

INTEGRATING AUTHENTIC AND MODEL-BASED ASSESSMENT IN ENGLISH LANGUAGE LEARNING: A COMPREHENSIVE EVALUATION FRAMEWORK FOR HIGHER EDUCATION

Saiyidinal Firdaus
Univesitas Negeri Jakarta
Saiyidinalfirdaus1995@gmail.com

ABSTRACT

Current debates in higher education English language assessment center on the tension between pedagogically valued authentic assessment and the growing demand for transparency, defensibility, and program-level accountability. While authentic tasks are widely promoted for representing real-world language use, their capacity to support valid inferences and systematic evaluation remains contested. This study aims to develop and empirically examine an integrated evaluation framework that combines authentic assessment and model-based assessment within a higher education English language program. Using a qualitative case study design with a program evaluation orientation, data were collected from authentic assessment tasks, student performance artifacts, analytic scoring rubrics, and program evaluation documents. Data were analyzed through construct alignment analysis, evidence modeling based on Evidence-Centered Design, and argument-based validation to examine how assessment evidence supports defensible interpretations and program-level decision-making. The findings show that authentic tasks generate rich and multidimensional language performance evidence; however, such evidence becomes evaluatively meaningful only when embedded within explicit construct definitions, evidence models, and validity arguments. Partial alignment was observed at the task and course levels, while coherence at the program level remained limited due to the absence of systematic integration. The study offers a practical contribution by providing a comprehensive framework that enables institutions and educators to align classroom-based authentic assessment with program-level evaluation and decision-making processes. This framework supports the development of assessment systems that are pedagogically meaningful, inferentially defensible, and institutionally actionable in increasingly complex and data-driven higher education contexts.

Keywords: *Authentic assessment; evidence-centered design; argument-based validity; program evaluation; higher education English learnig*

Higher education English language learning is being pushed in two directions at once. On the one hand, universities are under increasing pressure to demonstrate that students can use English in consequential, real-world academic and professional tasks (e.g., presenting research, negotiating meaning in teams, producing discipline-specific writing). This has reinvigorated interest in authentic assessment—assessment that asks learners to perform contextualized tasks resembling target-domain language use and that can foster higher-order and transferable competencies. Recent syntheses in higher education consistently show that authentic assessment is associated with the development of 21st-century capabilities (e.g., problem solving,

collaboration) while also raising practical concerns around implementation, workload, and staff capacity (Vlachopoulos & Makri, 2024; Ajjawi et al., 2024; Boud & Bearman, 2024; Zhan et al., 2025; Fawns et al., 2025). On the other hand, universities are simultaneously facing demands for comparability, defensibility, and actionable evidence at scale—demands intensified by digital learning environments, expanding student diversity, and growing expectations for accountability. This second pressure points to the need for model-based assessment approaches that make the reasoning from performance to claims about ability explicit and testable, rather than relying on implicit judgment alone. Evidence-centered approaches and argument-based validity traditions offer structured ways to articulate constructs, specify evidence, and justify score interpretation and use, increasingly including AI-mediated assessment contexts (Newton et al., 2021; Nguyen et al., 2023; Dai et al., 2024; Zapata-Rivera et al., 2024; Kubsch et al., 2024). This study is particularly important in the current higher education context, where institutions are increasingly required to demonstrate accountability, transparency, and defensible assessment practices at program level. However, existing assessment approaches remain fragmented, often failing to systematically connect authentic performance with explicit inferential reasoning and evaluative decision-making.

Despite their shared goal—credible inferences about learners' English ability—authentic and model-based assessment traditions have often developed in parallel. Within higher education, authentic assessment scholarship has advanced from listing “real-world tasks” to more critical conceptions of authenticity that foreground relationality, inclusion, academic integrity, and the limitations of treating authenticity as a universal solution (Boud & Bearman, 2024; Ajjawi et al., 2024; Zhan et al., 2025; Fawns et al., 2025; Hu et al., 2025). In digital contexts, assessment design is now understood as inherently sociotechnical: digital tools can enhance feedback and efficiency, but they can also reshape what is valued, what is visible as “evidence,” and whose participation is supported (Bearman et al., 2023; Buchan et al., 2024; Boud & Bearman, 2024; Hu et al., 2025; Nieminen et al., 2025). At the same time, model-based assessment work—particularly evidence-centered design (ECD) and argument-based validation—has refined methods for aligning tasks, scoring, and claims, and for evaluating whether the inferences drawn from assessment data are warranted (Newton et al., 2021; Nguyen et al., 2023; Tubis, 2023; Dai et al., 2024). Yet, in many university English programs, authentic assessment is implemented as a set of preferred formats (projects, portfolios, presentations) without a fully explicit evidentiary logic, while model-based assessment is sometimes operationalized as psychometric rigor without sufficient attention to authenticity, learner experience, and target-domain consequentiality.

A second gap concerns program evaluation. In higher education language education, assessment innovations are frequently evaluated at the level of a single course or instrument, while decisions and accountability operate at the level of programs and institutions. Program-level assessment requires coherence across courses, sequencing across time, and evidence that supports improvement and decision-making—not merely “grading.” (Charlton & Newsham-West, 2024; Krooi et al., 2024; Hadzhikoleva et al., 2025). Program evaluation studies in tertiary English contexts show the value of comprehensive evaluation models such as CIPP (Context–Input–Process–Product) for diagnosing strengths and improvement points, but they also highlight how evaluation findings can remain disconnected from day-to-day assessment design and classroom evidence practices (Wakid et al., 2024; Imsa-ard, 2025; Huang, 2025). In other words, programs may gather abundant data (test scores, surveys, course pass rates) without a framework that (a) links authentic performance to defensible construct claims, (b) organizes evidence across micro (task), meso (course), and macro (program) levels, and (c) supports continuous improvement.

A third gap is capacity and interpretability among stakeholders. Higher education reforms repeatedly identify language assessment literacy (LAL) as crucial for staff and institutions, particularly as assessment becomes more digital and data-rich. Contemporary scholarship characterizes LAL as situated and evolving; universities increasingly need context-specific competence frameworks that include digital language assessment and ethical use of assessment information (Vogt et al., 2024; Lam & Gayton, 2025; Zhang et al., 2025; Liu et al., 2025). Without such capacity, authentic assessment risks inconsistency and contested standards, while model-based assessment risks becoming technocratic and opaque to teachers and learners.

These gaps have become more urgent with the rapid uptake of AI and analytics in language education. Reviews of automated scoring and emerging AI-assisted assessment tools show accelerating innovation alongside persistent validity, fairness, and transparency challenges (Jones et al., 2023; Liu et al., 2025; Li et al., 2025). Recent work proposes using evidence-centered frameworks to guide AI-based assessments—precisely because they force explicit alignment among constructs, tasks, evidence, and inference (Viswanathan et al., 2022; Kubsch et al., 2024; Zapata-Rivera et al., 2024; Barno & Phelps, 2025). However, higher education English programs still lack an integrative framework that positions AI and digital evidence as *supporting* (not replacing) principled authenticity, validity arguments, and program evaluation needs.

Against this background, the goal of this study is to develop a comprehensive evaluation framework for higher education English language learning that integrates authentic assessment and model-based assessment within a program-evaluation architecture. Specifically, this study proposes an approach that (1) treats authentic tasks as the primary site where target-domain performance is elicited and learning is supported, consistent with learning-oriented assessment emphases on feedback, engagement, and learner development; (2) treats model-based logic (e.g., ECD and argument-based validation) as the mechanism that makes claims, evidence, scoring, and consequences explicit and evaluable; and (3) embeds this integration into program-level evaluation cycles (e.g., CIPP-informed decision points and program-level assessment sequencing) so that evidence supports improvement, accountability, and equity across a degree pathway. By doing so, the study aims to contribute a framework that is simultaneously *instructionally meaningful* (authentic, learning-oriented), *inferentially defensible* (model-based validity logic), and *organizationally usable* (program evaluation and decision-making).

In sum, this study responds to a central challenge in contemporary higher education language education: how to design assessment systems that are authentic enough to represent consequential English use, rigorous enough to justify interpretations and decisions, and coherent enough to evaluate and improve programs over time. Building on recent advances in authentic assessment scholarship, learning-oriented assessment, language assessment literacy, argument-based validation, evidence-centered design, AI-enabled assessment, and program evaluation, the study positions “integration” not as mixing formats, but as aligning tasks, evidence models, validity arguments, and evaluation decisions into a single, auditable framework tailored to higher education English language learning.

Grounded in the proposed integration of authentic and model-based assessment within a program evaluation perspective, this study is guided by the following research questions: (1) How can authentic assessment tasks in higher education English language programs be systematically aligned with explicit evidence models and validity arguments to support defensible inferences about students’ language ability?; (2) In what ways does an integrated authentic–model-based assessment framework enhance coherence across task, course, and program levels within a comprehensive program evaluation structure?; (3) What types of assessment evidence (task

performance, scoring, feedback, and outcomes) are most informative for program-level decision-making when organized through an integrated evaluation framework?; (4) How does the proposed framework address persistent challenges in higher education English assessment, including consistency, transparency, and stakeholder interpretability, while maintaining pedagogical authenticity?. These questions collectively operationalize the study's central aim: to move beyond fragmented assessment practices toward an evaluative framework that supports learning, validity, and institutional decision-making simultaneously.

Despite sustained advances in authentic assessment, learning-oriented assessment, and model-based validation over the past decade, higher education English language programs continue to face a critical structural gap: the absence of an integrative framework that connects authentic performance, explicit inferential logic, and program-level evaluation in a coherent and usable system. Existing studies tend to privilege either pedagogical authenticity—often at the expense of transparent evidentiary reasoning—or psychometric and model-based rigor—often detached from the realities of classroom practice and curricular sequencing. Moreover, program evaluation research in language education has largely treated assessment data as inputs or outcomes, rather than as theoretically grounded evidence generated through principled assessment design.

This study makes a theoretical, methodological, and practical contribution by positioning integration as a *design logic* rather than a compromise between traditions. Theoretically, it advances applied linguistics and language assessment scholarship by demonstrating how authentic assessment can be reconceptualized as evidence-generating practice within an explicit validity argument, thereby bridging long-standing divides between sociocognitive, usage-oriented views of language use and model-based inferential frameworks. Methodologically, it extends program evaluation in language education by embedding evidence-centered and argument-based assessment logic within a comprehensive evaluation architecture, enabling systematic alignment across micro-level tasks, meso-level courses, and macro-level program decisions. Practically, the framework offers higher education institutions a defensible yet pedagogically meaningful approach to assessment reform—one that supports accountability, learning, and continuous improvement in increasingly digital and AI-mediated environments.

By articulating how authentic tasks, evidence models, and evaluation decisions can be aligned within a single framework, this study responds directly to current demands for transparency, rigor, and relevance in higher education English language assessment, and provides a scalable model for research, policy, and practice in applied linguistics contexts globally.

METHOD

1. Location

This study adopts a qualitative case study design with a program evaluation orientation, aimed at examining how authentic and model-based assessment can be integrated within a higher education English language program. The case study approach enables an in-depth exploration of assessment practices within a bounded institutional context, while the program evaluation perspective provides a structured lens for analyzing how assessment evidence is generated, interpreted, and used for decision-making across task, course, and program levels. This design is particularly appropriate for addressing complex, context-dependent assessment phenomena where pedagogical practices, inferential reasoning, and institutional evaluation intersect.

This study was situated in a higher education English language program at a comprehensive university, where English functions as a compulsory academic subject across multiple disciplines. The program is designed to support students' academic literacy, oral communication, and

discipline-relevant language use over a multi-semester sequence. The institutional context was selected purposively because it had recently initiated curriculum reform emphasizing authentic assessment practices, while simultaneously facing program-level demands for accountability, coherence, and evidence-based evaluation. This contextual positioning further supports the study's qualitative case study and program evaluation-oriented design, allowing for an in-depth and context-sensitive analysis of assessment integration.

2. Participants

Participants comprised undergraduate students enrolled in intermediate to advanced English courses, as well as English language instructors and program coordinators directly involved in assessment design and implementation. Student participants represented diverse disciplinary backgrounds (e.g., social sciences, education, engineering, and business), reflecting the program's cross-faculty role and ensuring ecological validity with respect to authentic academic language use. Instructors were selected based on their active role in implementing project-based, performance-oriented assessments and their involvement in assessment moderation and program review processes.

3. Data Collection Techniques

Data were collected through four main techniques designed to capture assessment practices, performance evidence, and evaluative reasoning within the integrated framework: 1) document analysis of assessment tasks, rubrics, and program materials; 2) performance assessment analysis of student outputs from authentic tasks; 3) rubric analysis to examine construct representation and scoring criteria; and 4) program evaluation document review, including syllabi, assessment maps, and moderation records. These techniques enabled systematic triangulation across multiple sources of evidence, ensuring that both pedagogical practices and evaluative processes were comprehensively captured.

A purposive sampling strategy was employed to ensure alignment between the research objectives and the data sources. At the student level, samples of assessed performances were drawn from core authentic tasks embedded in the curriculum, including academic presentations, project-based written reports, and integrated skills tasks. These tasks were selected because they were consequential for course outcomes and explicitly designed to approximate target-domain academic language use. At the instructor and program level, participants were selected to represent different roles in the assessment system (task designers, raters, and program decision-makers), allowing the study to capture multiple perspectives on assessment evidence and evaluation use. This sampling strategy supports analytic generalization rather than statistical representativeness, consistent with design-oriented and evaluation-focused research in applied linguistics.

4. Instruments

Multiple instruments were employed to capture assessment practices, performance evidence, and evaluative reasoning in line with the integrated framework. The primary instruments consisted of authentic assessment tasks, analytic scoring rubrics, evidence and claim specifications, and program evaluation documentation.

First, authentic assessment tasks functioned as the central evidence-eliciting instruments. These tasks were embedded in regular coursework and required students to perform extended, contextualized language use aligned with academic target domains (e.g., presenting research-based arguments, synthesizing sources in discipline-related writing, and participating in collaborative problem-solving). Task specifications documented task purpose, target constructs, expected

language functions, and conditions of performance to ensure transparency and consistency across courses.

Second, analytic scoring rubrics were used to evaluate student performance. Rubrics were developed or refined to align with the constructs articulated in the framework, including dimensions such as communicative effectiveness, disciplinary appropriateness, linguistic control, and strategic language use. Each rubric criterion was explicitly linked to claims about language ability, thereby operationalizing model-based assessment principles within authentic task contexts. Rater guidelines and exemplars were provided to support consistent interpretation and scoring.

Third, evidence and claim specifications were developed as part of the model-based component of the framework. These specifications articulated the intended inferences from observed performance to underlying language abilities, identifying what counts as relevant evidence, how it is observed, and how it supports particular claims. This documentation drew on evidence-centered design logic and argument-based validation traditions, enabling systematic examination of the alignment among tasks, scoring, and interpretive use.

Finally, program evaluation instruments included course syllabi, assessment maps, moderation records, and program review documents. These artifacts were used to trace how assessment evidence was aggregated and interpreted at the program level, and how it informed curriculum decisions, quality assurance processes, and improvement initiatives. Together, these instruments enabled triangulation across performance, scoring, and evaluative decision-making.

5. Data Analysis

Data analysis followed a multi-stage, theory-informed procedure designed to operationalize the integration of authentic and model-based assessment within a program evaluation framework. The analysis was iterative and cyclical, reflecting the evaluative orientation of the study.

In the first stage, authentic assessment tasks and rubrics were analyzed qualitatively to examine construct alignment. Task specifications, rubric criteria, and instructional objectives were compared to identify correspondences and gaps between intended constructs and observable performance features. This stage focused on determining whether tasks genuinely elicited the kinds of language use claimed by the program and whether rubrics provided sufficient resolution to capture relevant evidence.

In the second stage, student performance data were analyzed using rubric-based scoring results and qualitative performance descriptors. Patterns of performance across tasks and courses were examined to identify consistencies and variations in evidence quality. Rather than treating scores as endpoints, the analysis interrogated what aspects of performance were being foregrounded as evidence and how these supported or constrained inferential claims about language ability.

The third stage involved constructing and examining validity arguments. Drawing on argument-based validation principles, claims, warrants, and backing were articulated for key assessment uses at both course and program levels. Evidence from task performance, scoring practices, and moderation processes was evaluated to assess the plausibility and robustness of these arguments. Potential threats to validity—such as construct underrepresentation, rater inconsistency, or task-specific effects—were identified and documented.

In the fourth stage, assessment evidence was synthesized within a program evaluation perspective. Using an evaluation logic informed by context–input–process–product reasoning, assessment data were examined in relation to program goals, instructional processes, and student

outcomes. This stage focused on how assessment evidence was used (or could be used) to inform curriculum alignment, instructional improvement, and decision-making across the program.

Finally, cross-source triangulation was conducted to enhance trustworthiness. Findings from performance analysis, rubric alignment, validity argument construction, and program documentation were compared to identify convergent and divergent patterns. Throughout the analysis, analytic memos were maintained to document interpretive decisions and to ensure transparency in how conclusions were derived from the data.

The methodological procedures outlined above were designed not merely to document assessment practices, but to generate analytically robust evidence for identifying recurring patterns in how authentic and model-based assessment interact across task, course, and program levels. By examining the alignment among task design, scoring rubrics, evidence specifications, and program evaluation artifacts, the analysis yielded a set of theoretically grounded themes that capture (a) how authenticity is operationalized as evidence-eliciting practice, (b) how inferential claims about language ability are constructed and constrained through model-based logic, and (c) how assessment evidence is aggregated and mobilized for program-level decision-making. Accordingly, the Findings section is organized thematically rather than instrumentally, reporting cross-cutting patterns that emerge from triangulated analyses of performance data, scoring practices, and evaluative documentation. These themes illuminate both points of convergence—where authentic assessment and model-based reasoning mutually reinforce validity and learning—and points of tension, where misalignment limits interpretability or evaluative use. In this way, the findings directly reflect the integrated methodological design and provide an empirical foundation for the subsequent discussion of theoretical implications, validity arguments, and program evaluation outcomes.

FINDINGS

This section reports the findings of the study thematically, reflecting the integrated analytical procedures described in the Method section. Drawing on triangulated analyses of authentic task specifications, scoring rubrics, student performance evidence, and program-level evaluation artifacts, four interrelated themes emerged. Together, these themes explicate how authentic and model-based assessment operate in practice, where alignment is achieved, and where tensions persist across task, course, and program levels.

1. Authentic Tasks as Primary Sites of Evidence Generation

Analysis of task specifications and student performance data revealed that authentic assessment tasks functioned as the principal sites for eliciting consequential English language use. Across courses, tasks were explicitly designed to approximate academic target domains, such as synthesizing scholarly sources, presenting discipline-relevant arguments, and collaborating in problem-solving contexts. These tasks generated rich, multidimensional performance evidence that extended beyond discrete linguistic features. Table 1 summarizes the dominant task types and the kinds of language evidence they elicited.

Table 1. Authentic Task Types and Elicited Evidence

Task Type	Target Domain	Primary Evidence Generated
Research-based oral presentation	Academic discourse	Argumentation, stance-taking, discourse organization
Project-based written report	Discipline-specific writing	Source integration, lexical precision, cohesion
Integrated skills project	Collaborative academic tasks	Interactional competence, strategic language use
Reflective learning portfolio	Metacognitive engagement	Awareness of language development and strategies

Performance excerpts illustrate how these tasks elicited complex language use aligned with program goals. For example, an excerpt from an oral presentation task shows evidence of both disciplinary framing and strategic language use:

“This study demonstrates that community-based learning not only enhances engagement but also reshapes students’ academic identities, particularly when reflective practices are embedded systematically.”

Such excerpts demonstrate that authenticity was not limited to surface realism but involved meaningful engagement with academic genres and epistemic practices. However, analysis also revealed variability in how explicitly tasks articulated their evidentiary purpose. In some cases, tasks emphasized “real-world relevance” without clearly specifying which aspects of performance constituted evidence for particular language constructs, foreshadowing alignment challenges addressed in subsequent themes. In response to Research Question 1, these findings indicate that authentic assessment tasks can support alignment with evidence models when their evidentiary purpose is explicitly articulated through clear construct–evidence relationships. However, the observed variability suggests that such alignment remains inconsistent in current practice, limiting the systematic interpretation of performance evidence.

2. Alignment Between Authentic Performance and Model-Based Claims

A central focus of the analysis was the degree to which authentic task performance aligned with explicit claims and evidence models articulated through rubrics and assessment documentation. Findings indicate partial but uneven alignment across courses.

Rubric analysis showed that most criteria corresponded broadly to intended constructs (e.g., communicative effectiveness, linguistic accuracy, disciplinary appropriateness). However, the inferential logic linking observed performance to claims about underlying language ability was often implicit rather than explicitly articulated. Table 2 presents an alignment analysis between task evidence and inferential claims.

Table 2. Task–Evidence–Claim Alignment

Construct Claim	Observable Evidence in Tasks	Alignment Strength
Academic argumentation ability	Thesis clarity, logical progression, stance markers	Strong
Disciplinary language use	Genre conventions, terminology	Moderate
Interactional competence	Turn-taking, responsiveness	Weak–Moderate
Strategic language use	Repair strategies, audience adaptation	Weak

While claims related to academic argumentation were consistently supported by observable evidence, constructs such as interactional competence and strategic language use were less systematically operationalized. For instance, although group projects generated interactional data, rubrics often foregrounded content and accuracy, marginalizing interactional evidence. This misalignment is illustrated in the following scoring comment:

“Good ideas and clear structure, but group interaction was not considered in scoring.”

These findings indicate that while authentic tasks generated rich evidence, the model-based mechanisms necessary to justify inferential claims were not uniformly embedded, limiting the defensibility of some assessment interpretations. These findings further address Research Question 1 by demonstrating that alignment is achieved only when rubric criteria explicitly operationalize construct claims and systematically capture relevant evidence from authentic performance. The absence of such explicit inferential links highlights the need for more rigorous evidence modeling within authentic assessment practices.

3. Coherence and Inconsistency Across Course-Level Assessment Practices

At the course level, analysis revealed both coherence and fragmentation in how assessment evidence was interpreted and used. Courses within the same program employed similar task formats (e.g., presentations, reports), suggesting curricular alignment. However, differences emerged in scoring emphasis, feedback practices, and interpretation of rubric criteria. Table 3 illustrates cross-course variation in rubric application.

Table 3. Cross-Course Variation in Rubric Emphasis

Course	Dominant Scoring Focus	Secondary Focus
Academic Speaking	Fluency and organization	Accuracy
Academic Writing	Source use and cohesion	Lexical range
Integrated English Skills	Task completion	Language form

Student feedback excerpts further reveal inconsistencies in evaluative messaging:

“In this course, grammar is very important, but in the previous course, the lecturer focused more on ideas.”

Such variation suggests that although authentic assessment was widely adopted, the absence of a shared model-based framework limited coherence across courses. This finding directly addresses the problem identified in the introduction regarding fragmented implementation of authentic assessment without program-level inferential alignment. Regarding Research Question 2, the findings suggest that an integrated authentic–model-based assessment framework enhances coherence by establishing shared constructs, aligned evidence models, and consistent evaluative criteria across courses. However, the current fragmentation observed across course-level practices indicates that such coherence cannot be achieved without systematic integration at the program level.

4. Program-Level Evaluation and the Use of Assessment Evidence

At the program level, findings indicate that assessment evidence was abundant but under-theorized in evaluative decision-making. Program documentation demonstrated systematic data collection (scores, pass rates, moderation records), yet the connection between authentic performance evidence and program decisions was often indirect. Table 4 summarizes how assessment evidence was used at different evaluation levels.

Table 4. Assessment Evidence Use Across Evaluation Levels

Level	Type of Evidence Used	Primary Purpose
Task	Performance artifacts	Grading and feedback
Course	Aggregated scores	Course improvement
Program	Pass rates, averages	Accountability reporting

Notably, performance-based evidence rarely informed discussions of construct coverage, longitudinal development, or equity across cohorts. Program coordinators tended to rely on numerical summaries rather than examining the quality of authentic performance evidence. This limitation is reflected in the following program review excerpt:

“Overall pass rates indicate satisfactory achievement, although detailed analysis of task performance was not conducted.”

These findings demonstrate that without explicit evidence models and validity arguments, authentic assessment data remain difficult to mobilize for meaningful program evaluation—confirming the need for the integrated framework proposed in this study. In relation to Research Question 3, the findings indicate that performance-based evidence—particularly qualitative task outputs and rubric-informed interpretations—emerges as the most informative for program-level decision-making when systematically organized within an explicit evaluative framework. These results highlight the importance of moving beyond aggregated numerical indicators toward richer, evidence-based interpretations of student performance.

Taken together, the findings address the problem formulations outlined in the introduction. Authentic assessment tasks effectively elicited consequential language use, confirming their pedagogical value. However, the absence of consistently articulated model-based inferential logic limited alignment, coherence, and evaluability. While partial integration was observed at the task and course levels, full integration across program-level evaluation remained underdeveloped. Overall, in response to Research Question 4, the integrated framework addresses persistent challenges of consistency, transparency, and stakeholder interpretability by making the inferential logic of assessment explicit, while preserving the pedagogical authenticity of task design.

Thematic analysis thus confirms that authenticity alone is insufficient to support defensible and actionable assessment systems in higher education English programs. Instead, authenticity must be embedded within explicit evidence models and evaluation frameworks to realize its full potential for learning, validity, and institutional decision-making. These findings provide the empirical foundation for the Discussion section, where they are interpreted in relation to contemporary theories of assessment, validity, and program evaluation.

DISCUSSION

The study set out to address a persistent and under-theorized problem in higher education English language education: how authentic assessment, widely promoted for its pedagogical and ecological value, can be rendered *inferentially defensible, coherent across programs, and evaluatively actionable* through integration with model-based assessment and program evaluation frameworks. The findings demonstrate that while authentic tasks successfully elicit consequential language use aligned with academic target domains, authenticity alone does not guarantee validity, coherence, or evaluability. Rather, authenticity must be embedded within explicit evidentiary and evaluative logics if assessment systems are to support learning, justify interpretations, and inform program-level decision-making.

From a theoretical perspective, the findings complicate simplified notions of authenticity that equate “real-world tasks” with assessment quality. Consistent with contemporary critiques, authenticity in this study emerged not as a fixed property of tasks but as a *relational and functional construct*, realized through how tasks generate interpretable evidence for particular claims. The rich language performances observed across presentations, project-based writing, and integrated tasks confirm long-standing arguments in learning-oriented and sociocognitive assessment literature that authentic tasks can capture higher-order language use, strategic competence, and disciplinary discourse practices. However, the uneven articulation of evidentiary purpose across tasks reveals a central tension identified in the introduction: when authenticity is prioritized without explicit construct modeling, the resulting evidence becomes pedagogically rich but inferentially fragile. This finding reinforces the argument that authenticity must be theorized not

only in relation to target-domain resemblance but also in relation to *evidence relevance* within a validity argument.

The integration of Evidence-Centered Design (ECD) offers a critical lens for interpreting this tension. ECD conceptualizes assessment as a chain of reasoning linking tasks, observations, evidence, and claims. The findings show that, in practice, authentic tasks often fulfilled the task-model component of ECD, but the evidence and claim models remained partially implicit. Constructs such as academic argumentation were well supported by observable features and rubric criteria, whereas interactional competence and strategic language use were weakly operationalized. This asymmetry suggests construct underrepresentation—not because tasks failed to elicit relevant behavior, but because the assessment system lacked mechanisms to recognize and formalize such behavior as evidence. Theoretically, this underscores a key contribution of ECD to authenticity debates: ECD does not constrain authenticity, but rather *disciplines it*, ensuring that rich performances can be meaningfully interpreted rather than reduced to impressionistic judgments.

Argument-based validity further illuminates the implications of these findings. The analysis revealed that many assessment interpretations relied on unarticulated warrants—shared assumptions about what scores “mean” rather than explicitly justified inferences. While this tacit reasoning may suffice for grading at the course level, it becomes problematic when assessment results are aggregated for program evaluation or accountability. The absence of explicit claims, warrants, and rebuttals limited stakeholders’ ability to scrutinize validity, identify threats (e.g., rater variability, task specificity), or defend decisions. This finding directly addresses the research problem identified in the introduction regarding defensibility and transparency in higher education language assessment. From a validity theory standpoint, the study supports a shift from viewing validity as a property of instruments to viewing it as a *property of arguments embedded in institutional practices*. Authentic assessment, when disconnected from argument-based validation, risks becoming normatively appealing but analytically opaque.

The findings also speak directly to learning-oriented assessment (LOA). Authentic tasks in this study clearly supported engagement, meaningful feedback, and learner reflection, aligning with LOA principles that emphasize assessment *for* and *as* learning. Student reflections indicate heightened awareness of language use, task demands, and disciplinary expectations. However, inconsistencies in rubric interpretation and feedback emphasis across courses undermined longitudinal learning coherence. From an LOA perspective, this fragmentation limits learners’ ability to build cumulative understanding of quality and progress across a program. Theoretically, the study extends LOA by demonstrating that learning orientation must be supported by *program-level alignment mechanisms*, not only task-level feedback practices. Without shared evidence models and construct continuity, learning-oriented assessment remains localized and vulnerable to inconsistency.

Program evaluation theory provides the final integrative lens for interpreting the findings. The study confirms that higher education English programs routinely collect large volumes of assessment data, yet struggle to translate this data into evaluative knowledge that supports improvement and decision-making. The dominance of aggregate scores and pass rates in program review reflects what evaluation scholars have long described as a “measurement–use gap.” The findings demonstrate that this gap is not merely technical but conceptual: when assessment evidence is not generated through explicit construct and evidence models, it cannot easily be repurposed for evaluative reasoning. By contrast, the integrated framework proposed in this study positions authentic assessment as a source of *diagnostic, interpretable evidence* that can inform

context–process–product evaluations. This reframes program evaluation not as an external audit function, but as a continuation of principled assessment design.

Taken together, the findings confirm several of the study's initial problem formulations while challenging others. The pedagogical value of authentic assessment is strongly confirmed; its sufficiency as a standalone solution is clearly disconfirmed. The assumption that model-based assessment necessarily undermines authenticity is also disconfirmed; instead, model-based reasoning emerges as a necessary condition for scaling authenticity beyond individual classrooms. What remains unresolved—and constitutes a limitation of the study—is the extent to which such integration can be institutionalized given constraints of time, assessment literacy, and organizational culture. While the framework demonstrates conceptual and methodological viability, its sustainability depends on professional development, shared ownership, and institutional support—factors not fully examined in this study.

In sum, this discussion positions the study's findings as a bridge between pedagogical and evaluative traditions in applied linguistics. By empirically demonstrating how authenticity, ECD, argument-based validity, learning-oriented assessment, and program evaluation intersect in practice, the study advances a more integrated understanding of assessment as *design, inference, learning, and governance*. This integrated perspective provides the theoretical grounding for the subsequent articulation of the study's theoretical and methodological contributions and its implications for future research and policy in higher education English language assessment.

This study makes a substantive contribution to applied linguistics and language assessment by reconceptualizing the relationship between authentic assessment and model-based assessment as one of *theoretical complementarity rather than methodological tension*. Theoretically, it advances authenticity debates by demonstrating that authentic tasks do not constitute valid assessment evidence by virtue of realism alone, but by their positioning within explicit evidence models and validity arguments. By integrating authenticity with Evidence-Centered Design and argument-based validity, the study bridges sociocognitive, learning-oriented perspectives on language use with inferential theories of assessment, offering a principled account of how rich performance can support defensible claims. Methodologically, the study extends program evaluation research in language education by embedding assessment design and validation logic within an evaluative framework that operates across task, course, and program levels. This integration enables assessment evidence to be both instructionally meaningful and institutionally actionable, addressing long-standing concerns about fragmentation, opacity, and limited use of assessment data in higher education English program.

CONCLUSION

This study set out to develop and empirically examine a comprehensive evaluation framework that integrates authentic and model-based assessment in higher education English language learning. The findings demonstrate that authentic tasks successfully elicit consequential, discipline-relevant language use and support learning-oriented assessment practices. However, the study also shows that authenticity alone is insufficient to ensure validity, coherence, or evaluability at scale. Without explicit evidence models and articulated validity arguments, rich performance data remain difficult to interpret, aggregate, and use for program-level decision-making.

By integrating Evidence-Centered Design and argument-based validity within a program evaluation perspective, this study provides a framework that aligns assessment design, inferential reasoning, and evaluative use. The framework illustrates how authentic tasks can function as systematic evidence-generating mechanisms rather than isolated pedagogical innovations. In doing so, it addresses persistent challenges in higher education English assessment related to

transparency, consistency, and accountability, while preserving the pedagogical strengths of authenticity and learning-oriented assessment.

Several limitations must be acknowledged. The study was conducted within a single institutional context, and implementation relied on existing assessment literacy and organizational structures. As such, the generalizability of the framework depends on contextual adaptation and institutional support. Future research should examine longitudinal implementation across diverse higher education settings, investigate professional development models that support integrated assessment design, and explore how digital and AI-assisted assessment tools can be incorporated within explicit evidence and validity frameworks without undermining authenticity or fairness.

In conclusion, this study argues for a shift in how assessment is conceptualized in higher education English programs—from a collection of tasks and scores to a coherent system of evidence, inference, and evaluation. Such a shift is essential if assessment is to support learning, justify decisions, and sustain program improvement in increasingly complex and accountable educational environments.

REFERENCES

- Ajjawi, R., Tai, J., Dollinger, M., Dawson, P., Boud, D., & Bearman, M. (2024). From authentic assessment to authenticity in assessment: broadening perspectives. *Assessment & Evaluation in Higher Education*, 49(4), 499–510. <https://doi.org/10.1080/02602938.2023.2271193>
- Barno, E., & Phelps, G. (2025). Using a Multi-Agent System and Evidence-Centered Design to Integrate Educator Expertise Within Generated Feedback. *Education Sciences*, 15(10), 1273. <https://doi.org/10.3390/educsci15101273>
- Bearman, M., Nieminen, J. H., & Ajjawi, R. (2023). Designing assessment in a digital world: an organising framework. *Assessment & Evaluation in Higher Education*, 48(3), 291–304. <https://doi.org/10.1080/02602938.2022.2069674>
- Boud, D., & Bearman, M. (2024). The assessment challenge of social and collaborative learning in higher education. *Educational Philosophy and Theory*, 56(5), 459–468. <https://doi.org/10.1080/00131857.2022.2114346>
- Buchan, M.C., Bhawra, J. & Katapally, T.R. (2024). Navigating the digital world: development of an evidence-based digital literacy program and assessment tool for youth. *Smart Learn. Environ*, 11(8). <https://doi.org/10.1186/s40561-024-00293-x>
- Charlton, N., & Newsham-West, R. (2024). A conceptual model for program-level assessment. *Higher Education Research & Development*, 43(8), 1721–1736. <https://doi.org/10.1080/07294360.2024.2364094>
- Dai, D. W., Vu, T., Knoch, U., Lim, A. S., Malone, D. T., & Mak, V. (2024). Expanding Kane's argument-based validity framework: What can validation practices in language assessment offer health professions education?. *Medical education*, 58(12), 1462–1468. <https://doi.org/10.1111/medu.15452>
- Fawns, T., Bearman, M., Dawson, P., Nieminen, J. H., Ashford-Rowe, K., Willey, K., Press, N. (2025). Authentic assessment: from panacea to criticality. *Assessment & Evaluation in Higher Education*, 50(3), 396–408. <https://doi.org/10.1080/02602938.2024.2404634>
- Hadzhikoleva, S., Hadzhikolev, E., Gaftandzhieva, S., & Pashev, G. (2025). A conceptual framework for multi-component summative assessment in an e-learning management system. *Frontiers in Education*, 10. <https://doi.org/10.3389/feduc.2025.1656092>

- Hu, A., Liu, Q., & Daniel, B. (2025). Digital Technologies in Authentic Assessment in Higher Education: A Systematic Literature Review and Narrative Synthesis. *Sage Open*, 15(3). <https://doi.org/10.1177/21582440251357198>
- Huang, S.-H. (2025). Evaluating the English for General Purposes (EGP) program at a Taiwanese University: A CIPP (context, input, process, and product) model study. *Evaluation and Program Planning*, 112, 102662. <https://doi.org/10.1016/j.evalprogplan.2025.102662>
- Imsa-ard, P. (2025). Learning-oriented assessment and L2 argumentative writing: effects on writing ability and academic resilience. *Discov Educ* 4, 396. <https://doi.org/10.1007/s44217-025-00779-x>
- Jones, Daniel & Cheng, Liying & Tweedie, Gregory. (2023). Automated Scoring of Speaking and Writing: Starting to Hit its Stride. *Canadian Journal of Learning and Technology*. 48. 1-22. 10.21432/cjlt28241.
- Krooi, M., Whittingham, J., & Beusaert, S. (2024). Introducing the 3P conceptual model of internal quality assurance in higher education: A systematic literature review. *Studies in Educational Evaluation*, 82, 101360. <https://doi.org/10.1016/j.stueduc.2024.101360>
- Kubsch, Marcus., Grimm, Adrian., Neumann, Knut., & Drachsler, Hendrik. (2024). Using Evidence-Centered Design to Develop an Automated System for Tracking Students' Physics Learning in a Digital Learning Environment. *Uses of Artificial Intelligence in STEM Education*, 11, 230—249. <https://doi.org/10.1093/oso/9780198882077.003.0011>
- Lam, D. M. K., & Gayton, A. M. (2025). '6.5 sounds a bit better than 6.0?': a case for embedding language assessment literacy in university teaching staff professional development. *Higher Education Research & Development*, 44(4), 976–991. <https://doi.org/10.1080/07294360.2024.2439854>
- Li, J., Huang, J., & Sheeran, T. (2025). ChatGPT4o as an AI Peer Assessor in EFL Speaking Classrooms: Examining Scoring Reliability and Feedback Effectiveness. *Sage Open*, 15(3). <https://doi.org/10.1177/21582440251369938>
- Liu, C., Hwang, G., Yu, P., Tu, Y., & Wang, Y. (2025). Effects of an automated corrective feedback-based peer assessment approach on students' learning achievement, motivation, and self-regulated learning conceptions in foreign language pronunciation. *Educational Technology Research and Development*, 73(4), 2403-2424. <https://doi.org/10.1007/s11423-025-10484-z>
- Liu, Xu Jared, Wang, Jingwen and Zou, Bin. (2025). Evaluating an AI speaking assessment tool: score accuracy, perceived validity, and oral peer feedback as feedback enhancement. *Journal of English for Academic Purposes*, 75, 101505. [doi:10.1016/j.jeap.2025.101505](https://doi.org/10.1016/j.jeap.2025.101505).
- Newton, S., Edwards, D., Usselman, M., Hernandez, D., Helms, M., Alemdar, M., & Rutstein, D. (2021). Utilizing Evidence-Centered Design to Develop Assessments: A High School Introductory Computer Science Course. *Frontiers in Education*, 6. <https://doi.org/10.3389/feduc.2021.695376>
- Nguyen, T. M. H., Gu, P., & Coxhead, A. (2023). Argument-based validation of Academic Collocation Tests. *Language Testing*, 41(3), 459-505. <https://doi.org/10.1177/02655322231198499>
- Nieminen, Juuso & Yan, Zi & Boud, David. (2025). Self-assessment design in a digital world: centring student agency. *Assessment & Evaluation In Higher Education*. 50. 10.1080/02602938.2025.2467647.
- Tubis, A. A. (2023). Digital Maturity Assessment Model for the Organizational and Process Dimensions. *Sustainability*, 15(20), 15122. <https://doi.org/10.3390/su152015122>

- Viswanathan, N., Meacham, S., & Adedoyin, F. F. (2022). Enhancement of online education system by using a multi-agent approach. *Computers and Education: Artificial Intelligence*, 3, 100057. <https://doi.org/10.1016/j.caeai.2022.100057>
- Vlachopoulos, Dimitrios., & Makri, Agoritsa. (2024). A systematic literature review on authentic assessment in higher education: Best practices for the development of 21st century skills, and policy considerations. *Studies in Educational Evaluation*, 83. <https://doi.org/10.1016/j.stueduc.2024.101425>
- Vogt, K., Bøhn, H., & Tsagari, D. (2024). Language assessment literacy. *Language Teaching*, 57(3), 325–340. doi:10.1017/S0261444824000090
- Wakid, M., Sofyan, H., Widowati, A., & Zaida Ilma, A. (2024). Learning-oriented assessment: a systematic literature network analysis. *Cogent Education*, 11(1). <https://doi.org/10.1080/2331186X.2024.2366075>
- Zapata-Rivera, D, Forsyth, C, Graf, E, & Jiang, Y. (2024). Designing and Evaluating Evidence-Centered Design based Conversations for Assessment with LLMs. *Proceedings of EDM 2024 Workshop: Leveraging Large Language Models for Next Generation Educational Technologies*, 1—9. <https://par.nsf.gov/biblio/10591735>.
- Zhan, Y., Boud, D. & Du, Z. (2025). Designing for authentic assessment: a scoping review. *High Educ.* <https://doi.org/10.1007/s10734-025-01588-9>
- Zhang, J., Yu, G., & Browne, W. (2025). Teachers' language assessment literacy: Exploring its construct and contextual factors. *Studies in Educational Evaluation*, 87, 101525. <https://doi.org/10.1016/j.stueduc.2025.101525>