

Capturing Feedback on Live EFL Academic Presentations: A Design-Based Study of a Web-Based Tool

Muhaimin Abdullah^{1*}, Ahmad Munir¹, Syafi'ul Anam¹, Lies Amin Lestari¹, Widyastuti¹, Chester B. Esnara²

¹ Universitas Negeri Surabaya, ² Benguet State University

*Corresponding author: muhaiminabdullah@unesa.ac.id

HOW TO CITE: | ABSTRACT:

Muhaimin Abdullah, Ahmad Munir, Syafi'ul Anam, Lies Amin Lestari, Widyastuti, Chester B. Esnara (2026). Capturing Feedback on Live EFL Academic Presentations: A Design-Based Study of a Web-Based Tool. International Journal of Special Education, 41(24s), 1374 - 1387

Feedback on oral presentations in English as a foreign language (EFL) courses is usually delivered orally and immediately, which makes it difficult for students to retain and for assessors to record while still observing a performance. This study aimed to design, refine, and evaluate a web-based tool for capturing scores and written comments during live academic presentations in a blended EFL academic speaking course at an Indonesian state university. Following design-based research, a needs analysis preceded two design cycles in which two pre-service teaching interns (TIs) assessed 22 undergraduates. Data were drawn from document analysis, non-participant observation, semi-structured interviews, a focus group discussion, a student questionnaire, response-time measurements, and automated page audits, and were integrated in a joint display. In the first cycle, repeated data entry and submission delays disrupted the interval between presenters, and the TIs bypassed the analytic rubric because criterion scores were judged too general to convey specific weaknesses. The second cycle introduced roster retrieval, a single form for assessors and students, and a holistic score with narrative comments; mean response times fell by 33% for retrieval and 16% for submission. The TIs reported a lighter workflow, and students valued revisiting written comments before later presentations. Five design principles are proposed. The findings show that tool design relocates, rather than resolves, the tension between rubric-based standardization and real-time judgment in L2 speaking assessment, which has implications for classroom-based speaking assessment, language assessment literacy in teacher education, and the design of technology for language education.

COPYRIGHT STATEMENT:

Copyright: © 2026 Authors.
Open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Keywords: EFL speaking assessment; Feedback literacy; Oral presentations; Technology-enhanced feedback; Teaching Interns

Introduction

Academic presentations are a central genre of oral communication in higher education and a key means through which students are socialized into academic discourse (Morita, 2000). In English as a foreign language (EFL) programs, presentations are widely used both to develop and to assess speaking ability, yet assessing them

is demanding. A presentation is ephemeral: the assessor must observe, judge, and record within the time the performance lasts, often for many students in one session. Rating spoken performance in real time places substantial demands on raters' attention and judgment (Bejar, 2012; Luoma, 2004), and the feedback that

follows is usually delivered orally, immediately after the talk.

Feedback is among the most powerful influences on learning, yet its effects vary with the information it carries and with whether learners act on it (Hattie & Timperley, 2007). Research in higher education therefore treats feedback as a process in which students make sense of information and use it to improve subsequent work (Carless & Boud, 2018; Nicol & Macfarlane-Dick, 2006). Feedback that students cannot retrieve, understand, or apply has limited value: students report that comments are often too general, phrased in unfamiliar terms, delivered too late, or unavailable when needed (Jonsson, 2013; Winstone et al., 2017). Staff and students also locate effective feedback differently, with staff emphasizing timing and mode and students emphasizing usable, detailed, and personalized comments (Dawson et al., 2019), and instructors identify workload as a major barrier to providing such feedback (Paris, 2022).

Second language research offers a complementary perspective. Studies of oral corrective feedback have concentrated on brief corrective moves within classroom interaction (Lyster & Ranta, 1997), and learners' engagement with feedback has been conceptualized as cognitive, behavioural, and affective (Ellis, 2010). Applying this framework to presentations, Wang et al. (2025) found that EFL students noticed and understood most of their teacher's oral feedback and responded by correcting language errors, revising slides, and improving delivery, although they also ignored some of it. Experienced EFL teachers, in turn, combine immediate and delayed feedback on presentations and regard individual feedback as more specific but more time-consuming than collective feedback (Wang et al., 2018). Feedback on extended monologic performances, in which comments address content, organization, language, and delivery together, therefore poses problems of retention and workload that interactional feedback research has rarely addressed.

A further complication concerns assessment criteria. Rubrics make expectations explicit and can support reliable scoring and self-assessment, particularly when they are analytic (Jonsson & Svingby, 2007; Panadero & Jonsson, 2013). In L2 speaking assessment, analytic marking has shown higher reliability and measurement precision than holistic marking (Li et al., 2023), although analytic

scales require more time and attention from raters (Luoma, 2004). At the same time, decomposing a holistic appraisal into criterion scores can produce anomalies (Sadler, 2009), and experienced assessors differ in which criteria they attend to and how they weight them (Bloxham et al., 2016). During a live presentation, these tensions are compounded by time, because an analytic scale requires several decisions for each performance within the brief interval before the next presenter begins.

Technology offers ways to make feedback persistent. In blended courses, a written record can extend a face-to-face assessment event into the online space, and students rate digital feedback modes favorably for detail, personalization, and usability (Ryan et al., 2019). Research on technology-supported feedback on L2 speaking has examined mobile-assisted peer feedback (Ding & Zhu, 2025; Wu & Miller, 2020), teacher electronic feedback (Kataoka et al., 2024), virtual reality tasks with system-generated feedback (van Ginkel et al., 2019), multimodal systems that analyze recorded presentations (Ochoa & Dominguez, 2020), automated scoring of spoken language (Zechner & Hsieh, 2024), and AI-generated feedback on oral presentations (Ayaz Yılmaz & Er, 2026). These approaches rely mainly on recordings, peers, or automated analysis, whereas platforms that store and organize teacher feedback have been studied mainly in L2 writing (Zhang & Hyland, 2018).

Little is therefore known about how lightweight tools can support human assessors in capturing feedback during live L2 presentations, or about the design trade-offs that emerge under the time pressure of real-time judgment. The question matters particularly for novice assessors. Language assessment literacy, the knowledge and skills that teachers need to design, use, and interpret assessments, develops gradually and through practice (Fulcher, 2012), and pre-service EFL teachers' assessment practices are still forming (Xie, 2021). Recent work on teacher feedback literacy also treats the design of workable feedback processes, including the management of workload, as part of teachers' expertise (Boud & Dawson, 2023; Carless & Winstone, 2023).

This study addressed this gap through design-based research (DBR) in a blended EFL academic speaking course in Indonesia, where feedback on in-class presentations had previously been given only orally. In collaboration with two

pre-service teaching interns (TIs), a web-based tool for capturing scores and written comments was designed, implemented, and refined across two cycles. The study aimed to identify the requirements for such a tool, to trace how its design changed in use, and to derive design principles for capturing feedback during live L2 speaking assessment. Three research questions (RQs) guided the study:

RQ1. What requirements for capturing feedback on live EFL academic presentations emerged from the analysis of the course context?

RQ2. How did the feedback-capture tool change across two design cycles, and what tensions shaped these changes?

RQ3. How did the TIs and students experience the refined tool?

2. Methodology

2.1 Research Design

DBR was adopted because the purpose was to develop a workable tool in an authentic course and to derive transferable design knowledge, not

to estimate an intervention effect. DBR combines iterative cycles of analysis, design, enactment, and evaluation conducted in collaboration with practitioners (Amiel & Reeves, 2008; Design-Based Research Collective, 2003), and it is increasingly used to study technology-supported and blended learning (Hoadley & Campos, 2022). Following the generic model of McKenney and Reeves (2019), the study comprised an analysis and exploration phase (the needs analysis) and two cycles of design, enactment in live presentation sessions, and evaluation (Table 1). Each cycle served as a tryout in the authentic setting, and its evaluation focused on practicality, that is, whether assessors could use the tool during live sessions, and on the tool's perceived value to students. Effects on learning outcomes were outside the scope of the study. To make the design logic explicit, each requirement was linked to a design feature and to the mechanism through which that feature was intended to work, following conjecture mapping (Sandoval, 2014). The findings of both cycles were then synthesized into design principles.

Table 1. Phases, Research Questions, Activities, and Data Sources

Phase	RQ	Activities	Data sources
Needs analysis	RQ1	Analysis of the course and of existing feedback practice (throughout the even semester in 2025)	Syllabus and assessment guidelines; non-participant observation; semi-structured interviews and a focus group discussion with the TIs and lecturers (n =2); student interviews (n = 22)
Cycle 1	RQ2	Design and tryout of the initial tool in 4 in-class teaching sessions.	Non-participant observation; debriefing interviews with the TIs; response times; page audits; interface inspection
Cycle 2	RQ2, RQ3	Design and tryout of the initial tool in 5 live teaching sessions.	Non-participant observation; debriefing interviews with the TIs; student interviews; questionnaire; response times; page audits; interface inspection

Note. TI = teaching intern; RQ = research question.

2.2 Context and Participants

The study took place in an Academic Speaking course in the undergraduate English Education program of a state university in Indonesia during the even semester of the 2024/2025 academic year. The syllabus aims to develop students' ability to plan, organize, and deliver academic presentations on issues in English language teaching, with attention to organization, visualization, grammar, and pronunciation.

Presentations were the main assessed task. The course was blended: presentations took place in face-to-face sessions, whereas learning materials and, after the intervention, written feedback were available on the course website.

All 22 undergraduates from the 2023 cohort enrolled in the class participated (15 females, 7 males; aged 18-21; with an estimated CEFR B2 English proficiency). The two TIs were Master's students in English Education completing a

teaching internship; they taught and assessed the class under the direct supervision of a senior course coordinator. Course lecturers ($n=2$) contributed to the needs analysis. The assessment tool was developed by the research team, who acted solely as non-participant observers and had no direct instructional or evaluative authority over the TIs or the undergraduates in the course.

2.3 The Feedback-Capture Tool

The tool was custom-built with HTML, CSS, JavaScript, and PHP, with a Google Sheets spreadsheet as the data store, and it was hosted on the course website alongside the learning materials (developed and maintained by the first author). The spreadsheet was chosen because it stores records in the cloud, is accessible from any device, and allowed the TIs and lecturers to view records directly. The initial and refined designs are reported in Section 3 because they are outcomes of the design process.

2.4 Data Collection

Data collection was matched to the purpose of each phase (Table 1). In the needs analysis, the course syllabus and institutional assessment guidelines were analyzed to establish course aims and assessment requirements. A researcher conducted non-participant observation of presentation sessions, recording field notes on how assessors observed, scored, recorded, and delivered feedback. Semi-structured interviews and a focus group discussion with the TIs and lecturers explored existing scoring and feedback practices, and student interviews explored students' experiences of feedback and their understanding of the assessment criteria.

In both cycles, non-participant observation of live sessions focused on the assessors' workflow in the interval between presenters, including data entry, use of the rubric, waiting time after submission, and the delivery of oral feedback. After each cycle, semi-structured debriefing interviews with the TIs addressed the tool's usability, the workflow, and problems encountered. After Cycle 2, student interviews addressed how students retrieved and used the written comments. A total of 9 sessions were observed (4 in Cycle 1 and 5 in Cycle 2). Interviews with the TIs were conducted individually, while student interviews were conducted in small focus groups of 3-4 students. Each interview lasted between 20-30 minutes and was conducted in Indonesian to allow

participants to express their experiences nuancedly.

A researcher-developed questionnaire was administered after Cycle 2. It contained 15 items rated on a five-point agreement scale ranging from 1 (strongly disagree) to 5 (strongly agree), seven items on interface and use and eight on perceived contribution to learning. In the item wording, "reflective feedback" referred to the TIs' written comments in the tool. The questionnaire was administered digitally via a secure online form at the end of Week 10 and the responses were completely anonymous. The content of the items was reviewed for clarity by two independent experts in applied linguistics prior to administration. The internal consistency, estimated using McDonald's omega, was satisfactory, $\omega = .82$ for interface and use, and $\omega = .85$ for perceived contribution to learning. With 22 respondents, no factor-analytic validation was attempted, and the results are reported descriptively.

Three technical measures complemented these sources. First, response times for data retrieval (GET) and for the submission of scores and comments (POST) were recorded with the Network panel of Chrome Developer Tools. The network response times were averaged over 50 automated requests per type (GET and POST) for each cycle to ensure stable measurements. Second, automated page audits were run with Google PageSpeed Insights, which applies Lighthouse audits on an emulated mid-range mobile device with throttled network conditions, for the assessor and student pages in Cycle 1 (February 2025) and for the unified form in Cycle 2 (June 2025). Third, the inputs required for each assessment were counted from screenshots of the assessor forms.

2.5 Data Analysis

Interview transcripts and field notes were analyzed through thematic analysis (Braun & Clarke, 2006), combining deductive codes derived from the design requirements with inductive codes that captured unanticipated issues. Questionnaire items were summarized with means and standard deviations, and technical measures with descriptive statistics. The findings were then integrated in a joint display that set qualitative and quantitative evidence for each design principle side by side, which allowed convergence and divergence across sources to be examined. Trustworthiness was supported by triangulation across data sources and participant

groups and by reporting each design change together with the evidence that prompted it. Member checking was conducted by sharing the interview transcripts and the preliminary summary of findings with the TIs to confirm that their experiences were accurately represented.

2.6 Ethical Considerations

All participants gave informed consent, and participation did not affect course grades. In this study, the participants are identified by role and number, and the screenshots in this article contain no personal information.

3. Results

3.1 Requirements for Capturing Feedback (RQ1)

The document analysis confirmed that presentations were the main assessed task.

Table 2. Requirements, Initial Design Features, and Intended Mechanisms

Requirement	Evidence from the needs analysis	Initial feature	Intended mechanism
R1. Fast capture during live performance	TIs struggled to give timely, individualized feedback in tightly scheduled sessions	One web form for scores and comments, usable on a phone or laptop	Assessors record feedback without losing track of the presentation
R2. Persistent, retrievable feedback	Feedback was oral and impromptu; students could not record their weaknesses	Comments stored centrally; student retrieval page	Students revisit comments before the next presentation
R3. Visible criteria	Students were uncertain about the criteria; no shared scoring tool existed	Rubric with four criteria and descriptors in the form	Scores and comments are linked to shared criteria

Note. TI = teaching intern.

3.2 Cycle 1: Initial Design and Its Enactment (RQ2)

The initial assessor page contained fields for assessor ID, class, student ID, and student name; a selector for the rating scale (0–4, 0–5, or 0–10); the four rubric criteria (content and organization,

Observation and interviews revealed three issues. First, no shared scoring tool existed, and lecturers and TIs reported that scores could differ across assessors. Second, feedback was delivered orally and impromptu after each presentation, so students had few opportunities to receive written formative input. Third, no digital system supported the delivery of feedback, which led to fragmented experiences and inconsistent follow-up. Many students were uncertain about the assessment criteria and found it difficult to identify their strengths and weaknesses, partly because oral comments could not easily be recorded. The TIs described consistent, timely, and individualized feedback during tightly scheduled sessions as their main challenge. Table 2 presents the resulting requirements, the initial design features, and their intended mechanisms.

language use, delivery, and visual aids), each with a descriptor and score buttons; and a comment field accepting up to 1,000 words (Figure 1). Records were stored with the selected criterion descriptors and an overall score, and students retrieved their score and comments on a separate page immediately after presenting.

Figure 1. Assessor Page in Cycle 1, Shown With the Three Selectable Rating Scales (0–4, 0–5, and 0–10)

The figure displays three side-by-side screenshots of the 'Presentation Rubric App' interface, each showing a different 'Preferred Rating Scale' selected at the top: 0-4, 0-5, and 0-10. Each interface includes the following sections:

- Assessor Information:** Fields for Assessor ID, Class, Student ID, and Student Name.
- Content and Organization:** A description of the criterion and a rating scale (0-4, 0-5, or 0-10).
- Language Use:** A description of the criterion and a rating scale (0-4, 0-5, or 0-10).
- Delivery:** A description of the criterion and a rating scale (0-4, 0-5, or 0-10).
- Visual Aids:** A description of the criterion and a rating scale (0-4, 0-5, or 0-10).
- Feedback & Recommendation:** A text area for comments.
- Submit:** A blue button at the bottom.

Most students responded positively to receiving written comments, but the TIs identified three problems. The first was repeated administrative entry. Before scoring each student, the TIs typed the assessor ID, class, student ID, and student name, which consumed the interval between presenters (Table 3).

I think the biggest issue was the administrative overload. Having to repeatedly enter the same student and assessor details felt unnecessary and time-consuming. (TI 2)

The second problem was waiting time. Both TIs described the delay after submission, which they perceived as about five seconds, as disruptive. Measured mean response times were 3.68 s for retrieval and 2.98 s for submission (Table 3), well above the one-second limit within which users' flow of thought remains uninterrupted (Nielsen, 1993).

The third problem concerned the rubric. The TIs tended to disregard the criterion scores, explaining that they and the students had already internalized the criteria, and judged criterion scores too general to communicate specific weaknesses, which they addressed in comments instead.

I think, providing sub-scores for each aspect in the rubric is too general for the students. What students really need is specific clarification, which I usually provide in the written feedback. (TI 1)

A Cycle 1 comment record illustrates this account. The comment addressed organization, language use, delivery, and visual aids in turn, for example: "Your delivery was confident and clear, though varying your tone and incorporating more eye contact could make it more engaging." The criteria thus remained present in the TIs' comments even when the criterion scores were treated as redundant.

Table 3. Interface and Technical Indicators of the Tool in Each Cycle

Indicator	Cycle 1	Cycle 2
Inputs displayed on the assessor form	11	8
Student name typed by the assessor	Yes	No (retrieved)
Criterion-level decisions per student	4	0

Indicator	Cycle 1	Cycle 2
Mean response time, GET (s)	3.68 (.45)	2.45 (.32)
Mean response time, POST (s)	2.98 (.38)	2.50 (.35)
Lighthouse performance, mobile	91 (assessor page); 97 (student page)	96 (unified form)
Lighthouse accessibility, mobile	68 (assessor page); 79 (student page)	79 (unified form)
Lighthouse best practices, mobile	100	100
Lighthouse performance / accessibility, desktop	Not audited	97 / 96

Note. Inputs were counted from the assessor forms (Cycle 1: four identification fields, a scale selector, four criterion ratings, a comment field, and a submit button; Cycle 2: four identification fields, a retrieval button, a score field, a comment field, and a submit button). Response times were recorded in the Network panel of Chrome Developer Tools ($n = 50$ requests per cell). Lighthouse scores range from 0 to 100 and come from single PageSpeed Insights runs in Cycle 1 and in Cycle 2.

3.3 Cycle 2: Refined Design and Its Enactment (RQ2)

The redesign aimed to reduce the actions required during scoring, because scoring took place while students were presenting (Table 4). A single form replaced the separate assessor and

student pages (Figure 2). Assessors entered a token matching the master spreadsheet, together with their assessor ID, the student ID, and the class, and the tool retrieved the assessor's and student's names. Entering an assessor ID enabled writing, whereas leaving the field blank opened a read-only view that listed all of a student's entries chronologically. Criterion scoring was replaced by one holistic score (0–100) and a free-text comment field. To ensure transparency despite the absence of sub-scores on the interface, the detailed assessment rubric was published on the course website, and the TIs briefly reviewed the criteria with the students at the start of each presentation session. As Table 3 shows, the form displayed eight inputs instead of 11, the student's name no longer had to be typed, and no criterion-level decisions were required.

Figure 2. Assessor View and Student View of the Unified Form in Cycle 2

Assessor Page

Enter Assessor ID (For assessor only - Case Sensitive)

Student ID

Class

Assessor Name

Fetch Student Details

Assessor: Teaching Intern 1

Student Name: Student 1

Enter notes

Enter score (0-100)

Submit Score

Student Page

Enter Assessor ID (For assessor only - Case Sensitive)

Student ID

Class

Student Name

Fetch Student Details

Assessor: Teaching Intern 1

Student Name: Student 1

NO.	SCORE	NOTES
1	85	1. You did a great job using visuals to support your points. It would be even better if you explained each visual more thoroughly to help everyone understand its relevance.
2	90	2. The way you introduced the topic was engaging. It caught my attention right away. You can add more examples to make the content even more interesting for the audience.
3	85	3. Your use of technology in the presentation was impressive. You showed how tools can be used in English teaching clearly. Try to use a bit more variety in your examples to make it more engaging.
4	85	4. Your presentation was very clear and well-organized. You used the slides effectively to explain concepts. Next time, try to make a table about each slide and compare the following:

Table 4. Cycle 1 Problems and the Corresponding Cycle 2 Design Changes

Problem in Cycle 1	Evidence	Change in Cycle 2
Repeated entry of assessor and student details	TI interviews; observation; interface inspection	Token verification; names retrieved from the master spreadsheet after a student ID is entered
Separate assessor and student pages	Design team decision	One form; an assessor ID enables writing, and a blank field gives read-only access
Criterion scores bypassed as too general	TI interviews; comment records	One holistic score (0–100) with free-text comments
Waiting time after each request	TI interviews; response times	Request-processing script simplified

Note. TI = teaching intern.

After the request-processing script was simplified, mean response times fell to 2.45 s for retrieval and 2.50 s for submission, reductions of 33% and 16%. Lighthouse performance on mobile rose from 91 for the Cycle 1 assessor page to 96 for the unified form, and best-practices scores remained at 100. Mobile accessibility rose from 68 to 79 but stayed below the 90–100 band that Lighthouse classifies as good, although it reached 96 in desktop mode. The delay was still noticeable, but the TIs judged it acceptable: "okay now" (TI 1) and "acceptable after the update" (TI 2). Both reported entering scores and comments without the earlier administrative burden. The tool did not replace oral feedback. Instead, TI 2 used the captured notes to structure a post-session debrief:

During presentation sessions, I typed all important points on the system and send them to the database. After the presentation session, I extracted the core points and delivered them orally, sometimes demonstrated them to all the students. (TI 2)

3.4 Experiences of the Refined Tool (RQ3)

The TIs described the refined tool as more efficient than their earlier practice:

Previously, it was quite time-consuming to evaluate each presentation manually because I had to deliver immediate oral feedback, while at the same time, I had to write down everything as a reference for evaluation for the students. (TI 1)

These accounts reflect perceptions, as the time spent on each assessment was not measured. Students' accounts formed three themes. The first was persistence: written comments could be revisited, unlike oral comments that were easily forgotten.

... really helpful because it shows me exactly which areas that I need to improve in the form that is accessible 24/7. (Student 1)

... feedback was commonly given verbally and I needed to write them down briefly. Sometimes I would forget it. Now, I can go back to the feedback whenever I need it, and that helps me focus on specific parts to improve before my next presentation. (Student 2)

The second theme was access across devices. Students used the tool on phones and laptops and described it as "visually simple but clear" (Student 3).

I found the website layout very clean and well-organized. It was easy to locate the information I needed, and moving between pages felt so intuitive and self-explanatory. (Student 4)

The third was perceived support for reflection, as specific comments prompted students to consider how they spoke rather than only the score. Comments on strengths were valued as much as comments on weaknesses.

What I appreciated most was how the feedback pointed out my specific strengths. It made me realize what I was already doing well, and that gave me more confidence for the next speaking tasks. (Student 5)

The system made me stop and think about how I spoke, not just the score. It pushed me to reflect on what I needed to develop, not just passively accept the grade. (Student 6)

Students also expressed reservations. Some preferred traditional oral feedback, and others recommended the tool "but maybe with some

improvements" (Student 8), whereas some were unreservedly positive:

I prefer this system It felt more personal, and I would definitely recommend it for future speaking classes. (Student 7)

The questionnaire results were consistent with these accounts (Table 5). All item means exceeded the scale midpoint, and the section means were 4.11 for interface and use and 4.13

for perceived contribution to learning. The lowest-rated items concerned error-free operation ($M = 3.59$) and preference for the tool over traditional feedback ($M = 3.91$). Follow-up discussions indicated that these glitches were not inherent flaws in the app's code, but rather intermittent connectivity issues with the university's Wi-Fi network, which occasionally caused delays in fetching student records.

Table 5. Student Questionnaire Responses ($n = 22$)

Item	<i>M</i>	<i>SD</i>
<i>Interface and use (section mean = 4.11)</i>		
The page layout facilitates easy access to information.	4.41	.50
Navigation between pages is smooth and intuitive.	4.27	.46
The website is responsive across various devices (laptop, smartphone, tablet).	4.09	.53
The page loading time is reasonably fast.	4.00	.62
The interface was visually clear and easy to understand.	4.18	.59
The website operates without errors or technical glitches.	3.59	.73
The system presents evaluation data in a clear and easy-to-understand format.	4.23	.61
<i>Perceived contribution to learning (section mean = 4.13)</i>		
Reflective feedback helps me understand my strengths in speaking.	4.27	.46
The scoring criteria are clear and understandable.	4.18	.59
I feel I improve my academic speaking after receiving reflective feedback.	4.05	.58
Reflective feedback is detailed and relevant to my performance.	4.14	.64
Reflective feedback encourages me to reflect on my speaking skills.	4.09	.53
The system supports my learning in the Academic Speaking class.	4.23	.61
I prefer this system over traditional feedback methods.	3.91	.75
I would recommend this system for future speaking courses.	4.14	.64

Note. Five-point agreement scale.

3.5 Integrated Findings and Design Principles

Table 6 integrates the evidence for each design principle. The sources converged on the value of persistent comments, on the disruptive effect of administrative steps, and on the acceptability of the reduced delay. Three divergences also emerged. The first concerns technical reliability: Lighthouse best-practices scores were perfect in both cycles, yet error-free operation was the lowest-rated questionnaire item, which shows that laboratory audits cannot capture run-time failures such as unstable connections. The second

concerns accessibility: students rated responsiveness across devices favorably ($M = 4.09$), whereas the mobile accessibility audits identified issues that may affect users with disabilities, who were not represented in the sample. The third concerns criteria: students were uncertain about the criteria before the intervention, yet the item on the clarity of scoring criteria received a mean of 4.18 after Cycle 2. The questionnaire cannot show whether this reflects the rubric used in Cycle 1, criteria communicated in class, or acquiescence, so the criteria remain the least resolved element of the design.

Table 6. Joint Display of Evidence for the Five Design Principles

Design principle	Qualitative evidence	Quantitative and technical evidence	Integration
DP1. Keep administrative work out of the moment of judgment.	Repeated data entry disrupted Cycle 1; no burden reported in Cycle 2	Inputs 11 to 8; name typing removed	Convergent
DP2. Prioritize specific comments during the performance, and make criteria visible through other channels.	Criterion scores bypassed; comments structured by criteria; students uncertain about criteria before the intervention	Criterion decisions 4 to 0; clarity of criteria $M = 4.18$	Partly divergent; criteria visibility unresolved
DP3. Make feedback persistent and cumulative across occasions.	Students valued revisiting comments before the next presentation	Understanding of strengths $M = 4.27$; support for learning $M = 4.23$	Convergent; actual use not observed
DP4. Treat response time as part of the assessment workflow.	Delay disruptive in Cycle 1 and acceptable in Cycle 2	GET -33% , POST -16% ; loading time $M = 4.00$; error-free operation $M = 3.59$	Convergent on speed; divergent on reliability
DP5. Use written capture to support, not replace, oral debriefing.	Notes structured a post-session oral debrief; some students preferred oral feedback	Preference over traditional feedback $M = 3.91$	Convergent

Note. DP = design principle. The principles are grounded in one course context and are offered as conjectures for further cycles.

4. Discussion

4.1 Relocating the Standardization Tension in L2 Speaking Assessment

The most consequential finding was the TIs' decision to bypass criterion scoring, which the redesign followed (DP2). The decision is consistent with critiques of preset criteria: analytic criteria cannot fully capture holistic judgments of complex performances (Sadler, 2009), experienced assessors use criteria variably (Bloxham et al., 2016), and holistic scales are faster to apply in real time (Luoma, 2004). It also conflicts with evidence that analytic marking improves the reliability of L2 speaking scores (Li et al., 2023) and that explicit criteria support transparency and self-assessment (Jonsson & Svingby, 2007; Panadero & Jonsson, 2013). The TIs' belief that the criteria had been internalized was not verified, and the intuitive judgments of novices may be less consistent than they assume (Bloxham et al., 2016). The Cycle 2 design therefore did not resolve the tension between standardization and practicality; it relocated it. Criterion scoring left the moment of judgment, while the need to make criteria visible to learners remained. The criterion-structured comments observed in Cycle 1 suggest a conjecture for the

next cycle: criteria can be communicated before and after performances through a shared rubric and exemplars, while comments captured during the performance are tagged by criterion. Because the refined design relies on comments, their quality becomes the main determinant of its value and should be examined directly.

4.2 Language Assessment Literacy and Teacher Feedback Literacy

The TIs' choices also illuminate how novice language teachers manage assessment in practice. Abandoning criterion scores reflected pragmatic judgments about workload and time, which are central to teacher feedback literacy (Boud & Dawson, 2023; Carless & Winstone, 2023), but it also bypassed principles of reliable rating that are part of language assessment literacy (Fulcher, 2012). Involving interns in the design of assessment tools, as in this study, made these trade-offs visible and therefore discussable with supervisors (Xie, 2021). Teacher education programs could use such design episodes to examine whether intuitive judgments are consistent across students and how written comments relate to the criteria of a speaking rubric. Because the tool stores scores and comments for every presentation, it could also

support calibration: interns and supervisors could rate the same recorded presentations, compare their scores and comments, and discuss discrepancies, which is a common basis for rater training in performance assessment (Eckes, 2015).

4.3 Protecting the Interval Between Presenters

Administrative steps and waiting time both occupied the short interval between presenters. Because live rating draws heavily on attention (Bejar, 2012), actions that contribute nothing to judgment compete with it. Moving identity data to the server side (DP1) and shortening response times (DP4) addressed the pragmatic constraints that shape feedback practice (Carless & Winstone, 2023; Paris, 2022). Response times nevertheless remained above one second (Nielsen, 1993), and the spreadsheet back end, chosen for accessibility, limited further gains. Practicality in live assessment therefore appears to depend less on absolute speed than on whether delays fit the rhythm of the session.

4.4 Persistence, Dialogue, and the Role of Automation

Students valued returning to written comments before their next presentation. This addresses barriers such as forgetting feedback and being unable to access it when needed (Jonsson, 2013; Winstone et al., 2017) and turns a face-to-face assessment event into an online resource (DP3). In terms of Ellis's (2010) framework, persistence creates conditions for cognitive and behavioral engagement across presentations, although the data show that students valued this possibility, not that they used it, because access and uptake were not analyzed. The tool also did not replace oral feedback: captured notes structured an oral debrief, and some students preferred oral feedback. This hybrid routine (DP5) is consistent with students' more positive ratings when they receive several feedback modes (Ryan et al., 2019), with the emphasis on dialogue in good feedback practice (Nicol & Macfarlane-Dick, 2006), and with EFL teachers' combined use of immediate and delayed feedback on presentations (Wang et al., 2018). Automated and AI-based systems increasingly provide feedback on presentations and spoken performance (Ayaz Yılmaz & Er, 2026; Ochoa & Dominguez, 2020; Zechner & Hsieh, 2024). The present findings suggest a complementary role for lightweight human-in-the-loop capture, which preserves

teachers' judgments of content and argument and makes them available to learners after the session.

4.5 Implications for EFL Teachers and Teacher Educators

Several practical implications follow for classroom-based assessment of L2 speaking. First, tools for live assessment should be judged by the number of actions they require between performances rather than by the number of features they offer, because every additional field competes with the assessor's attention to the speaker. Second, written comments should remain retrievable alongside earlier comments, so that learners can compare feedback across presentations and prepare the next one; a brief prompt asking learners to note one change for their next presentation could make this use visible. Third, criteria can be made visible outside the moment of judgment, for example by publishing the rubric and annotated exemplars on the course site, so that holistic scoring during the performance does not leave learners uncertain about expectations. Fourth, captured notes can structure a short oral debrief after each session, combining the persistence of written feedback with the dialogue that learners value.

4.6 Limitations and Further Research

The study involved one class, two TIs, and one semester, so transferability depends on the similarity of contexts. The researchers designed and evaluated the tool, and the TIs were co-designers, which may bias accounts toward positive evaluations. The evidence on experience consists of perceptions, and the questionnaire was researcher-developed and descriptive. Learning outcomes were not evaluated: scores assigned during the semester were not used as outcome measures because the scoring format changed between cycles, the raters were the tool's users, and no comparison group existed (Shadish et al., 2002). Students' actual use of feedback was not observed. Finally, students retrieved records with a token shared by the class and a student ID, and write access depended on one assessor identifier, so a student who knew a peer's ID could view that peer's feedback. Future versions require individual authentication and data handling consistent with ethical guidance on educational data (Slade & Prinsloo, 2013) and with Indonesia's Personal Data Protection Law (Law No. 27 of 2022). Subsequent cycles could rate recorded presentations with independent raters blind to the time point, using many-facet Rasch measurement to separate learner ability

from rater severity and task difficulty (Eckes, 2015), and could trace stored comments across presentations to examine learners' uptake.

5. Conclusion

This design-based study developed a web-based tool for capturing feedback on live EFL academic presentations in a blended course. Over two cycles, the tool moved from an analytic, form-heavy design to a minimal workflow in which assessors record a holistic score and narrative comments that students can revisit online. The study contributes five design principles for capturing feedback during live L2 speaking assessment. It also shows that tool design relocates, rather than removes, the tension between rubric-based standardization and real-time judgment, which makes the communication of criteria and the quality of comments central concerns for classroom-based speaking assessment. For language teachers and teacher educators, the central lesson is that tools for live assessment should protect the assessor's attention during the performance and move everything else, including administration and the communication of criteria, to the moments before and after it.

References

- Amiel, T., & Reeves, T. C. (2008). Design-based research and educational technology: Rethinking technology and the research agenda. *Journal of Educational Technology & Society*, 11(4), 29–40.
- Ayaz Yılmaz, H., & Er, G. (2026). AI-generated versus teacher feedback: A comparative study on EFL learners' public speaking anxiety and oral presentation performance. *Language Learning & Technology*, 30(1), 1–20. <https://doi.org/10.64152/10125/73698>
- Bejar, I. I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice*, 31(3), 2–9. <https://doi.org/10.1111/j.1745-3992.2012.00238.x>
- Bloxham, S., den Outer, B., Hudson, J., & Price, M. (2016). Let's stop the pretence of consistent marking: Exploring the multiple limitations of assessment criteria. *Assessment & Evaluation in Higher Education*, 41(3), 466–481. <https://doi.org/10.1080/02602938.2015.1024607>
- Boud, D., & Dawson, P. (2023). What feedback literate teachers do: An empirically-derived competency framework. *Assessment & Evaluation in Higher Education*, 48(2), 158–171. <https://doi.org/10.1080/02602938.2021.1910928>
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. <https://doi.org/10.1191/1478088706qp063oa>
- 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>
- Carless, D., & Winstone, N. (2023). Teacher feedback literacy and its interplay with student feedback literacy. *Teaching in Higher Education*, 28(1), 150–163. <https://doi.org/10.1080/13562517.2020.1782372>
- Dawson, P., Henderson, M., Mahoney, P., Phillips, M., Ryan, T., Boud, D., & Molloy, E. (2019). What makes for effective feedback: Staff and student perspectives. *Assessment & Evaluation in Higher Education*, 44(1), 25–36. <https://doi.org/10.1080/02602938.2018.1467877>

Acknowledgements and Funding Information

This research was fully supported by the Institute for Research and Community Service (LPPM) of Universitas Negeri X.

Conflict of Interest

The authors declare no conflicts of interest.

Author Contributions

Muhaimin Abdullah: Conceptualization, Methodology, Software, Project administration, Writing – original draft. **Ahmad Munir:** Formal analysis, Data curation. **Syafi'ul Anam:** Investigation, Validation, Writing – review & editing. **Lies Amin Lestari:** Investigation, Resources, Visualization. **Widyastuti:** Supervision, Conceptualization, Writing – review & editing. **Chester B. Esnara:** Supervision, Conceptualization, Writing – review & editing. All authors have read and agreed to the published version of the manuscript.

Declaration of Generative AI Use

During the preparation of this work, the authors used Grammarly for grammar checking purposes. AI was not used for data analysis, generation of ideas, or drawing conclusions.

- Design-Based Research Collective. (2003). Design-based research: An emerging paradigm for educational inquiry. *Educational Researcher*, 32(1), 5–8. <https://doi.org/10.3102/0013189X032001005>
- 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, Article 10. <https://doi.org/10.1186/s40862-024-00314-9>
- Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments* (2nd ed.). Peter Lang.
- Ellis, R. (2010). Epilogue: A framework for investigating oral and written corrective feedback. *Studies in Second Language Acquisition*, 32(2), 335–349. <https://doi.org/10.1017/S0272263109990544>
- Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly*, 9(2), 113–132. <https://doi.org/10.1080/15434303.2011.642041>
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112. <https://doi.org/10.3102/003465430298487>
- Hoadley, C., & Campos, F. C. (2022). Design-based research: What it is and why it matters to studying online learning. *Educational Psychologist*, 57(3), 207–220. <https://doi.org/10.1080/00461520.2022.2079128>
- Jonsson, A. (2013). Facilitating productive use of feedback in higher education. *Active Learning in Higher Education*, 14(1), 63–76. <https://doi.org/10.1177/1469787412467125>
- 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>
- Kataoka, Y., Thamrin, A. H., & Van Meter, R. (2024). The effect of teacher electronic feedback on Japanese SFL basic learners' speaking improvement. *Journal of Information Processing*, 32, 331–345. <https://doi.org/10.2197/ipsjip.32.331>
- Li, S., Wen, T., Li, X., Feng, Y., & Lin, C. (2023). Comparing holistic and analytic marking methods in assessing speech act production in L2 Chinese. *Language Testing*, 40(2), 249–275. <https://doi.org/10.1177/02655322221113917>
- Luoma, S. (2004). *Assessing speaking*. Cambridge University Press.
- Lyster, R., & Ranta, L. (1997). Corrective feedback and learner uptake: Negotiation of form in communicative classrooms. *Studies in Second Language Acquisition*, 19(1), 37–66. <https://doi.org/10.1017/S0272263197001034>
- McKenney, S., & Reeves, T. C. (2019). *Conducting educational design research* (2nd ed.). Routledge.
- Morita, N. (2000). Discourse socialization through oral classroom activities in a TESL graduate program. *TESOL Quarterly*, 34(2), 279–310. <https://doi.org/10.2307/3587953>
- Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, 31(2), 199–218. <https://doi.org/10.1080/03075070600572090>
- Nielsen, J. (1993). *Usability engineering*. Academic Press.
- Ochoa, X., & Dominguez, F. (2020). Controlled evaluation of a multimodal system to improve oral presentation skills in a real learning setting. *British Journal of Educational Technology*, 51(5), 1615–1630. <https://doi.org/10.1111/bjet.12987>
- 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>
- Paris, B. (2022). Instructors' perspectives of challenges and barriers to providing effective feedback. *Teaching and Learning Inquiry*, 10. <https://doi.org/10.20343/teachlearning.10.3>
- Ryan, T., Henderson, M., & Phillips, M. (2019). Feedback modes matter: Comparing student perceptions of digital and non-digital feedback modes in higher education. *British Journal of Educational Technology*, 50(3), 1507–1523. <https://doi.org/10.1111/bjet.12749>
- Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & Evaluation in Higher Education*, 34(2), 159–179. <https://doi.org/10.1080/02602930801956059>
- Sandoval, W. (2014). Conjecture mapping: An approach to systematic educational design research. *Journal of the Learning Sciences*, 23(1), 18–36. <https://doi.org/10.1080/10508406.2013.778204>
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

- Slade, S., & Prinsloo, P. (2013). Learning analytics: Ethical issues and dilemmas. *American Behavioral Scientist*, 57(10), 1510–1529. <https://doi.org/10.1177/0002764213479366>
- van Ginkel, S., Gulikers, J., Biemans, H., Noroozi, O., Roozen, M., Bos, T., van Tilborg, R., van Halteren, M., & Mulder, M. (2019). Fostering oral presentation competence through a virtual reality-based task for delivering feedback. *Computers & Education*, 134, 78–97. <https://doi.org/10.1016/j.compedu.2019.02.006>
- Wang, B., Yu, S., & Teo, T. (2018). Experienced EFL teachers' beliefs about feedback on student oral presentations. *Asian-Pacific Journal of Second and Foreign Language Education*, 3, Article 12. <https://doi.org/10.1186/s40862-018-0053-3>
- Wang, B., Yu, S., Zheng, Y., & Teo, T. (2025). Student engagement with teacher oral feedback in EFL university classrooms. *Language Teaching Research*, 29(5), 1994–2018. <https://doi.org/10.1177/13621688221105772>
- Winstone, N. E., Nash, R. A., Rowntree, J., & Parker, M. (2017). 'It'd be useful, but I wouldn't use it': Barriers to university students' feedback seeking and recipience. *Studies in Higher Education*, 42(11), 2026–2041. <https://doi.org/10.1080/03075079.2015.1130032>
- Wu, J. G., & Miller, L. (2020). Improving English learners' speaking through mobile-assisted peer feedback. *RELC Journal*, 51(1), 168–178. <https://doi.org/10.1177/0033688219895335>
- Xie, Q. (2021). Developing preservice English teachers' assessment literacy through design-based research. *Journal of Education for Teaching*, 47(3), 460–465. <https://doi.org/10.1080/02607476.2021.1906155>
- Zechner, K., & Hsieh, C. N. (2024). Automated scoring and feedback for spoken language. In M. D. Shermis & J. Wilson (Eds.), *The Routledge international handbook of automated essay evaluation* (pp. 141–160). Routledge. <https://doi.org/10.4324/9781003397618-10>
- Zhang, Z., & Hyland, K. (2018). Student engagement with teacher and automated feedback on L2 writing. *Assessing Writing*, 36, 90–102. <https://doi.org/10.1016/j.asw.2018.02.004>