

30 KEMBARA.pdf

by Eko Edi

Submission date: 13-Aug-2026 10:01PM (UTC+0700)

Submission ID: 3011987532

File name: 30_KEMBARA.pdf (560.34K)

Word count: 6438

Character count: 43254



Transforming Indonesian language learning evaluation through OBE: A multisite mixed-methods study of speaking assessment in higher education

(Transformasi evaluasi pembelajaran Bahasa Indonesia berbasis OBE: Studi mixed methods multisitus tentang penilaian keterampilan berbicara di perguruan tinggi)

4

Muhlis Fajar Wicaksana
Universitas Veteran Bangun
Nusantara, Indonesia
muhlisfajarwicaksana@gmail.com

Nurratri Kurnia Sari
Universitas Veteran Bangun
Nusantara, Indonesia
nuurratrikurniasari@gmail.com

25 **dewi Kusumaningsih**
Universitas Veteran Bangun
Nusantara, Indonesia
dewikusumaningsih71@gmail.com

4

Nanci Risky Rimadani
Universitas Veteran Bangun
Nusantara, Indonesia
nancirimadani@gmail.com

Ly Aminas
Sekolah Bersepadu Musa-Asiah
(SEPAMA), Kamboja
aminasly85@gmail.com

*Corresponding author: Muhlis Fajar Wicaksana | email: muhlisfajarwicaksana@gmail.com

Article History Accepted: June 02, 2025 Revised: October 23, 2025 Available Online: October 31, 2025

Abstract: Evaluation of Indonesian language learning in an Outcome-Based Education (OBE) context requires more than grading or collecting digital assignments; it requires traceable alignment among learning outcomes, tasks, criteria, evidence, feedback, and improvement. This convergent multisite mixed-methods needs-assessment study investigated the transformation of speaking evaluation across seven higher-education institutions in Central Java. Participants comprised 10 lecturers of Speaking Skills courses in Indonesian Language and Literature Education programs, and 132 students, from whom 140 questionnaires were distributed; eight incomplete responses were excluded. Data were collected through semi-structured interviews, reviews of course plans, rubrics, speaking tasks, and digital evidence, and a ten-item Likert questionnaire. The questionnaire was content-validated by three experts and piloted with 30 students. Quantitative data were summarized using frequencies, means, standard deviations, and positive-response percentages, while qualitative evidence was examined through cross-site thematic comparison and integrated in a joint display. The overall questionnaire mean was 3.93/5. Rubric clarity received the highest mean ($M = 4.10$), and the importance of OBE-aligned technology-based evaluation received the strongest positive response (77.27%; $M = 4.09$). Experience using digital platforms for speaking-task submission was lowest (61.36%; $M = 3.75$). Lecturer and document evidence showed that authentic tasks, linguistic and nonlinguistic criteria, and OBE terminology were present, but digital tools remained concentrated on communication, submission, publication, and storage. Outcome mapping, rater calibration, timely feedback, longitudinal evidence, analytics, and ethical governance were not yet systematic. The study formulates a six-stage framework for transforming Indonesian language speaking evaluation through OBE: outcome alignment, authentic tasks, analytic rubrics, digital evidence, feedback and reflection, and outcome analytics for continuous improvement. The framework is a design proposition requiring validation and field testing.

Keywords

Digital transformation, Indonesian language learning evaluation, Mixed methods, Outcome-Based Education, Speaking assessment

Abstrak: Evaluasi pembelajaran Bahasa Indonesia dalam konteks Outcome-Based Education (OBE) tidak cukup dilakukan melalui pemberian skor atau pengumpulan tugas secara digital, tetapi juga membutuhkan keterlacakan antara capaian pembelajaran, tugas, kriteria, bukti performa, umpan balik, dan tindak lanjut perbaikan. Penelitian analisis kebutuhan mixed methods konvergen dan multisitus ini menguji transformasi evaluasi keterampilan berbicara di tujuh perguruan tinggi di Jawa Tengah. Partisipan terdiri atas 10 dosen pengampu mata kuliah Keterampilan Berbicara pada Program Studi Pendidikan Bahasa dan Sastra Indonesia dan 132 mahasiswa dari 140 kuesioner yang didistribusikan; delapan kuesioner yang tidak lengkap dikeluarkan dari analisis. Data dikumpulkan melalui wawancara semi-terstruktur, telaah RPS, rubrik, tugas berbicara, bukti digital, serta kuesioner Likert sepuluh butir. Kuesioner divalidasi isinya oleh 3 pakar dan diuji coba pada 30 mahasiswa. Data kuantitatif dianalisis menggunakan frekuensi, rerata, simpangan baku, dan persentase respons positif, sedangkan data kualitatif dianalisis melalui perbandingan tematik lintas situs dan diintegrasikan dalam joint display. Rerata keseluruhan kuesioner adalah 3,93/5. Kejelasan rubrik memperoleh rerata tertinggi ($M = 4,10$), sedangkan pentingnya evaluasi berbasis teknologi yang selaras dengan OBE memperoleh respons positif paling kuat (77,27%; $M = 4,09$). Pengalaman menggunakan platform digital untuk mengumpulkan tugas berbicara merupakan aspek terendah (61,36%; $M = 3,75$). Data dosen dan dokumen menunjukkan adanya tugas autentik, kriteria linguistik dan nonlinguistik, serta terminologi OBE, tetapi teknologi masih dominan digunakan untuk komunikasi, pengumpulan, publikasi, dan penyimpanan tugas. Pemetaan capaian, kalibrasi penilai, umpan balik tepat waktu, dokumentasi longitudinal, analitik, dan tata kelola etis belum dilakukan secara sistematis. Penelitian ini merumuskan

kerangka enam tahap transformasi evaluasi berbicara berbasis OBE, yaitu penyalarsan capaian, tugas autentik, rubrik analitik, bukti digital, umpan balik dan refleksi, serta analisis capaian untuk perbaikan berkelanjutan. Kerangka tersebut merupakan proposisi desain yang masih memerlukan validasi dan uji lapangan.

Kata Kunci *Transformasi digital, Evaluasi pembelajaran Bahasa Indonesia, Mixed methods, OBE, Penilaian keterampilan berbicara*

How to Cite Wicaksana, M. F., Sari, N. K., Kusumaningsih, D., Rimadani, N. R., & Aminas, L. (2025). Transforming Indonesian language learning evaluation through OBE: A multisite mixed-methods study of speaking assessment in higher education. *KEMBARA: Jurnal Keilmuan Bahasa, Sastra, dan Pengajarannya*, 11(2), 1135–1147. <https://doi.org/10.22219/kembara.v11i2.45482>



Copyright©2025, Muhlis Fajar Wicaksana, Nurratri Kurnia Sari, Dewi Kusumaningsih, Nanci Risky Rimadani, & Ly Aminas

This is an open-access article under the CC-BY-3.0 license.



INTRODUCTION

Evaluation in Indonesian language **learning** is not merely the assignment of scores; it is a systematic process for determining whether students can use Indonesian accurately, coherently, appropriately, and responsibly in academic and public communication. This requirement is especially visible in speaking courses within Indonesian Language and Literature Education programs, where presentations, interviews, discussions, storytelling, speech delivery, and public communication provide evidence of both linguistic competence and communicative performance. Accordingly, evaluation needs to represent linguistic dimensions—such as diction, grammatical accuracy, coherence, pronunciation, and fluency—and nonlinguistic dimensions, including confidence, eye contact, gesture, posture, audience engagement, and contextual appropriateness (Hughes & Reed, 2017; Purwadi et al., 2021).

The demand for defensible evaluation becomes stronger in an Outcome-Based Education (OBE) environment. OBE requires constructive alignment among program learning outcomes (CPL), course learning outcomes (CPMK), learning activities, assessment tasks, criteria, evidence, feedback, and subsequent improvement decisions (Biggs & Tang, 2011). A speaking score alone is insufficient unless the evaluation demonstrates which outcome was measured, what task elicited the evidence, how the performance was judged, and how the evidence was used to improve student learning and the curriculum. Validity, reliability, transparency, fairness, and educational consequences therefore remain central to OBE-based evaluation (AERA et al., 2014; Kane, 2013).

Digital technologies have expanded how Indonesian language speaking performance can be elicited, recorded, reviewed, and documented. Video platforms, social media, messaging applications, and learning-management systems can preserve audiovisual evidence, enable repeated observation, broaden authentic communication contexts, and support portfolios and feedback. Nevertheless, digital submission is not equivalent to the digital transformation of evaluation. A video uploaded through YouTube, TikTok, WhatsApp, or an institutional platform remains only a digital artifact when outcome mappings are unclear, rubrics are inconsistently interpreted, feedback is delayed, and performance data are not used for continuous improvement (Gikandi et al., 2011; Nieminen et al., 2023).

Authenticity and evaluative quality provide two complementary foundations for this transformation. Speaking evaluation is more authentic when tasks resemble meaningful communication, students understand the expected performance standards, and evidence is interpreted in relation to genuine communicative demands (Hasrol et al., 2022; Menggo & Gunas, 2022). Analytic rubrics can make criteria transparent and strengthen consistency, but only when descriptors represent the intended construct, assessors interpret them similarly, and results generate actionable feedback (Jonsson & Svingby, 2007; Panadero & Jonsson, 2013; Anisah & Amreta, 2023). Therefore, the transformation of Indonesian language learning evaluation must integrate OBE alignment, authentic performance, sound judgment, digital evidence, and improvement-oriented feedback.

Previous research has examined authentic speaking assessment, digital media for speaking

instruction, and students' perceptions of technology-supported assessment. However, fewer studies have investigated the evaluation system as an interconnected chain across multiple Indonesian higher-education sites. In particular, limited empirical attention has been given to whether speaking tasks, rubrics, digital evidence, feedback, and program-level decisions are coherently aligned with CPL and CPMK. This gap is consequential because the visible adoption of technology may conceal an evaluation process that remains administratively digital but pedagogically and psychometrically conventional.

This study addresses the gap through a multisite mixed-methods needs assessment of speaking evaluation in Indonesian Language and Literature Education programs. Its contribution is twofold. Empirically, it identifies the current practices, perceptions, and maturity gaps that characterize the transition toward OBE-based evaluation. Conceptually, it translates the integrated findings into a six-stage framework connecting outcome alignment, authentic speaking tasks, analytic rubrics, secure digital evidence, feedback and reflection, and outcome analytics for continuous quality improvement. The framework is proposed as a development direction rather than as a validated digital or artificial-intelligence system.

Three research questions guide the study: (1) How is the evaluation of Indonesian language speaking learning planned and implemented across the participating institutions, particularly in relation to tasks, rubrics, OBE alignment, and digital platforms? (2) How do students perceive speaking-task frequency, outcome relevance, rubric transparency, fairness, digital experience, and readiness for technology-supported evaluation? (3) What design gaps and principles should inform the transformation of OBE-aligned Indonesian language speaking evaluation in higher education?

16

METHOD

Research Design

The study employed a convergent multisite mixed-methods needs-assessment design (Creswell & Clark, 2018). The unit of analysis was the evaluation practice used in Speaking Skills courses within Indonesian Language and Literature Education programs. The qualitative strand used lecturer interviews and course documents to examine evaluation planning, OBE alignment, task authenticity, rubric use, feedback, digital practices, and implementation constraints. The quantitative strand used a cross-sectional student questionnaire to describe perceptions of current evaluation practices and readiness for technology-supported evaluation. Both strands were analyzed separately and integrated during interpretation through a joint display.

Sites and Participants

The study was conducted at seven higher-education institutions in Central Java: Universitas Sebelas Maret, Universitas Islam Negeri Raden Mas Said Surakarta, Universitas Muhammadiyah Purworejo, Universitas Jenderal Soedirman, Universitas PGRI Semarang, Universitas Pancasakti Tegal, and Universitas Veteran Bangun Nusantara. The qualitative participants were 10 lecturers who taught Speaking Skills courses across the seven sites. Lecturers were selected purposively using three criteria: they were responsible for a speaking course, had at least three years of teaching experience, and had implemented OBE-based learning. The student respondents were enrolled in or had completed a Speaking Skills course. A total of 140 questionnaires were distributed; data screening identified 132 complete questionnaires that met the predefined completeness criterion. Eight questionnaires containing missing responses in key items were excluded, resulting in a usable-response rate of 94.3%.

Table 1
Questionnaire Blueprint for OBE-Based Indonesian Language Speaking Evaluation

Construct	Dimension	Indicator	Item	Scale
Speaking evaluation practice	Assessment planning	Frequency of speaking assignments	1	Likert 1–5
Speaking evaluation practice	Assessment planning	Relevance of speaking tasks to learning outcomes	2	Likert 1–5
Speaking evaluation practice	Assessment criteria	Clarity of the assessment rubric	3	Likert 1–5
Speaking evaluation practice	Assessment criteria	Student understanding of assessment criteria	4	Likert 1–5
Speaking evaluation practice	Assessment implementation	Objectivity and fairness of lecturer assessment	5	Likert 1–5
Digital transformation of evaluation	Technology use	Experience using digital platforms for assessment	6	Likert 1–5
Digital transformation of evaluation	Technology use	Ease and comfort in using digital platforms	7	Likert 1–5
Digital transformation of evaluation	Digital-assessment perception	Efficiency and objectivity of digital evaluation	8	Likert 1–5
Digital transformation of evaluation	Digital-assessment perception	Readiness to use a digital evaluation system	9	Likert 1–5
Digital transformation of evaluation	Digital-assessment perception	Importance of OBE-aligned digital evaluation development	10	Likert 1–5

Instruments and Data Sources

The qualitative data were collected through semi-structured lecturer interviews and document review. Interview domains covered lecturers' understanding of evaluation transformation, digital-platform use, alignment with CPL and CPMK, infrastructure, technical constraints, ¹⁸ implementation, feedback quality, workload, and readiness for further development. The semi-structured ²¹ format allowed follow-up questions, clarification, and exploration of experiences emerging during the interviews. Interviews were recorded with participant permission and transcribed into written form. The document review examined available RPS documents, speaking-task designs, assessment rubrics, and evidence of digital implementation in order to contextualize and triangulate lecturer accounts.

The student questionnaire was developed from theoretical constructs concerning speaking ⁵ evaluation, OBE, and the digital transformation of learning evaluation. Ten statements were rated on a five-point Likert scale from 1 (strongly disagree) to 5 (strongly agree). Content validity was reviewed by three experts: one specialist in educational evaluation, one specialist in Indonesian language education, and one specialist in OBE-based curriculum. They assessed construct alignment, indicator relevance, wording clarity, and item readability. Their comments led to simplified wording, consistent OBE terminology, and refined digital-evaluation indicators. The revised questionnaire was piloted with 30 students who had completed the Speaking Skills course at Universitas Veteran Bangun Nusantara. The pilot focused on wording clarity, readability, construct–indicator fit, and respondent comprehension. Respondent-level pilot data were not retained in the available archive; consequently, an auditable Cronbach's alpha coefficient could not be recalculated or reported.

Data Analysis and Integration

The 132 complete questionnaires were analyzed descriptively. Item frequencies were converted into means, sample standard deviations, and positive-response percentages, defined as the combined proportion of agree and strongly agree responses. Because respondent-level records were not available in the archived dataset, the analysis did not estimate internal consistency, test causal effects, or compare groups. Interview and document evidence were organized through cross-site

thematic comparison around OBE alignment, authentic tasks, digital use, rubric quality, feedback, infrastructure, documentation, readiness, and governance. Integration compared quantitative patterns with qualitative and documentary evidence in a joint display and used convergence, complementarity, and discrepancy to formulate design requirements for transforming Indonesian language speaking evaluation.

Ethical Considerations

Before data collection, the research team submitted institutional research-permission letters through the Institute for Research and Community Service (LPPM) of Universitas Veteran Bangun Nusantara and obtained authorization from the heads of the participating study programs. Lecturers and students received information about the study objectives, data-collection procedures, expected benefits and minimal risks, confidentiality, and their right to discontinue participation at any time without academic or professional consequences. Informed consent was obtained before questionnaire completion and before each interview was recorded.

Participation was voluntary. Survey, interview, and document records were de-identified, and reporting used aggregated quantitative results and cross-site themes rather than personally identifiable statements. Research data were stored in secure, researcher-controlled files and were not disclosed to third parties without participant permission. Student audiovisual evidence was treated as confidential research material and was not reproduced in this article. The available archive did not contain a separate ethics-committee approval or exemption number; accordingly, the manuscript reports the documented institutional-permission and informed-consent procedures without claiming formal committee approval.

RESULTS AND DISCUSSION

Results

Student Questionnaire Results

All 10 questionnaire items contained 132 complete responses. Across the items, the mean score was 3.93 on a five-point scale, and the average positive-response rate was 68.94%. The pattern indicates generally favorable perceptions of speaking evaluation and OBE-aligned digital development, while also showing that students' direct digital-assessment experience was less established than their recognition of its importance.

Table 2
Student Perceptions of OBE-Based and Digitally Supported Speaking Evaluation

Questionnaire item	M	SD	Positive response
1. Frequent speaking assignments	4.05	0.92	72.73%
2. Speaking tasks are relevant to communication and learning outcomes	3.98	0.93	70.45%
3. Clear assessment rubric	4.10	0.88	73.48%
4. Understand assessment criteria	3.94	0.93	68.94%
5. Fair and objective lecturer assessment	3.91	0.99	66.67%
6. Used digital platforms for speaking-task assessment	3.75	1.14	61.36%
7. Comfortable demonstrating speaking skills digitally	3.77	0.95	65.15%
8. Digital evaluation is more objective and efficient	3.86	0.87	68.18%
9. Ready to use a digital speaking-evaluation system	3.83	0.90	65.15%
10. OBE-aligned technology-based evaluation is important	4.09	0.88	77.27%

The highest item mean concerned rubric clarity ($M = 4.10$, $SD = 0.88$), while the strongest

positive-response percentage concerned the importance of OBE-aligned technology-based evaluation (77.27%; $M = 4.09$, $SD = 0.88$). Frequent speaking assignments (72.73%; $M = 4.05$) and the relevance of tasks to communication and learning outcomes (70.45%; $M = 3.98$) were also positively perceived. These results indicate that students recognized task frequency, outcome relevance, and transparent criteria as important foundations of Indonesian language speaking evaluation.

The lowest positive response concerned experience using digital platforms for speaking-task assessment (61.36%; $M = 3.75$, $SD = 1.14$). Comfort in demonstrating speaking skills digitally and readiness to use a digital evaluation system were both positive but more moderate (65.15%). Thus, students' conceptual support for digital transformation was stronger than their consistent experience of a mature digital evaluation process.

Lecturer and Document Themes

Cross-site evidence from 10 lecturers and the available course documents yielded five interconnected themes. First, the participating programs had established foundations for authentic and OBE-oriented evaluation. Lecturers used individual presentations, group discussions, job-interview simulations, field interviews, speeches, storytelling, and speaking content for YouTube or TikTok. These tasks provided opportunities to observe Indonesian language use in academic, professional, and public communication contexts, although explicit task-to-CPL/CPMK mappings were not consistently documented.

Second, digital tools were widely used, but their evaluative functions remained predominantly administrative and distributive. Google Meet and WhatsApp supported communication and online sessions, while YouTube, Instagram, and TikTok enabled production, publication, and submission of speaking performances. The evidence did not show an integrated architecture linking outcome indicators, scoring records, feedback, portfolios, and program-level analytics.

Third, lecturers recognized both linguistic and nonlinguistic dimensions of speaking performance. Linguistic criteria included diction, language style, sentence effectiveness, logic, coherence, pronunciation, and fluency; nonlinguistic criteria included articulation, intonation, eye contact, gesture, facial expression, posture, audience engagement, and confidence. However, criterion weighting, performance exemplars, assessor calibration, moderation, and inter-rater evidence were not systematically documented.

Fourth, feedback and workload constrained the formative function of evaluation. Large classes and limited instructional time restricted detailed and timely comments. Although digital recordings could be replayed, the available evidence did not demonstrate a consistent cycle connecting rubric results, student reflection, revision or follow-up tasks, and monitoring of improvement. Fifth, governance requirements—including privacy, accessibility, academic integrity, data access, and human oversight—had not yet been formalized as an integrated evaluation policy.

Table 3
Cross-Site Themes in the Transformation of Indonesian Language Speaking Evaluation

Theme	Current practice	Transformation requirement
OBE-oriented planning and authentic tasks	Presentations, discussions, interviews, simulations, and public digital content	Explicit mapping from tasks and evidence to CPL/CPMK indicators and performance standards
Digitalization of evaluation	Communication, recording, submission, publication, and storage	Integrated evidence repository, metadata, audit trail, portfolio, and outcome analytics
Criteria and assessor judgment	Linguistic and nonlinguistic criteria are recognized	Behavioral descriptors, weighting, exemplars, rater training, calibration, moderation, and reliability evidence
Feedback and improvement	Lecturer comments and replayable audiovisual evidence	Timely multimodal feedback, student reflection, feed-forward tasks, and progress tracking

Theme	Current practice	Transformation requirement
Ethical governance	Procedures are fragmented or not systematically reported	Validity, fairness, accessibility, privacy, academic integrity, secure access, and human oversight

Integrated Findings

The integration revealed a readiness–maturity gap in the evaluation system. Students valued speaking practice, clear rubrics, and technology-supported evaluation, while lecturers had already implemented authentic tasks and familiar digital platforms. Nevertheless, the evidence chain required by OBE—outcomes, tasks, criteria, judgment, digital evidence, feedback, analytics, and improvement—was only partially connected. The central transformation challenge was therefore not the absence of technology but the absence of a coherent evaluation architecture.

Table 4
Joint Display of Student Perceptions and Cross-Site Evaluation Practices

Quantitative pattern	Qualitative/document evidence	Integrated inference
Rubric clarity received the highest item mean ($M = 4.10$).	Lecturers used linguistic and nonlinguistic criteria.	Criteria are visible, but defensible judgment requires weighting, exemplars, calibration, moderation, and reliability evidence.
OBE-aligned technology-based evaluation was strongly endorsed (77.27%).	All sites used one or more digital platforms.	Readiness for transformation is high, but technology is still used mainly for communication, production, and submission.
Digital-platform assessment experience was the lowest item (61.36%).	Platform use varied and was not integrated into a common evaluation workflow.	Uneven experience may constrain consistent implementation across programs and institutions.
Perceived fairness was positive but lower than rubric clarity (66.67%).	No systematic moderation or inter-rater evidence was documented.	Transparency does not by itself ensure reliability, fairness, or comparable standards.
Digital comfort and readiness reached 65.15%.	Lecturers reported workload, infrastructure, confidence, and student anxiety as constraints.	Implementation requires learner orientation, assessor development, infrastructure, and ethical governance.

Proposed Framework for Transforming OBE-Based Indonesian Language Speaking Evaluation

The integrated gaps informed a six-stage framework. The cycle begins with explicit alignment of CPL, CPMK, indicators, and performance standards; continues through authentic Indonesian language speaking tasks and analytic rubrics; captures secure multimodal digital evidence; provides feedback, learner reflection, and feed-forward; and aggregates evidence for outcome analytics and continuous quality improvement. Validity, reliability, fairness, accessibility, privacy, academic integrity, secure access, and human oversight operate across all stages. The framework is a design proposition derived from the needs assessment and therefore requires expert validation, co-design, field testing, and evaluation of rater and learning outcomes.

Framework for Transforming OBE-Based Indonesian Language Speaking Evaluation

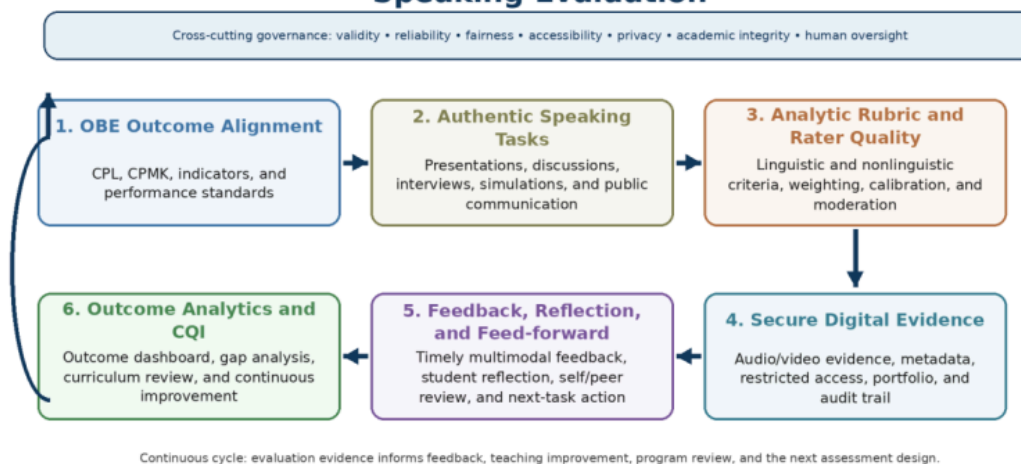


Figure 1. Proposed Framework for Transforming OBE-Based Indonesian Language Speaking Evaluation

Discussion

Digitalization Has Not Yet Transformed Indonesian Language Learning Evaluation

The findings portray a transitional evaluation ecosystem. Digital platforms have expanded where and how students demonstrate Indonesian speaking performance, but the dominant practices remain task distribution, recording, uploading, publication, communication, and storage. This pattern confirms that technological adoption should be judged by its contribution to evaluative quality, not by the mere presence of digital media. A transformed system needs to make outcomes traceable, preserve credible evidence, support consistent judgment, accelerate feedback, and connect course results to program improvement (Nieminen et al., 2023; Atmazaki et al., 2021).

YouTube, TikTok, Instagram, WhatsApp, and Google Meet can enrich authentic contexts and preserve performances for repeated observation. These affordances are valuable for Indonesian language learning because speaking competence includes verbal accuracy, discourse organization, audience awareness, and embodied communication. However, replayable video does not automatically produce valid evaluation. The construct, task specifications, rubric, assessor process, use of scores, and consequences for students must remain explicit (AERA et al., 2014; Kane, 2013). Digitalization becomes transformation only when these components operate as an integrated evaluation system.

OBE Alignment Requires a Traceable Evaluation Evidence Chain

The participating courses already contained important OBE foundations: authentic performance tasks, frequent speaking opportunities, assessment criteria, and the use of CPL and CPMK terminology. However, OBE is not demonstrated by terminology alone. A defensible evaluation chain should connect each outcome indicator to task specifications, observable evidence, rubric criteria, scoring decisions, feedback, and curriculum-improvement actions. Without this traceability, scores remain isolated course records rather than evidence of outcome attainment.

The proposed framework operationalizes constructive alignment as a continuous cycle. Outcome statements determine the communication performance to be elicited; authentic tasks create relevant evidence; analytic rubrics structure judgment; secure digital records preserve the evidence; feedback and reflection support improvement; and aggregated analytics identify outcome gaps. The cycle returns evidence to teaching and curriculum decisions, repositioning evaluation from an endpoint to a mechanism for continuous quality improvement (Biggs & Tang, 2011; Boud & Falchikov, 2006).

Student Readiness Exceeds Evaluation-System Maturity

Student responses were generally favorable, particularly regarding rubric clarity, task frequency, outcome relevance, and the importance of technology-supported evaluation. At the same time, direct experience with digital assessment and comfort with digital performance were comparatively lower. This distinction means that acceptance cannot be equated with uniform competence or readiness. Institutions need to provide orientation, low-stakes practice, accessible platforms, technical support, and accommodations for performance anxiety before using digital speaking evaluation for high-stakes decisions.

The positive perception of authentic speaking tasks is consistent with evidence that performance-based assessment can make communicative competence more visible and meaningful (Menggo & Gunas, 2022). Nevertheless, perception data cannot establish validity, reliability, or equity. Student experience should inform design, while technical quality must be examined through instrument validation, rater studies, moderation, evidence of consequences, and review of whether the evaluation supports learning rather than merely documenting performance.

Rubrics, Feedback, and Assessor Quality Are the Core Evaluation Infrastructure

Recognizing linguistic and nonlinguistic criteria provides a strong basis for evaluating Indonesian speaking performance. The next priority is not simply converting the rubric into a digital form but strengthening the quality of judgment. Criteria need behavioral descriptors, weights appropriate to the task purpose, annotated performance exemplars, and shared interpretation through assessor calibration. Research on rubrics shows that transparency can support learning, but consistency depends on clear criteria and disciplined scoring processes (Jonsson & Svingby, 2007; Panadero & Jonsson, 2013; Caesarine & Setyaningsih, 2023).

Feedback is equally central to evaluation transformation. Large classes make individualized feedback difficult, yet digital evidence can support timestamped comments, audio or video feedback, guided self-review, peer review, and forward tasks. Feedback becomes educationally effective when students understand it, evaluate its relevance, and use it in subsequent performance (Carless & Boud, 2018). Therefore, the system should preserve not only scores but also feedback, student responses, and evidence of improvement across tasks.

Digital Transformation Must Remain Ethical and Human-Supervised

The data concerned commonly used digital platforms and perceptions of technology; they did not test an operational artificial-intelligence system. AI-assisted functions—such as speech recognition, automated feature extraction, feedback suggestions, or learning analytics—should therefore be treated as a future development option rather than as a finding of this study. Their educational value would depend on construct representation, accuracy for diverse Indonesian accents and speaking contexts, explainability, privacy, and the avoidance of automated decisions that disadvantage students.

Any AI-supported evaluation should remain under professional human oversight and should be evaluated against the same validity, reliability, fairness, and consequential standards applied to other assessment practices. UNESCO emphasizes human agency, transparency, data protection, inclusion, and institutional accountability in educational AI (Miao & Holmes, 2023; UNESCO, 2021). In this framework, technology supports evidence management and professional judgment; it does not replace the lecturer's responsibility to interpret performance contextually and ethically (Rodiyah & Hartiningsih, 2023; Ruaidah et al., 2024; Uswati et al., 2022).

Pedagogical and Institutional Implications

At course level, lecturers should map each speaking task to specific CPMK indicators, provide analytic rubrics and exemplars before performance, conduct low-stakes calibration, and use self- and peer-evaluation to develop students' evaluative judgment. Feedback should be linked to a subsequent speaking task so that improvement is observable. At program level, institutions should develop a

secure evidence architecture instead of relying on public social-media links as the primary archive. The architecture should preserve consent, access controls, metadata, scoring records, feedback, outcome mappings, and longitudinal evidence while enabling program-level aggregation.

Professional development should prioritize outcome-based task design, speaking-construct validity, rubric construction, assessor moderation, accessible evaluation, digital feedback, and ethical data governance. Technology procurement should follow pedagogical and measurement requirements rather than determine them. The proposed framework should proceed to a subsequent research phase involving expert validation, co-design with lecturers and students, prototype development, usability testing, inter-rater studies, and comparative evaluation of feedback quality, student uptake, and outcome decisions.

Limitations

The study has several limitations. First, respondent-level questionnaire and pilot-test records were not available in the archived dataset; consequently, the quantitative strand was limited to descriptive analysis and an auditable reliability coefficient could not be reported. Second, complete interview transcripts and an exact inventory of documents from every site were not available for reanalysis, so qualitative reporting emphasizes cross-site themes rather than site-level comparison or frequency claims. Third, the seven institutions and 10 lecturers were purposively selected, and the findings do not estimate the prevalence of these practices across all Indonesian higher education. Fourth, student perceptions do not establish the technical quality or learning effectiveness of the evaluation system. Fifth, the archive did not include a separate ethics-committee approval or exemption number. Finally, the proposed framework was derived from needs-assessment evidence and literature but has not yet undergone expert validation, prototype testing, rater-reliability analysis, or impact evaluation.

CONCLUSION

The transformation of Indonesian language learning evaluation through OBE is underway but remains incomplete across the seven participating institutions. Speaking courses have incorporated authentic tasks, linguistic and nonlinguistic criteria, OBE terminology, and multiple digital platforms. Students generally valued task frequency, outcome relevance, rubric clarity, and technology-supported evaluation. Nevertheless, digital practices were still concentrated on communication, recording, submission, publication, and storage rather than on a traceable evaluation chain. Outcome mapping, assessor calibration, systematic feedback, longitudinal documentation, analytics, and ethical governance require further development. The study therefore proposes a six-stage framework linking CPL/CPMK alignment, authentic Indonesian speaking tasks, analytic rubrics and assessor quality, secure digital evidence, feedback and reflection, and outcome analytics for continuous improvement. Validity, reliability, fairness, accessibility, privacy, academic integrity, and human oversight operate across the entire cycle. For implementation, institutions should first strengthen evaluation design and assessor quality, then establish secure evidence and feedback workflows, and only afterward test advanced analytics or AI-supported functions. This sequence positions technology as infrastructure for defensible, learning-oriented professional judgment rather than as a substitute for that judgment.

6DECLARATIONS	
Author contributions	: All authors contributed substantively to the conception and design of the study, instrument development, data collection, analysis and interpretation, manuscript drafting, critical revision, and approval of the final version. Ly Aminas additionally contributed to methodological review, comparative interpretation, international literature review, and critical revision from an international perspective.
Funding statement	: This research was supported by the Directorate of Research and Community Service (DPPM), Ministry of Higher Education, Science, and

	Technology of the Republic of Indonesia, under the 2025 research-grant scheme. The present article is an additional scholarly output and was not designated as the mandatory grant deliverable.
3 Conflict of interest	The authors declare that they have no conflicts of interest that could have influenced the conduct of the study, data analysis, or publication of this article.
Ethics and consent	The study was conducted under institutional research permissions processed through LPPM Universitas Veteran Bangun Nusantara and the participating study programs. Participation was voluntary and informed consent was obtained before questionnaire completion and recorded interviews. Data were anonymized and stored securely. A separate ethics-committee approval or exemption number was not available in the archived documentation.
Data availability	Participant-level data are not publicly available because they contain confidential educational and interview information. Aggregated results are reported in the article; access to additional de-identified materials may be considered by the corresponding author subject to institutional authorization and participant-consent requirements.

REFERENCE

- Ajjawi, R., & Boud, D. (2017). Researching feedback dialogue: An interactional analysis approach. *Assessment & Evaluation in Higher Education*, 42(2), 252–265. <https://doi.org/10.1080/02602938.2015.1102863>
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.
- Anisah, G., & Amreta, M. Y. (2023). Pengembangan instrumen asesmen as learning berbasis proyek untuk pembelajaran menyimak dan berbicara monologis dan dialogis yang bermuatan karakter Bhineka Tunggal Ika. *KEMBARA: Jurnal Keilmuan Bahasa, Sastra, dan Pengajarannya*, 9(1), 76–88. <https://doi.org/10.22219/kembara.v9i1.22701>
- Atmazaki, A., Ramadhan, S., Indriyani, V., & Nabila, J. (2021). Dialogic-interactive media: Alternative learning media to improve speaking skills. *KEMBARA: Jurnal Keilmuan Bahasa, Sastra, dan Pengajarannya*, 7(2), 286–296. <https://doi.org/10.22219/kembara.v7i2.16402>
- Biggs, J., & Tang, C. (2011). *Teaching for quality learning at university* (4th ed.). Open University Press.
- Boud, D., & Falchikov, N. (2006). Aligning assessment with long-term learning. *Assessment & Evaluation in Higher Education*, 31(4), 399–413. <https://doi.org/10.1080/02602930600679050>
- Boud, D., & Molloy, E. (2013). Rethinking models of feedback for learning: The challenge of design. *Assessment & Evaluation in Higher Education*, 38(6), 698–712. <https://doi.org/10.1080/02602938.2012.691462>
- Brookhart, S. M. (2018). *Appropriate criteria: Key to effective rubrics*. ASCD.
- Brown, H. D., & Abeywickrama, P. (2019). *Language assessment: Principles and classroom practices* (3rd ed.). Pearson.
- Caesarine, R. D., & Setyaningsih, Y. (2023). The Values of Local Wisdom in the Oral Tradition of Healing Spells for the People of Osing, Banyuwangi Regency: An Anthropolinguistic Study. *KEMBARA: Jurnal Keilmuan Bahasa, Sastra, dan Pengajarannya*, 9(2), 668–680. <https://doi.org/10.22219/kembara.v9i2.26056>
- Carless, D. (2015). *Excellence in university assessment: Learning from award-winning practice*. Routledge.
- Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, 43(8), 1315–1325. <https://doi.org/10.1080/02602938.2018.1463354>

- Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). Sage.
- Djonaidi, S., Wahyuni, N., & Nova, F. (2021). Pengaruh media poster digital dalam pembelajaran daring pada masa pandemi terhadap kemampuan berbicara mahasiswa. *JINOTEP*, 8(1), 38–46. <https://doi.org/10.17977/um031v8i12021p038>
- Gikandi, J. W., Morrow, D., & Davis, N. E. (2011). Online formative assessment in higher education: A review of the literature. *Computers & Education*, 57(4), 2333–2351. <https://doi.org/10.1016/j.compedu.2011.06.004>
- Hasrol, S. B., Zakaria, A., & Aryadoust, V. (2022). A systematic review of authenticity in second language assessment. *Research Methods in Applied Linguistics*, 1(3), 100023. <https://doi.org/10.1016/j.rmal.2022.100023>
- Hidayati, N. A., & Darmuki, A. (2021). Penerapan model Auditory Intellectual Repetition untuk meningkatkan kemampuan berbicara pada mahasiswa. *Jurnal Educatio*, 7(1), 252–259. <https://doi.org/10.31949/educatio.v7i1.959>
- Hughes, R., & Reed, B. S. (2017). *Teaching and researching speaking* (3rd ed.). Routledge. <https://doi.org/10.4324/9781315692395>
- Jang, E. E., & Sawaki, Y. (2025). Advancing language assessment for teaching and learning in the era of the artificial intelligence revolution: Promises and challenges. *Language Testing*, 42(4), 361–368. <https://doi.org/10.1177/02655322251348685>
- Jeon, J., Lee, S., & Choe, H. (2023). Beyond ChatGPT: A conceptual framework and systematic review of speech-recognition chatbots for language learning. *Computers & Education*, 206, 104898. <https://doi.org/10.1016/j.compedu.2023.104898>
- Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity, and educational consequences. *Educational Research Review*, 2(2), 130–144. <https://doi.org/10.1016/j.edurev.2007.05.002>
- Kamaruddin, A., Fitria, N., & Patmasari, A. (2021). Needs analysis-based ESP course design for vocational high school students. *KEMBARA: Jurnal Keilmuan Bahasa, Sastra, dan Pengajarannya*, 7(2), 222–231. <https://doi.org/10.22219/kembara.v7i2.17591>
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73. <https://doi.org/10.1111/jedm.12000>
- Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- Menggo, S., & Gunas, T. (2022). College students' perceptions on performance-based assessment use in boosting speaking ability. *International Journal of Language Education*, 6(4), 423–436. <https://doi.org/10.26858/ijole.v6i4.22910>
- Miao, F., & Holmes, W. (2023). *Guidance for generative AI in education and research*. UNESCO.
- Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis: A methods sourcebook* (3rd ed.). Sage.
- Nieminen, J. H., Bearman, M., & Ajjawi, R. (2023). Designing the digital in authentic assessment: Is it fit for purpose? *Assessment & Evaluation in Higher Education*, 48(4), 529–543. <https://doi.org/10.1080/02602938.2022.2089627>
- Noviani, M., & Hasan, M. A. K. (2023). Problematika dan solusi pembelajaran keterampilan berbicara pada mahasiswa Program Studi Pendidikan Bahasa Arab. *Rayah Al-Islam*, 7(1), 245–259. <https://doi.org/10.37274/rais.v7i1.662>
- Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review*, 9, 129–144. <https://doi.org/10.1016/j.edurev.2013.01.002>
- Purwadi, A. J., Yulistio, D., & Kurniawan, R. (2021). Peningkatan keterampilan berbicara mahasiswa melalui praktik berpidato dan bercerita menggunakan video pembelajaran. *Silampari Bisa*, 4(2), 373–397. <https://doi.org/10.31540/silamparibisa.v4i2.1480>

- Rodiyah, S. F., & Hartiningsih, S. (2023). Investigating young learners' oral fluency through language immersion programs in the home country: A case study. *KEMBARA: Jurnal Keilmuan Bahasa, Sastra, dan Pengajarannya*, 9(2), 718–734. <https://doi.org/10.22219/kembara.v9i2.27972>
- Ruaidah, S., Wahyudi, & Nurchasanah. (2024). Strategy for influencing through Indonesian persuasive motivator. *KEMBARA: Jurnal Keilmuan Bahasa, Sastra, dan Pengajarannya*, 10(2), 455–470. <https://doi.org/10.22219/kembara.v10i2.33276>
- Rukmini, D., & Saputri, L. A. D. E. (2017). The authentic assessment to measure students' English productive skills based on the 2013 curriculum. *Indonesian Journal of Applied Linguistics*, 7(2), 263–273. <https://doi.org/10.17509/ijal.v7i2.8128>
- Sambell, K., McDowell, L., & Montgomery, C. (2013). *Assessment for learning in higher education*. Routledge.
- Tai, J., Ajjawi, R., Boud, D., Dawson, P., & Panadero, E. (2018). Developing evaluative judgement: Enabling students to make decisions about the quality of work. *Higher Education*, 76, 467–481. <https://doi.org/10.1007/s10734-017-0220-3>
- UNESCO. (2021). *Recommendation on the ethics of artificial intelligence*.
- Uswati, T. S., Nesi, A., & Wagiran. (2022). Implementasi evaluasi CIPP untuk pembelajaran daring teknik kewirausahaan: Studi kasus di IAIN Syekh Nurjati Cirebon. *KEMBARA: Jurnal Keilmuan Bahasa, Sastra, dan Pengajarannya*, 8(2), 288–302. <https://doi.org/10.22219/kembara.v8i2.21314>
- Wicaksana, M. F. (2021). Program Guru Ayo Menulis: Upaya peningkatan kompetensi profesional guru melalui penyusunan instrumen penilaian autentik. *Jurnal Terapan Abdimas*, 6(1), 43–48.
- Wicaksana, M. F., Suwandi, S., Winarni, R., & Ngadiso. (2019). The problems in applying authentic assessment for speaking skills in junior high school. *Indonesian Language Education and Literature*, 4(2), 226–239. <https://doi.org/10.24235/ileal.v4i2.3294>

ORIGINALITY REPORT

6%

SIMILARITY INDEX

6%

INTERNET SOURCES

5%

PUBLICATIONS

2%

STUDENT PAPERS

PRIMARY SOURCES

1	ejournal.umm.ac.id Internet Source	1%
2	etasr.com Internet Source	1%
3	kuey.net Internet Source	<1%
4	journal.assyfa.com Internet Source	<1%
5	academicinspired.com Internet Source	<1%
6	pajar.ejournal.unri.ac.id Internet Source	<1%
7	papers.iafor.org Internet Source	<1%
8	Submitted to Universitas Katolik Musi Charitas Student Paper	<1%
9	educationalchallenges.org.ua Internet Source	<1%
10	sidut88.blogspot.com Internet Source	<1%
11	jurnal.globalscients.com Internet Source	<1%
12	paeale.conferenceservice.jp Internet Source	<1%
13	repository.anu.ac.ke Internet Source	<1%
14	sciedupress.com Internet Source	<1%

15	Huseyin ATES. "Human-Centered GenAI Feedback Design in Higher Education: A Multisite Experiment on Direct, Reflective, and Hybrid Approaches to Scientific Argumentation", Springer Science and Business Media LLC, 2026 Publication	<1 %
16	econference.usd.ac.id Internet Source	<1 %
17	ojs.unm.ac.id Internet Source	<1 %
18	conference.herdsa.org.au Internet Source	<1 %
19	euraseans.com Internet Source	<1 %
20	ojs.scipub.de Internet Source	<1 %
21	Atta Gebril. "Learning-Oriented Language Assessment - Putting Theory into Practice", Routledge, 2021 Publication	<1 %
22	Auli Toom, Pyhältö Kirsi, Lynn McAlpine, Jan van Tartwijk. "The Routledge International Handbook of Learning and Instruction in University Education", Routledge, 2026 Publication	<1 %
23	Chafit Ulya, Nugraheni Eko W, Yant Mujiyanto. "MUATAN PENDIDIKAN ANTIKORUPSI DALAM MATA KULIAH PENGKAJIAN DAN APRESIASI PUISI DI JAWA TENGAH DAN YOGYAKARTA", Indonesian Language Education and Literature, 2016 Publication	<1 %
24	Rathburn, Terri R.. "Investigating Connections Between Cognitive Coaching, Student Agency, and Achievement in Ninth Grade Algebra 1.", University of South Carolina Publication	<1 %
25	e-journal.unipma.ac.id Internet Source	<1 %

Exclude quotes

On

Exclude matches

Off

Exclude bibliography

On