assignments due in 5 days... no real learning.'
Shallow outcomes	Common complaint	'System values quantity over quality.'
Weakened mental powers	A shared feeling	'No mental space left to care about integration.'

A finding that was not anticipated in the design of the original research emerged from the qualitative data. Students find themselves overwhelmed by the sheer volume of assignments they must complete, and this load undermines any given assessment, integrated / not, in terms of its perceived quality due to its weight. The workload theme emerged unsolicited in a substantial portion of the qualitative data, appearing in 38 of 64 open-ended responses and all five interviews. The fact that participants raised this issue without prompting indicates its salience as a structural concern. Quantitative data further supported this concern: When asked about workload, 41 students (64.1%) agreed / strongly agreed that *'the number of assignments makes it difficult to engage deeply with any single assessment'* ($M = 4.16$, $SD = 0.84$). Besides, 38 students (59.4%) agreed / strongly agreed that *'deadlines are often compressed at the end of the semester, leaving insufficient time for meaningful engagement'* ($M = 4.22$, $SD = 0.79$). One interviewee (P01) said, *'I have six courses this semester, and each course has at least three assignments. That means I am doing anywhere from eighteen to thirty assignments in one semester. I am in a hurry to get everything done. I have no time to think about any single assessment, no matter how well prepared it is. I am fatigued. You could hand me the perfect integrated evaluation, and I would still lack the mental energy to engage with it.'* Another interviewee (P04) said, *'The issue involves both the number of tasks and the deadlines. For the first two months, nothing is due. Then suddenly in the last two weeks, everything is due at once. I have had weeks where seven large assignments are due in five days. I cannot do good work under those conditions. I pull all-nighters until 2 AM just to finish, with no real learning taking place.'* A questionnaire respondent (Q56) added: *'With so many assignments, no one does well on anything. Because of this rush to complete grading, lecturers give generic feedback. Students feel rushed, and while the work gets finished, it stays shallow. Integration will not help when everyone feels exhausted and hurried. The system values quantity over quality.'* Another student explained, *'I want to do well on integrated assessments. Give me clear rubrics and integrated feedback. Right now, my focus is simply on making it through the semester and turning everything in on time. When I feel overwhelmed by papers, I have no mental space left to care about integration.'* The idea, then, is that if lecturers want students to engage with integrated assessment, they first need to reduce some of the workload. In other words, they need to lower the number of assignments. A questionnaire respondent (Q62) added: *'I have spoken with my friends, and they share the same experience. Let us be honest. This is a national issue in higher education. Any attempt to improve assessment quality will fail without addressing student workload, because students will be too busy drowning to gain any benefit.'* It highlights the fact that unless workload-integration tension is managed, any integrated assessment model will be dysfunctional and ineffective. Quantitative data supported the workload concern (64.1% agreeing that assignments hinder deep engagement; 59.4% agreeing that deadlines are compressed).

Qualitative data explained this concern by describing the overwhelming volume of assignments (18-30 per semester), compressed deadlines, and the lack of mental energy to engage meaningfully with any single assessment. The datasets converge on the finding that workload is a structural constraint that undermines assessment quality. Complementarity is evident; the quantitative data show the prevalence of the concern, while the qualitative data reveal the mechanisms through which workload undermines engagement (fatigue, time pressure, superficial completion).

5. Discussion

The results from this study expose a paradox that lies at odds with dominant views about language teacher assessment in the literature. Although much of the research conducted on this issue appeared to conceptualize it as a problem of integrated assessment models not being used (Giraldo, 2021). The data collected through this study support that pre-service English teachers are already participating in these holistic assessments. What students are reporting as missing is not integration per se, but the quality, transparency of that integration, and purposeful integration. The implication of this shows questioning the assumption that assessment fragmentation is normative across contexts at a level not substantiated by empirical evidence (Afshar & Movassagh, 2021). This refers to the examples bringing out this disconnectedness between task and rubric as ‘coupling feedback’ teachers with unravelled rubrics, stressing that although teacher educators have responded to calls for integration by creating integrated tasks, they have not simultaneously reconfigured their rubrics (which leads to a conundrum where students see no correlation between what they are being asked to do and how their performance is gauged). This study has identified the presence-quality paradox, which supports the argument introduced by (Pastore & Andrade, 2019) that assessment literacy goes beyond creating assessments to being able to communicate assessment criteria transparently to students. When rubrics separate language and teaching into two columns, when feedback addresses each domain in isolation, the integration that is inherent to the task disappears for students, negating the formative promise of assessment. This results in a system of more integrated assessment for students who report feeling less integrated than ever, generating implications for how teacher education programs assess their own practices surrounding assessment.

The rubric clarity deficit described in this study constitutes an obstacle to the development of integrated teaching competence. That students receive separate scores for the language and pedagogy dimensions in current rubrics that do not indicate how linguistic choices affect effectiveness pedagogically (Faez et al., 2021). When rubrics treat language and teaching separately, they send the message to pre-service teachers that these two dimensions are separate, a complete contradiction to everything about how language works in classroom instruction. In this study, students want more rubrics / more expansive rubrics, and they wanted rubrics that explicitly show the link between language and teaching. They were interested in whether the use of complex sentences decreases the clarity of instructions / whether simplification increases student participation. This result underscores the need for a holistic rather than merely analytic checklist rubric design to connect criteria so that they integrate language with pedagogy more closely aligned with actual classroom practices (Scarino, 2013). Pre-service teachers need to know what to teach and how their choices of language impact the success of their teaching (Lindahl, 2019). Without rubrics that make these links obvious, students are forced to guess how to improve and often focus on the wrong areas of their performance because the rubric does not give guidance on the relationship between language and pedagogy. The near-universal call for integrated rubrics by students in this study should be a signal to teacher educators that present rubric designs are not meeting the demands of learners.

The feedback integration gap identified in this study shows a similar pattern of fragmentation that appears to have been under-appreciated in the literature. Students reported receiving separate feedback from different lecturers / in disparate sections of feedback forms, with no comments addressing how language use enables or constrains effective teaching (Xu & Brown, 2016). Feedback that compartmentalizes language and pedagogy implies an unsupported view of the independence of these two dimensions, whereas teaching in the classroom disobeys this separation. Students in this study expressed a nuanced sense of what they required, feedback that explains that their directions were unclear not due to poor use of vocabulary but rather because their sentence structures were too complex for the students' level. Feedback of this nature necessitates that evaluators draw relationships between pedagogical function and linguistic form, a capacity that might also have to be explicitly entrenched in teacher educators (Yan & Fan, 2021). Although teacher educators tend to have strong content knowledge, many lack the confidence/skill to create assessments that reflect the language-pedagogy nexus (Loeneto et al., 2022; Mistar, 2014) and this shortfall applies more generally to feedback practices. This study finds that, even when assessors provide unambiguous feedback within each domain, they regularly neglect to mention how the domains operate in practice and leave students to piece together these critical links themselves.

This study describes the authenticity problem as another instance of a systematic tension between theory and practice in teacher education that has been reported across other Indonesian contexts (Hunaepi & Suharta, 2024). Students said that many of their assessment tasks were written exams in which they had to rote-learn theories about teaching, and the peer teaching demonstrations were done with peers who complied with what they wanted, rather than reflecting a real classroom. The conclusion is especially alarming due to the expectations of the Merdeka Curriculum, which expects teachers to continue with independent approaches in teaching and evaluation (Priawasana & Subiyantoro, 2024). Assessments that are more about rote regurgitation than functional application of material leave pre-service teachers under-prepared for the unpredictability of a real (potentially chaotic) classroom. The students in this study expressed a strong preference for simulation-based assessments that also incorporated actual classroom stimuli such as disengaged students, technology failures and questions that were unanticipated (Munandar & Newton, 2021). Previous research on authentic assessment in Indonesian contexts has consistently found large class sizes, limited resources, and cultural interaction patterns to impede the implementation of authentic assessment (Sabtiawan et al., 2019; Mariappan & Osman, 2023), but this study suggests that students themselves see value in simulation-based tasks as an authentic mode of assessing their preparedness for the classes they will face. Misalignment between task performance in student assessments and actual teaching practice generates an image of unprepared teachers that jeopardizes the development of their confidence and professional identity (Siahaan & Subekti, 2021), leading them to believe they had been trained only to pass exams, but not to teach.

Those features, in an integrated assessment, dual-skill rubrics, simulation-based tasks, portfolio-based assessment, manageable workload and embeddedness of reflective practice, did represent a vision declaring the recommendation for reforming assessments that fits well with contemporary Indonesian educational policy (Mufanti et al., 2024). Dual-skill rubrics alleviate the transparency problem by exposing the integration to learners. Simulation-based assignments try to solve the authenticity problem by asking students to show integrated competence in authentic settings (Darling-Hammond, 2017). Sato (2014) said that time of actual collection and multiple performances are required to assess their reliability and validity, therefore portfolio-based assessment settles these problems compared with a single high-stakes activity. Embedding reflective practice within assessment responds to the risk inherent in the developmental aim of enabling students to self-evaluate and self-improve teaching endeavours (Korthagen, 2017). These features

align with the principles of Outcome-Based Education (OBE), which has been implemented at Indonesian higher education to meet the requirements of the Indonesian Qualification Framework (Aminah et al., 2025). Yet, the present study shows that principles of OBE have mostly been integrated into curriculum design with much less focus on how assessment practices set those principles in motion (Asbary & Nurhayati, 2024). Students are requesting assessment practices aligned with the integrated, authentic, and developmental tenor of OBE, but assessments continue to be incoherent and single-event structures. This disconnect between what policies aspire to achieve and the realities of teacher education presents an implementation gap that will need to be addressed if national qualification frameworks are ever to achieve their aim with respect to on-the-ground teacher preparation.

The workload-integration tension, an emergent finding and one of the structural constraints, has been under-appreciated in the integrated assessment literature. Some of the students mentioned feeling overwhelmed because they had too many assignments to cope with, having multiple deadlines compressed at the end of the semester, and being left insufficient time and energy to take due care engaged with any one assessment (Abdillah & Fithriani, 2023). This finding implies that even the best-crafted integrated assessment model will not make a difference if students lack the cognitive and emotional capacities to interact with it effectively. Excessive workload is a well-known issue in Indonesian higher education (Multamasatika et al., 2023), where previous studies have found that institutional challenges such as large class sizes, time constraints on teaching agendas and activities, less-favourable facilities for the conduct of educational practice, and lack of social contexts in the workplace contribute to university teachers' perceptions of job stress. However, the current study is situating itself within this literature and goes further by exposing how workload can directly ruin assessment quality as perceived by students. One of the many issues is that when students are rushing to do twenty / thirty assignments in one semester, they are not able to respond with quality pieces of work, and they cannot be benefitted adequately from feedback (Abdillah & Fithriani, 2023). If students are far too worn out to tell/need integration when the number of assessment tasks is very high, integration does not matter. This result is noteworthy for program design because it points to the fact that improving assessment quality demands the more challenging work of designing better rubrics and tasks, as well as a decrease in assessments and a spread of deadlines throughout a semester. Efforts to improve integrated assessment will be carried out in an environment that systematically reduces their effectiveness unless the structural constraint is broken, a point well recognized and highlighted by students themselves in their responses.

6. Conclusion

This study is aimed at finding out the attitudes of pre-service English teachers towards the differences between what integrated assessment exists and how it happens in their teacher education program. The findings show a paradox that contradicts current beliefs in the language testing literature. Students in this study claimed that they had already experienced evaluation tasks combining language and teaching skills, contrary to the prevalent narrative of artificial dualism. All students reported being routinely assessed in both domains, having faced integrated assessments, and knowing how their language skills directly inform their classroom practice. Furthermore, the majority retorted that their assessments did not at all feel separate and disconnected. But despite this appearance of integration, students repeatedly reported that the integration fell short of being meaningful/adequate. The quality gap was identified in three dimensions. The first that appeared was a rubric clarity deficit: there is a demand for rubrics that allow for a demonstrated relationship between language and teaching to be assessed together rather than in separate columns. Students indicated that current analytic rubrics, which cleave language criteria from pedagogy criteria, actively sustain the very divide that integrated tasks are meant to eliminate. Second, it is found that the gap

between feedback on language and pedagogical content knowledge shows that students were getting feedback on specific aspects of each, but not on how they are related in practice in school. They were almost unanimous on wanting feedback that explains, pay attention, what the construction of their sentences was that affected whether they explained instructions clearly, what vocabulary choices made it easier / harder for the students to understand. Third, the issue of authenticity that emerged from this study showed how students viewed present assessments as lacking in authenticity. They are perceived to be highly theoretical and scripted for written examinations, which measure rote learning memory, and problematic teaching demonstrations where students are role-played with compliant peers, ineffective in preparing them for the unpredictable nature of real classrooms.

This study also found five characteristics that students care about when determining what constitutes an educationally meaningful integrated assessment. The students asked for: dual-skill rubrics (integrated criteria), simulation-based assessment tasks that include real classroom triggers such as disengaged students and technical glitches, portfolio-based assessment, the basis for capturing growth over time rather than one-off high-stakes activities, a manageable workload allowing deep engagement with each assessment, and reflective practice embedded within assessment inclusive of self-assessment and commentary. Another unexpected but required discovery was the workload-integration stress: students were overwhelmed by the final number of assignments needing to be completed in compressed deadlines at semester's end. What students said very clearly was that if they are too tired and time-pressured to be engaged, no amount of integrated assessment design is going to make a difference. Highly relevant to program design, this finding implies that improvements in assessment quality will require better rubrics and tasks themselves, but also fewer total assessments and a more even distribution of deadlines throughout the semester. This conclusion has implications for program design. Improving assessment quality involves establishing better rubrics and activities, as well as lowering the total number of assessments and spacing out due dates more equally throughout the semester. The study concludes that the problem in language teacher assessment is not the lack of integrated assessment, but the quality of integration as perceived by learners. Closing the gap in quality requires simultaneous attention to rubric design, feedback practices, authenticity, workload and reflective practice. Future research is needed to see how these findings generalize to other Indonesian teacher education contexts and the impact of interventions aimed at diminishing the quality gaps identified in this study.

References

- Abdillah, M. G., & Fithriani, R. (2023). Indonesian EFL Pre-service Teachers' Experiences: Revealing English Instructional Challenges during Teaching Practice. *REiLA: Journal of Research and Innovation in Language*, 5(1). <https://doi.org/10.31849/reila.v5i1.13228>
- Afshar, H. S., & Movassagh, H. (2021). Towards a critical language teacher identity: Contributions of a critical teacher education course. *Applied Research on English Language*, 10(2). <https://doi.org/10.22108/are.2021.127277.1682>
- Aminah, S., Alfa Krisnadhi, A., & Nizar Hidayanto, A. (2025). Ontological Framework for the Analysis of Outcome-Based Curriculum in Higher Education. *IEEE Access*, 13. <https://doi.org/10.1109/ACCESS.2025.3542881>
- Asbari, M., & Nurhayati, W. (2024). Outcomes-Based Education in Indonesian Higher Education: Empowering Students' Learning Competencies. *International Journal of Social and Management Studies*, 5(5).
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2). <https://doi.org/10.1191/1478088706qp063oa>

- Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative research in sport, exercise and health*, 11(4), 589-597.
- Creswell, J. W., & Clark, V. P. (2007). *Mixed methods research*. Thousand Oaks, CA.
- Darling-Hammond, L. (2017). Teacher education around the world: What can we learn from international practice? *European Journal of Teacher Education*, 40(3). <https://doi.org/10.1080/02619768.2017.1315399>
- Faez, F., Karas, M., & Uchihara, T. (2021). Connecting language proficiency to teaching ability: A meta-analysis. *Language Teaching Research*, 25(5). <https://doi.org/10.1177/1362168819868667>
- Field, A. (2024). *Discovering statistics using IBM SPSS statistics*. Sage Publications Limited.
- Giraldo, F. (2021). A reflection on initiatives for teachers' professional development through language assessment literacy. *Profile: Issues in Teachers' Professional Development*, 23(1). <https://doi.org/10.15446/profile.v23n1.83094>
- Hunaepi, H., & Suharta, I. G. P. (2024). Transforming Education in Indonesia: The Impact and Challenges of the Merdeka Belajar Curriculum. *Path of Science*, 10(6). <https://doi.org/10.22178/pos.105-31>
- Korthagen, F. (2017). Inconvenient truths about teacher learning: towards professional development 3.0. *Teachers and Teaching: Theory and Practice*, 23(4). <https://doi.org/10.1080/13540602.2016.1211523>
- Lindahl, K. (2019). Teacher Language Awareness Development and Its Implications for New Educators. *New Educator*, 15(2). <https://doi.org/10.1080/1547688X.2018.1526356>
- Loeneto, B. A., Alwi, Z., Ernalida, E., Eryansyah, E., & Oktarina, S. (2022). Teacher Education Research and Development in Indonesia: Preparing Educators for the Twenty-First Century. In *Handbook of Research on Teacher Education: Innovations and Practices in Asia*. https://doi.org/10.1007/978-981-16-9785-2_10
- Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing research*, 35(6), 382-386.
- Macaro, E., Curle, S., Pun, J., An, J., & Dearden, J. (2018). A systematic review of English medium instruction in higher education. *Language Teaching*, 51(1). <https://doi.org/10.1017/S0261444817000350>
- Mariappan, V. A., & Osman, K. (2023). A Systematic Literature Review of Authentic Assessment in K-12 ESL/EFL Education. *Malaysian Journal of Social Sciences and Humanities (MJSSH)*, 8(5). <https://doi.org/10.47405/mjssh.v8i5.2303>
- Mistar, J. (2014). Teaching English as a foreign language (TEFL) in Indonesia. In *Teaching English to the World: History, Curriculum, and Practice*. <https://doi.org/10.4324/9781410612861>
- Mufanti, R., Carter, D., & England, N. (2024). Outcomes-based education in Indonesian higher education: Reporting on the understanding, challenges, and support available to teachers. *Social Sciences and Humanities Open*, 9. <https://doi.org/10.1016/j.ssaho.2024.100873>
- Multamasatika, D. F., Akbar, K. A., Hartanti, R. I., & Kallawicha, K. (2023). Work factors affect work stress among lecturers: a study in Indonesia. *Nevrologiya, Neiropsikhiatriya, Psikhosomatika*, 15(4). <https://doi.org/10.14412/2074-2711-2023-4-12-16>
- Munandar, M. I., & Newton, J. (2021). Indonesian EFL teachers' pedagogic beliefs and classroom practices regarding culture and interculturality. *Language and Intercultural Communication*, 21(2). <https://doi.org/10.1080/14708477.2020.1867155>
- Pastore, S., & Andrade, H. L. (2019). Teacher assessment literacy: A three-dimensional model. *Teaching and Teacher Education*, 84. <https://doi.org/10.1016/j.tate.2019.05.003>
- Priawasana, E., & Subiyantoro, S. (2024). Evaluating the K-13 Versus Merdeka Curriculum : Impacts on Primary, Junior, and Senior High School Education in Indonesia. *Jurnal Kependidikan: Jurnal Hasil Penelitian Dan Kajian Kepustakaan Di Bidang Pendidikan, Pengajaran Dan Pembelajaran*, 10(3). <https://doi.org/10.33394/jk.v10i3.12060>

- Sabtiawan, W. B., Yuanita, L., & Rahayu, Y. S. (2019). Effectiveness of Authentic Assessment: Performances, Attitudes, and Prohibitive Factors. *Journal of Turkish Science Education*, 16(2). <https://doi.org/10.12973/tused10272a>
- Sato, M. (2014). What Is the Underlying Conception of Teaching of the edTPA? *Journal of Teacher Education*, 65(5). <https://doi.org/10.1177/0022487114542518>
- Scarino, A. (2013). Language assessment literacy as self-awareness: Understanding the role of interpretation in assessment and in teacher learning. *Language Testing*, 30(3). <https://doi.org/10.1177/0265532213480128>
- Siahaan, S. M. Y., & Subekti, A. S. (2021). Pre-Service English Teachers' Professional Identity Development: A Case Study. *Metathesis: Journal of English Language, Literature, and Teaching*, 5(2). <https://doi.org/10.31002/metathesis.v5i2.4231>
- Vygotsky, L. S., & Cole, M. (1978). *Mind in society: Development of higher psychological processes*. Harvard University Press.
- Xu, Y., & Brown, G. T. L. (2016). Teacher assessment literacy in practice: A reconceptualisation. *Teaching and Teacher Education*, 58. <https://doi.org/10.1016/j.tate.2016.05.010>
- Yan, X., & Fan, J. (2021). 'Am I qualified to be a language tester?': Understanding the development of language assessment literacy across three stakeholder groups. *Language Testing*, 38(2). <https://doi.org/10.1177/0265532220929924>