Danijela Ljubojević[1]
Institute for Educational Research, Belgrade
Mirjana Daničić[2]
University of Belgrade, Faculty of Philology, Belgrade

Original scientific paper
UDC: 371.26:811.111'243
http://doi.org/10.5937/IstrPed2601022L

==========

ASSESSING ORAL PROFICIENCY IN YOUNG EFL LEARNERS: A STUDY OF EIGHTH-GRADE PUPILS' PERFORMANCE AT THE SERBIAN NATIONAL ENGLISH COMPETITION[3]

Abstract: The present study examines construct and predictive validity of oral assessment used at the 2024 Serbian National Competition in the English language for eighth-grade pupils. The study descriptively and critically evaluates the assessment of oral proficiency among top EFL learners at the end of compulsory education. A total of seventy-four competitors participated in the final speaking exam. Their performances were assessed holistically and analytically across four criteria: coherence and interaction, lexical accuracy, grammatical accuracy, and pronunciation. Descriptive statistics and correlational analysis were employed to compare holistic and analytic scores and to identify which component most strongly influenced overall ratings. The findings indicate that most candidates achieved scores between 17 and 20 points, suggesting a potential ceiling effect that may limit discrimination among top-level performances. The strongest predictors of the overall speaking score were coherence, interaction and lexical accuracy, whereas grammar and pronunciation played a secondary, though still meaningful, role. As the written test components showed a very loose relation with speaking performance (indicating weak predictive validity of the written exam for oral proficiency), the results of the study call for the refinement of performance descriptors and rubrics to enhance construct validity and differentiation of proficiency levels. These findings suggest that the speaking test functions as a reliable global measure of advanced oral proficiency and adds distinct value to the competition's assessment system, but that the analytic criteria and written components may require certain refinement if finer-grained distinctions among top-level performers are desired. The study contributes empirical evidence to the discussion of nationwide standards for a reliable EFL oral proficiency assessment.

Keywords: EFL assessment, English language competition, oral proficiency, performance descriptors, reliability, validity.

Introduction

Pupils in Serbia learning English as a foreign language are typically immersed in an oral rather than a written culture. From early childhood, they are exposed to English through songs, nursery rhymes, cartoons, films, and other audiovisual media. Most begin learning English in pre-school, where instruction emphasizes listening and speaking activities. Formal learning begins in the first grade of elementary school (age 7) and continues to the eighth grade, the final year of compulsory education. Speaking, as a fundamental linguistic skill, is initially developed for grammatical and lexical accuracy and later for communicative effectiveness. As Slobin (1996) notes, language use involves "thinking for speaking", highlighting the cognitive dimension of oral

[1] danijela.ljubojevic@ipi.ac.rs; https://orcid.org/0000-0002-4337-4884
[2] mirjana.danicic@fil.bg.ac.rs; https://orcid.org/0000-0001-9152-0877
[3] This research was funded by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia (Contract No. 451-03-33/2026-03/ 200018).

communication. Accordingly, communication and pragmatic competence become focal objectives in the upper elementary grades. Oral proficiency is therefore routinely assessed separately from other language skills, not only to evaluate school achievement but also in nationwide competitions. The particular relevance of assessing oral proficiency among eighth-grade pupils at the National English Language Competition lies in the fact that, for the elementary school pupils, this is the only round of the competition that includes oral component i.e. the Speaking part of the test. In the previous three rounds (school competition, municipal competition, city competition / regional competition), the test has only the written format. Moreover, not all pupils who advance to the national level of the competition are asked to take the Speaking part. Only those who attain a predetermined point threshold in the written part are invited to do so.

This study is conducted after analysing the data collected at the Serbian National Competition in the English language in April 2024[4] when pupils who achieved the first, second or third place at the regional competitions were invited to compete at the national level. Initially, there were 1,158 eight-graders qualified for the finals. They came to the Faculty of Philology in Belgrade to take part in the National Competition organized by the Society of Foreign Languages and Literatures (SFLL). Furthermore, at this highest level of competing, not all the competitors earned the right to take the Speaking part. Only those pupils who achieved the minimum required number of points in each part of the written test were invited to attend the Speaking part which measures their oral performance. A certain threshold of points in the written test is required to qualify for the oral part (the final part of the assessment), in this way ensuring that only top pupils will reach the ultimate competition level: 25 points in the grammar test (out of 30), 6 points in the reading comprehension test (out of 8), and 5 points in the listening comprehension test (out of 7). In the 2024 competition finals, the total number of pupils who qualified for the oral test was 77.

This study focuses on the process and instruments of oral performance assessment, trying to offer insight into its validity and credibility and possibly form some guidance for making nationwide standards for EFL oral assessment. In line with this, the study addresses the following research questions:

> **RQ1** Construct validity: What is the internal structure of the speaking test used in the National English Competition, and to what extent do the analytic criteria represent distinct aspects of oral proficiency?
> **RQ2** Predictive validity: To what extent do learners' scores on the written components of the competition (reading, listening, and grammar) predict their performance on the speaking test?

By answering these questions, the study seeks to generate evidence that can inform the refinement of oral assessment procedures and rating scales in nationwide EFL competitions.

Literature Review

Research on the use of oral proficiency language tests for young learners, both in first and second language, has not been scarce. However, previous research on assessing oral proficiency in English has been limited in providing practical suggestions and guidelines for oral testing of L2 young learners. A few of the key references which put this study in context are described below. Grounded on Lin's (2022) definition of oral proficiency as "L2 learner's ability to speak their second language to ensure communicative objectives in real life settings", we perceive oral proficiency in English as a pupil's ability to understand spoken output and to speak English not

[4] At the time of conducting the study and writing the paper, these are the most recent data available to the authors as the competition was cancelled in 2025 due to turbulent social circumstances which resulted in interrupted teaching process in a number of schools across the country.

only in the classroom but beyond it. Secondly, our purpose to investigate validity and reliability of oral proficiency assessment lies upon the desire to check if the assessment measures what it is supposed to measure. We perceive validity as "evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (Messick, 1989, p. 13). Three aspects are particularly relevant in the present context. First, content validity refers to the alignment between the oral tasks and the curricular and CEFR-based descriptors for the end of compulsory education. Second, construct validity concerns the internal structure of the analytic scale and the extent to which the four criteria (coherence and interaction, lexical accuracy, grammatical accuracy, and pronunciation) represent distinguishable dimensions of oral proficiency. Third, criterion-related validity is explored through the relationship between oral scores and performance on the written components of the competition (reading, listening, and grammar). We perceive reliability as another crucial element of assessment. Having acknowledged that it could easily be compromised "when judges are not properly trained or calibrated for specific scoring tasks, or not adhering to scoring protocols" (Elizondo-Gonzalez, 2025, p. 471), the authors of this study conceptualise reliability as the consistency of scores across raters and rating criteria. In this study, inter-rater reliability (IRR) is examined at the level of examination boards, while internal consistency is explored through the correlations among analytic scores and their relation with the overall speaking score. Lastly, in this paper we investigate the oral achievement of eighth-graders and for the purpose of this research we tend to call them young EFL learners. More precisely, they are 13–14 years old. It is not easy to synthesise all age-related research on second-language acquisition as different angles of interest are applied in various studies. Nevertheless, we here use the term to refer to the legal fact that they are younger than eighteen (hence not adult) and that they are at the end of their compulsory education which means that they have acquired their language skills throughout the elementary-school education. We use it rather as a generic than scholarly term, very aware of the need to make a distinction of the various age groups in foreign-language acquisition (Ellis, 2014).

It may also be important to explain some of the key words we use in the study. When analysing the performance descriptors and checking their validity, we use the terms of holistic grade and analytic grade. The holistic assessment assigns a single score to the tasks and focuses on the overall accomplishment of the tasks (Brown, 2004) whereas the analytic assigns a separate score for each tested language component. Building on this context, the present study focuses on the validity and reliability of the oral component of the National Competition in English. More specifically, it examines both the internal structure of the analytic rating scale and the relationship between written and oral performance among top-achieving 8th-grade learners.

Theoretical background

Fulcher, Davidson and Kemp (2011) distinguish two dominant traditions in developing rating scales for speaking tests. The measurement-driven approach constructs scales by ordering descriptors through statistical modelling, with meaning derived from the psychometric hierarchy rather than from observable performance. While efficient and generalisable, this approach often yields abstract descriptors that are weakly connected to actual language use and may oversimplify the construct of speaking. In contrast, the performance data–based method begins with detailed observations of real test-taker performance i.e. transcribing samples, analysing discourse, and deriving descriptors directly from linguistic behaviours. This approach produces richer, context-sensitive descriptions of interactional competence, grounding the scale in authentic performance features. For the purposes of the present study, which seeks to examine construct validity and understand how analytic dimensions of the Serbian National Competition rubric reflect underlying speaking constructs, the performance data–based perspective is more appropriate. It allows us to evaluate whether the existing analytic categories (coherence, lexical accuracy, grammar,

pronunciation) genuinely represent the competencies observed in candidates' oral output, and whether the scale captures the complexity of communicative performance rather than relying on abstract psychometric ordering.

The form of the oral test in the final round of the national competition in English is based on the findings of many previous studies (Ferrari, 2012; Michel, 2011; Tavakoli, 2016 etc.) which have shown that L2 learners tend to speak more fluently in dialogic tasks than in monologic. At the Speaking part of the test, the pupil faces a three-member examination panel: one interlocutor (interviewer) and two examiners. This structure enables the interviewer to focus on guiding the pupil to use as much language as possible in a short time. However, the interlocutor's role is double – to interview the pupil and to assess their performance. The pupil is talking only to the interlocutor (interviewer) while examiners sit aside (unlike the interviewer who is sitting face-to-face to the pupil). The ten-minute-long oral examination consists of three parts which are set in an interview format. The oral test begins with a very brief description of what the exam looks like and a short warm-up conversation with the interlocutor on topics of general interest but the pupil is often asked to respond to a personal question (so as to discourage answers "I don't know", "I'm not sure"). The topics are grouped into three umbrella categories – people, things, places, each branching out into various prompts for questions the interviewer asks (e.g. best friend, favourite teacher, a beloved family member, or favourite school subject, pastime, hobby, pets, sport, film, music, or school, home, holidays). In this phase the pupil's communication competence is evaluated, as well as ability to produce coherent spoken output. In the second stage, the pupil is given a picture and asked to describe it and talk about it. If it turns out to be necessary, the interlocutor provides back-up prompts to the pupils. It is expected that in this phase the pupil shows their fluency, range of vocabulary, knowledge of syntax and grammar as they are given liberty to perceive and interpret the picture in their own ways. In the (last) third phase, the interlocutor asks a couple of questions which are loosely related to the picture. The aim of the questions is to prompt pupils to give reasons for their opinions and to justify and explain them so that their pragmatic competence can be assessed. Nevertheless, the questions are formed in the way not to intimidate the pupil but to encourage them to talk and to give to the assessors an accurate picture of oral ability. The pupils are examined in a non-stressful, friendly way with the goal to show what they can do in/ with a spoken language. The interviewer's role is to mitigate pupil's anxiety to the fullest extent possible.

When the pupil has finished their oral exam and left the room, the examiners and the interlocutor take five minutes to measure the pupil's achievement and produce a final evaluation – first individually and then as a panel of examiners, they provide the final score on a one-to-twenty-point grading scale. In this way, they take the holistic approach to assessment along with the objective criteria. Examiners and interlocutor use assessment sheets to take notes on the pupil's performance. A large body of data is entered into the sheet since the assessment is divided into four main segments – coherence and interaction (which can bring to the pupil the maximum of 6 points), lexical accuracy (the same maximum number of points), grammatical accuracy (which can bring to the pupil the maximum of 5 points), and pronunciation (which can bring the maximum of 3 points). The maximum number of points per evaluator is 20. Each segment of the assessment sheet requires the assessor to enter the instances and details of positive (+) and negative (–) performance. The performance descriptors used by the assessors are descriptive in nature and targeted at evaluating individual pupil's skills in EFL: speaking, listening, pronunciation, core language skills (grammar and vocabulary), communication and pragmatic competence. Communication competence is here understood as the ability to use communication strategies and to overcome communication problems. Pragmatic competence is tested through creating communication situations that would challenge the pupil to employ a range of speech acts and vocabulary. The descriptors list what a successful candidate can do in the said four categories (grammar, vocabulary, communication competence and pragmatic competence). For instance, the grammar descriptors say that the successful candidate can "use Present Simple to describe

the picture", "use Present Continuous to say what people in the picture are doing" etc. The vocabulary descriptors say, for example, that the successful candidate can "use phrases like "In the middle of the photo, there is…". The communicative and pragmatic criterion grid also includes instructions on what the pupil *can* do (not what they cannot), for example, can "use words and phrases to give reasons for their opinion", or can "justify their opinion(s)". The pupil's performance is assessed based on the quality of spoken output i.e. number of grammatically, lexically, semantically and pragmatically error-free structures.

Methodology

This research followed a quantitative design, using a structured dataset of candidates' scores to examine the psychometric properties of the speaking assessment used at the Serbian National Competition in English. The study focused on analysing the internal structure of the analytic rating scale, the relations among its components, and the extent to which performance on the written test predicted oral proficiency outcomes. To address these aims, the dataset was subjected to descriptive statistics, correlation analyses and distributional checks (skewness and kurtosis), enabling a systematic evaluation of the validity and reliability of the oral assessment procedure.

Participants

Seventy-seven pupils from seventy-one elementary schools across Serbia qualified for the speaking test based on their written-exam results. Seventy-four pupils (34 males, 40 females; aged 13–14 years) decided to take part in the oral examination. Participants represented fourteen of the sixteen national school districts, ensuring broad geographical coverage.

Instrument and Test Design

The oral component of the National Competition test is designed to assess pupils' ability to participate in age-appropriate spoken interaction on familiar topics (*see* Appendix 1), in line with the Grade 8 curriculum and CEFR A2+/B1 descriptors. Each candidate is examined individually by a three-member board consisting of one interlocutor and two assessors. The oral examination lasts approximately ten minutes and follows a three-part interview format informed by research on dialogic task performance (Ferrari, 2012; Michel, 2011; Tavakoli, 2016). The examination includes three sections:

1. **Interview/Warm-up:** short personal questions on familiar topics.
2. **Picture Description:** description and interpretation of an image.
3. **Discussion:** opinion and justification questions related to the picture.

In the first phase (warm-up), the interlocutor asks short personal questions about familiar people, activities and places (e.g., family members, friends, favourite school subjects, hobbies, school or local area) in order to put the candidate at ease and elicit spontaneous interaction. In the second phase (picture description), the candidate is given a photograph depicting an everyday scene (e.g., people learning, helping at home, doing sports or spending time together) and is asked to describe and comment on what they can see. If necessary, the interlocutor provides brief prompts to support extended production. In the third phase (discussion), the interlocutor asks a small number of opinion and justification questions loosely related to the picture topic, encouraging candidates to explain and support their views so that their communicative and pragmatic competence can be assessed.

All three examiners independently rate the candidate's performance using a standardised analytic scale comprising four criteria: coherence and interaction (0–6 points), lexical accuracy (0–6),

grammatical accuracy (0–5) and pronunciation (0–3) (*see* Appendix 2). The maximum total per rater is 20 points. For each criterion, descriptors specify what successful candidates can do in terms of discourse management, vocabulary range, grammatical control and clarity of pronunciation, with an emphasis on functional communicative ability. After the candidate leaves the room, examiners briefly confer to review their notes and assign a final overall speaking score on the same 0–20 scale. The analytic scores and the final overall score are recorded on the rating sheet and serve as the basis for the present analysis.

Data Collection and Analysis

Data were extracted from seventy-four completed rating sheets. Descriptive statistics (mean, standard deviation, frequency distribution) were computed for each analytic category and the holistic score. Correlation analysis (Pearson's *r*) was used to examine relationships between holistic and analytic scores to evaluate internal consistency and construct validity.

For RQ1 (construct validity), descriptive statistics were calculated for each analytic criterion and for the overall speaking score. Pearson product–moment correlations were computed among the four analytic scores and between each analytic score and the overall speaking score in order to examine the internal structure of the rating scale and the extent of overlap between the criteria. In addition, an exploratory analysis of the correlation patterns was used to evaluate whether the scale behaves as a multidimensional instrument or as a predominantly unidimensional measure of oral proficiency.

For RQ2 (predictive validity), Pearson correlations were calculated between the speaking score and the written components of the competition (reading, listening, and grammar test). These analyses were used to investigate whether performance on the written tests meaningfully predicts oral proficiency among this high-achieving group of learners.

Data processing was conducted in Microsoft Excel 365 using built-in statistical functions and the Data Analysis ToolPak for computing descriptive statistics and Pearson correlations.

Ethical Considerations

The study used anonymized archival data with permission from the Society of Foreign Languages and Literatures, ensuring compliance with data-protection and ethical research standards.

Results

Sample characteristics and score distribution

Out of seventy-seven candidates who were invited to attend the Speaking part of the National Competition in English based on their written test results, a total of 74 candidates participated, consisting of 34 males and 40 females. They represented 71 primary schools from 43 municipalities across 14 school districts, providing a representative sample of pupils from diverse educational and regional backgrounds. Since there are 16 school districts in the country, plus the Group for professional and pedagogical supervision in Novi Pazar, it can be noted that three districts did not have their representatives at the National Competition.

Accordingly, the data were gathered from 74 assessment sheets filled out by the examination board for each candidate. The distribution of points achieved by candidates is shown below (Figure 1). The highest possible score was 20 points. Thirteen candidates (17.6%) achieved the

maximum of 20 points, 17 candidates (22.97%) scored 19 points, 9 candidates (12.16%) scored 18 points, and 18 candidates (24.32%) obtained 17 points. Five candidates (6.76%) scored 15 points and two candidates (2.70%) scored 16 points. The remaining ten candidates (13.51%) scored below the pass mark of 15 points: six candidates scored 8.11 points (5.4%), two scored 13 points (2.70%), and two scored 12 points (2.70%), representing the lowest scores in the sample.

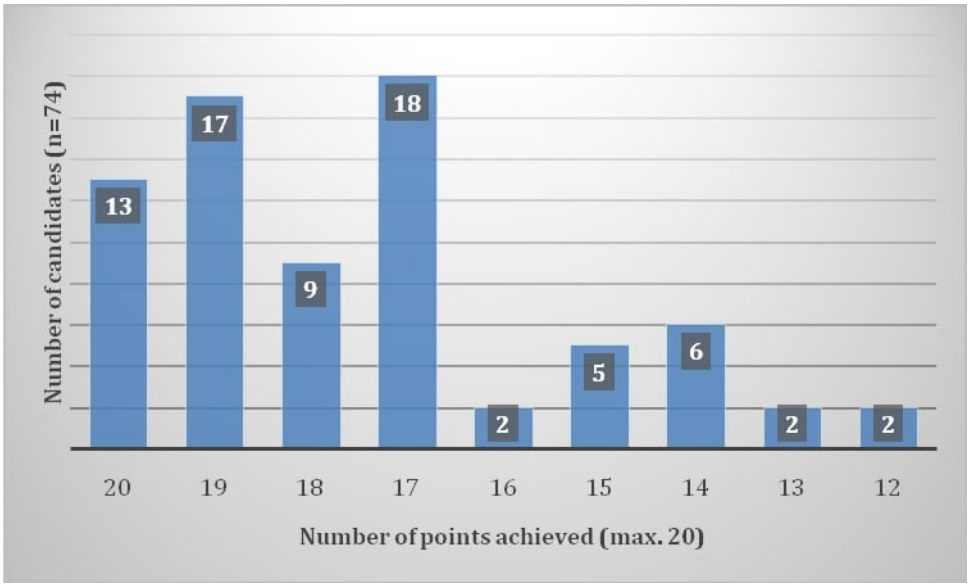

Figure 1. *Distribution of candidates and the points achieved in the Speaking part of the National Competition in English*

The distribution of scores in the Speaking part roughly follows a normal distribution, with the majority of candidates scoring between 17 and 20 points. However, a slight negative skew is observed, as higher scores are more frequent than lower ones, and the distribution shows a moderately peaked shape.

Descriptive statistics for all written and speaking scores are shown in Table 1. Reading and listening comprehension scores were very high, with means close to the respective maxima (M = 7.73 out of 8 and M = 6.77 out of 7), and marked negative skewness, reflecting strong ceiling effects on the written components. The grammar test also showed a high mean (M = 26.00 out of 30), although with somewhat greater spread (SD = 1.16). The speaking test showed substantially greater variability (M = 17.45, SD = 2.09) than the written components. This indicates that, although overall performance was high, the oral test was more effective in differentiating among candidates. Most scores fell within one standard deviation of the mean (approximately 15–20 points), suggesting that the speaking component provided meaningful dispersion even within this high-achieving cohort.

The reading comprehension and listening comprehension scores showed a strong negative skew (skewness = -1.83; -2.15 respectively), indicating a pronounced ceiling effect. The majority of candidates achieved near-maximum scores, with very few lower scores. Such a distribution suggests limited discriminative power of the test among high-performing learners. When it comes to the grammar test, the distribution showed moderate positive skewness (skewness = 0.94), indicating that lower scores were more frequent, with a smaller number of higher-scoring outliers. This pattern suggests that the test provided some degree of differentiation, although the distribution deviated from normality. Skewness and kurtosis analyses showed that two written

components (reading comprehension and listening comprehension) suffer from severe ceiling effects, limiting their discriminative power. In contrast, the grammar test in the written part and overall speaking score fall within acceptable psychometric ranges. Among the analytic speaking criteria, coherence and interaction showed the strongest ceiling tendency, while pronunciation displayed the flattest distribution, offering the greatest variability.

Table 1. *Descriptive statistics for written and speaking scores*

	Reading	*Listening*	*Grammar Test*	*Speaking*	*Coherence and Interaction*	*Lexical Accuracy*	*Grammatical Accuracy*	*Pronunciation*
Mean	7.73	6.77	26.00	17.45	5.23	5.28	4.42	2.52
Standard Error	0.06	0.06	0.13	0.24	0.10	0.08	0.07	0.06
Median	8.00	7.00	26.00	17.67	5.33	5.33	4.33	2.67
Mode	8.00	7.00	25.00	19.00	6.00	6.00	5.00	3.00
Standard Deviation	0.53	0.51	1.16	2.09	0.85	0.70	0.58	0.48
Sample Variance	0.28	0.26	1.34	4.38	0.72	0.48	0.33	0.23
Kurtosis	2.55	3.91	-0.04	-0.03	1.08	-0.24	-0.12	-1.41
Skewness	-1.83	-2.15	0.94	-0.83	-1.28	-0.78	-0.67	-0.38
Range	2.00	2.00	4.00	8.33	3.33	2.67	2.33	1.33
Minimum	6.00	5.00	25.00	11.67	2.67	3.33	2.67	1.67
Maximum	8.00	7.00	29.00	20.00	6.00	6.00	5.00	3.00
Count	77.00	77.00	77.00	74.00	74.00	74.00	74.00	74.00

Note: Speaking criteria scores are means across three raters. Grammar test, reading, and listening are the three components of the written test at the National Competition.

Inter-rater reliability

The oral exam was conducted by a panel of examiners who assessed candidates based on coherence and interaction, lexical accuracy, grammatical accuracy, and pronunciation. Inter-rater reliability was evaluated at the level of individual examination boards (not individual examiners), as candidates were assessed by five different panels and not all raters scored all pupils. Each panel (board) consisted of three examiners who jointly evaluated the candidates on four analytic criteria: Coherence and Interaction, Lexical Accuracy, Grammatical Accuracy, and Pronunciation. Because the design was nested rather than fully crossed, reliability was estimated separately for each board using a two-way mixed-effects model with absolute agreement (ICC [3,1]). IRR was calculated separately for each board using a two-way mixed-effects model with absolute agreement (ICC [3,1]). Reliability coefficients ranged from moderate to excellent. Examination board 3 demonstrated moderate agreement (ICC = .64), while Examination boards 1, 4, and 5 showed good agreement (.76–.83). Examination board 2 reached excellent reliability (ICC = .94), indicating highly consistent judgments among its raters.

Across boards, the coefficients indicated consistently acceptable levels of inter-rater agreement. Coherence and Interaction demonstrated the highest reliability, generally falling within the good range, whereas Lexical Accuracy, Grammatical Accuracy, and Pronunciation showed moderate agreement across panels. These findings are consistent with previous research showing that

global communicative descriptors tend to yield higher inter-rater consistency than micro-linguistic features in oral proficiency assessment. Overall, the results suggest that the scoring procedure was sufficiently reliable for research purposes, while also indicating potential benefits of additional calibration to support more consistent judgment of fine-grained linguistic features.

Construct validity: internal structure of the analytic scale (RQ1)

To examine the internal structure of the analytic rating scale, descriptive statistics and correlations were calculated for the four analytic criteria (coherence and interaction, lexical accuracy, grammatical accuracy, pronunciation) and the overall speaking score. As shown in Table 1, mean scores on the analytic criteria were relatively high, again with pronounced negative skewness, indicating that most performances were rated positively across criteria.

When it comes to correlations, as shown in Table 2 the overall speaking score demonstrated strong positive correlations with coherence and interaction (r = .87) and lexical accuracy (r = .88), indicating that these two criteria contributed most substantially to candidates' final speaking ratings. Grammatical accuracy also correlated strongly with the overall score (r = .74), while pronunciation showed the weakest relationship (r = .65). These results suggest that examiners' holistic impressions were driven primarily by discourse management and lexical performance, with grammar and pronunciation playing comparatively smaller roles.

Table 2. *Intercorrelations among analytic speaking criteria and overall speaking score*

	Speaking overall score	*Coherence and interaction (6pt)*	*Lexical accuracy (6pt)*	*Grammatical accuracy (5)*	*Pronunciation (3pt)*
Speaking overall score	1.00				
Coherence and interaction (6pt)	0.87	1.00			
Lexical accuracy (6pt)	0.88	0.68	1.00		
Grammatical accuracy (5)	0.74	0.47	0.60	1.00	
Pronunciation (3pt)	0.65	0.45	0.47	0.31	1.00

Predictive validity: the relationship between written and oral performance (RQ2)

To examine predictive validity, correlations were calculated between the overall speaking score and the written components of the competition: reading comprehension, listening comprehension, and the grammar test (Table 3). All three correlations were very low, indicating no meaningful association between written and oral performance in this sample. Reading comprehension showed a small negative correlation with speaking (r = –.17), listening comprehension was essentially unrelated to speaking performance (r = –.03), and the grammar test displayed a similarly weak negative correlation with speaking (r = –.10). Overall, we can see that the written test components did not meaningfully predict candidates' oral proficiency.

Table 3. *Correlations between the overall speaking score and the written components of the competition*

	Reading	*Listening*	*Grammar Test*	*Overall Speaking Score*
Reading	1.00			
Listening	0.10	1.00		
Grammar Test	0.01	0.22	1.00	
Overall Speaking Score	-0.17	-0.03	-0.10	1.00

Discussion

The observed distribution, with a concentration of scores between 17 and 20, suggests that the oral exam was well within the proficiency range of most candidates, who were likely already high achievers. The slight negative skew and a moderate peak point to a possible ceiling effect, where the assessment may not have fully captured the range of performance among the best candidates. This calls for a potential revision of the speaking tasks and scoring rubrics to ensure finer distinctions in future competitions.

With regard to construct validity, the high intercorrelations among the analytic criteria and their strong association with the overall speaking score indicate that the rating scale operates in a predominantly holistic manner. This finding is consistent with previous research showing that, in practice, raters often base their judgements on a global impression of communicative effectiveness, even when using analytically framed scales (e.g., Brown, 2004; Ko, 2023). In the present context, coherence and interaction and lexical accuracy emerged as the strongest contributors to the overall score, suggesting that raters prioritise discourse management and lexical resources when differentiating among high-level performances. While this pattern supports the scale's ability to capture a meaningful construct of advanced oral proficiency, it also raises questions about the extent to which the four criteria truly function as separate dimensions in operational use.

The holistic speaking score (average of the three raters) correlated highly with coherence and interaction ($r = .87$) and lexical accuracy ($r = .88$), demonstrating that examiners placed the greatest weight on candidates' ability to structure discourse and use vocabulary appropriately. Grammatical accuracy showed a moderately strong correlation with the overall score ($r = .74$), suggesting that grammatical control supported, but did not dominate, raters' judgments. Pronunciation was the least associated with the holistic rating ($r = .66$), implying that while clarity of speech mattered, it exerted a smaller influence on overall proficiency judgments compared to higher-level communicative dimensions. Taken together, the pattern indicates that the oral assessment emphasized functional, communicative competence more than lower-level phonological accuracy.

The absence of a substantial relationship between written and oral scores should not be interpreted as evidence against the validity of the speaking test. It rather reflects the highly selective nature of the candidate pool and the ceiling effects observed in the written components. Once learners have reached a very high level of receptive and grammatical competence, additional variance in speaking ability is likely to depend more on interactional skills, fluency, strategic competence and affective factors than on further gains in reading, listening or discrete-point grammar knowledge. In this sense, the speaking test adds distinctive value to the assessment system by discriminating among top-performing pupils in ways that the written tests no longer can.

This pattern is partly explained by a pronounced ceiling effect in the written tests. Most candidates achieved near-maximum scores in reading, listening and grammar, reflecting their status as top performers who had already passed stringent cut-scores to qualify for the speaking exam. As a result, the written components offered little variance and thus limited potential to predict differences in oral proficiency. Within this selective, high-achieving cohort, the speaking test therefore appears to provide unique information that is not captured by the written components.

The inter-rater reliability indices obtained at the level of examination boards indicated acceptable to good agreement, particularly for the more global criterion of coherence and interaction. This

provides additional support for the reliability of the speaking scores and suggests that examiners shared a broadly similar understanding of what constitutes successful performance at this level.

Taken together, the findings of the study point to several practical implications for further refinement of the speaking component at the National Competition. First, the analytic rating scale would benefit from clearer and more discriminating descriptors, as the very high intercorrelations among criteria suggest that examiners primarily rely on global impressions rather than distinct dimensions of oral proficiency. Refining the scale could help capture finer differences among high-performing candidates. Second, although inter-rater reliability – the key component of the test's construct validity – was acceptable, additional examiner training and calibration, particularly for more fine-grained linguistic criteria, could strengthen consistency across examination boards. Finally, both the oral tasks and the written components could be made more discriminative, given the pronounced ceiling effects observed in the written tests and the concentration of speaking scores in the upper range. Adjustments of this kind would enhance the overall validity of the assessment and support more precise identification of top-level EFL learners.

Limitations and future research

Several limitations of this study should be acknowledged. First, the sample consisted exclusively of high-achieving 8th-grade learners who had already passed strict cut-scores in the written test components, which limits the generalisability of the findings to the broader population of EFL pupils. Second, the nested rating design, in which candidates were assessed by five different examination boards, constrained the scope of reliability analyses and did not allow for a fully crossed rater–candidate design. Moreover, it might be useful to have a study which would yield conclusions on IRR based on the interviewer's prompt choice in the Speaking part.

Future research could address these limitations by examining the performance of a wider range of learners across proficiency levels, by implementing many-facet Rasch or generalizability studies with fully crossed designs, and by triangulating competition scores with external speaking tests or classroom-based assessments. Such work would further clarify the role of national competitions within the wider ecosystem of EFL assessment in Serbia.

Conclusion

The paper demonstrates that the Speaking test at the Serbian National English Competition can be used as a reliable, meaningful measure of high-level oral proficiency among 8th graders, but that its analytic criteria and written-test components require refinement to enhance discrimination and construct validity. These findings are significant because they identify and address potential concerns regarding the suitability of the reading, listening and grammar components of the written test, as well as their limited selective and discriminative capacity. The primary objective of the study was to analyze and discuss the process and instruments of oral performance assessment by offering insight into its validity and credibility with the aim to enhance assessment standards with more discriminating performance descriptors. Additionally, the study underscores that examiner training and calibration for specific scoring tasks at the National Competition in English could strengthen consistency in assessment by solidifying inter-rater reliability, i.e. raters' consistency and precision. Finally, some adjustments in the nature of tasks, both oral and written, could improve the discriminative power of the overall test so that the top-achievers are easily recognized in a group of high-achievers.

Acknowledgement

The authors of this paper would like to thank the board members who participated in the competition.

Disclaimer (Artificial Intelligence)

This paper was not written or edited with assistance of AI tools, LLMs or AI-assisted technology.

References

Brown, H. D. (2004). *Language assessment: Principles and classroom practices*. Longman.

Ellis, G. (2014). 'Young learners': Clarifying our terms. *ELT Journal, 68*(1), 75–78. https://doi.org/10.1093/elt/cct062

Elizondo-Gonzalez, J. F. (2025). Optimizing inter-rater reliability in foreign language constructed-response assessments. *Research in Pedagogy, 15*(2), 471–482. https://doi.org/10.5937/IstrPed2502471F

Ferrari, S. (2012). A longitudinal study of complexity, accuracy and fluency variation in second language development. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 277–308). John Benjamins Publishing Company.

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing, 28*(1), 5–29. https://languagetesting.info/articles/store/PDTs2011.pdf

Ko, H. S. (2023). English oral proficiency measured by holistic and analytic assessments in dialogic and monologic tasks. *English Teaching, 78*(1), 63–82. https://doi.org/10.15858/engtea.78.1.202303.63

Lin, J. (2022). A structural relationship model for L2 oral proficiency, L2 interest, perceived importance of speaking, and out-of-class L2 contact. *Language Teaching Research, 29*(2), 700–725. https://doi.org/10.1177/13621688221074027

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education/Macmillan.

Michel, M. C. (2011). Effects of task complexity and interaction on L2 performance. In P. Robinson (Ed.), *Second language task complexity: Researching the cognition hypothesis of language learning and performance* (pp. 191–216). John Benjamins Publishing Company. https://doi.org/10.1075/tblt.2.12mic

Slobin, D. I. (1996). From "thought and language" to "thinking for speaking." In J. J. Gumperz & S. C. Levinson (Eds.), *Rethinking linguistic relativity* (pp. 70–96). Cambridge University Press.

Tavakoli, P. (2016). Fluency in monologic and dialogic task performance: Challenges in defining and measuring L2 fluency. *International Review of Applied Linguistics in Language Teaching, 54*(2), 133–150. https://doi.org/10.1515/iral-2016-9994

Appendix 1

Thematic Areas

1. Personal identity
2. Family and immediate social environment (friends, neighbours, teachers, etc.)
3. Geographical features
4. Serbia – my homeland
5. Housing – forms and habits

6. Living world – nature, pets, environmental protection, ecological awareness
7. History, temporal experience, and the perception of time (past–present–future)
8. School, school life, the school system, education and upbringing
9. Professional life (chosen or future profession), plans related to future employment
10. Young people – children and youth
11. Life cycles
12. Health, hygiene, disease prevention, treatment
13. Emotions, love, partnerships, and other interpersonal relationships
14. Transport and means of transportation
15. Climate and weather conditions
16. Science and research
17. Art (especially modern literature for young people; contemporary music; visual arts; dramatic arts)
18. Spiritual life; norms and values (ethical and religious principles); attitudes, stereotypes, prejudices, tolerance and empathy; caring for others
19. Customs and tradition, folklore, celebrations (birthdays, holidays)
20. Leisure time – entertainment, recreation, hobbies
21. Nutrition and gastronomic habits
22. Travel
23. Fashion and clothing
24. Sports
25. Verbal and non-verbal communication, conventions of behaviour and etiquette
26. Media, mass media, the internet and social networks
27. Living abroad, contacts with foreigners, xenophobia

(Excerpt from the Curriculum Regulation for Grade 8 of Primary School)

Appendix 2

Oral Component at the National Foreign Language Competition

The oral exam assesses a pupil's ability to participate in spoken interaction on familiar topics related to a wide range of age-appropriate situations, events, and experiences, aligned with the curriculum for the corresponding grade level. During the oral part of the competition (the Speaking part), the pupil draws one of the topics prepared by the examination committee and then prepares their presentation for no longer than five minutes. The pupil begins with a monologue, after which one of the examination board members (the interviewer) enters into interaction with the candidate by asking several questions related to the topic. The oral exam lasts approximately 10 minutes per pupil.

In the Speaking part of the competition, each board member evaluates the following achievements:

a) **Coherence of presentation, conversational ability, situational and content appropriateness** (maximum 6 points):

The pupil describes elements related to the assigned topic, answers simple questions from familiar areas, expresses agreement or disagreement with others, provides basic information on familiar topics, and so on.

b) **Lexical accuracy** (maximum 6 points):

The pupil uses an adequate range of vocabulary that allows them to communicate fluently about familiar topics.

c) **Grammatical accuracy** (maximum 5 points):

The pupil forms comprehensible and grammatically correct sentences and correctly uses a variety of grammatical structures.

d) **Pronunciation** (maximum 3 points): The pupil uses sufficiently clear pronunciation and intonation that enable successful communication.

Each member of the oral examination committee independently evaluates the elements of the pupil's speaking competence. The final score for the oral exam represents the arithmetic mean of the total points awarded by all three committee members.

The maximum number of points a pupil can achieve on the oral part of the test is **20 points**.

(Excerpt from the Regulations of SFLL)

Biographical notes:

Danijela Ljubojević is a Research Associate at the Institute for Educational Research in Belgrade, Serbia. She holds a PhD in Applied Linguistics from the Faculty of Philology, University of Belgrade (2017), and has extensive teaching experience in English across different educational levels. Her research interests focus on foreign language teaching and learning, particularly the integration of digital technologies and artificial intelligence in language education. She also served as the lead coordinator for the implementation of national foreign language competitions within the Society for Foreign Languages and Literatures of Serbia from 2024 until 2026.

Mirjana Daničić is Assistant Professor of Translatology at the English Department of the Faculty of Philology, University of Belgrade where she gained her BA (1998), MA (2005) and PhD (2012) degrees. She has co-authored several publications in the EFL testing and authored a number of articles in academic journals and monographs. She has been actively engaged in a number of international projects in the areas of foreign language and translation teaching. She is a reviewer for the Accreditation Board for Serbian Higher-Education Institutions; a member of the commission "Translation, Interpreting and Related Technology" at the Standardization Institute of Serbia; and a chair of the Ministry of Education's commission for the EFL teacher licence exam.

Submitted/Received	Accepted
22 January 2026	23 February 2026