frontiers in Education

TYPE Original Research
PUBLISHED 28 May 2026
DOI 10.3389/feduc.2026.1828864

Check for updates

OPEN ACCESS

EDITED BY
Shamim Akhter,
INTI International University, Malaysia

REVIEWED BY
Ghadah Albarqi,
Taif University, Saudi Arabia
Karin Sari Saputra,
University of Indonesia, Indonesia

*CORRESPONDENCE
Suliman Abdelaty
✉ suliman.abdelaty@uob.edu.ly

RECEIVED 12 March 2026
REVISED 24 April 2026
ACCEPTED 07 May 2026
PUBLISHED 28 May 2026

CITATION
Abdelaty S (2026) From rubric to monitoring system: a CEFR-aligned approach to reliability and learning in online ESL speaking.
Front. Educ. 11:1828864.
doi: 10.3389/feduc.2026.1828864

COPYRIGHT
© 2026 Abdelaty. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

From rubric to monitoring system: a CEFR-aligned approach to reliability and learning in online ESL speaking

Suliman Abdelaty*

Faculty of Arts, El-Marj Campus, University of Benghazi, Benghazi, Libya

Online English as a Second Language (ESL) platforms have expanded rapidly over the past decade, yet the evaluation of learners' speaking ability on these platforms remains inconsistent, opaque, and under-researched. This study examined the CEFR-Aligned Monitoring System (CAMS)—a structured intervention comprising an analytic rubric, rater calibration, a written feedback protocol, and an in-session rubric routine—deployed in three commercial online ESL platforms. Using a mixed-methods quasi-experimental design, 214 adult ESL learners (aged 19–47) in Southeast Asia and the Middle East were tracked over 16 weeks between September 2024 and January 2025. Learners were grouped by their instructor's condition, forming a CAMS arm ($n = 112$) and a comparison arm ($n = 102$) assessed through conventional holistic ratings. Two research questions asked (a) whether CAMS improved inter-rater reliability among online ESL instructors relative to conventional practice, and (b) how learners and instructors experienced the shift from impressionistic to criterion-referenced evaluation. Pre- and post-intervention speaking scores were complemented by 38 learner interviews and four instructor focus groups ($n = 12$ instructors), analysed as separate datasets before convergence and divergence were identified. The study advances a conceptual shift from assessment as measurement to assessment as pedagogical monitoring in online learning environments. Instructors in the CAMS arm showed substantially higher inter-rater reliability than those in the comparison arm (ICC = .87 vs.61, $p < .001$), and learners in that arm showed larger gains than their comparison-arm counterparts across fluency, coherence, interaction management, and task fulfilment (composite $d = 0.56$; fluency and coherence $d = 0.74$). Learners valued the transparency of criterion-referenced feedback and several spontaneously adopted the rubric for self-monitoring in ways consistent with established models of self-regulated learning; instructors appreciated the structure but raised concerns about workload, task fit, and the limits of the phonological descriptors. Taken together, and read against the limits of the quasi-experimental design, the findings are best interpreted as system-level rather than rubric-level outcomes, and they extend the discourse on digital speaking assessment from measurement alone toward pedagogical monitoring embedded within a broader instructional ecology.

KEYWORDS
CEFR, inter-rater reliability, language learning technology, online ESL, oral proficiency assessment, pedagogical monitoring, rubric-based assessment, speaking evaluation

1 Introduction

The market for online English language learning has grown at a pace that few working in language education would have predicted a decade ago. Industry estimates place the sector's expansion from roughly USD 7.2 billion in 2019 to over USD 14.8 billion by 2024, initially accelerated by pandemic-era school closures but sustained since by improvements in videoconferencing infrastructure and shifting learner preferences (Ambient Insight, 2023; HolonIQ, 2024). Growth of this magnitude has outrun the development of assessment practices suited to digital environments, a pattern documented in broader accounts of technology-mediated language testing (Chapelle and Voss, 2016; Isbell and Kremmel, 2020). Grammar and reading items have migrated online without much resistance; speaking evaluation has not. Oral production requires real-time interaction, and the quality of judgments about a learner's spoken ability depends heavily on who is judging, on what criteria are being applied, and on how consistently those criteria are enforced across raters, sessions, and platforms (Luoma, 2004).

Within established language programmes, the Common European Framework of Reference for Languages (CEFR) has come to serve as something approaching a shared vocabulary for describing proficiency. First published by the Council of Europe in 2001, it organises language ability into six broad bands—A1 through C2—with illustrative descriptors for reception, production, interaction, and mediation. Its status owes to several features: an extended empirical calibration involving thousands of learners and hundreds of raters across Europe (North, 2000); a common metalanguage that could bridge national curricula without prescribing particular teaching methods (North, 2014); and deliberate flexibility in adaptation to diverse educational settings, which eased its uptake by major testing organisations such as Cambridge English, IELTS, and the Goethe-Institut (Council of Europe, 2001, p. 21). The 2020 Companion Volume subsequently expanded the descriptor inventory, most notably by adding scales for online interaction and phonological control that are, at least in principle, well suited to tasks learners encounter on digital platforms (Council of Europe, 2020).

A substantial body of research has examined CEFR-aligned evaluation procedures in various assessment contexts. In high-stakes testing, studies have analysed how CEFR descriptors function as rating-scale anchors. Wisniewski (2017), for example, found that trained raters applying CEFR-aligned analytic categories in Germany showed considerable variability across dimensions of spoken performance, with grammar ratings sometimes diverging from fluency ratings by an entire band. In writing assessment, Harsch and Martin (2012) showed that CEFR-referenced rubrics could improve scoring consistency when raters received adequate calibration. In automated speaking assessment, research on SpeechRater (Zechner et al., 2009) and comparable systems has demonstrated that machine-scorable features align only partially with the full range of CEFR descriptors, particularly for interactive and pragmatic competences (Galaczi and Taylor, 2018; Xi, 2010). Few studies, however, have examined whether structured CEFR-aligned evaluation translates into improved measurement quality in live online instruction. The small body of relevant work (Kang et al., 2019; Nakatsuhara et al., 2017, 2021; Sato and Loewen, 2019) has examined adjacent questions—rater training, videoconferenced speaking tests, and online corrective feedback—but none has investigated the implementation of a structured CEFR-aligned monitoring system in the day-to-day practice of commercial online tutoring platforms.

This gap matters for several intersecting reasons, all of them specific to speaking assessment in online environments. First, many online ESL platforms market their courses using CEFR levels without making public how those classifications are derived, a concern echoed in critical analyses of CEFR uptake across the language-teaching industry (Alderson, 2007; Figueras, 2012; Papageorgiou et al., 2019). A learner labelled "B1" on one platform may have completed a rigorous oral assessment; another "B1" learner on a different platform may have been placed by a multiple-choice test that never sampled speaking at all. Unlike reading or writing, where standardised instruments are widely adopted in online contexts, speaking assessment on these platforms lacks consistent tools because it requires real-time human judgment under time pressure (Nakatsuhara et al., 2017). Second, the instructors on these platforms are typically freelance tutors whose training in language assessment varies widely, and whose employment conditions (session-by-session pay, minimal professional development, high turnover) limit opportunities for systematic calibration (Sato and Loewen, 2019). Kang et al. (2019) showed that even modest differences in rater training can produce substantial divergence in oral proficiency scores. The problem is especially acute for speaking, because the evaluative decisions cannot be revisited in the way that written work can be re-read and re-scored. Third, learners themselves increasingly expect transparency about their spoken proficiency; survey evidence indicates a consistent preference for specific, criterion-referenced feedback over global impressionistic judgments (Nielson, 2011; Panadero and Jonsson, 2013), and learners routinely report wanting to understand what "intermediate" actually means for their capacity to sustain conversation, manage turn-taking, and make themselves understood.

Previous studies employing CEFR-aligned rubrics have yielded mixed results that reinforce the need for further investigation. In high-stakes contexts, CEFR-referenced scales have been used primarily for writing (e.g., Harsch and Martin, 2012) and integrated skills (Turner and Purpura, 2016), with fewer studies examining their use in speaking under formative or low-stakes conditions. Where CEFR-aligned speaking rubrics have been implemented, outcomes have depended heavily on rater training: studies with extensive calibration (e.g., Cambridge English, the Goethe-Institut) have reported acceptable reliability, while studies with less trained raters have documented significant scoring variability (Fulcher, 2004; Wisniewski, 2017). Most of this work, crucially, has been situated in face-to-face examination contexts rather than in the technology-mediated, formative assessment environments characteristic of commercial online tutoring.

1.1 Reframing the contribution: from rubric to monitoring system

A further motivation for this study is conceptual. Much of the existing literature on CEFR-aligned assessment treats the rubric as

a scoring instrument whose role is to improve measurement accuracy. In online learning environments, however, assessment functions less as a discrete measurement occasion and more as a continuous pedagogical monitoring process embedded within ongoing instruction (Black and Wiliam, 1998; Carless, 2015; Carless and Boud, 2018). In one-to-one online tutoring, the instructor simultaneously teaches and evaluates, and the evaluative information generated during a session feeds directly into the next instructional move. From this perspective, introducing a structured evaluation tool is not merely a matter of standardising scores; it is an intervention in the monitoring ecology of the platform, altering the content of feedback, the transparency of evaluative criteria, and the learner's capacity for self-regulation (Nicol and Macfarlane-Dick, 2006; Sadler, 1989).

This reframing has two consequences for the present study. First, the intervention is not a rubric alone. It is a system with four interdependent components: (a) a CEFR-aligned analytic speaking rubric, (b) a structured rater training and calibration workshop, (c) a written feedback protocol organised by rubric category, and (d) a session-level routine for discussing rubric categories with learners. Throughout the manuscript, this integrated arrangement is referred to as the CEFR-Aligned Monitoring System (CAMS). Treating the intervention as a system acknowledges that reliability and learner outcomes cannot be attributed to the rubric in isolation, and it brings the study into closer alignment with contemporary learning-oriented assessment perspectives. Second, outcome interpretation must accommodate both measurement evidence (reliability, proficiency gains) and evidence about the pedagogical ecology of online tutoring (feedback quality, learner self-regulation, instructor workload, interactional dynamics). The study therefore takes a deliberately ecological view of online assessment rather than focusing solely on psychometric outcomes.

1.2 Research questions and contribution

Two research questions guided the inquiry:

RQ1. Does the introduction of a structured CEFR-aligned monitoring system improve inter-rater reliability among online ESL instructors compared with conventional holistic assessment?

RQ2. How do learners and instructors experience and interpret the shift from impressionistic to criterion-referenced oral evaluation in an online setting, and how do their accounts converge and diverge?

The term *impressionistic evaluation* denotes the practice of assigning global, undifferentiated judgments of learner performance ("good," "needs work") without reference to explicit criteria or defined proficiency levels—an approach documented as prevalent among online ESL tutors (Kang et al., 2019). *Criterion-referenced evaluation*, by contrast, involves systematically rating performance against predefined descriptors anchored to established proficiency levels. The aim of this study is not to argue that the CEFR is the only viable framework—credible alternatives exist (e.g., ACTFL, Trinity)—but to test whether one careful operationalisation of it, embedded within a coherent monitoring system, can make online speaking assessment both more dependable and more transparent for the people involved.

The study contributes to the literature in four ways. First, it provides empirical evidence on the reliability of a CEFR-aligned speaking monitoring system within commercial online ESL tutoring, a context that has received limited scholarly attention. Second, it combines reliability analysis with separately conducted learner and instructor qualitative analyses, so that the distinct perspectives of each stakeholder group are preserved before points of convergence and divergence are identified. Third, it develops an analytical framework that links structured feedback, learner engagement, and self-regulated learning behaviour, thereby foregrounding the pedagogical mechanisms through which evaluation intersects with learning. Fourth, it offers practical implications for platform designers and programme coordinators who wish to implement transparent, criterion-referenced speaking evaluation in technology-mediated environments.

2 Literature review

2.1 The CEFR as a proficiency framework

The CEFR's influence on language education has grown progressively from its origins as a Council of Europe policy instrument. North (2000) documented the framework's development through the empirical calibration of descriptors across Swiss and European contexts, establishing the six-level scale (A1–C2) through a combination of qualitative descriptor development and Rasch modelling. This empirical foundation, together with the framework's deliberate flexibility, facilitated its widespread adoption. North (2014) traced how curricula in more than 40 countries have adopted or adapted the CEFR, while major international examinations—the Cambridge English suite, IELTS, and the Goethe-Zertifikat among them—have been explicitly mapped to its levels.

A crucial distinction, however, is that the CEFR's descriptors were never intended to function as test specifications—detailed blueprints that prescribe specific task types, scoring procedures, and psychometric targets for a given examination. A test specification defines the operational parameters of an assessment instrument (Bachman and Palmer, 2010), whereas the CEFR provides illustrative "can-do" statements describing what learners are typically able to do at each level. The Council of Europe has stated this distinction more than once (Council of Europe, 2001, p. 21), and it matters because it explains both the framework's adaptability and the difficulties that arise when CEFR descriptors are operationalised as scoring rubrics without careful local calibration.

Criticism of the CEFR has been persistent and substantive, and these critiques directly informed the present study's design. Fulcher (2004) questioned whether the six-level scale had been empirically validated or merely imposed as an administrative convenience, raising concerns about its construct validity for assessment purposes. Hulstijn (2007) observed that the framework treated linguistic proficiency as a unitary construct, whereas evidence from psycholinguistics points toward dissociable components—a concern particularly acute for

speaking, where fluency, accuracy, and interactional competence may develop at different rates. Wisniewski (2017) provided empirical support for this concern, showing that CEFR-level classifications assigned by trained raters in Germany varied considerably depending on the analytic category being applied. The 2020 Companion Volume addressed some of these issues by expanding the descriptor inventory, particularly for phonological control and online interaction (Council of Europe, 2020), but the extent to which the revised descriptors improve scoring consistency in practice remains an open empirical question.

2.2 Speaking assessment in digital contexts

Research on speaking assessment in technology-mediated channels has developed along two broadly distinct tracks. The first has examined automated speech evaluation systems—SpeechRater (Zechner et al., 2009) and Versant (Pearson, 2011) being the most prominent—which use speech recognition and natural language processing to produce proficiency scores without human raters. These systems offer impressive scalability, but they have drawn sustained criticism for narrowing the construct of speaking ability to features amenable to machine scoring: pronunciation accuracy, speech rate, and vocabulary diversity, at the expense of discourse management, pragmatic appropriacy, and interactional competence (Galaczi and Taylor, 2018; Xi, 2010). Recent developments have sought to address some of these limitations; Xu et al. (2021), for example, demonstrated acceptable reliability for certain automarker features of spoken performance, while acknowledging persistent difficulty in evaluating interactional competence through automated means.

The second track, more directly relevant to the present study, has investigated human-rated speaking assessment delivered over video or audio platforms. Nakatsuhara et al. (2017) reported that paired speaking tests administered via videoconferencing produced interaction patterns comparable to face-to-face conditions, though test-takers described heightened anxiety about technical disruptions. Building on that work, Nakatsuhara et al. (2021) applied Many-Facet Rasch Measurement to compare IELTS Speaking Test scores across face-to-face and videoconferencing modes and found no statistically significant differences in test-taker scores, although the videoconferencing mode elicited fewer clarification requests, suggesting subtle differences in interactional dynamics. These studies, importantly, were situated in high-stakes examination contexts involving trained examiners and standardised scoring rubrics. Neither addressed how evaluation might be structured for online tutoring contexts where brief sessions with freelance instructors—rather than formal examinations with certified raters—constitute the primary assessment occasion.

2.3 Assessment as a pedagogical monitoring process

A useful way to understand the role of evaluation in online ESL tutoring is to locate it within the broader literature on formative assessment and learning-oriented assessment. Black and Wiliam (1998), in their influential review of classroom assessment, argued that formative assessment exerts its strongest effects when evaluative information is used to regulate learning rather than to certify attainment. Sadler (1989) specified three conditions for formative assessment to improve learning: learners must understand the standard being aimed for, must be able to compare current performance against that standard, and must be able to take action to close the gap. Nicol and Macfarlane-Dick (2006) later reframed these conditions as seven principles of good feedback practice, all of which position feedback as an instrument of self-regulation rather than a mere report card. Carless (2015) coined the term *learning-oriented assessment* to describe evaluation designed primarily to generate information that feeds forward into learning, and Carless and Boud (2018) extended this into the concept of "feedback literacy," the capacity of learners to make sense of and act on evaluative information.

In online tutoring, these ideas are not merely pedagogically useful; they are practically inescapable. Sessions are short (25–50 min), frequent (weekly or more often), and shaped by ongoing relationships between instructor and learner. Evaluation is not separated from instruction—it is threaded through every exchange. For this reason, the appropriate unit of analysis in examining evaluation on these platforms is not a single rating event but a monitoring process that unfolds across sessions. This view is consistent with recent research on assessment in digital learning environments, which has argued for moving beyond measurement frameworks toward frameworks that capture the continuous, iterative nature of digital evaluation (Timmis et al., 2016; Whitelock, 2010).

Framing the intervention as a monitoring system carries an important methodological implication. If the system comprises a rubric, training, calibration, feedback protocols, and instructional routines—all interdependent—then observed effects cannot plausibly be attributed to the rubric alone. The present study therefore acknowledges this interdependence explicitly and treats CAMS as the unit of intervention rather than the rubric.

2.4 An analytical framework: feedback, engagement, and learning

If a structured monitoring system is to influence learning, it does so through identifiable behavioural pathways. The present study draws on three converging literatures to propose a working analytical framework linking structured feedback to behavioural engagement and, in turn, to learning outcomes.

The first component is *feedback internalisation*. Research on formative feedback indicates that for evaluative information to influence performance, it must be understood, accepted, and acted upon (Hattie and Timperley, 2007; Shute, 2008). Structured rubrics support internalisation because they make evaluative criteria explicit and observable. Panadero and Jonsson (2013) meta-analytic review found that rubric transparency was associated with improved self-regulation and performance, though effects depended on whether learners were given opportunities to engage with the rubric as a thinking tool rather than merely a scoring instrument.

The second component is *behavioural engagement* in the form of self-regulated learning (SRL). Zimmerman (2000) model of self-regulation posits a three-phase cycle of forethought, performance, and self-reflection, all of which are activated when learners have clear standards, informative feedback, and an environment that supports goal-setting. In language learning specifically, Andrade and Heritage (2017) argued that rubrics can mediate all three SRL phases by making the criteria for success visible before a performance, available during a performance, and available again for reflection afterwards.

The third component is *learning outcome formation*. When structured feedback drives targeted practice, and targeted practice accumulates across sessions, the resulting learning is both directed (focused on identified weaknesses) and observable (detectable on the dimensions for which feedback was given). In speaking specifically, the features most responsive to this kind of targeted attention appear to be fluency, coherence, and interaction management—areas where learners can consciously adjust practice in response to feedback. Pronunciation, by contrast, is often less responsive because its improvement depends on perceptual and articulatory processes that are not straightforwardly adjusted through rubric-level awareness alone (Derwing and Munro, 2015).

This framework—feedback internalisation → behavioural engagement → directed practice → learning outcomes—informs both the interpretation of quantitative findings (why the gains are uneven across subscales) and the reading of the qualitative data (what learners do with structured feedback once they receive it). It also foregrounds the conceptual claim that a monitoring system is not only a measurement device but a mediating element in the pedagogical chain linking instruction and learning.

2.5 Online ESL platforms and assessment practices

The commercial online ESL sector has attracted surprisingly little scholarly attention given its scale. Platforms such as italki, Preply, Cambly, and Engoo collectively serve many millions of learners, yet published research on their assessment practices remains sparse. This is a significant lacuna, because these platforms are now the primary site of speaking practice and evaluation for a large and growing population of language learners, and assessment quality on these platforms shapes learner progress and satisfaction in ways that are largely invisible to the research literature Sato and Loewen (2019) examined corrective feedback in online tutoring sessions but did not extend their analysis to summative or formative speaking assessment. Nielson (2011), surveying learners using early online language courses, found that most wanted more structured feedback on their oral production, though the technology then available constrained what could realistically be offered. More directly relevant to the present study, Kang et al. (2019) demonstrated that rater training significantly influences oral proficiency scores, with even modest differences in calibration producing divergent ratings—findings that underscore the importance of structured monitoring systems in contexts where instructor training is highly variable.

The disconnect between what online platforms market and how oral performance is actually assessed motivated the present study. Across the three platforms examined here, CEFR levels were attached to instructor profiles, placement outcomes, and course tiers, yet the underlying assessment procedures were not described in public documentation. The consequences of such unreliable speaking assessment at scale—misplaced learners, misallocated instructional time, erosion of learner trust—are substantial.

2.6 Gaps in the literature

Four gaps warranted investigation. First, CEFR validation studies have been conducted almost exclusively in high-stakes testing contexts (e.g., Cambridge examinations, DELF/DALF; Harsch and Martin, 2012; Wisniewski, 2017) rather than in the low-to-medium-stakes formative environments characteristic of online tutoring. Second, research on inter-rater reliability in online speaking assessment (Nakatsuhara et al., 2017, 2021) has not examined whether CEFR-aligned rubrics perform differently from holistic scoring under the specific constraints of online tutoring—short sessions, diverse instructor backgrounds, and pressure to provide real-time feedback. Third, while the learner experience of online instruction has been studied in general terms (Nielson, 2011; Sato and Loewen, 2019), learner perceptions of CEFR-aligned evaluation specifically—as distinct from general satisfaction with online courses—remain largely unexplored. Fourth, and most importantly for the framing adopted here, the theoretical relationship between structured evaluation and the broader pedagogical monitoring ecology of online platforms, encompassing feedback, engagement, self-regulation, and instructional adaptation, has not been empirically investigated.

3 Methodology

3.1 Research design

The study employed a mixed-methods quasi-experimental design (Creswell and Plano Clark, 2018). A fully experimental design involving random assignment of individual learners was impractical: learners on commercial platforms choose their own instructors, and instructors could not reasonably be asked to switch assessment methods mid-course. Conditions were therefore formed through instructor self-selection: instructors who agreed to implement the CEFR-Aligned Monitoring System (CAMS) constituted the treatment arm, while those who elected to continue with their habitual assessment practices constituted the comparison arm. Learners were classified into the treatment or comparison arm on the basis of their instructor's condition rather than through random allocation. The design carries the limitations of quasi-experimentation, which are addressed in Section 8, but it reflects the operational realities of conducting research within commercial educational settings. The design is also consistent with recent methodological guidance on applied educational research, which recognises that strict randomisation is often neither possible nor desirable in authentic learning environments (Shadish et al., 2002).

3.2 The CEFR-aligned monitoring system (CAMS)

CAMS is the intervention evaluated in this study. Rather than treating the rubric as an isolated instrument, CAMS integrates four components designed to function together as a pedagogical monitoring system.

3.2.1 Component 1: CEFR-aligned analytic speaking rubric

The rubric was developed over four months by the author in collaboration with two assessment specialists who were not otherwise involved in the study. The development drew on multiple sources: the CEFR's Spoken Production and Spoken Interaction descriptors from both the original 2001 framework and the 2020 Companion Volume (Council of Europe, 2001, 2020); established speaking assessment scales, including the Cambridge English Proficiency examination and the IELTS Speaking band descriptors (Taylor and Galaczi, 2011); and principles of analytic rubric construction (Fulcher, 2003; Luoma, 2004). The rubric operationalises these sources across five analytic categories: (a) Range and Accuracy of vocabulary and grammar, (b) Fluency and Coherence, (c) Phonological Control, (d) Interaction Management, and (e) Task Fulfilment. Each category was scored on a 0–5 scale anchored to CEFR band descriptors, with scores of 1–2 corresponding to A-level performance, 3–4 to B-level performance, and 5 to C1-level performance. This structure was chosen because the participant population ranged from A2 to B2, and the 0–5 scale provided sufficient granularity to discriminate within this range while remaining practical for instructors to apply during or immediately after tutoring sessions. An abridged summary of the rubric appears in Appendix A. Piloting with 40 recorded speaking samples yielded acceptable internal consistency ($\alpha = .83$) and an initial intraclass correlation coefficient (ICC) of .79 among three raters following a norming session.

3.2.2 Component 2: rater training and calibration workshop

Treatment-group instructors attended a six-hour online workshop before the intervention began. The workshop included detailed explanations of each descriptor, collaborative scoring of recorded samples, calibration exercises against benchmark scores, and a question-and-answer segment. Each instructor independently scored the same set of five recordings and then compared their ratings with facilitator benchmarks. Persistent discrepancies were discussed in plenary. This component served to standardise how the rubric was applied across instructors rather than merely informing them that it existed.

3.2.3 Component 3: structured feedback protocol

Treatment-group instructors followed a written feedback protocol after each session. Feedback was organised by rubric category (one paragraph per category), delivered through the platform's messaging system, and designed to describe current performance, indicate the next developmental target, and suggest one concrete action the learner could take before the next session. The protocol added approximately five to eight minutes of post-session work—an operational load examined further in Sections 4.3 and 5.5.

3.2.4 Component 4: in-session rubric routine

Treatment-group instructors discussed at least one rubric category in depth with the learner during every four-session cycle. The intent was to make evaluative criteria explicit to learners rather than keeping them as the private possession of the instructor, thereby supporting the feedback internalisation pathway described in Section 2.4.

These four components were conceptualised as a single, interdependent system. The study therefore claims no ability to disaggregate the effect of any one component from the others. The research questions are framed accordingly—in terms of CAMS as a whole rather than the rubric in isolation.

3.3 Participants

A total of 214 adult ESL learners participated, recruited from three online platforms selected because they offered one-on-one video tutoring sessions of 25–50 min, employed instructors from at least two nationalities, and agreed to cooperate with the research team. For confidentiality reasons mandated by the platforms' legal departments, the platforms are referred to as Platform A, Platform B, and Platform C.

Learners ranged from 19 to 47 years old ($M = 28.2$, $SD = 6.1$). The majority were employed adults studying English for career advancement, though a sizeable minority (approximately 22%) were university students. First languages included Arabic, Thai, Vietnamese, Bahasa Indonesia, and Mandarin, among others. Pre-test speaking levels, established through an independent baseline assessment described below, ranged from A2 to B2, with participants clustered primarily at B1. The two groups did not differ significantly on age, gender distribution, first-language background, or pre-test proficiency level (all $ps > .05$; see Table 1).

Twenty-six instructors participated: 14 in the CAMS condition and 12 in the comparison condition. Instructors were based in the Philippines ($n = 11$), South Africa ($n = 6$), the United Kingdom ($n = 5$), and the United States ($n = 4$). Teaching experience ranged from 1 to 14 years ($M = 5.8$, $SD = 3.2$). Nine held master's degrees in TESOL or applied linguistics; the remainder held bachelor's degrees—typically in education or English literature—supplemented by TEFL/TESOL certificates.

3.4 Outcome instruments

3.4.1 Standardised elicitation protocol

Both groups completed speaking tasks from the same item bank at Weeks 1 and 16. The protocol comprised three task types calibrated to the participants' proficiency range: (a) a one-minute monologue describing a personal experience (A2–B1 target), (b) a two-minute opinion discussion with a follow-up question from the assessor (B1–B2 target), and (c) a simulated transactional interaction (e.g., negotiating a schedule change

TABLE 1 Participant demographics by group.

Characteristic	CAMS group (n = 112)	Comparison (n = 102)	Total (N = 214)	p
Mean age (*SD*)	28.4 (6.3)	27.9 (5.8)	28.2 (6.1)	.56
Female (%)	63 (56.3%)	55 (53.9%)	118 (55.1%)	.73
L1				.95
Arabic	41 (36.6%)	37 (36.3%)	78 (36.4%)	
Thai	29 (25.9%)	28 (27.5%)	57 (26.6%)	
Vietnamese	24 (21.4%)	22 (21.6%)	46 (21.5%)	
Other	18 (16.1%)	15 (14.7%)	33 (15.4%)	
Pre-test CEFR level				.92
A2	34 (30.4%)	31 (30.4%)	65 (30.4%)	
B1	52 (46.4%)	49 (48.0%)	101 (47.2%)	
B2	26 (23.2%)	22 (21.6%)	48 (22.4%)	

p values derived from independent-samples t-tests (age) and chi-square tests (gender, L1 background, and pre-test CEFR level).

with a colleague). Task versions were counterbalanced across pre-test and post-test to minimise practice effects, and all sessions were video-recorded with participant consent. Pre- and post-test recordings were scored independently by two trained external raters blind to group assignment.

3.4.2 Semi-structured interview guide

A 14-item interview protocol explored learner perceptions of assessment clarity, fairness, and its influence on learning behaviour. The protocol was developed with reference to established frameworks for investigating assessment perceptions in language education, drawing on Panadero and Jonsson (2013) work on rubric-mediated self-assessment and Carless and Boud (2018) framework for feedback literacy. Questions included prompts such as "When you receive feedback on your speaking, what information is most useful to you?" and "Has the way your speaking is evaluated affected how you practise outside of class?" The guide was piloted with five learners not included in the main sample, and minor wording adjustments were made to address comprehension difficulties.

3.4.3 Instructor focus group protocol

A separate 12-item focus group protocol examined instructor experiences of applying the rubric, producing structured feedback, integrating evaluative routines into sessions, and managing workload. The protocol included prompts about perceived task fit, challenges in adapting rubric categories to conversational tutoring, and reflections on the feedback-production burden.

3.5 Procedure

The study unfolded in three phases. During Phase 1 (Weeks 1–2), all participants completed the pre-test speaking protocol, which was scored independently by two trained external raters blind to group assignment. Instructors in the treatment condition then attended the six-hour online training and calibration workshop that constitutes Component 2 of CAMS.

Phase 2 (Weeks 3–15) constituted the intervention period. Treatment-group instructors used CAMS during every tutoring session, including the rubric itself, the structured written feedback protocol, and the in-session rubric routine. Comparison-group instructors continued with their normal assessment approach, which typically involved brief end-of-session verbal comments (e.g., "Your grammar was good today; try to speak faster") and an occasional written note.

Phase 3 (Weeks 16–18) involved post-test administration, interviews, and focus groups. The same external raters who scored the pre-tests scored the post-tests, again blind to condition. Semi-structured interviews were conducted with 38 learners (20 from the CAMS group, 18 from the comparison group), selected through stratified purposive sampling to ensure representation across proficiency levels, first-language backgrounds, and platforms. Four focus groups, comprising three instructors each (two CAMS, two comparison), were held via Zoom.

3.6 Data analysis

3.6.1 Quantitative analysis

Quantitative analysis proceeded in two steps. Inter-rater reliability was assessed using two-way random-effects intraclass correlation coefficients [ICC(2,k)] for both the external raters' pre- and post-test scores and the instructors' session-level evaluations. ICCs were computed using SPSS (version 28), with 95% confidence intervals estimated via Fisher's r-to-z transformation. Improvement in speaking performance was analysed through repeated-measures ANCOVA, with group (CAMS vs. comparison) as the between-subjects factor, time (pre vs. post) as the within-subjects factor, and pre-test score, first-language background (dummy-coded with Arabic as the reference category), and platform as covariates. Because the within-subjects factor "time" had only two levels, the sphericity assumption did not apply and was not tested (Field, 2018). The

remaining assumptions—normality of residuals, homogeneity of variance (Levene's test), homogeneity of regression slopes, and linearity of the covariate–outcome relationship—were evaluated and no violations were detected. Effect sizes are reported as Cohen's d and partial eta-squared (η^2_p). A Holm–Bonferroni correction was applied across the five subscale comparisons to control family-wise error.

3.6.2 Qualitative analysis

Qualitative data were analysed using reflexive thematic analysis (Braun and Clarke, 2021). Consistent with Reviewer 2's methodological guidance and with current best practice in mixed-methods research (Moss, 2016), learner interview data and instructor focus group data were treated as *separate datasets* at the coding and theme-development stage. This decision reflects the substantively different positions of the two groups: learners as recipients of evaluation and consumers of feedback, instructors as producers of evaluation and mediators of feedback. Collapsing the two at the outset would have risked flattening distinctions that are pedagogically consequential.

Analysis proceeded in four stages. In Stage 1, all interviews and focus groups were transcribed verbatim by a professional transcription service. All data were in English, which was the shared language of instruction; no translation was required. In Stage 2, the author conducted line-by-line open coding of the learner dataset in NVivo 14, generating codes inductively. Codes were organised into candidate themes through an iterative process of grouping, comparing, and refining. In Stage 3, the same procedure was conducted independently for the instructor dataset, with a separate codebook developed from scratch rather than imported from the learner analysis. In Stage 4, the two sets of candidate themes were brought into dialogue through a cross-case comparison designed specifically to identify points of convergence and points of divergence. This comparison was treated as a substantive analytical task in its own right, not a mechanical overlay.

Consistent with Braun and Clarke (2021) framing of reflexive thematic analysis as an interpretive rather than a coding-reliability enterprise, intercoder agreement was not pursued; within this paradigm, coding is understood as an evolving analytic judgement shaped by the researcher's positionality rather than as a measurement to be calibrated against a second coder. Trustworthiness was instead cultivated through four reflexive TA strategies (Braun and Clarke, 2021). First, a reflexive journal was maintained throughout coding and theme development, documenting assumptions, emergent patterns, and shifts in interpretation; it was consulted regularly to interrogate how the researcher's position as both a university-based applied linguist and an active online ESL tutor was shaping analytic attention. Second, the transcripts were re-engaged across successive coding passes, with candidate themes refined against the full dataset rather than finalised prematurely. Third, peer debriefing was conducted with two colleagues in applied linguistics who were not otherwise involved in the study; they read a subset of coded transcripts and candidate thematic maps and challenged emerging interpretations, prompting closer attention to disconfirming cases and to the distinct analytic warrant of each dataset. Fourth, an audit trail was kept that documented key decisions about code construction, theme boundaries, and the rationale for preserving the separation between the learner and instructor analyses until the cross-case comparison stage. Final themes were named through an iterative process that moved between the qualitative findings and the quantitative results in the interpretive stage.

3.7 Ethical considerations

The study received ethical approval from the Institutional Review Board of the University of Benghazi, Faculty of Arts, El-Marj Campus. All participants provided written informed consent. Learner identities were anonymised in transcripts and reporting, and platform names were withheld per contractual agreement. Instructors were compensated at their normal session rate for any additional time spent on rubric-related activities. Data were stored on encrypted university servers and will be retained for five years in accordance with institutional policy.

4 Results

4.1 Inter-rater reliability

The primary reliability question—RQ1—asked whether CAMS would improve scoring consistency among instructors. Table 2 presents ICC values for the external raters' scores alongside the instructors' session-level ratings.

External raters scored both groups with comparable consistency at pre-test (ICC = .84 and .82, respectively), which was expected given that they applied the same scoring criteria to all participants. The critical comparison involved instructor session ratings: how consistently instructors within each condition rated their own learners' speaking during regular tutoring sessions. Here, the difference was substantial. CAMS-group instructors achieved an ICC of.87 [95% CI (.82,.91)], while comparison-group instructors produced an ICC of.61 [95% CI (.49,.71)]. The difference was statistically significant ($z = 4.83$, $p < .001$). Two interpretive points warrant emphasis. First, ICC(2,k) reflects the reliability of the mean rating averaged across all raters in each condition (14 instructors in the CAMS group, 12 in the comparison group), rather than the agreement between any two individual raters. The high ICC in the CAMS group therefore indicates that the mean of CAMS instructors' ratings would be

TABLE 2 Intraclass correlation coefficients for speaking scores.

Measure	CAMS group	Comparison	Difference (z)
External raters—Pre-test	.84 [.78,.89]	.82 [.75,.87]	0.41
External raters—Post-test	.86 [.80,.90]	.81 [.74,.87]	1.12
Instructor session ratings	.87 [.82,.91]	.61 [.49,.71]	4.83***

ICC(2,k) values with 95% confidence intervals in brackets. ***$p < .001$. z-tests based on Fisher's r-to-z transformation.

highly reproducible, whereas the mean of comparison-group ratings would show considerably more variability across replications. Second, the observed reliability difference cannot be attributed to any single component of CAMS; it reflects the combined effect of the system as implemented. This is the pattern anticipated by the monitoring-system framing introduced in Section 1.1: reliability here is a property not of the rubric but of the coordinated arrangement of rubric, calibration, feedback protocol, and in-session routine operating together.

4.2 Speaking performance gains

Both groups improved from pre-test to post-test, but the CAMS group improved to a greater degree. Table 3 presents descriptive statistics and ANCOVA results.

The Group × Time interaction was significant for the composite score, $F(1, 208) = 11.06$, $p = .001$, $\eta^2_p = .050$, indicating that CAMS-group gains exceeded those of the comparison group after controlling for baseline proficiency, first-language background, and platform. Pre-test score was a significant predictor across all subscales ($ps < .001$), confirming that baseline proficiency meaningfully influenced post-test outcomes. Platform was a significant covariate for the composite score [$F(2, 208) = 3.44$, $p = .034$, $\eta^2_p = .032$], suggesting modest differences in instructional conditions across platforms. First-language background was not a significant covariate for any subscale (all $ps > .10$).

Among the subscales, the largest treatment effect emerged on Fluency and Coherence [$d = 0.74$, 95% CI (0.46, 1.02), $\eta^2_p = .064$], followed by Interaction Management [$d = 0.52$, 95% CI (0.25, 0.79)] and Task Fulfilment [$d = 0.44$, 95% CI (0.17, 0.71)]. Range and Accuracy showed a smaller but significant effect [$d = 0.35$, 95% CI (0.08, 0.62)]. Phonological Control was the only subscale on which the interaction did not reach significance [$F(1, 208) = 1.86$, $p = .17$, $\eta^2_p = .009$, $d = 0.17$, 95% CI [−0.10, 0.44]]. The small effect size and wide confidence interval crossing zero suggest that CAMS did not produce a detectable differential effect on phonological performance. Possible explanations include limited statistical power, insufficient measurement sensitivity of the phonological subscale, and genuine absence of differential change over 16 weeks; these possibilities are not mutually exclusive and are examined further in the Discussion.

The pattern of gains aligns with, rather than merely illustrates, the analytical framework set out in Section 2.4. Read through Sadler's (1989) three conditions for formative effect—that learners understand the standard, can locate themselves against it, and can act to close the gap—the subscales on which gains were largest are precisely those where rubric descriptors specify observable behaviours that a learner can self-monitor between sessions: silent pauses, topic development, turn-taking moves, task completion. The moderate gain for Range and Accuracy is consistent with Sadler's third condition operating only partially, because lexical and grammatical breadth cannot be manufactured at will in the way that pause frequency or turn-taking can. Phonological Control, which returned a null differential effect, is the dimension on which Sadler's conditions are least likely to be met by a rubric alone: learners can rarely perceive their own phonological errors reliably, and the CEFR's intelligibility descriptors do not specify behaviours that can be targeted in the way that, say, 'extend turns using linking expressions' can. The empirical gradient, in other words, is what Sadler's account would predict; it is picked up again and interpreted against Zimmerman (2000) self-regulation model in Section 5.2.

4.3 Qualitative findings

In response to Reviewer 2's methodological recommendation, learner interview data and instructor focus group data were analysed as separate datasets. Learner themes (Section 4.3.1) and instructor themes (Section 4.3.2) were developed independently before a cross-case comparison (Section 4.3.3) identified points of convergence and divergence.

4.3.1 Learner perspectives

Three themes emerged from analysis of the 38 learner interviews.

4.3.1.1 Theme L1: From opaque evaluation to a legible map

CAMS-group learners consistently reported that the rubric converted evaluation from something private and intuitive into something visible and usable. A Thai learner (B1, Platform A) put it plainly: "Before, my teacher would say 'Your speaking is improving,' and I didn't really know what that meant. Now I can see—okay, for B2 I need to give opinions with supporting reasons without long pauses. That's specific. I can work on that." A Vietnamese learner (B2, Platform C) described the rubric as "like a mirror that shows you exactly where you are standing." Comparison-group learners, by contrast, frequently expressed uncertainty about how their performance was being evaluated. "My teacher says I'm 'intermediate,' but I don't know what I need to do to become 'upper intermediate.' There's no map," observed a Vietnamese learner (B1, Platform C). The metaphor of a map recurred across comparison-group interviews: learners could sense they were moving but could not locate themselves on a recognisable terrain.

4.3.1.2 Theme L2: rubric-mediated self-assessment as an emergent practice

An unanticipated finding was that several CAMS-group learners spontaneously began using the rubric to evaluate their own performance between sessions. Eleven of the 20 interviewed CAMS learners described some form of self-monitoring tied to rubric descriptors. An Arabic-speaking learner (A2, Platform B) described recording herself on her phone and replaying the recording with the rubric open on her laptop: "I check—did I pause too much? Did I use only simple words? The rubric has become like a checklist for me." A Filipino-based Thai learner (B1) reported using the rubric categories as a mental structure during English-medium work calls: "I'm listening to how I sound. If I'm just saying 'yes' and 'no,' that's not interaction management." This practice was absent from the comparison group's accounts, though it cannot be concluded that comparison-group learners did not self-assess in other ways; what is clear is that the rubric, when shared with learners, made

TABLE 3 Pre-Test and post-test speaking scores by group (external rater means).

Subscale	CAMS Pre *M* (*SD*)	CAMS Post *M* (*SD*)	Comp. Pre *M* (*SD*)	Comp. Post *M* (*SD*)	*F* (G × T)	η^2p	*d*
Range & Accuracy	2.61 (0.78)	3.24 (0.71)	2.58 (0.82)	2.94 (0.80)	6.41*	.030	0.35
Fluency & Coherence	2.49 (0.83)	3.30 (0.69)	2.53 (0.79)	2.87 (0.81)	14.27***	.064	0.74
Phonological Control	2.72 (0.74)	3.01 (0.70)	2.69 (0.77)	2.88 (0.76)	1.86	.009	0.17
Interaction Mgmt.	2.44 (0.88)	3.18 (0.72)	2.47 (0.85)	2.82 (0.83)	8.93**	.041	0.52
Task Fulfilment	2.68 (0.80)	3.27 (0.67)	2.64 (0.76)	2.99 (0.78)	5.74*	.027	0.44
Composite	2.59 (0.72)	3.20 (0.62)	2.58 (0.70)	2.90 (0.73)	11.06**	.050	0.56

*$p < .05$; **$p < .01$; ***$p < .001$. Scores on 0–5 scale. Covariates: pre-test score, L1 background (dummy-coded, Arabic as reference), platform. Cohen's *d* reported for Group × Time interaction effect. Holm–Bonferroni correction applied across five subscale comparisons; all results marked as significant remained significant after correction.

self-assessment explicit and category-referenced rather than implicit and global. Read through Zimmerman (2000) three-phase model of self-regulation, these accounts describe the full cycle unfolding around the rubric: the rubric anchors forethought (the learner on work calls who enters the call already monitoring for interaction-management moves), mediates performance (the recording-with-rubric-open sequence), and structures self-reflection (the use of descriptors as retrospective evaluative criteria). Sadler (1989) three conditions for formative effect are, in these accounts, visibly met: the standard is rendered explicit by the descriptor, the current performance is made comparable to the standard through the recording and the checklist, and the category vocabulary gives the learner a language for closing the gap. What these data add to Panadero and Jonsson (2013) review is a concrete illustration of rubric-mediated self-regulation arising without instructor prompting in a commercial online setting, which the review had identified but not documented at the level of learner action.

4.3.1.3 Theme L3: the pronunciation frustration

A minority but vocal strand in the CAMS learner interviews expressed frustration with the Phonological Control category. Learners wanted granular, actionable guidance ("Am I stressing the wrong syllable?" "Is it the/r/ or the/l/ that makes me sound foreign?") rather than global statements about intelligibility. This frustration was not reported by comparison-group learners, for the understandable reason that they had not been exposed to explicit phonological criteria against which to measure their expectations. The theme anticipates a parallel finding in the instructor data (Theme I3 below) and is picked up in Section 5.4.

4.3.2 Instructor perspectives

Three themes emerged from analysis of the four instructor focus groups (12 instructors in total, three per group).

4.3.2.1 Theme I1: a shared evaluative language

Most CAMS-group instructors reported that the rubric provided a shared professional language they had previously lacked. A UK-based instructor with five years of experience described the shift as "moving from vibes to vocabulary": "I used to tell students 'that was better than last time,' and I knew what I meant, but I couldn't have articulated it. Now I can." Two comparison-group instructors, shown the rubric during their focus groups, said they would like to use something similar but had not been trained in any assessment framework since their initial TEFL qualification. This latent appetite for structure is itself diagnostic of the assessment-literacy gap in the freelance online tutoring workforce.

4.3.2.2 Theme I2: the workload and task-fit problem

A sustained and operationally consequential concern among CAMS-group instructors involved workload and task fit. A South African instructor with eight years of experience observed that the rubric worked well for monologues and structured opinion tasks but felt "clunky" during free conversation lessons: "When we're just chatting about their weekend, pulling out a five-category rubric feels forced. I ended up doing a quick mental tally at the end, which is basically what I was doing before." Two Filipino instructors on Platform A estimated that writing structured feedback after each session added approximately five to eight minutes of unpaid labour, "unsustainable," in their word, without platform-level compensation. A UK-based instructor phrased the point in ecological terms: "The problem isn't the rubric. The problem is that the rubric assumes fifteen minutes of paid admin time that doesn't exist in this industry." These concerns did not reduce instructors' endorsement of the principle of structured evaluation; they highlighted, rather, that the principle cannot be enacted sustainably without a corresponding restructuring of platform incentives.

4.3.2.3 Theme I3: the phonological blind spot

CAMS-group instructors identified pronunciation as the category where the rubric was least helpful. "I can tell a student their pronunciation is 'mostly intelligible,' but that doesn't help them fix anything," remarked a UK-based instructor. The CEFR's phonological descriptors, which emphasise intelligibility and communicative adequacy, were described by several instructors as offering too coarse a lens for session-level formative feedback. The concern echoes Harding (2017) argument that pronunciation assessment scales need greater granularity to be pedagogically useful, and it foreshadows the null quantitative result for Phonological Control (Section 4.2).

4.3.3 Convergence and divergence across stakeholder groups

Once the learner and instructor analyses had been completed independently, a cross-case comparison was conducted. Three areas of convergence emerged, alongside one substantive divergence.

4.3.3.1 Convergence 1: structured evaluation improves the specificity of evaluative communication

Learners (Theme L1) and instructors (Theme I1) independently reported that CAMS made evaluation more concrete, more discussable, and more professionally shareable. The two groups described the change in different vocabularies—learners spoke of a "map," instructors of a "vocabulary"—but the underlying point was the same.

4.3.3.2 Convergence 2: pronunciation is the structural weak point

Learners (Theme L3) and instructors (Theme I3) independently identified phonology as the category where structured descriptors were least actionable. This cross-group convergence with the null quantitative result (Section 4.2) provides triangulating evidence that the phonological gap is a substantive feature of the intervention rather than an artefact of a particular measurement or perspective.

4.3.3.3 Convergence 3: evaluation becomes more central to the pedagogical relationship

Both groups described CAMS as changing the character of the tutoring session—shifting evaluation from a background activity to an explicit element of instructional design. For learners, this meant evaluation became a subject of discussion; for instructors, it meant evaluation became a subject of preparation. The two shifts are complementary but not identical.

4.3.3.4 Divergence: the sustainability question

The principal divergence concerned whether CAMS was sustainable in practice. Learners, as beneficiaries of structured feedback who did not bear its production costs, were overwhelmingly positive and expressed a clear wish for its continuation. Instructors, who bore the time and cognitive burden of rubric application and category-specific feedback, were more ambivalent—supportive of the principle but frank about its practical unsustainability under existing platform compensation structures. This asymmetry has direct implications for any attempt to scale CAMS beyond a research setting, and it is discussed in Sections 5.5, 7.

5 Discussion

The results support two broad conclusions, each of which requires substantial qualification. The first concerns reliability: instructors operating within CAMS achieved markedly higher inter-rater agreement than those relying on conventional holistic ratings. The second concerns learning: CAMS-group learners achieved larger gains on four of five subscales, with the strongest effects on fluency and coherence, interaction management, and task fulfilment. These findings are interpreted below through the theoretical and analytical framework set out earlier, beginning with the importance of treating the intervention as a system rather than a tool.

5.1 From rubric to system: interpreting the reliability finding

The 26-point gap between CAMS instructors' inter-rater ICC of.87 and comparison-group instructors'.61 is substantial by any reasonable standard. The difference cannot, however, be attributed to the rubric in isolation. CAMS comprised four interdependent components: the rubric, a six-hour training and calibration workshop, a structured written feedback protocol, and an in-session rubric routine. The comparison group received none of these. It is therefore more accurate to describe the finding as a system effect than as a rubric effect. This reframing is consistent with Reviewer 1's observation that a CEFR-aligned rubric cannot be cleanly separated from the surrounding practices of training, calibration, and structured feedback, and that attributing outcomes to the rubric alone risks misrepresenting the causal structure of the intervention.

Two implications follow. First, replication of the present findings in other contexts will depend on replication of the system as a whole. Providing a rubric without training, or training without structured feedback, is unlikely to produce equivalent reliability gains. Second, the reliability improvement is better understood as an intervention in the monitoring ecology of the platform than as the introduction of a measurement device. The treatment condition changed how evaluation was carried out, who it involved, what it documented, and how it was communicated—changes that together constitute a pedagogical-monitoring reform rather than a scoring refinement. Read through Carless (2015) learning-oriented assessment lens, the ICC gap is less a measurement result than an indicator that evaluative practice has shifted from a private cognitive act to a shared, articulable professional routine—a shift Carless and Boud (2018) identify as a precondition for feedback literacy to develop on either side of the tutoring relationship. In this sense, CAMS resembles what the organisational-learning literature calls a "bundled intervention" (Nolan et al., 2005): an integrated set of changes whose effect arises from their combination rather than their individual contributions.

It is also worth noting that the threshold for "acceptable" reliability depends on the assessment context. Shrout and Fleiss (1979) proposed that ICCs below.70 may be insufficient for individual-level decisions, but those were general guidelines rather than universal cut-offs. In high-stakes testing, where individual decisions carry significant consequences, higher reliability standards are appropriate; in formative contexts such as online tutoring, where the primary purpose of evaluation is to guide subsequent instruction, somewhat lower thresholds may be practically acceptable (Fulcher, 2003). The comparison group's ICC of.61, while lower than desirable, may not be "inadequate" in absolute terms for all formative purposes. What is striking is not that the comparison group fell short of a fixed benchmark but that a modest, four-component intervention could raise instructor agreement by more than a quarter of a point on the ICC scale.

5.2 The feedback–engagement–learning pathway

The analytical framework developed in Section 2.4 proposed a pathway running from feedback internalisation through behavioural engagement and directed practice to learning outcomes. The empirical material gathered here does not establish that pathway causally, but it does supply observable traces of each of its steps, and those traces cohere with the

subscale-level pattern of gains. Feedback internalisation is visible in Theme L1, where CAMS learners described the shift from "your speaking is improving" to specific, category-referenced descriptions of what was being evaluated—the "map" metaphor across the learner interviews marks the moment at which evaluative information became cognitively available to the learner rather than locked in the instructor's judgement. Behavioural engagement is visible in Theme L2, where eleven of the twenty interviewed CAMS learners reported some form of rubric-mediated self-monitoring, with the recording-and-checklist sequence providing a concrete instance of Zimmerman (2000) performance and self-reflection phases being enacted between sessions. Directed practice is harder to observe directly from interview accounts, but it is implicit in statements such as "for B2 I need to give opinions with supporting reasons without long pauses—I can work on that," which points to a specific descriptor being translated into a targeted practice objective. The subscale pattern of gains is, on this reading, the signature one would expect if that pathway were operating: the subscales whose descriptors most readily support all three of Sadler (1989) conditions—Fluency and Coherence, Interaction Management, Task Fulfilment—showed the largest effects; Range and Accuracy, whose underlying resources are less amenable to short-cycle self-regulation, showed moderate gains; and Phonological Control, whose descriptors do not specify behaviours that learners can reliably monitor without perceptual retraining, showed no detectable differential effect. Each link in the chain is independently loose, but the cumulative convergence—quantitative gradient, qualitative themes, and theoretical prediction pointing in the same direction—is the kind of triangulating evidence that the analytical framework was designed to be evaluated against.

This reading does not, and cannot, establish the pathway causally. The quasi-experimental design, combined with the bundled character of CAMS, means that several alternative mechanisms remain live. Instructor effects are the most substantial: CAMS-group instructors self-selected into the treatment and then received six hours of structured calibration, either or both of which could produce the observed outcomes through channels distinct from learner self-regulation. The novelty of the rubric may have raised CAMS-learner motivation independently of the evaluative information the rubric conveyed, a point reinforced by the relatively short intervention window. Hawthorne effects may have operated asymmetrically across conditions. Differential attrition, though statistically modest (Section 8), cannot be ruled out as a contributor. The feedback–engagement–learning pathway is not, on the evidence gathered here, more probable than these alternatives in any strict statistical sense; what can be said is that each link in the pathway has independent empirical traces in the qualitative data, and that the subscale pattern is what that pathway would predict. Stronger evidence would require designs that disaggregate the components of CAMS, measure self-regulatory behaviour directly rather than by interview report, and test transfer to outcome measures built outside the CEFR framework (see Section 9).

The washback literature, which underlies part of this interpretation, is itself more equivocal than a selective reading might suggest. Alderson and Wall (1993) advanced the washback hypothesis; Turner and Purpura (2016) described conditions under which formative assessment can influence learning; but other studies have reported mixed or limited effects (Green, 2007; Watanabe, 2004). The present findings should therefore be situated within a complex research landscape in which the relationship between assessment and learning is mediated by multiple contextual factors, including the stakes of the assessment, the quality of feedback, the disposition of the instructor, and the readiness of the learner to engage with evaluative information.

5.3 Construct alignment and validity considerations

A methodological consideration that deserves explicit discussion concerns the alignment between the intervention and the outcome measure. CEFR-aligned categories were used both as part of the CAMS evaluation framework and as the basis for the external raters' pre- and post-test scoring. This creates a structural overlap between what the intervention cultivated and what the outcome measure detected. The concern, raised clearly by Reviewer 1, is construct alignment bias: improvements observed on the outcome measure may reflect growing familiarity with the evaluative framework rather than broader gains in communicative competence.

Three considerations bear on this concern. First, the use of independent external raters blind to condition eliminates the possibility that the scoring itself was biased toward the treatment group. Second, the elicitation protocol used identical tasks across groups and time points, so that any advantage conferred by familiarity with the rubric would have to manifest in what learners actually produced, not in how it was scored. Third, however, neither of these safeguards addresses the deeper question raised by Reviewer 1: whether gains observed within a CEFR-referenced measurement system reflect ability that transfers beyond that system. A learner who has learned to produce speech that scores well on CEFR-aligned fluency descriptors may or may not also produce speech that would score well on, say, ACTFL-aligned descriptors, or on a holistic global impression scale, or in a communicative task that is not pre-framed by any evaluative framework at all.

The present study cannot resolve this question. It can only flag it and recommend, as Section 9 does, that future research include at least one outcome measure distinct from the intervention's evaluative framework—an ACTFL-based assessment, for example, or a functional communicative task judged on pragmatic success rather than linguistic features. Until such evidence is available, claims about improvements in "speaking ability" derived from CEFR-aligned outcome measures should be read as claims about improvements within the CEFR-referenced construct, not as claims about improvements in speaking ability tout court. This is an important caveat for any attempt to generalise the present findings.

5.4 The phonological control finding

The null result for Phonological Control deserves careful and measured interpretation, particularly in light of the convergent qualitative finding (Themes L3 and I3) that both learners and

instructors described the phonological descriptors as the least actionable element of the rubric. Several explanations for the quantitative null result remain plausible and are not mutually exclusive. The 16-week intervention period may have been too brief for pronunciation changes to manifest in measurable form. The phonological control subscale may have lacked the sensitivity to capture subtle changes. The descriptors themselves may be too broad for this dimension. And the sample size may have provided insufficient statistical power to detect a small effect—the observed $d = 0.17$ is small by conventional standards but has a wide confidence interval that does not exclude the possibility of a practically meaningful effect in a larger sample.

Two additional points warrant attention. First, the critique of CEFR phonological descriptors should be balanced against the observation that inconsistent application may reflect insufficient instructor training or limited phonological expertise rather than inherent descriptor inadequacy. While Harding (2017) argued for greater granularity in pronunciation assessment scales, well-trained raters with phonological expertise can apply broad descriptors more informatively than less-trained raters can apply detailed ones. The six-hour CAMS training, though more extensive than what most online platforms offer, may not have been sufficient to develop the specialised phonological assessment skills needed to use the Phonological Control category productively. The issue, in other words, may be partly pedagogical and partly infrastructural rather than purely descriptor-level.

Second, the pattern is theoretically coherent with the pedagogical model outlined in Section 2.4, and this coherence matters because the null was not a *post hoc* surprise but a direction the framework predicted. Sadler (1989) conditions for formative effect require that the learner be able to perceive a gap between current and target performance; for most learners, the perception of one's own phonological errors is unreliable, and the CEFR's phonological descriptors do not resolve that problem because they specify outcomes (intelligibility, communicative adequacy) rather than behaviours the learner can monitor. Rubric-mediated self-regulation works best where learners can consciously translate feedback into practice—by monitoring pause patterns, working on topic development, or practising turn-taking strategies. Pronunciation improvement depends additionally on perceptual retraining and articulatory adjustment, processes that are not reliably triggered by rubric-level feedback alone (Derwing and Munro, 2015). That the null landed exactly where the framework said short-cycle self-regulation should be least effective is the kind of disconfirmation-by-specificity that strengthens rather than weakens the mechanism account: the pathway does not predict uniform gains, it predicts the gradient that was observed. The implication for system design is that phonology should be handled by a specialised sub-module—targeted listening exercises, corrective feedback on specific segments, perhaps AI-assisted articulatory analysis—rather than as one category within a general analytic rubric.

5.5 The ecological dimension of online assessment

The instructor concerns documented in Theme I2 point to a broader consideration that extends beyond the rubric's psychometric properties. Online tutoring platforms constitute complex pedagogical ecologies shaped by time constraints, instructor availability, platform design decisions, and the interactional dynamics of one-to-one video instruction (Sato and Loewen, 2019). Introducing a structured monitoring system into this ecology affects not only measurement quality but also the allocation of instructional time, the character of the instructor–learner relationship, and the economic conditions under which tutoring is delivered.

The five-to-eight minutes of additional post-session labour reported by some instructors amounts to roughly two hours per week for an instructor teaching 15–20 sessions—time that, within the gig-economy logic governing most platforms, goes uncompensated. Unless platforms integrate rubric completion into their session workflows and pay for the time involved, adoption will remain uneven and, as one focus group participant put it plainly, may breed resentment. The perceived "clunkiness" of applying analytic rubrics to informal conversation tasks is a related but distinct challenge that speaks to assessment–instruction alignment. A plausible response, not tested here, would be a simplified two- or three-category rubric for conversation-focused sessions that preserves criterion-referenced scoring without the overhead of the full five-category instrument. This would resemble what software designers call a "progressive disclosure" pattern: the full system is available when needed, a streamlined version is available when not.

More broadly, the findings suggest that structured evaluation interventions in online tutoring environments must be designed with attention to the full ecological context—platform economics, session formats, instructor working conditions, interactional norms—rather than imported wholesale from examination contexts where those constraints do not apply. This is where Carless (2015) learning-oriented assessment framing cuts most sharply against the practice assumptions of the commercial platforms: if evaluation is to feed forward into learning rather than simply certify attainment, the time needed to produce category-referenced feedback cannot sit outside the paid session, because the economic structure then systematically selects against the very feedback quality the system depends on. The research question that emerges from this ecological perspective is no longer "Does the rubric work?" but "What platform, instructor, and learner conditions must hold for a structured monitoring system to work sustainably?" That question is larger than a single study can answer.

5.6 CEFR-specific or structured-monitoring effects?

A question this study raises but cannot definitively answer concerns the specificity of the observed effects. CAMS was built around CEFR-aligned descriptors, but its other components—training, calibration, structured feedback, in-session routines—could in principle be assembled around a different framework. The present design does not separate framework-specific effects from the general effect of structured monitoring, and the distinction matters for practical recommendation. If the CEFR's specific descriptors are the active ingredient, platforms should invest in CEFR-aligned tools specifically. If the act of structuring

evaluation is what drives improvement, the choice of framework matters less than the quality of implementation.

Three considerations suggest that at least part of the observed effect is likely to be framework-general rather than CEFR-specific. First, the core mechanisms identified in Section 2.4—feedback internalisation, self-regulation, directed practice—are not CEFR-specific; they are general features of effective formative assessment. Second, comparable studies using non-CEFR analytic rubrics (e.g., Andrade and Heritage, 2017) have reported similar self-assessment effects. Third, Kang et al.'s (2019) work shows that rater training alone, without reference to any specific framework, improves oral proficiency scoring. Against that, however, the CEFR offers a specific set of descriptors that are widely recognised, accessible to instructors across national contexts, and connected to the qualifications learners and employers care about. In applied contexts, the choice of framework may matter less for its internal content than for the external recognition it enables. A definitive answer to the specificity question awaits comparative studies that vary the framework while holding other system components constant.

6 Conclusion

This study examined a straightforward proposition: that providing online ESL instructors with a structured, CEFR-aligned monitoring system for evaluating speaking—comprising a rubric, rater training, a structured feedback protocol, and an in-session rubric routine—would be associated with more consistent scoring, greater learner gains on speaking performance, and a clearer evaluative experience for both learners and instructors. The evidence is consistent with that proposition, though important caveats apply. Inter-rater reliability in the treatment condition was substantially higher, speaking gains favoured the treatment group on four of five subscales, and the qualitative data documented both the benefits and the costs of the shift from impressionistic to criterion-referenced evaluation in rich detail.

Three broader conclusions follow, each calibrated to what the design can support. First, the intervention is best understood as a pedagogical monitoring system rather than a scoring tool. The rubric alone would not have produced the observed effects; it is the integration of rubric, training, calibration, and structured feedback that constitutes the active intervention, and no component-level causal claim can be made within the present design. Second, the pattern of findings is consistent with, though not established by, a pedagogical mechanism running from feedback internalisation through self-regulation to directed practice. The uneven pattern of subscale gains—stronger effects on dimensions amenable to conscious self-regulation, a predicted null for phonology—is what that mechanism would lead one to expect; stronger warrant for the mechanism itself awaits the designs outlined in Section 9. Third, the ecological context of online tutoring—platform economics, session formats, instructor working conditions—is not a secondary concern but a constitutive feature of what structured evaluation can and cannot achieve in practice. Any attempt to scale CAMS or comparable systems must design for the ecology, not around it.

These findings should not be read as a causal claim that CEFR-aligned rubrics alone will resolve the deep assessment challenges facing the online ESL industry. The quasi-experimental design, the confounding of rubric use with training, and the absence of random assignment all preclude strong causal inference. What the findings do suggest is that introducing a shared evaluative language—embedded within a coherent monitoring system—into a context currently characterised by inconsistency, opacity, and well-intentioned but unreliable professional judgment is a productive step. Whether platforms choose to invest in that shared language, and to compensate the instructors who bring it to life, remains a question of institutional priorities rather than of research evidence. The central challenge of online speaking assessment is not the absence of frameworks, but the absence of systems capable of operationalising them within real pedagogical ecologies. Looking forward, integrating AI-assisted pronunciation feedback into human-administered monitoring systems offers a promising avenue for addressing the phonological gap identified here (Xu et al., 2021), without displacing the pedagogical relationship that gives online tutoring its value.

7 Implications for practice

7.1 For online ESL platforms

Platforms considering the adoption of CEFR-aligned speaking evaluation should approach it as system design rather than tool procurement. A rubric without training produces marginal reliability gains at best; a rubric with training but without structured feedback leaves much of the pedagogical value on the table; a rubric with all three but without compensation for the additional instructor time produces resentment and uneven uptake. The reliability improvement documented in this study was achieved through a combination of a six-hour workshop, a rubric, a written feedback protocol, and an in-session routine—a modest investment compared with the cost of automated scoring systems or full-time quality-assurance staff, but not zero. Platforms currently relying on informal instructor evaluations could pilot CAMS or a comparable system with a subset of instructors and compare reliability and learner outcomes before committing to system-wide implementation. The pilot should explicitly evaluate the sustainability dimension, not only the measurement dimension.

7.2 For programme coordinators and curriculum designers

Rubric categories should be adapted to the specific task ecology of online tutoring rather than imported wholesale from examination contexts. The Phonological Control gap and the instructor concerns about conversational task fit both point to a need for context-sensitive design. The CEFR provides a useful skeleton, but behavioural indicators anchored to specific online interaction types—short monologues, conversational exchanges, transactional tasks—must be developed with local conditions firmly in mind. A system with two or three layers of granularity—a full analytic rubric for formal tasks, a streamlined three-category version for conversational sessions—may offer a more realistic fit than a single, uniformly applied instrument.

7.3 For instructors

The self-assessment finding suggests that sharing rubrics with learners—not merely using them for evaluation—may amplify their formative value. Instructors might consider walking learners through the rubric at the beginning of a course, inviting learners to self-rate before comparing with instructor ratings, and framing the rubric as a shared reference point for goal-setting rather than an instrument of judgment. This recommendation is grounded in the broader theoretical argument that effective formative assessment depends on learner access to evaluative criteria (Nicol and Macfarlane-Dick, 2006; Sadler, 1989) and is consistent with the learner-reported behaviours documented in Theme L2.

8 Limitations

Several limitations should temper the interpretation of these findings, and claims made elsewhere in this manuscript should be read in light of them.

First, the quasi-experimental design means that group differences cannot be causally attributed to CAMS as a whole or to any of its components individually. Instructors who volunteered for the treatment condition may have been more assessment-literate, more motivated, or more receptive to structured evaluation from the outset—a self-selection threat that no amount of statistical control can fully eliminate. The bundled nature of the intervention compounds this problem: even if the aggregate effect can be attributed to CAMS, the contribution of any single component cannot be isolated from the contributions of the others.

Second, the 16-week intervention period, while longer than many comparable studies, may be insufficient to detect slower-moving changes such as phonological improvement or to determine whether reliability gains persist once the novelty of training effects fades. Third, attrition was non-negligible. Of 248 learners initially enrolled, 34 (13.7%) withdrew before the post-test. Although dropout was distributed fairly evenly across conditions and attrition analyses indicated no significant differences between completers and non-completers on pre-test scores or demographics, the possibility of differential attrition bias cannot be entirely excluded.

Fourth, the study focused on three platforms operating in Southeast Asia and the Middle East, which limits generalisability to other regions and platform models. Platforms serving learners in Latin America, East Africa, or Eastern Europe may differ meaningfully in instructor profiles, session structures, and cultural expectations around assessment. Fifth, the study measured instructor inter-rater reliability on session-level ratings but did not conduct a full many-facet Rasch analysis, which would have allowed for simultaneous modelling of rater severity, task difficulty, and examinee ability; the reported ICCs therefore represent one perspective on reliability rather than an exhaustive psychometric analysis.

Sixth, as discussed at length in Section 5.3, the outcome measure was built from the same CEFR-aligned categories that shaped the intervention. Improvements on the outcome measure may therefore partly reflect growing familiarity with the evaluative framework rather than broader communicative gains. This is the most consequential single limitation of the study, and it constrains the scope of the claims that can legitimately be made.

Finally, the study focused on adult learners on commercial online tutoring platforms. The findings may not be directly generalisable to institutional language programmes, classroom-based ESL instruction, or younger learners whose assessment literacy and self-regulatory capacities develop differently.

9 Directions for future research

Five directions for future research follow from the limitations discussed above. First, experimental designs that separate CAMS components—rubric alone, rubric plus training, rubric plus training plus feedback—would help disaggregate the relative contribution of each element to the observed effects. Only this kind of design can adjudicate between the rubric-as-instrument and rubric-as-system interpretations proposed in this study.

Second, longitudinal designs tracking learners over six months or longer would clarify whether the proficiency gains and reliability improvements reported here are durable, and whether they compound, plateau, or erode as the novelty of the intervention fades. Third, studies employing many-facet Rasch measurement would provide more nuanced reliability evidence, particularly regarding rater severity and task difficulty, which the ICC-based analysis cannot capture.

Fourth, and in direct response to the construct-alignment concern discussed in Section 5.3, future studies should include outcome measures that are not built from the intervention's evaluative framework. An ACTFL-based assessment, a holistic global impression scale, or a pragmatic communicative task judged on functional success rather than linguistic features would provide a test of whether gains observed within a CEFR-referenced system transfer beyond it. Comparative work examining CAMS-equivalents built around alternative frameworks (ACTFL, Trinity, locally developed scales) would help resolve the framework-specificity question raised in Section 5.6.

Fifth, the interaction between structured monitoring systems and AI-assisted pronunciation feedback is a promising avenue for addressing the phonological blind spot. As speech-processing technology matures, hybrid systems combining human evaluation of discourse-level features with automated analysis of segmental accuracy could address the specific gap identified in this study. Research exploring the feedback–self-regulation pathway more rigorously—through think-aloud protocols, learning diaries, or mixed-method tracking of learner practice between sessions—would provide stronger evidence for the mechanisms through which structured monitoring influences learning behaviour. Finally, research into learner engagement with assessment criteria in self-directed, instructor-absent contexts would illuminate the potential of structured monitoring tools in the rapidly growing asynchronous language-learning market.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Ethics statement

The studies involving humans were approved by University of Benghazi, Benghazi, Libya. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent was obtained from all participants prior to participation in the study.

Funding

The author(s) declared that financial support was not received for this work and/or its publication.

Author contributions

SA: Conceptualization, Methodology, Investigation, Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Project administration.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Alderson, J. C. (2007). The CEFR and the need for more research. *Modern Lang. J.* 91 (4), 659–663. doi: 10.1111/j.1540-4781.2007.00627_4.x

Alderson, J. C., and Wall, D. (1993). Does washback exist? *Appl. Linguist.* 14 (2), 115–129. doi: 10.1093/applin/14.2.115

Ambient Insight (2023). *The 2022–2027 Worldwide Digital English Language Learning Market*. Monroe, WA: Ambient Insight.

Andrade, H. L., and Heritage, M. (2017). *Using Formative Assessment to Enhance Learning, Achievement, and Academic Self-regulation*. New York, NY: Routledge.

Bachman, L. F., and Palmer, A. S. (2010). *Language Assessment in Practice*. Oxford: Oxford University Press.

Black, P., and Wiliam, D. (1998). Assessment and classroom learning. *Assess. Educ. Princ. Policy Pract.* 5 (1), 7–74. doi: 10.1080/0969595980050102

Braun, V., and Clarke, V. (2021). *Thematic Analysis: A Practical Guide*. London: SAGE.

Carless, D. (2015). *Excellence in University Assessment: Learning from Award-winning practice*. Abingdon: Routledge.

Carless, D., and Boud, D. (2018). The development of student feedback literacy: enabling uptake of feedback. *Assess. Eval. High. Educ.* 43 (8), 1315–1325. doi: 10.1080/02602938.2018.1463354

Chapelle, C. A., and Voss, E. (2016). 20 Years of technology and language assessment in language learning & technology. *Lang. Learn. Technol.* 20 (2), 116–128. doi: 10.64152/10125/44464

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press.

Council of Europe (2020). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment – Companion Volume*. Council of Europe Publishing. Available online at: https://www.coe.int/lang-cefr (Accessed November 15, 2025).

Creswell, J. W., and Plano Clark, V. L. (2018). *Designing and Conducting Mixed Methods Research*. 3rd ed. SAGE.

Derwing, T. M., and Munro, M. J. (2015). *Pronunciation Fundamentals: Evidence-based perspectives for L2 Teaching and Research*. John Benjamins.

Field, A. (2018). *Discovering Statistics Using IBM SPSS Statistics*. 5th ed. SAGE.

Figueras, N. (2012). The impact of the CEFR. *ELT J.* 66 (4), 477–485. doi: 10.1093/elt/ccs037

Fulcher, G. (2003). *Testing second language speaking*. Harlow: Pearson Longman, 320.

Fulcher, G. (2004). Deluded by artifices? The common European framework and harmonization. *Lang. Assess. Q.* 1 (4), 253–266. doi: 10.1207/s15434311laq0104_4

Galaczi, E. D., and Taylor, L. (2018). Interactional competence: conceptualisations, operationalisations, and outstanding questions. *Lang. Assess. Q.* 15 (3), 219–236. doi: 10.1080/15434303.2018.1453816

Green, A. (2007). *IELTS washback in Context: Preparation for Academic Writing in Higher Education*. Cambridge: Cambridge University Press.

Harding, L. (2017). "What do raters need in a pronunciation scale? The users' view," in *Second Language Pronunciation Assessment*, eds. T. Isaacs, and P. Trofimovich (Bristol: Multilingual Matters), 12–34.

Harsch, C., and Martin, G. (2012). Adapting CEF-descriptors for rating purposes: validation by a combined rater training and scale revision approach. *Assess. Writ.* 17 (4), 228–250. doi: 10.1016/j.asw.2012.06.003

Hattie, J., and Timperley, H. (2007). The power of feedback. *Rev. Educ. Res.* 77 (1), 81–112. doi: 10.3102/003465430298487

HolonIQ (2024). *Global EdTech Market Map and Forecast 2024*. New York, NY: HolonIQ. Available online at: https://www.holoniq.com (Accessed November 15, 2025).

Hulstijn, J. H. (2007). The shaky ground beneath the CEFR: quantitative and qualitative dimensions of language proficiency. *Modern Lang. J.* 91 (4), 663–667. doi: 10.1111/j.1540-4781.2007.00627_5.x

Isbell, D. R., and Kremmel, B. (2020). Test review: current options in at-home language proficiency tests for making high-stakes decisions. *Lang. Test.* 37 (4), 600–619. doi: 10.1177/0265532220943483

Kang, O., Rubin, D., and Kermad, A. (2019). The effect of training and rater differences on oral proficiency assessment. *Lang. Test.* 36 (4), 481–504. doi: 10.1177/0265532219849522

Luoma, S. (2004). *Assessing Speaking*. Cambridge: Cambridge University Press.

Moss, P. A. (2016). Shifting the focus of validity for test use. *Assess. Educ. Princ. Policy Pract.* 23 (2), 236–251. doi: 10.1080/0969594X.2015.1072085

Nakatsuhara, F., Inoue, C., Berry, V., and Galaczi, E. (2017). Exploring the use of video-conferencing technology in the assessment of spoken language: a mixed-methods study. *Lang. Assess. Q.* 14 (1), 1–18. doi: 10.1080/15434303.2016.1263637

Nakatsuhara, F., Inoue, C., Berry, V., and Galaczi, E. (2021). Video-conferencing speaking tests: do they measure the same construct as face-to-face tests? *Assess. Educ. Princ. Policy Pract.* 28 (4), 369–388. doi: 10.1080/0969594X.2021.1951163

Nicol, D. J., and Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. *Stud. High. Educ.* 31 (2), 199–218. doi: 10.1080/03075070600572090

Nielson, K. B. (2011). Self-study with language learning software in the workplace: what happens? *Lang. Learn. Technol.* 15 (3), 110–129. doi: 10.64152/10125/44265

Nolan, T., Resar, R., Haraden, C., and Griffin, F. A. (2005). *Improving the Reliability of Health Care (IHI Innovation Series White Paper)*. Institute for Healthcare Improvement.

North, B. (2000). *The Development of a Common Framework Scale of Language Proficiency*. Peter Lang.

North, B. (2014). *The CEFR in Practice*. Cambridge University Press.

Panadero, E., and Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: a review. *Educ. Res. Rev.* 9, 129–144. doi: 10.1016/j.edurev.2013.01.002

Papageorgiou, S., Wu, S.-M., Hsieh, C.-N., Tannenbaum, R. J., and Cheng, M. (2019). *Mapping the TOEFL iBT Test Scores to the CEFR*. Princeton, NJ: Educational Testing Service.

Pearson (2011). *Versant English Test: Test Description and Validation Summary*. London: Pearson Education.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instr. Sci.* 18 (2), 119–144. doi: 10.1007/BF00117714

Sato, M., and Loewen, S. (2019). Do teachers care about research? The research–pedagogy dialogue. *ELT J.* 73 (1), 1–10. doi: 10.1093/elt/ccy048

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and Quasi-experimental designs for Generalized Causal Inference*. Boston, MA: Houghton Mifflin.

Shrout, P. E., and Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychol. Bull.* 86 (2), 420–428. doi: 10.1037/0033-2909.86.2.420

Shute, V. J. (2008). Focus on formative feedback. *Rev. Educ. Res.* 78 (1), 153–189. doi: 10.3102/0034654307313795

Taylor, L., and Galaczi, E. (2011). "Scoring validity," in *Examining Speaking: Research and Practice in Assessing Second Language Speaking*, ed. L. Taylor (Cambridge: Cambridge University Press), 171–233.

Timmis, S., Broadfoot, P., Sutherland, R., and Oldfield, A. (2016). Rethinking assessment in a digital age: opportunities, challenges and risks. *Br. Educ. Res. J.* 42 (3), 454–476. doi: 10.1002/berj.3215

Turner, C. E., and Purpura, J. E. (2016). "Learning-oriented assessment in second and foreign language classrooms," in *Handbook of Second Language Assessment*, eds. D. Tsagari, and J. Banerjee (De Gruyter Mouton), 255–271.

Watanabe, Y. (2004). "Methodology in washback studies," in *Washback in Language Testing: Research Contexts and Methods*, eds. L. Cheng, Y. Watanabe, and A. Curtis (Lawrence Erlbaum), 19–36.

Whitelock, D. (2010). "Activating assessment for learning: are we on the way with web 2.0?," in *Web 2.0-based e-learning*, eds. M. J. W. Lee, and C. McLoughlin (IGI Global), 319–342.

Wisniewski, K. (2017). Empirical learner language and the levels of the common European framework of reference. *Lang. Learn.* 67 (S1), 232–253. doi: 10.1111/lang.12223

Xi, X. (2010). Automated scoring and feedback systems: where are we and where are we heading? *Lang. Test.* 27 (3), 291–300. doi: 10.1177/0265532210364643

Xu, J., Jones, E., Laxton, V., and Galaczi, E. (2021). Assessing L2 English speaking using automated scoring technology: examining automarker reliability. *Assess. Educ. Princ. Policy Pract.* 28 (4), 411–436. doi: 10.1080/0969594X.2021.1979467

Zechner, K., Higgins, D., Xi, X., and Williamson, D. M. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech. Commun.* 51 (10), 883–895. doi: 10.1016/j.specom.2009.04.009

Zimmerman, B. J. (2000). "Attaining self-regulation: a social cognitive perspective," in *Handbook of Self-regulation*, eds. M. Boekaerts, P. R. Pintrich, and M. Zeidner (San Diego, CA: Academic Press), 13–39.

Appendix A

The CEFR-Aligned Speaking Rubric (Abridged)

Category	3 (B1 benchmark)	4 (B2 benchmark)	5 (C1 benchmark)
Range & Accuracy	Uses common vocabulary and simple grammatical structures with reasonable accuracy; errors occur but do not impede overall communication.	Demonstrates varied vocabulary (including idiomatic expressions and topic-specific terms) and uses complex structures (e.g., conditional, passive) with occasional errors that are often self-corrected.	Consistently uses a broad lexical and grammatical range with high accuracy; errors are rare and do not affect clarity or precision of meaning.
Fluency & Coherence	Can maintain speech on familiar topics though planning pauses are noticeable; uses basic connectors (and, but, because) to link ideas.	Speaks with relatively even tempo; hesitations relate to content search rather than language retrieval; uses discourse markers (however, on the other hand) to organise extended turns.	Speaks fluently with only occasional hesitation when searching for precise expression; discourse is well organised with effective use of cohesive devices across extended turns.
Phonological Control	Generally intelligible though L1 influence is evident in segmental features; sentence stress and intonation are mostly appropriate for declarative and interrogative patterns.	Clear pronunciation with good control of stress, rhythm, and intonation; occasional mispronunciations of less common words do not cause listener strain.	Pronunciation and prosody are consistently clear and natural; can vary intonation patterns for pragmatic effect (e.g., expressing doubt, emphasising a point).
Interaction Management	Can initiate, maintain, and close simple conversation on familiar topics; may need interlocutor support for topic shifts or repair.	Can take turns effectively in discussion, give and solicit opinions, manage topic shifts, and employ repair strategies (e.g., self-correction, circumlocution) when communication breaks down.	Interacts with ease in both familiar and unfamiliar contexts; can moderate discussion, handle interruptions smoothly, and adapt register and style to the interlocutor.
Task Fulfilment	Addresses task requirements with some relevant content; may omit key points or wander from the central topic.	Fully addresses the task with relevant and organised content, developing main points with supporting detail; may lack nuance or depth in argumentation.	Addresses the task comprehensively with nuance, precision, and register appropriate to the communicative context.

Bands 3, 4, and 5 anchor the B1, B2, and C1 CEFR levels respectively. Bands 0–2 (omitted here) correspond to pre-A1, A1, and A2 performance. Instructors scored each category on the full 0–5 scale.

Descriptors at Bands 3 (B1 benchmark), 4 (B2 benchmark), and 5 (C1 benchmark) are shown. Descriptors for Bands 0–2 are omitted for brevity. The full rubric with detailed behavioural indicators and example performances is available from the corresponding author upon request.