E·S·P·Today
JOURNAL OF ENGLISH FOR SPECIFIC PURPOSES AT TERTIARY LEVEL

Jasmina Đorđević*

Faculty of Philosophy
University of Niš, Serbia
jasmina.djordjevic@filfak.ni.ac.rs

RUBRICS IN THE ASSESSMENT OF EAP SPEAKING SKILLS SUPPORTED BY MOBILE ASSISTED LANGUAGE LEARNING

Abstract

Research on oral performance assessment in an English for Academic Purposes (EAP) learning context supported by Mobile Assisted Language Learning (MALL) is scarce. This research aimed to investigate whether rubrics may be expected to yield valid, reliable and replicable data when evaluating EAP speaking skills in remote classes supported by MALL. An action-based case study was conducted with a convenience sample of 93 students (from three different cohorts: 2019, 2020 and 2021) video-recording themselves delivering speeches with their smartphones in an English language university course including EAP writing and speaking skills as substantial requirements. The students' EAP speaking skills were evaluated over three years with descriptive, analytic, task-specific and teacher-created rubrics contextualised in real-world language. The rubrics were designed to include a rating scale, specific criteria, levels of achievement and descriptors. The results indicate that when implemented as a preparation and assessment tool, rubrics provide valid, reliable and replicable data for assessing EAP speaking skills in remote classes supported by MALL. The practical implications of these findings are significant since the data collected in this way can help language instructors evaluate the students' spoken communicative competence in remote classes that would otherwise be difficult to assess.

Key words

rubrics, assessment tool, EAP speaking skills, MALL, communicative competence.

* Corresponding address: Jasmina Đorđević, Faculty of Philosophy, Ćirila i Metodija 2, 18105 Niš, Serbia.

1. INTRODUCTION

Mobile Assisted Language Learning (MALL) is defined as "the use of smartphones and other mobile technologies in language learning, especially in situations where portability and situated learning offer specific advantages" (Kukulska-Hulme, 2020, p. 1). During the first decade of the new millennium, research dominating the field of MALL was mainly content-based and less focused on design issues (Kukulska-Hulme & Shield, 2008). During the second decade of the 2000s, the focus of MALL research shifted to design investigations (Wong & Looi, 2010). Little research is available on specific assessment procedures focusing on speaking skills (Kukulska-Hulme, 2020, 2021) and even less research seems to be available on how to collect valid, reliable and replicable data to evaluate EFL/EAP/ESP students' mastery of speaking skills in a MALL-supported context in actual classroom conditions. This research aims to fill that gap.

Changes in technology provide language teachers with new opportunities regarding assessment (Soo, 2023). The COVID-19 pandemic forced educators to consider different assessment alternatives and implement those "they deemed critical for maintaining the validity of their assessment", which is why they "used the challenge as an opportunity to upgrade the quality of their assessment" (Muhammad & Ockey, 2021, p. 54). The research presented here started in May 2020, amid the COVID-19 pandemic, to assess student speaking skills in remote English classes supported by MALL with the help of rubrics. The course includes English for general academic purposes (EGAP) focusing on a general academic register (see Hulme, 2021 regarding details about EGAP). For the sake of simplicity, the acronym EAP is used throughout this study. The students are expected to reach a C1.2 level of competence in General English and EAP. While the CEFR recognises six levels of language proficiency (A1, A2, B1, B2, C1 and C2) "supplementary descriptors and sublevels are a desideratum, particularly in tertiary language education" (Berger, 2020, p. 85). The faculty where the present research was conducted, implements sublevels as a finer gradation to tailor the courses to the learners' specific needs (B1.1 – 1st semester, B1.2 – 2nd semester…, C1.2 – 6th semester and each level is specified in the respective course syllabi).

A starting point in this research was Brunfaut's (2023) argument that technology-enhanced language assessment today is mostly focusing on large-scale (English) proficiency testing suggesting that "[f]uture opportunities lie therefore in the feasibility of, for example, small-scale, classroom testing, and testing in a variety of world languages and for different purposes and needs" (2023, p. 20). Therefore, the aim of the small-scale three-year action-based case study presented here, which relied on a convenience sample of 93 students from three different cohorts (2019, 2020 and 2021), is not to propose rubrics as an approach in large-scale language testing because that would require far more elaborate research to verify the validity of rubrics for an accurate interpretation of the learners' language ability. This study is meant to show that speaking as a productive skill in EAP in everyday classroom

conditions can be measured with quantitative data from descriptive, analytic, task-specific and teacher-created rubrics contextualised in real-world language in remote classes supported by MALL providing a numerical assessment and evaluation of the actual product – the student's speech. Given that assessing rubrics over three years allows for observing their consistency and adaptability across different student groups, this study addresses the critical area of speaking production in language education, where objective assessment is challenging. Following an overview of assessment in MALL and a review of rubrics in language assessment, the article will present evidence that specifically designed rubrics are a practical classroom solution that can yield valid, reliable and replicable data serving the objective assessment of the student's oral communicative competence in remote EAP classes supported by MALL.

1.1. Assessment in MALL

Assessment in MALL and how to implement it are not a novelty in scholarly research (García Laborda et al., 2014; Hwang & Chen, 2013; Rezaee et al., 2020; Samaie et al., 2018; Tarighat & Khodabakhsh, 2016). García Laborda et al. (2014) developed an online testing system (PAULEX, "PAU en Lenguas Extranjeras"), proving that mobile-based assessment works in high-stakes testing, such as a university entrance exam. Rezaee et al. (2020) investigated the application of mobile-based dynamic assessment on the development of learners' oral fluency in the Iranian EFL context; Hwang and Chang (2021) explored what the impact of a mobile concept mapping system would be in the context of bi-directional peer-assessment; Samaie et al. (2018) relied on Bonk and Ockey's (2003) rating scale, while Tarighat and Khodabakhsh (2016) implemented dynamic assessment, a learner-based assessment which enabled them to consider individual differences among learners during the assessment procedure. Each of these assessment strategies proved numerous advantages.

Nevertheless, quite a substantial number of investigations presenting empirically collected data on the success of MALL when teaching and practising speaking lack a presentation of the assessment procedure they relied on when testing their students' spoken performance (e.g., Abugohar et al., 2019; Azlan et al., 2019; Chen Hsieh et al., 2017; Sun et al., 2017; Wong et al., 2010). All these authors predominantly explored the formal context of learning and how to deliver content to learners, i.e., traditional education paradigms. They also showed that MALL contributes to better language learning in general. Yet, none provided insight into how the improvement was formally measured, so it remains unclear what assessment tool these studies relied on to gather necessary information based on which the identified progress in the learning process was evaluated.

For instance, Abugohar et al. (2019) relied on a survey to analyse the opinions of higher education EFL teachers who perceived smartphone applications as having

a positive impact on their students' fluency. However, the authors failed to report a specific procedure testifying how they determined the positive impact. In a study on Active Learning, Azlan et al. (2019) determined that learners actively engaging and constructing meaning in real-world tasks stimulated their motivation and increased engagement. As reported, MALL provided a lively and fun environment that was perceived as less stressful (Azlan et al., 2019). The participants were asked to complete a task-based speaking performance and record it. The recordings were then sent to their parents (not teachers) to provide student performance feedback. Yet, formal measuring, assessment or evaluation techniques are not reported. Similarly, Sun et al. (2017) confirmed that MALL provided a comfortable atmosphere for learners to express their thoughts in oral communication. The formal assessment data were again based on a survey distributed to the respondents, relying solely on their feedback. Wong et al. (2010) conducted their study by drawing on observation and interactions with the teacher and the students throughout the study and post-interviews. An apparent reference regarding which assessment tool they used is missing. Since the main focus of the study was on the students' meaning-making skills based on how they contextualised the idioms they were expected to learn, the least a reader would expect is that the study relied on some formal measuring scale. If it did, there is no reference to it in the article.

This lack of precise assessment strategies in MALL was also confirmed by Chen and Lin (2023), who reviewed 11 peer-reviewed articles published between 2015 and 2021 on MALL assessment. Their review showed that the articles featured assessment tools focusing on vocabulary, grammar and pronunciation while higher-level skills (e.g., inferencing, comprehension monitoring and awareness of text structure) were not included, skills essential for EAP. The review further showed that most of the assessment tools presented in the articles did not rely on user-centred and interactive features, nor did they include a wider variety of tasks which could target more complex skills.

Similar to Bazhutina and Tsepilova (2024), the term "assessment" in this study refers to the implementation of language competence and focuses on learner performance and its analysis. An assessment tool that could fill the gap identified by Chen and Lin (2023) are rubrics, in particular when implemented to assess speaking skills in MALL. Rubrics have been tested in multiple investigations related to language learning in different contexts (for a more detailed review, see Dawson, 2017). When assessing speaking, rubrics have been used primarily to test formal aspects of speech production (fluency, clarity, articulation, sentence structure, grammar, etc.). This study shows that specifically designed rubrics can yield valid, reliable and replicable data in remote classes supported by MALL to assess student communicative competence in EAP. This aspect of oral performance is challenging to evaluate in the context of MALL due to the lack of face-to-face real-time communication. Therefore, it is difficult to evaluate more complex higher-level skills (Chen & Lin, 2023; Fulcher, 2017; Schreiber et al., 2012).

1.2. Rubrics as an assessment tool in EFL

In their re-evaluation of rating scales, i.e., scoring rubrics, Knoch et al. (2021) relied on Fulcher (2012), who argued that performance-driven scales are constructed based on real-world data and corpora, which is why they are more in line with real-world language use. Unlike that, measurement-driven scales, usually adapted from the Council of Europe Framework of References (CEFR) descriptors (Council of Europe, 2020), have been criticised for lacking an empirical and theoretical foundation (Knoch et al., 2021). If rubrics are to provide validity, the observed performance should be translated to a score with real-world meaning. Therefore, the alignment between scale criteria and real-world language use must be strong (Fulcher, 2012). Fulcher (2017) also suggested that rating scale validation is possible only if the test goals, tasks and rating criteria are contextualised. If this is not the case, rubrics will have the quality of generalizability but not precision and predictive strength. Performance data-based scales may be assumed to be most adequate as they can be empirically derived and operationalised in various ways to enable a smooth rating process (Knoch et al., 2021). Finally, Dawson (2017) suggested that task-specific rubrics are probably best if the rubrics have to apply to a specific instance of assessment in a particular context. Moreover, if rubrics combine task-specific and teacher-created rubrics, they should yield the best possible results.

According to Rezaei and Lovorn (2010), rubrics can be holistic, analytical or combined. Holistic rubrics are product-oriented and analytical rubrics consist of separate scales for multiple traits, so they can provide a set of scores rather than just one score. According to Schreiber et al. (2012), rubrics can be designed as rating scales or as descriptive rubrics. The former includes a list of key competencies and a scale to demonstrate a degree or level of aptitude. The performance levels can be numeric (e.g., scores within a specific range), descriptive (e.g., good, fair, poor), indicate the presence or absence of a specific behaviour (e.g., often, sometimes, rarely), or rely on criteria that the rater can define. Criticism of rating scale rubrics stresses subjectivity and a lack of clarity. Descriptive or analytic rubrics are more detailed and reliable. They include brief descriptions of the expected performance level for each score within a category. The descriptors "spell out the performance standards for each outcome or competency on the rubric", which ensures an explicitly definable "difference between an advanced and a proficient performance" (Schreiber et al., 2012, p. 212). Rubrics are quantifiable since raters first score individual aspects of a student's performance and then calculate the average scores within each category and for the entire performance.

Rubrics design has to align with the rater's needs in the assessment of their students' performance. For a basic rubric, the rater first defines the categories they want to evaluate. These are then aligned along several different scores. The categories depend on the expectations and components of an assignment. For instance, if the students' formal oral performance is assessed, the categories focus

on fluency, clarity, articulation, sentence structure, grammar, etc. If the students' communicative competence is evaluated (for instance, in EAP), the categories are more specific, such as the introduction of the topic, organisation of thoughts, providing examples, verbal delivery, etc. In either case, objectivity will be achieved if the rater defines distinguishable descriptions of the components at each expected level, i.e., relies on descriptive or analytic rubrics (Fulcher, 2012, 2017; Schreiber et al., 2012). During the assessment, the rater allocates the most appropriate descriptor and score to each predefined category in the rubrics sheet.

In case more detailed items within a category are needed, the items are categorised into groupings, which reflect various competencies expected from the students. An example of such detailed rubrics is the Public Speaking Competence Rubric (PSCR) developed by Schreiber et al. (2012) for assessment in the communication discipline. The PSCR assessed a total of 11 dimensions or outcomes. Each competency was measured on a 5-point scale. The authors additionally selected five performance levels, with corresponding scores and descriptors, to provide more precision in the feedback to their students. They arranged the performance levels from best to substandard (i.e., advanced, proficient, basic, minimal and deficient).

However, rubrics are not a magic tool. Their reliability has often been challenged because raters may still be guided by their overall impression of their students' performance despite carefully designed criteria (Knoch, 2009). When rubrics are given to the students before they complete an assignment, student performance can be increased significantly as the rubrics serve as a set of instructions with clarified expectations and components of the assignment. In that way, students are more aware of their learning process and progress and improve their work through timely and detailed feedback (Stevens & Levi, 2023).

2. MATERIALS AND METHODS

Validity is a fundamental topic in large-scale language testing. Its purpose is to show "how logical and true interpretations and decisions are made based on scores (or in general data) from assessments" (Giraldo, 2020, p. 195). In other words, a test is valid if it measures what it has to measure and nothing more (Brown & Abeywickrama, 2010). Stakeholders in testing systems have to rely on validity because otherwise, the test system would be useless.

However, in classroom language testing, interpretations and decisions based on curriculum objectives and learning outcomes predicted in the syllabi are also crucial, if not more important (Messick, 1989). Giraldo (2020) argues that validity for classroom testing should be a modification of the definitions proposed by the American Educational Research Association (AERA), American Psychological Association and National Council on Measurement in Education (2014) suggesting that validity "in classroom language testing depends on how appropriate

interpretations and decisions are, based on the data from instruments used to activate the relevant language skills stated in a curriculum" (Giraldo, 2020, p. 197). Since validity is an abstract concept, Giraldo (2020) proposes that teachers make it practical by validating the tests they use for accurate interpretations and decisions.

Several other studies support and highlight the concept that the applicability of testing approaches in language education extends beyond statistical evidence. For instance, Brunfaut (2023) argues that data collection procedures must be adapted to the research questions, needs and priorities imposed by the actual research context and Im et al. (2019) suggest that effective validation requires diverse evidence and stakeholder involvement, thus moving beyond mere statistical analysis. Norouzian et al. (2019) show that reliance on *p*-values can be misleading and Mobärg (1997) argues that vocabulary testing should align with pedagogical approaches, indicating that statistical methods may be more appropriate in structural contexts, while teaching-based assessments are better suited for lexical contexts. Kilgarriff (2005) states that statistical evidence alone is not enough to prove the applicability of a testing approach in language education because language use is inherently non-random and does not follow a random distribution.

Collectively, these studies advocate for a multifaceted approach to validation in language education which is why the action-based case study presented here follows Giraldo's suggestion to use statistical evidence as a confirmation of how appropriate interpretations and decisions are, based on the data from instruments used to activate the relevant language skills stated in a curriculum (Giraldo, 2020). In other words, the instructor, i.e., researcher who conducted this study has made validity practical by validating the rubrics used for accurate interpretations and decisions related to real classroom conditions based on relevant statistical instruments (Giraldo, 2020).

2.1. Participants

Generalizability in this small-scale language research was not the primary goal, so random participant selection and random assignment were not a precondition (Turner, 2014). Therefore, over three years, a convenience sample of 93 university students (cohort 2019, cohort 2020 and cohort 2021 enrolled in the Serbian language and the German language departments) relied on digital video creation (DVC) with mobile (smart) phones in remote classes during the final semester of their four-semester English courses to practise their EAP oral performance (on the implementation of DVC in language teaching, see Alley-Young, 2017; Hafner & Miller, 2011; Han & Yi, 2019). The participants were duly informed and signed consent forms. Each year, DVC was introduced in the final semester of the four-semester English language course (second year of studies). DVC was meant to provide the students with an additional tool which would motivate them to practise their speaking skills for their final oral exam while helping them gain more confidence despite the remote context imposed by COVID-19.

The course syllabi predict that the EAP aspect is covered during all four semesters and it is based on general academic vocabulary (Hulme, 2021). The students are assigned writing tasks (e.g., essays, reports, analyses) and speaking tasks (discussions, project work, collaborative research, PPT presentations). They rely on material supplied by the language instructor but are also encouraged to search for sources independently.

2.2. Procedure

The rubrics designed for this case study were implemented for the assessment of the students' oral performance in three different instances, i.e., in May 2020, April 2021 and May 2022, as well as in one additional assessment in June 2021 (it will be presented in the section Results that this additional assessment was introduced to provide reliability evidence). In May 2020 (after the COVID-19 pandemic had started in March), the language instructor (the author of this article) asked the students to prepare a talk (relying on the benefits presented in Harmer, 2015) and record themselves in the form of a digital video while delivering the prepared talk. As online classes continued in 2021 and 2022, the practice of speaking was still limited, so the video recordings were introduced again in 2021 and 2022 to motivate the students to practise speaking. To ensure that all students used their mobile phones for the recordings, along with the recording they submitted in the Google Classroom, the students submitted screenshots of their phone display presenting the recording in the respective folder in their phones. The students were offered extra points to be added to their final grade as an incentive.

The instructions for the videos included the length of the video, the title of the talk, the structure, suggested resources, the submission deadline and the rubrics. The title of the talk was "A problem that concerns me very much is…" The students could choose any topic which would be a logical continuation of the proposed title. Some of their choices were pollution, exam anxiety, suicide, online teaching, domestic violence, etc. The students were expected to rely on the academic register used in the course, conduct relevant research and cite the resources at the end of their recordings.

2.3. Assessment

Apart from their purpose of measuring the students' oral performance as part of the continuous assessment conducted throughout the semester, the rubrics designed for this study served as a preparation tool based on which the students could organise and structure their speeches for the recordings. The CEFR recommendation (Council of Europe, 2001) was taken into account regarding the number of categories in a rubric (the recommendation is up to six categories) (Table 1).

CATEGORIES	1	2	3	4
Introduction of the topic	Provides a poor statement of what the concern is in one sentence only. The choice of vocabulary suggests limited knowledge of the topic.	Provides a poor statement of what the concern is in two sentences but they lack coherence. The choice of vocabulary suggests some knowledge of the topic.	Provides a fairly clear statement of what the concern is in about two sentences. The coherence is acceptable. The choice of vocabulary suggests a good command of the topic.	Provides a clear statement of what the concern is in two or more sentences. The sentences are coherent. The choice of vocabulary suggests a solid command of the topic.
Background details (scientific facts, evidence, examples, etc.)	Fails to provide background details or provides only one. There is no proper link to the topic.	Provides more than one background detail but does not link them properly to the topic. The presentation of the concern is rather weak.	Provides several background details but links them only loosely to the topic. The concern stated in the introduction is presented in a more or less solid way.	Provides several background details and links them properly to the topic giving thus a solid presentation of the concern stated in the introduction.
Presentation of the personal reasons	Fails to mention the personal reasons or provides only one.	Provides more than one personal reason for the concern but does not explain it properly.	Provides more than one personal reason for the concern, the explanation is proper but lacks an argumentation.	Provides more than one personal reason for the concern and the explanation of the reasons is based on a solid and proper argumentation.
Reference to the consequences in the future	The reference to the consequences in the future is not clearly stated.	The reference to the consequences in the future is supported by vague arguments.	The reference to the consequences in the future is explicitly stated but not providing convincing arguments.	The reference to the consequences in the future is clearly and explicitly stated based on a convincing argumentation.
Conclusion	The conclusion is expressed in a single sentence without summarizing the main points and purpose of the speech.	The conclusion is expressed in more than one sentence but fails to summarize the main points and purpose of the speech.	The conclusion is expressed in one or more than one sentence, summarizes the main points and purpose of the speech but fails to leave a lasting impression on the audience.	The conclusion is expressed in more than one sentence, summarizes clearly the main points and purpose of the speech and leaves a lasting impression on the audience by including an effective final remark.
Language (general impression based on vocabulary, grammar, fluency and coherence)	weak	satisfactory	good	very good

Table 1. The assessment rubrics used for the students' videos

Apart from the basic CEFR recommendation, the rubrics used in this research relied on several other criteria. Given that the communicative aspect of the task is the

primary focus of the evaluation based on the rubrics, an important recommendation to adhere to is Fulcher's (2017) suggestion to contextualise test goals, tasks and rating criteria in real-world language. This criterion was secured based on the fact that the rubrics included precise descriptors based on the vocabulary developed during the four-semester course and aligned with the EAP objectives and outcomes predicted in the course syllabi. Furthermore, Dawson (2017) stated that combining task-specific and teacher-created rubrics yields the best results. That is why the rubrics were descriptive and analytic, with clearly defined descriptors that measured the degree of accomplishment for each category on a scale from 1 to 4 (Fulcher, 2012, 2017; Schreiber et al., 2012). The categories were determined to meet the needs for a specific evaluation as determined by Fulcher (2012, 2017) and Schreiber et al. (2012) including aspects, such as introduction of the topic, organisation of thoughts, providing examples, verbal delivery, etc. To meet all the requirements outlined here, the rubrics included six categories: 1) Introduction of the topic (clear statement of what the concern was); 2) Background details (e.g., scientific facts, evidence, examples, etc.); 3) Detailed presentation of the personal reasons for the concern; 4) Reference to the consequences in the future; 5) Conclusion and 6) Language (see Table 1). To achieve objectivity, clearly distinguishable descriptions of the components at each expected level were defined (Fulcher, 2017). Following Dawson's (2017) recommendations for task-specific rubrics, the expected levels of accomplishment were based on the EAP learning goals, tasks and outcomes predicted in the course syllabi for each semester.

In addition, the rubrics evaluated the students' speech contextualised in real-world language (Fulcher, 2017) while including EAP vocabulary covered in the course. Since the primary purpose of the rubrics was to translate the observed performance to a score with real-world meaning, the focus was less on formal aspects of language, including grammar and syntax, as those were evaluated in the written part of the exams. In other words, the aim was to establish and assess the alignment between scale criteria and language use (Fulcher, 2012). That is why the students were explicitly told not to worry too much about language accuracy, as the primary assessment criterion would be their overall communicative competence as defined by Whyte (2019) who argues for "an expanded view of communicative competence which is more faithful to Hymes' (1972) original conception" (2019, p. 17). Whyte (2019) further states that communicative competence in Language for Specific Purposes includes three main types of knowledge (linguistic, pragmatic and content) and three main strategies for use (discourse, interaction and performance) while Hyland (2022) states that ESP is concerned with "communicative practices rather than more narrowly with specific aspects of language" (2022, p. 212) which is what this research relied on. Finally, the rubrics were teacher-created (Dawson, 2017) as the instructor carefully designed the expectation levels within each descriptor to facilitate an objective assessment of all the talks the students would deliver.

Regarding the issue of rater bias, large-scale language testing systems rely on different measures mitigating rater bias and rater variability. These measures

primarily include complex statistical measures and calibration (Rossi & Brunfaut, 2020) which are not always possible in everyday classroom conditions, nor are they needed. Factors such as rater/teacher language proficiency, classroom experience and knowledge about language assessment can significantly help mitigate bias (Kubota, 2018; Lumley, 2005).

The study presented here included a convenience sample attending English language classes at the Serbian public university where the study was conducted. Teaching staff recruitment procedures at the university follow the Law on Higher Education in Serbia (Zakon o visokom obrazovanju, 2023) according to which only a person with an appropriate professional, academic, scientific, i.e., artistic title acquired at an accredited study programme and an accredited higher education institution, as well as the teaching capacity, may be elected to the position of a teacher. In other words, every higher education language teacher in Serbia has the necessary experience and knowledge about language assessment (factors proposed by Kubota, 2018 and Lumley, 2005) because if they did not, they could neither teach nor assess. The instructor/first rater and author of this article has been a teacher for 30 years and the second rater (later included in the research, Section 3.3.) for 31 years. Both raters have had high scores (~ 4.5/5) in the categories of fairness and objectivity in the bi-annual student evaluations, an obligatory quality assurance criterion stipulated by the Bologna process in Serbia. Therefore, rater bias can be excluded as an issue influencing the collected data.

2.4. Data analysis

When referring to validity in the assessment of speaking, probably the most important type to consider is criterion-oriented validity (for the sake of simplicity, this term will be referred to as validity), as it helps the tester identify the "relationship between a particular test and a criterion to which we wish to make predictions" (Fulcher & Davidson, 2007, p. 4) or "the extent to which the 'criterion' of the test has been reached" (Brown, 2019, p. 24). It evaluates how well a test can predict a concrete outcome and is considered a key issue in any assessment. The criterion of validity of the rubrics in this study was established by comparing the students' scores based on the implemented rubrics with the number of points the students achieved in the oral part of their final exam during the last year of the study in 2022 (based on interview questions and allocated 30 out of 100 points). To ensure the internal validity of rubrics in all three instances measuring the students' oral performance (May 2020, April 2021 and May 2022), the same tool had to provide accurate results in three different contexts. Therefore, the descriptors within each category had to be as precise as possible, i.e., the same vocabulary within each descriptor and the same categories were supposed to be applicable for the evaluation of speaking in three contexts.

Regarding reliability, this study relied on the idea that "[w]henever a test is administered, the test user would like some assurance that the results could be replicated if the same individuals were tested again under similar circumstances. This desired consistency (or reproducibility) of test scores is called reliability" (Crocker & Algina, 1986, p. 105). Reliability can be seen as the degree to which an assessment tool can produce stable and consistent results. If paired with validity, reliability is another significant criterion indicating the value of an assessment tool. As Fulcher and Davidson (2007) pointed out, "it has always been stated in the language testing literature that without reliability there could be no validity" (2007, p. 31).

Finally, when evaluating the aspect of replicability, this study was based on Turner's (2014) viewpoint that in language education research, especially research with small samples where the research is conducted primarily in real classrooms, the notion of replicability is viewed differently than in controlled laboratory settings. This means that in language research, data collection procedures must be adapted to the research questions, needs and priorities imposed by the actual research context (Brunfaut, 2023). Therefore, replicability is not meant to provide evidence for the consistent application of criteria but to provide elements based on which the thoroughness of an original report can be checked. It is "an avenue for building knowledge through investigating the inevitable differences that occur in recreating an experiment in a non-controlled setting" (Turner, 2014, p. 29). Fulcher and Davidson (2007) assumed that "for a score to be meaningful and interpretable, the sum of the parts should be reproducible" (2007, p. 104), which supports the "desired consistency (or reproducibility) of test scores" that Crocker and Algina (1986, p. 105) pointed out as a precondition for reliability. Finally, according to Knoch et al. (2021), score generalizability refers to the extent to which the scores obtained from a rating scale can be generalised to other tasks, raters or contexts, which presupposes that the rating scale is valid and representative of the assessment construct across different tasks, raters and contexts. In that way, replicability will confirm validity.

When using rubrics to assess EAP speaking in a MALL-supported context, the criterion of replicability may be challenging to meet. Students cannot be expected to deliver the same performance on different occasions because the exact conditions of a previous study may not be replicated (Turner, 2014). Even if students memorised and repeated the exact words, the overall performances would not be the same (intonation, speed, pronunciation, etc. would differ). Thus, when a study in language education research is repeated, the same outcomes are not expected (Turner, 2014). Instead, the researcher anticipates "that the subtle or deliberate differences in the conditions of the original research environment and the new research environment will result in clearer outcomes" or, if that is not the case, in outcomes "that add to our collective knowledge by demonstrating how the differences impact the outcomes" (Turner, 2014, p. 13).

Fulcher and Davidson (2007) suggested that reliability estimates in language assessment are based on four assumptions: stability, discrimination, test length and homogeneity. In other words, these assumptions indicate that student abilities will not change dramatically over short periods, that the test will show whether an

individual student has achieved the planned outcome and that several "pieces of evidence need to be collected to ensure that a 'score' adequately shows what a test taker knows or can do" and that "the picture that the teacher creates of the learner is therefore very different and multifaceted" (Fulcher & Davidson, 2007, p. 31). If rubrics are used as an assessment tool in a MALL context to assess EAP speaking, the best measure to confirm reliability is test-retest reliability, or the administering of the same assessment tool twice over some time to the same group of individuals which is what was done within the study presented here.

3. RESULTS

3.1. The validity of rubrics

The rubrics were first applied in May 2020. The same rubrics were employed for a second time in April 2021 with a new group of students. The scores were based on a scale from 1 to 4 (the lowest being 1 and the highest 4). The scores were allocated relying on the rubrics in the respective final semesters, and when compared in three different contexts, the sum scores showed similar achievement levels (Table 2).

Value	**May 2020**	**April 2021**	**May 2022**
Mean	3.521	3.498	3.504
Shapiro-Wilk W	0.738	0.775	0.762
N	31	31	31

Table 2. Student scores testing the validity of rubrics

As Table 2 shows, the Shapiro-Wilk Test indicated a normal distribution for all three assessments (0.738, 0.775 and 0.762, respectively). The criterion of validity of the rubrics in this study was established based on the fact that in 2022, the students who participated in this study achieved an average of 23 points (out of 30) in the oral part of their final exam, which approximately corresponds to the average score of 3.504 (on a scale from 1 to 4) all the students achieved based on the rubrics during the study. The students' scores in the three instances are quite high (between 3 and 4), indicating similar results in the three groups with only little differences (3.521 vs. 3.498 vs. 3.504).

A Pearson's correlation was computed to test the validity of the rubrics in the three different contexts (Table 3).

	May 2020	April 2021	May 2022
May 2020	1	.686**	.514**
April 2021	.686**	1	.885**
May 2022	.514**	.885**	1

**. Correlation is significant at the 0.01 level (2-tailed).

Table 3. Pearson's correlation testing the validity of rubrics

3.2. The reliability of rubrics

Some students whose speaking skills were measured in April 2021 were assessed based on the same rubrics in their final oral exam in June 2021 (the additional assessment mentioned in the section Materials and Methods). The repeated use of the rubrics with 11 students from the group of students who used the rubrics in April 2021 provided the researcher with data enabling the comparison between their oral performance presented in two different contexts, once in the mid-term in April 2021 and a second time in the final oral exam in June 2021 (Table 4).

Value	**April 2021**	**June 2021**
Mean	3.643	3.592
N	11	11

Table 4. Student scores in the repeated measuring testing the reliability of rubrics

The results in Table 4 indicate that the 11 students who used the rubrics twice had a slightly weaker score in June 2021 (3.592) compared to April 2021 (3.643). That can be attributed to the fact that in June, they were preparing for an exam, which might have made them a little more nervous.

For the sake of confirmation, Pearson's correlation was computed for these two occasions, confirming that the correlation is significant ($p < 0.001$) (Table 5).

	April 2021	**June 2021**
April 2021	1	.967***
June 2021	.967***	1

***. Correlation is significant at the 0.001 level (2-tailed).

Table 5. Pearson's correlation testing the reliability of rubrics

As an additional precaution, the instructor asked a colleague to assess the videos of those 11 students who opted for the recordings in June 2021. Upon comparing their assessments, the two raters could establish that they had given identical average scores to 8 out of 11 students (Table 6).

Value	Rater 1 in June 2021	Rater 2 June 2021
Mean	3.592	3.613
N	11	11

Table 6. Student scores in the additional assessment performed by a second rater

The additional Pearson's correlation confirmed one more time that the correlation between the measurements performed by the two raters is significant ($p < 0.001$) (Table 7).

	Rater 1 in June 2021	Rater 2 June 2021
Rater 1 in June 2021	1	.947***
Rater 2 June 2021	.947***	1

***. Correlation is significant at the 0.001 level (2-tailed).

Table 7. Pearson's correlation testing the reliability of rubrics

To add further evidence, the scale reliability of the rubrics was tested based on Cronbach's Alpha (Table 8) to confirm internal consistency reliability, i.e., the degree to which items within a test are consistent in measuring the same construct (Fulcher & Davidson, 2007).

Measurement model	Number of items	Threshold	Cronbach's α
Category	6	.70	0.983

Table 8. Cronbach's Alpha testing the validity of rubrics

The results in Table 8 suggest a strong α coefficient of 0.983, which confirms the validity of rubrics according to significant scale reliability. The outcome is that the same rubrics in 2021 yielded similar data on repeated occasions, which proved that the rubrics met the reliability criterion as well.

3.3. The replicability of rubrics

A second assessor used the same rubrics to assess their students' oral performances in 2022, asking them to create and submit videos as homework assignments based on the same criteria and tasks. The comparison of the assessments provided insight into "the inevitable differences that occur in recreating an experiment in a non-controlled setting (Turner, 2014, p. 29) (Table 9).

Value	Rater 1/Group 1 2022	Rater 2/Group 2 2022
Mean	3.601	3.640
N	31	31

Table 9. Student scores testing the replicability of rubrics

An additional inter-rater reliability assessment was applied based on Cohen's Kappa coefficient (Table 10).

Statistics	Value
Cohen's Kappa	0.791

Table 10. Inter-rater reliability assessment

As indicated in Table 10, Cohen's Kappa coefficient of 0.791 indicates a strong level of agreement between the two raters in their assessments. This high level of reliability suggests that the ratings are consistent thereby enhancing the confidence in the results obtained. In addition, as stated in Section 2.3., inter-rater reliability, as well as the criterion of objectivity and fairness were also secured based on the fact that both raters have 30 years of experience in both teaching and assessment, that they teach the same course, follow the same syllabus and apply the same assessment criteria. The comparison of the results revealed that in both groups, the performances were very good, as seen in Table 9 (3.601 compared to 3.640). The conclusion is that the rubrics presented here successfully met the criterion of replicability as well.

4. DISCUSSION

As presented, the rubrics were first applied in May 2020, and then the same rubrics were employed for a second time in April 2021 with a new group of students. In April 2021, the students were given the alternative of creating videos with their mobile phones for their mid-term test instead of the usual oral presentation in front of the entire class. The student scores showed no considerable difference which can be justified based on several factors. First, the underlying language proficiency the students exhibited in their overall mastery was more or less consistent. In addition, the assessments measured the same constructs, i.e., oral performance and both assessments relied on the same level of difficulty. Finally, the students' motivation was high as the context was changed in the sense that the entire situation was more relaxed, and the stress level that the students usually experience when preparing oral presentations was lower (Đorđević, 2020). They knew they would not have to do the presentation in class, which most students do not feel comfortable with. They knew they would record the videos at home, perfect them and submit them when ready, which confirms the findings presented by Sun et al. (2017), whose research showed that MALL provided a comfortable atmosphere for learners to express their thoughts in oral communication. All these factors contributed that the scores reflect a more or less stable level of mastery which is a strong indicator of the scores being a reliable indicator of their speaking skill. The rubrics were used for a third time in May 2022 with a third group of students who were asked to create videos with their mobile phones as a homework assignment for extra points. The students were promised extra points as an additional incentive. The results confirm that the rubrics implemented in this study comply with the suggestions put forward by Chen and Lin (2023), Dawson (2017), Fulcher (2017) and Schreiber et al. (2012), thereby confirming that rubrics in MALL, if relying on user-centred and interactive features while including a variety of tasks, can target more complex skills, in particular skills needed for oral presentations in an EAP context.

Another important fact is that the rubrics were used as a preparation tool and as an assessment tool. As a preparation tool, the rubrics helped the students prepare following the predefined expectations. They later reported that they liked the opportunity and felt very motivated, while the stress level was considerably lower because the rubrics told them what the teacher would focus on in the assessment. In that way, the study showed that the implemented rubrics served as a set of instructions with clarified expectations and components of the assignment, which helped the students be more aware of their learning process and progress, thereby leading to the improvement of their work (Stevens & Levi, 2023). As an assessment tool, the rubrics provided the researcher with valuable data about the spoken performance of each student who submitted a video. In this way, the implemented rubrics proved to be a valuable assessment tool based on performance-driven scales while constructed based on real-world data and academic corpora, which is what Knoch et al. (2021) insisted on in their re-evaluation of rating scales.

The decision to conduct the additional assessment in June 2021 imposed itself because, due to the COVID-19 pandemic, many students could not travel/come to the faculty to take their oral exam, which is why they were offered the alternative to do the videos again for their oral exam in June 2021 (faculty management approved these exceptions as the pandemic was raging during that period). The students who opted for the video recordings in June 2021 were given a new topic while the conditions and the rubrics were the same. About a third of the students chose to prepare videos one more time, whereas the rest opted for the regular oral exam at the faculty. One more time, the rubrics proved to be an adequate solution because they were task-specific and were applied to a specific instance of assessment in a particular context (Dawson, 2017).

In the case of the three students whom the two raters had assessed with different average scores in June 2021 (Table 6), the performance discrepancy resulted from the different scores allocated to the category “language”. Apart from indicating that rubrics are not a perfect data collection tool, the discrepancy proves that if a category is not defined with precise descriptors, the rubrics are prone to deliver unreliable data. In other words, if the alignment between scale criteria and real-world language is not strong enough, the observed performance cannot be translated to a score with real-world meaning (Fulcher, 2012). The category language was the most loosely defined category given that the descriptors did not include detailed descriptions but only single words: “weak”, “satisfactory”, “good”, and “very good”. Since the rubrics aimed to focus on the students’ use of real-world language leaning on a general academic register, the content of the speeches and its overall delivery (Fulcher, 2012), the final category, “language”, does not diminish the value of the rubrics. If language as a category were to be assessed more formally, the descriptors in the rubrics would have to be far more detailed (Fulcher, 2017).

Some limitations have to be mentioned. The so-called observer’s paradox (Labov, 1972) indicates that the participants in this study were aware that their recorded talks would be assessed by their language instructor, so they must have

tried hard to deliver the best talk they could do. Such a talk is far from being natural. The aspect of control in a MALL context is also problematic because the researcher cannot control a remote situation. In the contexts presented here, the students prepared their videos at home, where only they could control every aspect of the setting. In addition, the stimuli introduced to test the participants were artificial, meaning that the study settings cannot be qualified as producing natural speech. The students would probably not speak in real life as they did in their videos. It may also be argued that the reliability of rubrics in the setting presented here could not be determined. However, despite a small sample, the same rubrics yielded more or less similar results in the repeated measurements indicating a certain trend, which may be assumed to be confirmed with a larger sample.

5. CONCLUSION

The action-based case study presented here illustrates how speaking as a productive skill in EAP can be measured with quantitative data from rubrics as a rating scale in remote classes supported by MALL providing a numerical assessment and evaluation of the actual product – the student's speech. The findings confirm that descriptive, analytic, task-specific and teacher-created rubrics contextualised in real-world language meet the validity, reliability and replicability criteria.

The implications of the results presented in this study indicate that specifically designed rubrics, if employed in remote classes supported by MALL and aimed at assessing students' oral performance and their overall mastery of EAP speaking skills, will be as objective as in the examples presented here. In addition, the results confirm that rubrics can be an equally valuable assessment tool in other remote English learning contexts (e.g., ESP, EMI) where technology is an essential requirement. The results also show that rubrics must be designed carefully with specific categories and precisely determined descriptors. If they do not rely on obvious qualitative differences, they fail to provide nuances to be measured numerically. Therefore, more empirical research is needed to explore the applicability of rubrics as an assessment tool for EAP speaking in other more diverse settings entirely relying on MALL.

[Paper submitted 16 May 2024]
[Revised version received 19 Aug 2024]
[Revised version accepted for publication 26 Sep 2024]

References

Abugohar, M. A., Yunus, K., & Rashid, R. A. (2019). Smartphone applications as a teaching technique for enhancing tertiary learners' speaking skills: Perceptions and practices. *International Journal of Emerging Technologies in Learning*, *14*(9), 74–92. https://doi.org/10.3991/ijet.v14i09.10375

Alley-Young, G. (2017). Creating digital videos in an ESL learning community to develop communication skills and content area knowledge. In S. Pixy Ferris & H. Wilder (Eds.), *Unplugging the classroom: Teaching with technologies to promote students' lifelong learning* (13–35). Chandos Publishing. https://doi.org/10.1016/B978-0-08-102035-7.00002-3

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Psychological Association.

Azlan, N. A. B., Zakaria, S. B., & Yunus, M. M. (2019). Integrative task-based learning: Developing speaking skill and increase motivation via Instagram. *International Journal of Academic Research in Business and Social Sciences*, *9*(1), 620–636. https://doi.org/10.6007/ijarbss/v9-i1/5463

Bazhutina, M. M., & Tsepilova, A. V. (2024). The development of CEFR-based descriptors for assessing engineering students' integrative ESP competence. *ESP Today*, *12*(1), 93–117. https://doi.org/10.18485/esptoday.2024.12.1.5

Berger, A. (2020). Specifying progression in academic speaking: A keyword analysis of CEFR-based proficiency descriptors. *Language Assessment Quarterly*, *17*(1), 85–99. https://doi.org/10.1080/15434303.2019.1689981

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, *20*(1), 89–110. https://doi.org/10.1191/0265532203lt245oa

Brown, H. D. (2019). *Language assessment: Principles and classroom practices* (3rd ed.). Pearson.

Brown, H. D., & Abeywickrama, P. (2010). *Language assessment: Principles and classroom practices*. Pearson Longman.

Brunfaut, T. (2023). Future challenges and opportunities in language testing and assessment: Basic questions and principles at the forefront. *Language Testing*, *40*(1), 15–23. https://doi.org/10.1177/02655322221127896

Chen, M. Y., & Lin, Y.-M. (2023). Mobile-assisted language assessment for adult EFL learners: Recommendations from a systematic review. In S. W. Chong & H. Reinders (Eds.), *Innovation in learning-oriented language assessment* (237–256). Palgrave Macmillan. https://doi.org/10.1007/978-3-031-18950-0_14

Chen Hsieh, J. S., Huang, Y.-M., & Wu, W.-C. V. (2017). Technological acceptance of LINE in flipped EFL oral training. *Computers in Human Behavior*, *70*, 178–190. https://doi.org/10.1016/j.chb.2016.12.066

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment.* Council of Europe. https://rm.coe.int/16802fc1bf

Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment: Companion volume*. Council of Europe Publishing. https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/16809ea0d4

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart, and Winston.

Dawson, P. (2017). Assessment rubrics: Towards clearer and more replicable design, research and practice. *Assessment and Evaluation in Higher Education*, *42*(3), 347–360. https://doi.org/10.1080/02602938.2015.1111294

Đorđević, J. (2020). Improved understanding of meanings of modal verbs in Legal English and increased motivation through Computer Assisted Language Learning. *Ibérica*, *39*, 295–318. https://doi.org/10.17398/2340-2784.39.295

Fulcher, G. (2012). Scoring performance tests. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (378–392). Routledge.

Fulcher, G. (2017). Criteria for evaluating language quality. In E. Shohamy (Ed.), *Language testing and assessment* (179–192). Springer.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment*: *An advanced resource book*. Routledge.

García Laborda, J., Magal Royo, T., Litzler, M. F., & Giménez López, J. L. (2014). Mobile phones for Spain's university entrance examination language test. *Educational Technology & Society*, *17*(2), 17–30. https://drive.google.com/file/d/1-AH0sFUZSi-SJgj1tfewxysrZUwNC_LA/view

Giraldo, F. D. (2020). Validity and classroom language testing: A practical approach. *Colombian Applied Linguistics Journal*, *22*(2), 194–206. https://doi.org/10.14483/22487085.15998

Hafner, C. A., & Miller, L. (2011). Fostering learner autonomy in English for science: A collaborative digital video project in a technological learning environment. *Language Learning & Technology*, *15*(3), 68–86. https://www.lltjournal.org/item/10125-44263/

Han, S., & Yi, Y. J. (2019). How does the smartphone usage of college students affect academic performance? *Journal of Computer Assisted Learning*, *35*(1), 13–22. https://doi.org/10.1111/jcal.12306

Harmer, J. (2015). *The practice of English language teaching* (5th ed.). Pearson Education ESL.

Hulme, A. (2021). Uncovering the principles behind EAP programme design: Do we do what we say we're going to do? *ESP Today*, *9*(2), 206–228. https://doi.org/10.18485/esptoday.2021.9.2.2

Hyland, K. (2022). English for specific purposes: What is it and where is it taking us? *ESP Today*, *10*(2), 202–220. https://doi.org/10.18485/esptoday.2022.10.2.1

Hwang, G.-J., & Chang, S.-C. (2021). Facilitating knowledge construction in mobile learning contexts: A bi-directional peer-assessment approach. *British Journal of Educational Technology*, *52*(1), 337–357. https://doi.org/10.1111/bjet.13001

Hwang, W.-Y., & Chen, H. S. L. (2013). Users' familiar situational contexts facilitate the practice of EFL in elementary schools with mobile devices. *Computer Assisted Language Learning*, *26*(2), 101–125. https://doi.org/10.1080/09588221.2011.639783

Im, G.-H., Shin, D., & Cheng, L. (2019). Critical review of validation models and practices in language testing: Their limitations and future directions for validation research. *Language Testing in Asia*, *9*(14), 1–26. https://doi.org/10.1186/s40468-019-0089-4

Kilgarriff, A. (2005). Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory*, *1*(2), 263–276. https://doi.org/10.1515/cllt.2005.1.2.263

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing*, *26*(2), 275–304. https://doi.org/10.1177/0265532208101008

Knoch, U., Deygers, B., & Khamboonruang, A. (2021). Revisiting rating scale development for rater-mediated language performance assessments: Modelling construct and contextual choices made by scale developers. *Language Testing*, *38*(4), 602–626. https://doi.org/10.1177/0265532221994052

Kubota, K. (2018). The potential of empirically derived rating scales for inexperienced raters: A comparative study on rating scales. *JLTA Journal*, *21*, 141–159. https://doi.org/10.20622/jltajournal.21.0_141

Kukulska-Hulme, A. (2020). Mobile-assisted language learning [Revised and updated version]. In C. A. Chapelle (Ed.), *The concise encyclopedia of applied linguistics*. Wiley & Sons, Ltd. https://doi.org/10.1002/9781405198431.wbeal0768.pub2

Kukulska-Hulme, A. (2021). Reflections on research questions in mobile assisted language learning. *Journal of China Computer-Assisted Language Learning*, *1*(1), 28–46. https://doi.org/10.1515/jccall-2021-2002

Kukulska-Hulme, A., & Shield, L. (2008). An overview of mobile assisted language learning: From content delivery to supported collaboration and interaction. *ReCALL*, *20*(3), 271–289. https://doi.org/10.1017/S0958344008000335

Labov, W. (1972). Some principles of linguistic methodology. *Language in Society*, *1*(1), 97–120. https://doi.org/10.1017/S0047404500006576

Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. Peter Lang.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., 13–103). Macmillan.

Mobärg, M. (1997). Acquiring, teaching and testing vocabulary. *International Journal of Applied Linguistics*, *7*(2), 201–222. https://doi.org/10.1111/j.1473-4192.1997.tb00115.x

Muhammad, A. A., & Ockey, G. J. (2021). Upholding language assessment quality during the Covid-19 pandemic: Some final thoughts and questions. *Language Assessment Quarterly*, *18*(1), 51–55. https://doi.org/10.1080/15434303.2020.1867555

Norouzian, R., de Miranda, M., & Plonsky, L. (2019). A Bayesian approach to measuring evidence in L2 research: An empirical investigation. *The Modern Language Journal*, *103*(1), 248–261. https://doi.org/10.1111/modl.12543

Rezaee, A. A., Alavi, S. M., & Razzaghifard, P. (2020). Mobile-based dynamic assessment and the development of EFL students' oral fluency. *International Journal of Mobile Learning and Organisation*, *14*(4), 511–532. https://doi.org/10.1504/IJMLO.2020.110789

Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing*, *15*(1), 18–39. https://doi.org/10.1016/j.asw.2010.01.003

Rossi, O., & Brunfaut, T. (2020). Raters of subjectively-scored tests. In J. I. Liontas (Ed.), *The encyclopedia of English language teaching.* Wiley Online Library. https://doi.org/10.1002/9781118784235.eelt0985

Samaie, M., Mansouri Nejad, A., & Qaracholloo, M. (2018). An inquiry into the efficiency of WhatsApp for self-and peer-assessments of oral language proficiency. *British Journal of Educational Technology*, *49*(1), 111–126. https://doi.org/10.1111/bjet.12519

Schreiber, L. M., Paul, G. D., & Shibley, L. R. (2012). The development and test of the public speaking competence rubric. *Communication Education*, *61*(3), 205–233. https://doi.org/10.1080/03634523.2012.670709

Soo, J. Y. (2023). Test design and validity evidence of interactive speaking assessment in the era of emerging technologies. *Language Testing*, *40*(1), 54–60. https://doi.org/10.1177/02655322221126606

Stevens, D. D., & Levi, A. J. (2023). *Introduction to rubrics: An assessment tool to save grading time, convey effective feedback, and promote student learning.* Routledge. https://doi.org/10.4324/9781003445432

Sun, Z., Lin, C.-H., You, J., Shen, H. J., Qi, S., & Luo, L. (2017). Improving the English-speaking skills of young learners through mobile social networking. *Computer Assisted Language Learning*, *30*(3–4), 304–324. https://doi.org/10.1080/09588221.2017.1308384

Tarighat, S., & Khodabakhsh, S. (2016). Mobile-assisted language assessment: Assessing speaking. *Computers in Human Behavior*, *64*, 409–413. https://doi.org/10.1016/j.chb.2016.07.014

Turner, J. L. (2014). *Using statistics in small-scale language education research: Focus on non-parametric data.* Routledge. https://doi.org/10.4324/9780203526927

Whyte, S. (2019). Revisiting communicative competence in the teaching and assessment of language for specific purposes. *Language Education & Assessment*, *2*(1), 1–19. https://doi.org/10.29140/lea.v2n1.33

Wong, L.-H., & Looi, C.-K. (2010). Vocabulary learning by mobile-assisted authentic content creation and social meaning-making: Two case studies. *Journal of Computer Assisted Learning*, *26*(5), 421–433. https://doi.org/10.1111/j.1365-2729.2010.00357.x

Wong, L.-H., Chin, C.-K., Tan, C.-L., & Liu, M. (2010). Students' personal and social meaning making in a Chinese idiom mobile learning environment. *Educational Technology & Society*, *13*(4), 15–26. https://drive.google.com/file/d/1dGc-DPuodit_G9BcDpJv-b4KwwMtSaAd/view

Zakon o visokom obrazovanju [Law on Higher Education] (2023). https://www.paragraf.rs/propisi/zakon_o_visokom_obrazovanju.html

JASMINA ĐORĐEVIĆ is an Associate Professor at the Faculty of Philosophy, University of Niš, Serbia. Her research explores multimodal/digital media discourse, CALL and the Study of Translation. During the last two years, her work has focused on investigations of artificial intelligence and its impact on computer-mediated communication. Her most recent publications include the monograph *Digital Media Discourse in Linguistic Research* (2022) and several articles in high-impact journals.