p-ISSN 2252-7427, e-ISSN 2541-3635
https://ieflj.uniku.ac.id/pub

p-ISSN 2252-7427 e-ISSN 2541-3635
Indonesian EFL Journal
AISEE
The Association of Indonesia Scholars of English Education

AN INNOVATION-BASED ASSESSMENT MODEL FOR BIPA: EXAMINING INTEGRATED LISTENING AND SPEAKING EVALUATION

Novi Haryanti[1], Ribut Wahyu Eriyanti[2], Hari Sunaryo[3]
[1,2,3]*Program Studi Doktor Pendidikan Universitas Muhammadiyah Malang*

Corresponding author's email: haryanti633@gmail.com

APA Citation: Haryanti, N., Eriyanti, R. W., & Sunaryo, H. (2026). An innovation-based assessment model for BIPA: Examining integrated listening and speaking evaluation. *Indonesian EFL Journal, 12*(1), 101-116. https://doi.org/10.25134/rpf7gb87

Received:07-09-2025 Accepted: 11-11-2025 Published: 30-01-2026

Abstract: This study proposes and empirically tests an innovation-based evaluation model for Bahasa Indonesia bagi Penutur Asing (BIPA) assessment by examining the effects of listening–speaking integration, authentic communication tasks, and assessment feedback quality on evaluation model quality. Employing a quantitative explanatory design, data were collected from BIPA learners and analyzed using Partial Least Squares Structural Equation Modeling. The results indicate that authentic communication tasks and assessment feedback quality have significant positive effects on evaluation model quality, whereas listening–speaking integration does not show a significant direct effect. The proposed model explains 63.5% of the variance in evaluation model quality, demonstrating strong explanatory power. These findings highlight that innovation in BIPA assessment is primarily driven by task authenticity and meaningful feedback rather than skill integration alone. This study contributes to communicative competence-oriented assessment by clarifying the relative roles of authenticity, feedback, and integrated skills in strengthening the pedagogical relevance and effectiveness of BIPA evaluation models.
Keywords: *BIPA assessment; evaluation model quality; listening–speaking integration; authentic communication tasks; assessment feedback quality.*

INTRODUCTION

The teaching and learning of Indonesian for Foreign Speakers (Bahasa Indonesia bagi Penutur Asing or BIPA) has expanded significantly in recent years due to increasing international mobility, academic exchange, tourism, and regional economic cooperation. This growth is closely linked to Indonesia's efforts to strengthen the position of Bahasa Indonesia as a language of regional and international communication, particularly within ASEAN and global academic networks (Fitria, 2023; Simanullang et al., 2024; Syafruddin et al., 2025). As BIPA programs continue to develop across universities, language centers, and cultural institutions worldwide, greater attention has been directed toward improving instructional quality, including curriculum design, teaching strategies, and assessment systems. Among these elements, assessment plays a central role in measuring learners' proficiency, informing instructional decisions, and ensuring alignment between learning outcomes and real communicative needs through valid and reliable practices that reflect contemporary trends in language education (Yang & Wang, 2025; Gao & Hishamuddin, 2024).

Despite this positive development, assessment practices in many BIPA programs remain largely traditional, particularly in listening and speaking evaluation. Recent studies indicate that these skills are often assessed separately through controlled and decontextualized tasks, such as multiple-choice listening tests or scripted oral presentations (Alam & Ahmad, 2025; Lestari et al., 2025). Such approaches fail to reflect how language is used in real communication, where listening and speaking operate simultaneously in interactive exchanges. Research in second and foreign language assessment consistently shows that isolated skill testing reduces construct validity and inadequately represents authentic communication processes, especially in contexts

requiring spontaneous interaction and meaning negotiation (Asy'ari Rahim & Hermawan, 2024; Ockey & Na, 2025; Xi, 2025). Consequently, existing BIPA assessment practices may not fully capture learners' communicative ability in real-life situations.

In response to these limitations, recent language assessment research increasingly emphasizes integrated listening–speaking assessment, authentic communication tasks, and learning-oriented feedback as core components of communicative competence-oriented evaluation. However, these components are often examined independently, with limited attention to how they jointly contribute to the overall quality of evaluation models (Skehan, 2024; Rosy et al., 2025). This fragmented approach has resulted in a lack of empirically tested evaluation models that systematically integrate skill integration, task authenticity, and feedback quality within a single assessment framework, particularly in BIPA contexts.

Authentic communication tasks represent one critical dimension of assessment innovation. Authenticity refers to the extent to which assessment tasks resemble real-world language use and require learners to interpret meaning, respond spontaneously, and negotiate communicative goals. Recent reviews suggest that assessment tasks with low authenticity tend to reduce learner engagement and limit the transfer of classroom learning to real communication contexts, especially in speaking-oriented assessment (Van Der Steen et al., 2023; Souzandehfar, 2024). Conversely, performance-based and authentic tasks have been shown to improve learners' communicative readiness and provide more valid representations of functional language ability (Gao & Hishamuddin, 2024; Shang et al., 2024). For BIPA learners who aim to use Indonesian in academic, professional, and social settings, assessment tasks that lack authenticity may diminish the pedagogical value of evaluation outcomes.

Alongside task design, assessment feedback quality has received increasing attention in recent language assessment research. Studies published after 2023 emphasize that feedback limited to numerical scores or brief evaluative comments provides little support for learners' communicative development and does not sufficiently guide improvement in listening and speaking performance (Maslucha et al., 2024; Tran & Ma, 2024). High-quality feedback is now understood as clear, specific, and focused on communicative performance, enabling learners to reflect on their abilities and regulate their learning more effectively (Veugen et al., 2024; Nurfaidah et al., 2024). Without meaningful feedback, assessment tends to function primarily as a summative tool rather than as a mechanism for learning improvement.

While feedback, authenticity, and integrated skills assessment have been widely discussed in recent literature, empirical studies that position assessment feedback quality as a key determinant of evaluation model quality in BIPA contexts remain limited. In particular, little attention has been given to how feedback interacts with authentic task design and listening–speaking integration to support communicative competence development (Elbaioumi Shaddad & Jember, 2024). Addressing this gap is essential for strengthening innovation-oriented assessment practices in BIPA programs.

Recent research strongly supports the use of integrated listening–speaking assessment as a more valid representation of oral communication ability, especially when embedded within authentic communicative tasks (Sadeghi & Bakhshi, 2025; Qasserras, 2025). However, recent scholarship also indicates that skill integration alone is insufficient to enhance assessment quality unless it is supported by appropriate task design, transparent assessment criteria, and pedagogically meaningful implementation (Xi, 2025; Rosy et al., 2025). This perspective suggests that listening–speaking integration should be viewed as a supportive component within a broader evaluation model that emphasizes authenticity and feedback quality.

From a methodological perspective, recent language education research increasingly applies Partial Least Squares Structural Equation Modeling (PLS-SEM) to examine complex relationships among latent constructs such as assessment quality, feedback effectiveness, authenticity, and skill integration (Angelelli et al., 2025; Sabol et al., 2023). PLS-SEM is particularly suitable for exploratory and innovation-based studies with moderate sample sizes, which are common in BIPA research contexts (Hair & Alamer, 2022; Subhaktiyasa, 2024).

In response to these theoretical and empirical gaps, this study proposes an innovation-based evaluation model for BIPA assessment that integrates listening–speaking assessment, authentic communication tasks, and assessment feedback quality. Evaluation model quality is positioned as the dependent variable, reflecting

the validity, reliability, practicality, and communicative alignment of assessment practices. Listening–speaking integration, authentic communication tasks, and assessment feedback quality are examined as key predictors influencing evaluation model quality using PLS-SEM.

Accordingly, the primary objective of this study is to examine the effects of listening–speaking integration, authentic communication tasks, and assessment feedback quality on the quality of BIPA evaluation models. The findings are expected to contribute theoretically to communicative competence-oriented assessment in BIPA, methodologically to the application of PLS-SEM in language assessment research, and practically to the development of more communicative, learner-centered, and feedback-driven assessment practices.

Assessment plays a central role in BIPA programs because it determines how learners' Indonesian language proficiency is measured and how instructional effectiveness is evaluated. Recent BIPA-focused studies indicate that assessment practices in many programs remain largely conventional, with a strong emphasis on discrete language skills and linguistic accuracy rather than communicative performance (Fitria, 2023; Alam & Ahmad, 2025). Listening and speaking are often assessed independently through standardized listening tests, oral presentations, or structured interviews that prioritize grammatical correctness and pronunciation accuracy. While such approaches offer practicality and scoring efficiency, they provide limited insight into learners' ability to engage in real-life communication using Indonesian.

Empirical evidence from BIPA classrooms further suggests that skill-separated assessment does not adequately represent learners' functional language ability. Studies conducted in Indonesian universities and language centers report that learners frequently struggle to transfer assessed skills to spontaneous interaction outside the classroom when assessment tasks are decontextualized (Halimah et al., 2023; Simanullang et al., 2024). These findings are consistent with recent language assessment research emphasizing that integrated performance tasks elicit more natural language use, interactional strategies, and communicative decision-making (Ockey & Na, 2025; Sadeghi & Bakhshi, 2025). In addition, classroom-based assessment models that integrate assessment with instruction have been shown to enhance learners' communicative performance and willingness to communicate, highlighting the value of more authentic and integrated assessment designs (Liu, 2024).

Another limitation of existing BIPA evaluation models is their strong emphasis on summative outcomes. Learners typically receive numerical scores with minimal explanation of their communicative strengths and weaknesses. Recent classroom-based studies indicate that such feedback practices reduce the formative function of assessment and limit learners' opportunities to reflect on their listening and speaking performance (Nurfaidah et al., 2024; Haryanti et al., 2025). Earlier research on authentic assessment in Indonesian educational contexts also emphasizes that without reflective and process-oriented evaluation practices, assessment tends to focus on outcomes rather than learning development (Eriyanti, 2018). As a result, assessment often functions primarily as a measurement tool rather than as a mechanism for communicative growth.

In response to these limitations, contemporary research on BIPA assessment increasingly calls for innovation-oriented evaluation models that integrate communicative competence principles. Innovation-based assessment emphasizes integrated skills, authentic task design, and meaningful feedback to ensure alignment between assessment practices and learners' real communicative needs (Fitria, 2023; Marden & Herrington, 2025). In line with collaborative and reflective assessment principles proposed by (Eriyanti, 2018), such models position assessment as an integral part of the learning process rather than a terminal activity. Consequently, innovation-based evaluation models are expected to improve evaluation model quality by enhancing validity, authenticity, and pedagogical relevance in BIPA contexts.

Integrated skills assessment refers to the evaluation of two or more language skills within a single task, reflecting how language is used in natural communication. In spoken interaction, listening and speaking are inherently interconnected, as speakers must comprehend incoming messages before producing appropriate responses. Recent assessment research highlights that evaluating these skills separately provides an incomplete representation of communicative competence and interactional ability, particularly in performance-based language use contexts (Xi, 2025; Ockey & Na, 2025). This perspective is

consistent with integrative learning principles, which emphasize that receptive and productive skills mutually reinforce each other during language development (Eriyanti, 2020).

Empirical studies published between 2023 and 2025 consistently demonstrate that integrated listening–speaking tasks improve construct validity by closely mirroring real communicative demands. Learners are assessed not only on comprehension or production in isolation but also on their ability to manage meaning, maintain interaction, and respond coherently in real time (Sadeghi & Bakhshi, 2025; Qasserras, 2025). These abilities are particularly relevant for BIPA learners who aim to use Indonesian in academic, professional, and social interactions.

Within BIPA contexts, recent studies report that integrated listening–speaking assessment supports a more accurate measurement of learners' functional language ability. Research conducted in BIPA programs at Indonesian universities shows that learners demonstrate stronger interactional competence and pragmatic awareness when assessed through integrated communicative tasks rather than isolated skill tests (Alam & Ahmad, 2025; Haryanti et al., 2025). These findings indicate that listening–speaking integration represents an important innovation in BIPA evaluation models.

By combining receptive and productive skills within a single assessment framework, evaluation models can better reflect authentic communication and align more closely with communicative competence principles. Evidence from integrative learning and task-based assessment research suggests that integrated listening–speaking activities enhance learners' real-time language use, interactional ability, and overall communicative performance, particularly when supported by appropriate task design and instructional alignment (Eriyanti, 2020; qizi & Nabieva, 2025). Therefore, listening–speaking integration is expected to enhance evaluation model quality.

Authenticity in language assessment refers to the extent to which assessment tasks reflect real-world language use and require learners to engage in meaningful communication. Recent research emphasizes that authenticity remains a critical yet challenging dimension of assessment design, particularly in listening and speaking evaluation, where spontaneous interaction and contextual interpretation are essential (Hains-Wesson et al., 2024**;** Gao & Hishamuddin, 2024). Authentic communication tasks require learners to interpret meaning, respond spontaneously, and negotiate communicative goals within realistic contexts.

In BIPA assessment, authenticity is particularly important because learners are expected to use Indonesian beyond the classroom in academic, professional, and social settings. However, recent BIPA-focused studies indicate that assessment tasks are still frequently designed for convenience and standardization, resulting in limited contextual relevance (Halimah et al., 2023; Simanullang et al., 2024). Learners assessed through highly controlled tasks often experience difficulty when faced with open-ended communicative demands in real interactions.

Recent empirical findings demonstrate that the use of authentic tasks in BIPA assessment enhances learner engagement, communicative confidence, and pragmatic competence. Studies conducted in BIPA speaking and listening courses show that role-plays, simulated conversations, and problem-solving discussions provide a more valid representation of learners' communicative ability than scripted responses (Marden & Herrington, 2025; Fitria, 2023). These findings support broader assessment research emphasizing that authenticity strengthens the validity and pedagogical meaningfulness of evaluation outcomes.

As an innovation-oriented component of BIPA evaluation, authentic communication tasks are therefore expected to improve evaluation model quality by aligning assessment performance with real communicative needs.

Assessment feedback quality is a crucial element in effective language assessment because it links evaluation results with learning improvement. Recent studies in foreign and second language education emphasize that feedback should be clear, specific, and focused on communicative performance rather than limited to numerical scores or general judgments. Descriptive feedback has been shown to support learners' awareness of strengths and weaknesses in fluency, interaction, and comprehension, thereby enhancing the formative function of assessment (Maslucha et al., 2024; Tran & Ma, 2024). Nevertheless, empirical evidence indicates that feedback in many language learning contexts, including settings comparable to BIPA classrooms, remains minimal and evaluative, offering limited guidance for listening and speaking development (Nurfaidah et al., 2024).

High-quality feedback in listening–speaking assessment addresses how learners interpret spoken input, respond meaningfully, manage

interaction, and apply appropriate communicative strategies. Recent BIPA studies report that descriptive and performance-oriented feedback supports learner reflection, autonomy, and communicative growth (Haryanti et al., 2025; Alam & Ahmad, 2025). Such feedback shifts assessment from a purely summative process toward a learning-oriented practice.

From an innovation-based perspective, assessment feedback quality represents a core determinant of evaluation model effectiveness. When feedback is systematically integrated into the evaluation framework, it enhances transparency, learner engagement, and instructional alignment. Therefore, assessment feedback quality is expected to positively influence evaluation model quality in BIPA programs (Ghaneiarani et al., 2024).

Based on the reviewed literature, this study proposes a conceptual framework in which evaluation model quality is positioned as the dependent variable influenced by three innovation-oriented assessment components: listening–speaking integration, authentic communication tasks, and assessment feedback quality. These components represent structural, contextual, and pedagogical dimensions of BIPA assessment innovation. Listening–speaking integration reflects assessment task design, authentic communication tasks capture alignment with real-world language use, and assessment feedback quality represents the learning-oriented value of evaluation outcomes. This framework aligns with recent assessment models that integrate multiple design elements to enhance theoretical coherence and overall evaluation quality in language assessment research (Pandey, 2025). The proposed relationships among variables are examined using Partial Least Squares Structural Equation Modeling (PLS-SEM) through SmartPLS, providing an empirical basis for understanding how innovation-based assessment components contribute to improving BIPA evaluation models.

METHOD

This study adopts a quantitative explanatory research design to examine the influence of listening–speaking integration, authentic communication tasks, and assessment feedback quality on evaluation model quality in BIPA assessment. This approach is appropriate because the study aims to test hypothesized relationships among latent constructs using statistical modeling, consistent with recent empirical research employing quantitative explanatory designs in educational and language assessment contexts (Li et al., 2024; Angelelli et al., 2025).

The explanatory design enables the identification of both the direction and strength of relationships among variables, providing empirical support for innovation-oriented evaluation models in BIPA contexts. Accordingly, this study is positioned as an innovation-based evaluation study, focusing on assessment practices that emphasize integrated language skills, authentic communication, and feedback-driven learning—dimensions increasingly recognized as central to communicative competence-oriented assessment (Gao & Hishamuddin, 2024; Rosy et al., 2025).

The participants consisted of 100 BIPA students, all foreign learners enrolled in formal BIPA programs at universities and language institutions. These learners were selected because they directly experienced listening and speaking assessments and were therefore able to provide informed perceptions of assessment integration, task authenticity, and feedback quality.

A purposive sampling technique was applied by targeting students who had completed listening and speaking assessments as part of their programs, ensuring adequate exposure to BIPA evaluation practices.

The sample size exceeds the minimum requirement for Partial Least Squares Structural Equation Modeling (PLS-SEM). Following the widely applied 10-times rule, the minimum sample size is calculated by multiplying ten by the maximum number of structural paths directed at the dependent variable (Hair et al., 2022). As evaluation model quality is predicted by three independent variables, the minimum required sample size is 30. Therefore, the use of 100 respondents is considered methodologically adequate and statistically robust, particularly in BIPA research contexts where large samples are often difficult to obtain (Subhaktiyasa, 2024).

Data were collected using a structured questionnaire developed specifically for BIPA students. A five-point Likert scale ranging from strongly disagree to strongly agree was used to measure learners' perceptions of assessment practices. This format is widely adopted in recent language assessment research due to its effectiveness in capturing perceptions of authenticity, feedback quality, and assessment effectiveness (Souzandehfar, 2024).

The questionnaire measured four latent constructs: listening–speaking integration,

authentic communication tasks, assessment feedback quality, and evaluation model quality. Items related to listening–speaking integration focused on assessment experiences requiring learners to process spoken input and respond orally in an integrated and interactive manner. Authentic communication task items assessed whether assessment activities reflected real-life communication contexts relevant to BIPA learners. Assessment feedback quality items examined the clarity, usefulness, and communicative relevance of feedback, while evaluation model quality items captured perceptions of effectiveness, fairness, transparency, and contribution to communicative competence development.

All items were adapted from established language assessment instruments and recent empirical studies, then contextualized for the BIPA learning environment to ensure content validity and relevance. The instrument was reviewed to ensure clarity and accessibility for foreign learners, following recent recommendations for questionnaire development in multilingual education contexts (Maslucha et al., 2024; Nurfaidah et al., 2024).

Data collection was conducted using an online questionnaire distributed via Google Forms. This platform was selected for its accessibility and widespread use in higher education and international learning contexts. The questionnaire link was shared through official communication channels commonly used by BIPA programs. Participants received a brief explanation of the study's purpose, the voluntary nature of participation, and instructions for completing the questionnaire. The questionnaire was written in clear and accessible language to ensure comprehension. Ethical considerations were carefully addressed: participation was anonymous, no identifying information was collected, informed consent was obtained electronically, and all data were securely stored and used solely for academic research purposes.

Data were analyzed using SmartPLS software (version 4) to conduct Partial Least Squares Structural Equation Modeling (PLS-SEM). PLS-SEM was selected due to its suitability for innovation-based and exploratory studies involving latent constructs and moderate sample sizes, which are common in BIPA research (Hair et al., 2022; Angelelli et al., 2025).

Measurement model evaluation assessed indicator reliability through outer loadings and convergent validity using Average Variance Extracted (AVE). Discriminant validity was examined to confirm construct distinctiveness. Structural model evaluation involved testing the proposed hypotheses using bootstrapping procedures.

In line with recent SEM reporting standards, path coefficients, effect sizes (f^2), and coefficients of determination (R^2) were reported to enhance robustness and transparency (Subhaktiyasa, 2024). In addition, 95% bootstrapped confidence intervals were examined to assess the robustness of the estimated path coefficients. Predictive relevance (Q^2) and model fit indices, including the Standardized Root Mean Square Residual (SRMR), were also reported to provide a comprehensive evaluation of the proposed model.

RESULTS AND DISCUSSION

Table1. *Outer loadings*

Indicators	Outer Loadings
Y1.1 Evaluation Model Quality	0.736
Y1.2 Evaluation Model Quality	0.714
Y1.3 Evaluation Model Quality	0.816
Y1.4 Evaluation Model Quality	0.838
Y1.5 Evaluation Model Quality	0.746
Y1.6 Evaluation Model Quality	0.777
X1.1 Listening–Speaking Integration	0.722
X1.2 Listening–Speaking Integration	0.753
X1.3 Listening–Speaking Integration	0.794
X1.4 Listening–Speaking Integration	0.806
X1.5 Listening–Speaking Integration	0.716
X1.6 Listening–Speaking Integration	0.768
X2.1 Authentic Communication Tasks	0.796
X2.2 Authentic Communication Tasks	0.895
X2.3 Authentic Communication Tasks	0.878
X2.4 Authentic Communication Tasks	0.859
X2.5 Authentic Communication Tasks	0.728
X2.6 Authentic Communication Tasks	0.768
X3.1 Assessment Feedback Quality	0.835
X3.2 Assessment Feedback Quality	0.842
X3.3 Assessment Feedback Quality	0.791
X3.4 Assessment Feedback Quality	0.913
X3.5 Assessment Feedback Quality	0.871

Table 1 shows the outer loadings of the four constructs, namely Evaluation Model Quality, Listening–Speaking Integration, Authentic Communication Tasks, and Assessment Feedback Quality. Strong relationships between the indicators and their respective latent constructs are demonstrated, as all indicators exhibit outer loading values above the recommended threshold of 0.70.

The outer loadings for Evaluation Model Quality range from 0.714 to 0.838, indicating that the indicators accurately represent students' perceptions of the quality of the BIPA evaluation model. These values confirm that the indicators

are appropriate in capturing essential aspects of evaluation effectiveness and clarity.

For Listening–Speaking Integration, the outer loading values span from 0.716 to 0.806, highlighting the reliability of the indicators in measuring the extent to which listening and speaking skills are assessed in an integrated manner. This finding supports the adequacy of the indicators in reflecting communicative interaction within BIPA assessment.

The indicators for Authentic Communication Tasks show strong outer loading values ranging from 0.728 to 0.895, demonstrating their effectiveness in representing authentic assessment practices. These values validate the importance of real-life communication tasks as a key component of innovation-based BIPA evaluation.

Similarly, Assessment Feedback Quality exhibits robust outer loading values between 0.791 and 0.913, reflecting the stability of this construct. The results indicate that the indicators reliably measure the clarity, usefulness, and relevance of feedback received by BIPA students.

Overall, the strong outer loading values across all constructs confirm the suitability of the measurement model. The results imply that, within the framework of this study, the indicators appropriately and fairly reflect their corresponding latent constructs, allowing the analysis to proceed to the structural model evaluation.

Table 2. *Construct reliability and validity*

Variable	Cronbach's Alpha	rho_A	Composite Reliability	AVE
Assessment Feedback Quality	0.905	0.910	0.929	0.725
Authentic Communication Tasks	0.899	0.905	0.923	0.667
Evaluation Model Quality	0.864	0.867	0.899	0.597
Listening–Speaking Integration	0.854	0.857	0.891	0.578

Table 2 presents the construct reliability and validity results for Assessment Feedback Quality, Authentic Communication Tasks, Evaluation Model Quality, and Listening–Speaking Integration. The Cronbach's alpha values for all constructs range from 0.854 to 0.905, indicating high internal consistency among the measurement items. These values exceed the recommended threshold of 0.70, confirming the reliability of the constructs.

The composite reliability values further support the dependability of the measurement model. Both rho_A and composite reliability values for all constructs exceed 0.85, with composite reliability ranging from 0.891 to 0.929. These results indicate that the indicators consistently measure their respective latent constructs.

Convergent validity is assessed using the Average Variance Extracted (AVE). The AVE values for all constructs range from 0.578 to 0.725, exceeding the minimum acceptable criterion of 0.50. This finding demonstrates that each construct explains more than half of the variance of its indicators, confirming adequate convergent validity.

Overall, these results confirm that the measurement model demonstrates satisfactory reliability and validity. The constructs are measured consistently and accurately, indicating that the data are robust and suitable for further structural model analysis.

Table 3. *Discriminant validity*

Variable	AFQ	ACT	EMQ	LSI
Assessment Feedback Quality (AFQ)	0.851			
Authentic Communication Tasks (ACT)	0.832	0.817		
Evaluation Model Quality (EMQ)	0.740	0.778	0.773	
Listening–Speaking Integration (LSI)	0.761	0.850	0.699	0.760

Table 3 presents the discriminant validity assessment for Assessment Feedback Quality, Authentic Communication Tasks, Evaluation Model Quality, and Listening–Speaking Integration using the Heterotrait–Monotrait ratio (HTMT). The results indicate that all HTMT values are below the recommended threshold of 0.90, suggesting that each construct is empirically distinct from the others.

Specifically, the HTMT values between Assessment Feedback Quality and Authentic Communication Tasks (0.832), Evaluation Model Quality (0.740), and Listening–Speaking Integration (0.761) remain within acceptable limits. Similarly, the relationships among Authentic Communication Tasks, Evaluation Model Quality, and Listening–Speaking Integration also show HTMT values below 0.90, indicating no issues of excessive overlap among constructs.

These findings confirm that the measurement model satisfies discriminant validity requirements. Each construct measures a different concept within the evaluation model for BIPA assessment, supporting the structural integrity of the model and ensuring that the variables capture distinct dimensions of the proposed framework.

Table 4. *R-square*

Variable	R Square	R Square Adjusted
Evaluation Model Quality	0.635	0.623

Table 4 presents the R-square and adjusted R-square values for the dependent variable, Evaluation Model Quality. The R-square value of 0.635 indicates that Listening–Speaking Integration, Authentic Communication Tasks, and Assessment Feedback Quality jointly explain 63.5% of the variance in Evaluation Model Quality. This suggests that the proposed predictors have a substantial explanatory power in explaining the quality of the evaluation model.

Furthermore, the adjusted R-square value of 0.623 confirms the stability of the model after accounting for the number of predictors included. Overall, these results demonstrate that the structural model has strong explanatory capability and is effective in capturing the relationships among the constructs within the context of BIPA assessment evaluation.

Table 5. *f-square*

Variable Relationship	f-square
Assessment Feedback Quality → Evaluation Model Quality	0.068
Authentic Communication Tasks → Evaluation Model Quality	0.118
Listening–Speaking Integration → Evaluation Model Quality	0.005

Table 5 presents the f-square values, which indicate the effect size of each predictor on Evaluation Model Quality. The results show that Authentic Communication Tasks have the strongest effect on Evaluation Model Quality, with an f-square value of 0.118, indicating a small to moderate effect. This suggests that the use of authentic, real-life communication tasks plays an important role in improving the quality of the evaluation model for BIPA learners.

Assessment Feedback Quality shows an f-square value of 0.068, which reflects a small effect size. Although its contribution is relatively modest, this finding indicates that constructive and clear feedback still contributes meaningfully to the improvement of the evaluation model.

In contrast, Listening–Speaking Integration demonstrates a very small effect on Evaluation Model Quality, with an f-square value of 0.005. This suggests that, while integration of listening and speaking skills is conceptually important, its direct contribution to the evaluation model quality in this study is limited when compared to the other predictors.

Overall, these findings highlight that Authentic Communication Tasks are the most influential predictor in the model, followed by Assessment Feedback Quality, whereas Listening–Speaking Integration shows a minimal direct effect. This pattern emphasizes the importance of task authenticity and feedback mechanisms in developing an effective evaluation model for Indonesian language learners.

Table 6. *Model fit*

Fit summary	Saturated Model	Estimated Model
SRMR	0.092	0.092

Table 6 presents the model fit evaluation using the Standardized Root Mean Square Residual (SRMR). The results show that the SRMR value for both the saturated model and the estimated model is 0.092. This value is slightly above the commonly suggested threshold of 0.08, indicating

that the overall model fit is acceptable but not optimal.

Although the SRMR value does not indicate a perfect fit, it still suggests that the discrepancy between the observed correlations and the correlations implied by the model remains within a tolerable range. In the context of exploratory and innovation-oriented evaluation studies, this level of model fit can still be considered reasonable. Therefore, the findings indicate that the proposed structural model is adequately aligned with the empirical data and can be used to interpret the relationships among the constructs with caution.

Table 7. *Direct effects*

Variable Relationship	Original Sample (O)	T Statistics	P Values
Assessment Feedback Quality → Evaluation Model Quality	0.289	3.030	0.003
Authentic Communication Tasks → Evaluation Model Quality	0.467	3.495	0.001
Listening–Speaking Integration → Evaluation Model Quality	0.082	0.757	0.449

Table 7 presents the direct effects among the variables in the proposed model. The results show that Assessment Feedback Quality has a positive and statistically significant effect on Evaluation Model Quality (original sample = 0.289, T-statistics = 3.030, p = 0.003). Since the T-statistics value exceeds the critical value of 1.96 and the p-value is below 0.05, this relationship is considered significant, indicating that better feedback quality contributes meaningfully to the quality of the evaluation model.

Similarly, Authentic Communication Tasks demonstrate a strong and significant positive effect on Evaluation Model Quality (original sample = 0.467, T-statistics = 3.495, p = 0.001). This finding suggests that the use of real-world, authentic communication tasks plays a major role in improving the effectiveness and relevance of BIPA evaluation models.

In contrast, Listening–Speaking Integration does not show a significant direct effect on Evaluation Model Quality (original sample = 0.082, T-statistics = 0.757, p = 0.449). The T-statistics value is below 1.96 and the p-value exceeds the 0.05 threshold, indicating that, in this model, listening–speaking integration alone does not directly influence evaluation model quality in a statistically significant way.

Overall, these findings indicate that Assessment Feedback Quality and Authentic Communication Tasks are the key predictors of Evaluation Model Quality, while Listening–Speaking Integration may require support from other instructional or assessment components to exert a meaningful impact. The results partially support the proposed hypotheses and highlight the importance of authenticity and feedback in innovation-based BIPA evaluation models. The bootstrapping analysis further indicates that the 95% confidence intervals support the observed significance and non-significance patterns of the structural paths.

This study examined an innovation-based evaluation model for BIPA by analyzing the effects of listening–speaking integration, authentic communication tasks, and assessment feedback quality on evaluation model quality. The results demonstrate that the proposed model has strong explanatory power, with the three predictors jointly explaining 63.5 percent of the variance in evaluation model quality. This level of explanatory strength indicates that innovation-oriented assessment components play a substantial role in shaping learners' perceptions of evaluation quality in BIPA contexts. These findings align with recent research trends that emphasize communicative, task-based, and learning-oriented assessment approaches as central to contemporary language assessment theory (Yang & Wang, 2025).

The measurement model evaluation confirms that all constructs demonstrate satisfactory reliability and validity. Outer loading values exceed recommended thresholds, indicating strong indicator–construct relationships. Composite reliability and AVE values further support internal consistency and convergent validity, while discriminant validity confirms that listening–speaking integration, authentic communication tasks, assessment feedback quality, and evaluation model quality represent empirically distinct constructs. These results indicate that the data provide a sound basis for interpreting the structural relationships among variables.

Structural model analysis reveals that authentic communication tasks and assessment feedback quality have significant positive effects on evaluation model quality, whereas listening–speaking integration does not demonstrate a

statistically significant direct effect. Among the predictors, authentic communication tasks emerge as the strongest determinant, suggesting that assessment designs that closely mirror real-world communicative demands are perceived by learners as more meaningful and effective (Vlachopoulos & Makri, 2024). The SRMR value indicates acceptable model fit for an exploratory and innovation-oriented study, supporting the adequacy of the proposed framework while leaving room for refinement in future research.

The effect of assessment feedback quality on evaluation model quality

The findings show that assessment feedback quality has a positive and statistically significant effect on evaluation model quality. This result reinforces contemporary perspectives in language assessment that position feedback as a core mechanism linking evaluation with learning improvement. In BIPA contexts, where learners often navigate unfamiliar linguistic and sociocultural environments, feedback plays a critical role in shaping learners' understanding of communicative performance and guiding subsequent learning efforts.

Recent research emphasizes that effective feedback in listening and speaking assessment should extend beyond numerical scores to include descriptive information related to fluency, comprehensibility, interactional strategies, and pragmatic appropriateness (Maslucha et al., 2024; Tran & Ma, 2024; Ding & Zhu, 2025). Such feedback enables learners to identify gaps between their current performance and communicative expectations, supporting reflection, self-regulation, and sustained improvement. Studies further indicate that learners value feedback that explicitly addresses real communicative challenges, such as responding appropriately in interaction and managing spoken discourse (Nurfaidah et al., 2024; Haryanti et al., 2025).

The significant effect observed in this study suggests that evaluation models incorporating high-quality feedback are perceived as more transparent, supportive, and pedagogically meaningful. This finding supports recent arguments that feedback should be treated as a core component of assessment design, rather than as a supplementary or post-assessment activity. Empirical evidence from recent EFL and BIPA-related studies confirms that learners benefit most from feedback that is communicative, detailed, and performance-oriented (Fahreza & Mahaputri, 2024). Within innovation-based BIPA evaluation, feedback therefore functions not only as a corrective mechanism but also as a tool for fostering communicative competence, learner autonomy, and engagement.

The effect of authentic communication tasks on evaluation model quality

Authentic communication tasks demonstrate the strongest and most significant effect on evaluation model quality, indicating that authenticity is a key driver of effective BIPA assessment. This finding is consistent with recent assessment research emphasizing that tasks resembling real-world language use provide more valid and meaningful representations of learners' communicative ability, particularly in speaking and interaction-focused assessment (Gao & Hishamuddin, 2024; Souzandehfar, 2024).

In BIPA contexts, learners are typically motivated by practical communicative goals, such as participating in academic discussions, engaging in social interaction, or using Indonesian in professional settings. Authentic tasks allow learners to demonstrate not only linguistic knowledge but also pragmatic awareness and sociocultural competence. Recent BIPA studies report that assessment tasks grounded in real-life scenarios increase learner engagement, confidence, and the perceived relevance of evaluation outcomes (Fitria, 2023; Marden & Herrington, 2025).

From an institutional perspective, the strong effect of authentic tasks suggests that authenticity can be operationalized even within large-scale or standardized BIPA assessments. Recent assessment frameworks propose the use of scenario-based tasks, standardized role-play prompts, and structured interactive simulations that preserve consistency while maintaining communicative realism (Rosy et al., 2025; Ockey & Na, 2025). By developing task banks based on common communicative situations encountered by BIPA learners, institutions can balance standardization with authenticity, ensuring reliability without sacrificing communicative relevance.

The effect of listening–speaking integration on evaluation model quality

In contrast to the other predictors, listening–speaking integration does not exhibit a significant direct effect on evaluation model quality. Although integrated skills assessment is widely promoted within communicative language

teaching, this finding suggests that integration alone may not be sufficient to enhance perceived evaluation quality in BIPA contexts.

One theoretical explanation is that listening–speaking integration functions as an enabling condition rather than an independent determinant of evaluation quality. Recent studies argue that integration becomes pedagogically meaningful only when supported by authentic task design, transparent assessment criteria, and formative feedback mechanisms (Xi, 2025; Qasserras, 2025). Without these supporting elements, learners may not perceive integrated tasks as fundamentally different from traditional assessment practices.

Empirical research further indicates that the effectiveness of integrated listening–speaking tasks depends on how responses are interpreted, scored, and linked to communicative outcomes (Ockey & Na, 2025). Studies on listening-to-speak tasks highlight the role of cognitive processing, strategy use, and interactional demands in shaping performance (Xu et al., 2025). Similarly, research on multimodal and task-based assessment demonstrates that integration embedded within rich communicative contexts yields stronger communicative outcomes than simple skill combination (Lai, 2024; Aliah & Natsir, 2025). These findings help explain why listening–speaking integration shows a limited direct effect in this study, while authenticity and feedback emerge as stronger predictors.

Implications for BIPA evaluation innovation

The findings suggest that innovation in BIPA evaluation should prioritize authentic communication tasks and assessment feedback quality, while positioning listening–speaking integration as a supportive component within a broader assessment framework. Evaluation models that emphasize authenticity and feedback are more likely to align assessment practices with learners' real communicative needs and expectations. Recent studies on authentic assessment in Indonesian language learning emphasize that assessments designed around real-world language use can enhance learner engagement, motivation, and competence development, making them more meaningful than traditional tests (Friyandini, 2025). Moreover, research on formative assessment literacy highlights how teachers' ability to deploy meaningful feedback and formative practices contributes to instructional alignment and supports continuous learning improvement (Kang & Lam, 2024). Empirical evidence also demonstrates that authentic assessment approaches enhance students' communicative skills and confidence, underscoring the pedagogical value of embedding real communicative tasks within evaluation models rather than treating them as add-ons (Arnold & Croxford, 2025). These implications suggest that innovation-based BIPA evaluation should intentionally integrate authentic tasks and high-quality feedback to support communicative competence and learner autonomy.

Pedagogical and practical implications

From a pedagogical perspective, BIPA instructors are encouraged to design assessment tasks that simulate realistic communication scenarios and provide structured, performance-oriented feedback. Such practices support learners' communicative development and enhance the perceived quality of evaluation. Recent research indicates that teachers' pedagogical competence is strongly linked to the effectiveness of authentic assessment practices, suggesting that professional development in assessment design is crucial for ensuring meaningful and reliable evaluations (Afdal & Sukenti, 2024). At the institutional level, BIPA programs can operationalize authenticity by developing standardized communicative task frameworks, training assessors to apply consistent scoring criteria, and embedding feedback mechanisms within assessment systems. Studies on classroom assessment culture highlight the importance of teacher assessment literacy and institutional support in implementing authentic, learning-oriented assessment practices aligned with communicative competence goals (Latif & Wasim, 2024). Evidence from higher education contexts also shows that lecturers perceive authentic assessment as relevant and effective, but that institutional expectations and resource constraints influence how such approaches are implemented, reinforcing the need for systemic planning and support (Lumintang & Kuswandono, 2025). These strategies allow institutions to maintain assessment reliability while ensuring alignment with communicative competence principles. The proposed evaluation model therefore offers a practical reference for institutions seeking to implement innovation-based assessment in diverse BIPA settings.

Limitations and future research

Despite its contributions, this study has several limitations. The use of self-reported questionnaire

data reflects learners' perceptions rather than objective communicative performance. In addition, the sample size and institutional context may limit the generalizability of the findings.

Future research should incorporate performance-based assessment data, classroom observation, or mixed-method designs to further validate the proposed model. Longitudinal studies may also explore how innovation-based evaluation models influence communicative competence development over time. Further investigation into moderating variables, such as learner proficiency level or digital assessment tools, would provide deeper insight into the dynamics of BIPA assessment innovation.

CONCLUSION

This study examined an innovation-based evaluation model for BIPA by analyzing the effects of listening–speaking integration, authentic communication tasks, and assessment feedback quality on evaluation model quality. The findings provide empirical support for strengthening communicative-oriented assessment practices in Indonesian for Foreign Speakers programs and contribute to the growing discussion on innovation in language assessment.

The results indicate that authentic communication tasks and assessment feedback quality have significant positive effects on evaluation model quality, with authentic communication tasks emerging as the strongest predictor. This finding confirms that BIPA evaluation models are most effective when assessment tasks closely reflect real-life communicative situations in which learners use Indonesian meaningfully. Authenticity enables evaluation to move beyond linguistic accuracy toward a more comprehensive measurement of communicative competence, aligning assessment practices with the actual goals of BIPA learning.

Assessment feedback quality is also shown to play a critical role in improving evaluation model quality. Clear, constructive, and timely feedback supports learners' understanding of their listening and speaking performance and promotes continuous improvement. This finding emphasizes that innovation in BIPA evaluation does not depend solely on task design, but also on how assessment results are communicated and used as part of the learning process.

In contrast, listening–speaking integration does not demonstrate a significant direct effect on evaluation model quality. This result suggests that integrating skills alone is insufficient to enhance assessment quality unless integration is supported by authentic communicative tasks and meaningful feedback. Although integrated skills remain central to communicative language teaching, their contribution to evaluation quality depends on pedagogically grounded implementation rather than technical combination.

Overall, this study highlights that innovation-based evaluation in BIPA should prioritize authenticity and feedback as core components. Beyond its empirical contribution, the proposed model offers broader implications for the development of more standardized, communicative, and internationally comparable BIPA assessment frameworks, particularly as BIPA programs continue to expand across diverse institutional and cultural contexts.

Based on the findings, several recommendations can be proposed for BIPA practice, institutional policy, and future research. In instructional practice, BIPA instructors are encouraged to design evaluation tasks that reflect realistic communication contexts, such as everyday interaction, academic discussion, and professional communication. Authentic tasks allow learners to demonstrate integrated listening and speaking abilities in ways that better represent real language use and strengthen assessment validity.

Assessment feedback should be systematically embedded within BIPA evaluation models. Feedback is recommended to be specific, understandable, and actionable, addressing not only linguistic accuracy but also fluency, comprehension, interactional strategies, and pragmatic appropriateness. Embedding feedback within evaluation practices enhances learner engagement and supports formative assessment processes.

At the institutional level, BIPA providers are encouraged to adopt innovation-based evaluation frameworks that align assessment objectives with communicative competence outcomes. Instructor development programs should strengthen assessment literacy, particularly in authentic task design, integrated assessment rubric development, and effective feedback delivery. These efforts can support greater consistency and quality across BIPA programs.

In relation to broader application, the proposed evaluation model can be piloted and validated across different BIPA proficiency levels to examine its suitability for beginner, intermediate, and advanced learners. In addition, future implementation may explore the use of digital and

technology-assisted assessment environments, such as online speaking tasks or computer-mediated interaction, to enhance scalability and accessibility. Future research is also encouraged to employ longitudinal or mixed-method designs to further examine how innovation-based evaluation models influence communicative development over time and contribute to the standardization of BIPA assessment at national and international levels.

REFERENCES

Afdal, A., & Sukenti, D. (2024). Enhancing authentic assessment through pedagogical competence: Insights from Indonesian language teachers in Riau. *AL-ISHLAH: Jurnal Pendidikan*, *16*(4), 5232–5242. https://doi.org/10.35445/alishlah.v16i4.6193

Alam, A., & Ahmad, A. I. (2025). Integrated language skills and its assessment: A literature review. *Asian Journal of Education and Social Studies*, *51*(8), 304–316. https://doi.org/10.9734/ajess/2025/v51i82242

Aliah, N., & Natsir, N. (2025). Development of integrated language skills assessment instruments for Korean as a foreign language learning in higher education. *International Journal of Educational Development*, *2*(2), 36–45. https://doi.org/10.61132/ijed.v2i2.288

Angelelli, M., Ciavolino, E., Ringle, C. M., Sarstedt, M., & Aria, M. (2025). Conceptual structure and thematic evolution in partial least squares structural equation modeling research. *Quality & Quantity*, *59*(3), 2753–2798. https://doi.org/10.1007/s11135-025-02071-4

Arnold, L., & Croxford, J. (2025). Is it time to stop talking about authentic assessment? *Teaching in Higher Education*, *30*(3), 735–743. https://doi.org/10.1080/13562517.2024.2369143

Asy'ari Rahim, L. M., & Hermawan, B. (2024). The integration of listening and speaking to develop students' communicative competence in Islamic boarding school. *IJEAL (International Journal of English and Applied Linguistics)*, *4*(2). https://doi.org/10.47709/ijeal.v4i2.4468

Ding, Y., & Zhu, J. (2025). Mobile based peer feedback in EFL speaking: Learners' motivation, behavioral engagement in feedback provision, and achievement. *Asian-Pacific Journal of Second and Foreign Language Education*, *10*(1), Article 10. https://doi.org/10.1186/s40862-024-00314-9

Elbaioumi Shaddad, A. R., & Jember, B. (2024). A step toward effective language learning: An insight into the impacts of feedback-supported tasks and peer-work activities on learners' engagement, self-esteem, and language growth. *Asian-Pacific Journal of Second and Foreign Language Education*, *9*(1), Article 39. https://doi.org/10.1186/s40862-024-00261-5

Eriyanti, R. W. (2018). Lesson study kolaboratif sebagai upaya peningkatan profesionalitas guru dalam pelaksanaan penilaian autentik. *Prosiding Seminar Nasional & Internasional*, *1*(1).

Eriyanti, R. W. (2020). Effectiveness of integrative learning model to develop student's language receptive skills. *The Asian EFL Journal*, *27*(1), 196–209.

Fahreza, A., & Mahaputri, R. A. (2024). EFL students' perception towards audio and written peer feedback experiences in online class. *Research and Innovation in Language Learning*, *7*(3), 65–81. https://doi.org/10.33603/rill.v7i3.10950

Fitria, T. N. (2023). Introducing Bahasa Indonesia bagi Penutur Asing (BIPA): Method and challenges of teaching Indonesian as a Foreign Language (IFL). *Inovasi Kurikulum*, *20*(2), 205–224. https://doi.org/10.17509/jik.v20i2.60374

Friyandini, R. (2025). Authentic assessment in Indonesian language learning: Theoretical review and its implications. *International Seminar on Humanity, Education, and Language*, *1*, 397–411. https://doi.org/10.21009/ishel.v1i1.57721

Gao, L., & Hishamuddin, A. (2024). A literature review of language assessment scale. *Asian Pendidikan*, *4*(1), 65–72. https://doi.org/10.53797/aspen.v4i1.7.2024

Ghaneiarani, S., Alavi, S. M., & Kaivanpanah, S. (2024). Enhancing writing ability of Iranian EFL learners through learning-oriented assessment: Peer and teacher feedback implementation. *Language Testing in Asia*, *14*(1), Article 39. https://doi.org/10.1186/s40468-024-00298-w

Hains-Wesson, R., & Roux, S. le. (2024). Bridging teacher knowledge and practice: Exploring authentic assessment across educational levels. *Education Sciences*, *14*(8), Article 894. https://doi.org/10.3390/educsci14080894

Hair, J., & Alamer, A. (2022). Partial Least Squares Structural Equation Modeling (PLS-SEM) in second language and education research: Guidelines using an applied example. *Research Methods in Applied Linguistics*, *1*(3), Article 100027. https://doi.org/10.1016/j.rmal.2022.100027

Hair, J. F., Ringle, C. M., Sarstedt, M., & Hult, G. T. M. (2022). A primer on partial least squares structural equation modeling (PLS-SEM). *International Journal of Research & Method in Education*, *38*(2), 220–221.

Halimah, H., Khaerunnisa, K., Liliana, L. M., & Zainal, Z. R. (2023). On analyzing the need of teaching material for the Indonesian language for foreigner speakers (BIPA) based on Betawi local wisdom through multimedia interactive. *International Journal of Language and*

Education Research, *5*(3), 103–133. https://doi.org/10.29329/ijler.2023.594.6

Haryanti, N., Eriyanti, R. W., & Sunaryo, H. (2025). Effectiveness of implementing productive skills assessment in BIPA learning. *Acuity: Journal of English Language Pedagogy, Literature, and Culture*, *10*(3). https://doi.org/10.35974/acuity.v10i3.4025

Kang, M., & Lam, R. (2024). Understanding university English instructors' assessment literacy: A formative assessment perspective. *Language Testing in Asia*, *14*(1), Article 52. https://doi.org/10.1186/s40468-024-00323-y

Lai, C. J. (2024). Examining the impact of multimodal task design on English oral communicative competence in fourth-grade content-language integrated social studies: A quasi-experimental study. *Asian-Pacific Journal of Second and Foreign Language Education*, *9*(1), Article 64. https://doi.org/10.1186/s40862-024-00289-7

Latif, M. W., & Wasim, A. (2024). Investigating EFL instructors' approaches to classroom-based assessment culture: An explanatory sequential mixed-method approach. *Language Testing in Asia*, *14*(1), Article 57. https://doi.org/10.1186/s40468-024-00332-x

Lestari, L. K. P., Rahmawati, F., Amelia, I., Utami, N. A., & Handayani, R. (2025). The implementation of authentic assessment in speaking tests for secondary school students. *SELL (Scope of English Language Teaching, Linguistics, and Literature) Journal*, *10*(2), 88–100. https://doi.org/10.31597/sl.v10i2.1231

Li, W., Zhang, X., Li, J., Yang, X., Li, D., & Liu, Y. (2024). An explanatory study of factors influencing engagement in AI education at the K-12 level: An extension of the classic TAM model. *Scientific Reports*, *14*(1), Article 13922. https://doi.org/10.1038/s41598-024-64363-3

Liu, Y. (2024). The effect of classroom-based assessment (CBA) on Chinese EFL learners' speaking performance, engagement, and willingness to communicate: Undergraduate students' perceptions. *Heliyon*, *10*(12), Article e32629. https://doi.org/10.1016/j.heliyon.2024.e32629

Lumintang, S. F., & Kuswandono, P. (2025). Indonesian EFL lecturers' positive perceptions and adaptive strategies in implementing authentic assessment for sustainable education: Mixed-methods study. *Journal of English Language Teaching Innovations and Materials (Jeltim)*, *7*(2), 38–59. https://doi.org/10.26418/jeltim.v7i2.96334

Marden, M. P., & Herrington, J. (2025). Designing authentic assessments for foreign language learning environments. *Journal of Education, Innovation and Communication*, *7*(2). https://doi.org/10.34097/jeicom-7-2-4

Maslucha, S., Firmansyah, M. A., Uswah, M., & Rachmatian, N. A. (2024). Current practices of corrective feedback in Indonesian secondary EFL classrooms: A systematic review. *IREELL: Indonesian Review of English Education, Linguistics, and Literature*, *2*(1), 79–101. https://doi.org/10.30762/ireell.v2i1.2819

Nurfaidah, S., Khairunnisa, U., Syukri, S., & Jamilarin, P. A. P. (2024). Exploring teachers' corrective feedback in EFL classrooms: Pre-service teachers' reflections. *Journal on English as a Foreign Language*, *14*(2), 523–544. https://doi.org/10.23971/jefl.v14i2.8863

Ockey, G. J., & Na, I. (2025). Assessing second language listening in the 21st century. *Reference Module in Social Sciences*. https://doi.org/10.1016/B978-0-323-95504-1.00449-X

Pandey, G. P. (2025). Designing effective language tests: A framework for integrating stages, processes, and best practices. *Jo-ELT (Journal of English Language Teaching)*, *12*(1), 62–74. https://doi.org/10.33394/jo-elt.v12i1.15274

Qasserras, L. (2025). Systematic review of effective teaching listening practices in ESL/EFL settings. *European Journal of English Language Teaching*, *9*(6). https://doi.org/10.46827/ejel.v9i6.5768

Qizi, N. R. T., & Nabieva, A. (2025). Integrating speaking and listening skills through task-based activities in EFL classrooms. *Modern Education and Development*, *27*(1), 393–399.

Rosy, F., Husna, N., Kulstum, U., & Farkhan, M. (2025). Assessing the authenticity and validity of tests in CLT-based English classrooms in West Java's formal education. *Journal on Language Teaching and Learning, Linguistics and Literature*, *13*(2), 5474–5486. https://doi.org/10.24256/ideas.v13i2.5474

Sabol, M., Hair, J., Cepeda, G., Roldán, J. L., & Chong, A. Y. L. (2023). PLS-SEM in information systems: Seizing the opportunity and marching ahead full speed to adopt methodological updates. *Industrial Management and Data Systems*, *123*(12), 2997–3017. https://doi.org/10.1108/imds-07-2023-0429

Sadeghi, K., & Bakhshi, N. (2025). Integrated listening/speaking skill assessment: The role of ambiguity tolerance, cognitive/metacognitive strategy use, and foreign language anxiety. *International Journal of Language Testing*, *15*(1), 77–90. https://doi.org/10.22034/ijlt.2024.452581.1338

Shang, Y., Aryadoust, V., & Hou, Z. (2024). A meta-analysis of the reliability of second language listening tests (1991–2022). *Brain Sciences*, *14*(8), Article 746. https://doi.org/10.3390/brainsci14080746

Simanullang, A., Batrisyia, A., Hidayah, E., & Harahap, S. H. (2024). Eksplorasi upaya internasionalisasi bahasa dan budaya Indonesia melalui program BIPA. *Journal on Education*, *6*(4), 20267–20275. https://doi.org/10.31004/joe.v6i4.6090

Skehan, P. (2024). Linking second language speaking task performance and language testing. *Language Teaching*, *57*(4), 501–516. https://doi.org/10.1017/s0261444823000344

Souzandehfar, M. (2024). New perspectives on IELTS authenticity: An evaluation of the speaking module. *International Journal of Language Testing*, *14*(1), 34–55. https://doi.org/10.22034/ijlt.2023.409599.1272

Subhaktiyasa, P. G. (2024). PLS-SEM for multivariate analysis: A practical guide to educational research using SmartPLS. *EduLine: Journal of Education and Learning Innovation*, *4*(3), 353–365. https://doi.org/10.35877/454ri.eduline2861

Syafruddin, Pratiwi, B., Izzati, A. N., & Sianipar, I. F. (2025). Teaching Indonesian for Foreign Speakers (BIPA): A bibliometric analysis. *Issues in Language Studies*, *14*(1), 55–73. https://doi.org/10.33736/ils.7551.2025

Tran, T. T. T., & Ma, Q. (2024). Online peer feedback training based on self-regulated learning in English as a foreign language writing: Perceived usefulness and students' engagement. *Studies in Educational Evaluation*, *83*, Article 101418. https://doi.org/10.1016/j.stueduc.2024.101418

Van Der Steen, J., Van Schilt-Mol, T., Van Der Vleuten, C., Desirée, & Brinke, J. T. (2023). Designing formative assessment that improves teaching and learning: What can be learned from the design stories of experienced teachers? *Journal of Formative Design in Learning*, *7*(2), 182–194. https://doi.org/10.1007/s41686-023-00080-w

Veugen, M. J., Gulikers, J. T. M., & den Brok, P. (2024). Secondary school teachers' use of formative assessment practice to create co-regulated learning. *Journal of Formative Design in Learning*, *8*(1), 15–32. https://doi.org/10.1007/s41686-024-00089-9

Vlachopoulos, D., & Makri, A. (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*, Article 101425. https://doi.org/10.1016/j.stueduc.2024.101425

Xi, X. (2025). Revisiting communicative competence in the age of AI: Implications for large-scale testing. *Annual Review of Applied Linguistics*, *45*, 200–221. https://doi.org/10.1017/s0267190525000078

Xu, Y., Yang, M., & Li, X. (2025). Investigating the relationships between listening strategies and speaking performance in integrated listening-to-speak tasks. *System*, *129*, Article 103586. https://doi.org/10.1016/j.system.2024.103586

Yang, Z., & Wang, P. (2025). Current status and research trend of English language assessment: A bibliometric analysis. *Language Testing in Asia*, *15*(1), Article 11. https://doi.org/10.1186/s40468-024-00317-w