doi.org/10.61425/wplp.2024.19.80.103

PERCEPTIONS OF EXAMINERS ON THE ROLE OF EXAMINER TRAINING IN ASSESSMENT IN THE ADVANCED LEVEL MATURA EXAM IN ENGLISH: A QUESTIONNAIRE STUDY IN THE HUNGARIAN CONTEXT

Éva Végh-Rupert
Eötvös Loránd University, Budapest
rupert.eva@gmail.com

Abstract: This paper investigates language assessment, specifically the impact of examiner training on the assessment process of the Hungarian advanced level Matura exam in English. The study employs a quantitative research methodology to evaluate how examiners' professional and theoretical background knowledge, along with their training, affect the consistency of the assessment of productive skills. The research is based on a questionnaire adapted from the Language Assessment Literacy Survey by Kremmel and Harding (2020). A diverse sample of 64 teacher–examiners from various professional backgrounds (secondary grammar schools and technical schools) and geographic regions in Hungary offers valuable insights into the assessment procedure of writing and speaking skills. The results show that most participants were older teachers, with an equal distribution of teaching experience and involvement in high–stakes exams. The study reveals the factors that influence language assessment in the advanced level Matura exam in English and highlights the complexity of language assessment, the role of examiner background knowledge and the necessity of continuous training and standardisation initiatives to ensure the reliability and validity of assessment results.

Keywords: assessment, examiner–training, A–level Matura exam in English, productive skills

1 Introduction

Given the fundamental role that assessment plays in our society today, it is of high importance to conduct an in–depth analysis of the constructs that are being evaluated, the procedures that are being utilised, and the wider societal and human impacts that testing has. Every aspect of life involves assessment, and they have an impact on educational and professional areas across all age groups (Bachman, 2000). A deeper understanding of the pedagogical frameworks and task designs that boost performance in language proficiency examinations has been promoted over the course of the last few decades due to advances in language instruction and educational assessment (Sahib & Stapa, 2021). The procedures of creating, implementing, and assessing tasks have also been modified because of these developments. Consequently, the variety of assessment possibilities has been expanded, and there has been a higher alignment with the targeted learning goals and construct validity (Bachman, 2000).

Every language test, be it the Hungarian advanced level Matura exam in English (A–level Matura exam) or a nationally or internationally accredited language exam, has its unique combination of exam tasks, the assessment and testing methodology, and whether the exam certificates are valid for life (like all nationally accredited language examinations in Hungary) or expire after a specific length of time (for example IELTS, TOEFL). The execution of the test process, the preparation, the evaluation, the performance feedback (if any), and the exam cost considerably impact which language exam candidates choose to take. In Hungary, language test providers were required by law to align their exams with the Common European Framework of Reference (CEFR) starting in 2007. Since then, all nationally recognised language exams have had to comply with the CEFR levels, and except for the A–level Matura exam qualify for re–accreditation every two years (see Oktatási Hivatal (2025) for legal details). Another notable distinction between accredited language exams and the A–level Matura exam is that while the former types demand a minimum of 40% in each skill (reading, writing, listening, and speaking) to succeed in the exam, the latter only requires a minimum of 12% in each of the four skills.

Although there have been empirical studies undertaken on several areas of language examinations and the CEFR (Dávid, 2012; Einhorn, 2015) and the A–level Matura exam (Együd et al., 2012; Tankó & Andréka, 2021), relatively little attention has been paid to how productive skills are assessed in the A–level Matura exam. Compared to the rigorous certification requirements of other language examinations, the A–level Matura exam merely requires examiner training for the written components in each exam session. What is more, the assessors of the speaking examinations are required to meet only one hour before the exam begins to discuss any questions or concerns (Oktatási Hivatal, 2024). Fulcher (2003) emphasized that effective speaking skill assessment requires well–defined evaluation rubrics and extensive and standardised training for assessors to ensure reliability and validity. He argued that variability in assessor training can significantly impact the consistency of speaking test scores (Fulcher, 2003). Standardised evaluation cannot be assured if most of the speaking test evaluation procedure depends solely on evaluation rubrics and assessors' expertise. Against this backdrop, the present research aims to explore the factors influencing the assessment of productive skills in the A–level Matura exam in English, with a particular focus on examiner training, given the current gaps in standardised training and evaluation practices compared to other language examinations. It also seeks to evaluate how examiner training impacts the reliability and validity of speaking and writing assessments in the context of the A–level Matura exam in Hungary.

2 Literature review

The present literature review will seek to give an overview of the role of examiner training by investigating the outcomes of research conducted so far on the difficulties of assessing productive skills, the impact of theoretical and professional background knowledge on examiners' understanding of the role of examiner training and the effect of examiner training on the evaluation of productive skills in A–level Matura exams.

In the study, the umbrella term '*examiner'* is used to refer to experts who participate in language examinations and A–level Matura exams. The examiners involved in the Hungarian A–level Matura exams do not construct or write test tasks; rather, they only assess the test papers submitted by students. In the evaluation of productive skills, which include writing and speaking tasks, the examiners execute the roles of interlocutors, raters, and assessors simultaneously. Therefore, professionals involved in the Hungarian A–level Matura language exams—who perform roles such as assessing, interacting, and eliciting responses—are all referred to as assessors, raters, or examiners, with no distinction in their titles. Also, *'examiner training*' may encompass any form of training related to the A–level Matura exam assessment procedures, whether they are organized concerning written or oral exams (Davies, 2006).

2.1 Assessment literacy

There is a growing concern regarding the role of assessment in language learning, as well as what teachers should know about language assessment theory and practice (Csépes, 2014). The term "language assessment literacy", introduced by Stiggins in 1991, has been used to describe the professional and theoretical background knowledge that teachers require to carry out language assessment effectively (Csépes, 2014). Also, as Csépes argued, while assessment traditionally focused on measuring students' achievements (assessment *of* learning), there is now a shift towards using assessment to enhance students' learning progress (assessment *for* learning). Although regulations (ITM Decree, 64/2021) outline additional competencies for teacher training programs, including motivating students through assessment, employing various forms of continuous and formative assessment, and evaluating learners' development based on individual needs by using criterion–referenced assessment, the content of language testing courses in Hungarian English teacher education programs needs revision. Since there is a growing emphasis on assessment *for* learning, English language educators and testing tutors should recognize the value of alternative assessment forms and update their syllabi accordingly (Csépes, 2014).

2.2 The difficulties of assessing speaking

Mead (1980) argued that evaluating oral communication abilities involves distinct measuring challenges. First, the assessment must elicit interactive and temporary behaviour, as communication proficiency is demonstrated via interactions with others. Secondly, measurement standards are reliant upon cultural and contextual elements. In some settings, some communication practices are acceptable, even preferred, whereas in others they are unsuitable. However, regardless of the complexity of communicative competence, the evaluation of speaking abilities must adhere to the standards of any assessment activity. Included in these standards are the reliability and validity of the measurement instruments. Additionally, assessment tools must be unbiased, and this is of special importance when assessing speaking abilities due to the cultural and situational aspects of communication competency. Lastly, measuring instruments must be practical, as it is necessary to design generally applicable assessment methodologies.

Luoma (2004) also confirmed that speaking is the most challenging linguistic ability to measure accurately and identified some more variables upon which evaluation may hinge, including the criteria used to evaluate performance and how these scales are interpreted by the assessor. Of course, proper task design, accurate and regular training of assessors and interlocutors, and time for assessors to evaluate and update their ratings can help solve some of the challenges of evaluating oral performance. Besides Luoma, Ginther (2012) also considered speaking the most challenging of the four language skills to evaluate because elicitation strategies, assessment scales, and assessor training must all be considered.

2.3 Familiarity with the Common European Framework of Reference (CEFR) at the school level

According to Bachman (1990) the primary goal of testing is to provide a systematic and objective method of collecting information about an individual's language ability for specific decision–making purposes. These tests aim to provide reliable and valid inferences about the test–taker's performance based on a well–defined construct. Scores for receptive skills, including listening and reading, are determined by the number of correct answers. In contrast, for productive skills, such as speaking and writing, scores are allocated based on predetermined criteria, including coherence, fluency, and grammatical accuracy. The total test score is the basis for assessing whether a candidate has passed or failed; it represents the total number of points obtained across all tasks and sections.

When a language exam is aligned with the Common European Framework of Reference for Languages (CEFR), the assessment process should involve an additional layer of evaluation. Beyond determining pass or fail status, it requires

assessing whether the test–taker has attained a specific CEFR proficiency level, such as B2. To perform this assessment effectively, exam panellists—committee members responsible for examination–related decisions—must possess a thorough understanding of the CEFR (Council of Europe, 2020). They should also be able to apply its descriptors accurately and consistently to ensure the validity of their judgments.

Standardization is another important aspect of the panel's responsibilities, ensuring fairness and consistency in decision–making. This formal process is not only critical for individual test–takers but also carries significant implications for policymakers, who rely on standardized assessment outcomes to inform educational and institutional decisions.

The influence of the CEFR extends beyond formal assessment, playing a transformative role in language instruction at the school level. A study by Sahib and Stapa (2021) highlighted that one of the CEFR's most significant contributions has been its impact on language assessment practices (Sahib & Stapa, 2021, p. 656). Although the primary focus of their study is the integration of the CEFR into classroom teaching, their findings emphasize the importance of CEFR–specific training for educators. Such training is essential to deepen teachers' understanding of the framework, refine their instructional methods, and improve their ability to assess students effectively in alignment with CEFR standards.

2.4 Examiner training, scoring consistency, and bias

Within language assessment, examiner bias and variability in scoring have emerged as critical challenges impacting reliability and fairness (Weigle, 1998), so training programs have been developed to address these issues and enhance inter–rater reliability (Brown, 2004).

Yan (2014) referred to three factors having an impact on test results: "experience (e.g., trained vs. novice examiners), rater type (e.g., form–oriented vs. content–oriented raters), and examiners' linguistic backgrounds (e.g., familiarity/unfamiliarity with examinees' linguistic features due to examiners' first language (L1) or L2)" (Yan, 2014, p. 503). In the proposed study, answers will be sought for the question of how professional and untrained examiners differ in their consistency and severity of assessing. Although the CEFR provides standardised rules for rating scales, allowing for the differentiation of language competency levels yet maintaining scoring reliability, examiners may understand the same scales differently, based on what type of examiners they are. Baker (2012) studied how raters use the scales to assess language performance, emphasising the importance of their opinions on the evaluation process. Understanding both the framework and the subjective components that influence examiners is essential for granting fair and consistent language assessment.

An examiner training programme that familiarises assessors with the levels and descriptors is a common method for enhancing examiner reliability (Dávid, 2012; Dimova, 2022). According to Fulcher and Davidson (2013), training can help examiners develop a shared understanding of the scale descriptors, which allows them to harmonise their approaches to the rating process and scale descriptors. Lumley and McNamara (1995) provided insights into the challenges of examiner variability, demonstrating that examiners' backgrounds, experiences, and subjective interpretations may influence scoring outcomes. They claimed that without adequate training, even well–constructed scoring scales cannot ensure consistency, as individual examiners often bring personal biases and varying conceptualizations of assessment criteria into the evaluation process. These biases can stem from elements such as attitudes, prior experiences, and performance expectations, potentially leading to differences in scoring. Lumley and McNamara (1995) emphasised the two goals of examiner training: improving reliability by aligning judgments with standardised criteria and addressing fairness concerns by reducing bias. Effective training programs, they argued, should not only focus on understanding the scale descriptors but also include interactive and reflective components to help examiners become aware of their tendencies and biases. Trainings can also help examiners become more aware of their own biases concerning the aspects of L2 speaking performances (Sundqvist et al., 2018). Examiners' interpretations and interactions with the scoring scales are influenced by several factors, such as the origins of their language, their knowledge of different accents, their educational and instructional opinions, their traditions and concepts, as well as their previous experiences with assessment and their understanding of the context (Dimova, 2022).

Doosti and Ahmadi Safa (2021) concluded that examiner training is one of the suggested methods for addressing inconsistency and using standardised scoring scales may effectively enhance agreement. Examiners in speaking exams are advised to ensure that all test–takers are subjected to identical settings for assessment to resolve testing–related ethical problems and improve test administration fairness. At the level of selection and planning, they must ensure uniformity in test administration, interview location, instructions, and test length. To improve examiner consistency and uniformity, it is recommended that assessors participate in brainstorming or training sessions before the exam to ensure that they apply the same scoring scale and criteria. This will help to achieve a score that more accurately reflects the examinee's genuine skill and is fair to them. As other researchers (Davis, 2016; Kang et al., 2019) also explained, the goal of training in terms of dependability is to minimise the differences between examiners and that training tends to standardise examiners' assessment practices and enhance the quality of their evaluations by defining the scoring criteria, minimising bias, and increasing reliability.

3 Method

3.1 Design of the study and the research questions

To investigate the impact of professional and theoretical background knowledge, namely assessment literacy, as well as examiner training on the consistency of assessing productive skills in the Hungarian A–level Matura exam in English, the quantitative research paradigm was used, and a questionnaire study was conducted. The questionnaire–based approach and the collected quantitative data were used to analyse the extent of knowledge the different groups of examiners had about the principles, theories, and practices of language assessment. Furthermore, it examined the impact of examiner training on assessment outcomes and the variables that influence the perceived significance of participating in examiner training.

After considering the theoretical background and the research aim, the following research questions were formed:

(1) How much self-perceived knowledge do the different groups of examiners in the sample have about the principles, theory, and practice of language assessment?

(2) How does self-perceived understanding the role and benefits of examiner training impact the assessment outcomes of productive skills in the A–level Matura exam?

(3) How does assessment literacy impact the examiners' understanding of the importance of examiner training?

3.2 Participants

The 5–point Likert–scale questionnaire was adapted from the Language Assessment Literacy Survey by Kremmel and Harding (2020). It was completed by 64 EFL teacher–examiners from diverse professional backgrounds, geographical locations, and age groups, recruited through snowball and convenience sampling techniques to maximise the representation of EFL examiners involved in A–level Matura exam assessment in Hungary.

The participants encompassed a wide age range, with the older age group being predominant; those aged 46–65 made up the majority of the sample, accounting for 70.3% of the participants, while those aged 27–45 constituted approximately 29.7% of the respondents. Regarding the length of teaching experience in English, 5 respondents, 7.8% of the total sample, reported having 1–10 years of experience. Eleven to 20 years of teaching experience was indicated by 18 respondents or 28.1% of the participants, and 25 of the participants had 21–30 years of teaching experience, which is 39.1% of the total group. There

were 16 respondents, 25% of the sample, with 31 or more years of teaching experience.

The questionnaire also asked about the participants' experience in A–level Matura exam assessment. The results show that 31 respondents—48.4% of the sample—had 1–14 years of experience, and 33 participants—51.6% of the sample—reported 15–27 years of high–stakes language exam assessment experience. The cut–off point was set at 14 years in the grouping process as professionals become eligible to apply for the master–teacher status after completing 14 years of teaching (Oktatási Hivatal, 2019).

The participants completed the questionnaire mostly from county seats (41 respondents, 64% of the participants). Out of the total number of respondents, 11 participants work both in Budapest and provincial towns (altogether 35% of the sample), and one participant chose 'other' as their place of work. The uneven distribution of participants across different locations limited the statistical power of the analyses and made it difficult to detect meaningful differences between groups therefore, these data were not considered representative in the data analysis.

Three workplace groups, or rather school–type–based examiner groups were formed. The first group consisted of 25 examiners from secondary grammar schools, the second group consisted of 32 participants employed in technical or vocational schools and the third group comprised seven higher education examiners. As the higher education examiners also held positions in secondary grammar schools, they were included in the first group. This resulted in the formation of two equal groups, each consisting of 32 participants.

The respondents were also grouped by position and assessment involvement. Group 1, with 39 respondents, included EFL teachers in their main positions. Examiners in the second group, 25 respondents, were not only language teachers but also professional examiners and test developers.

3.3 The instrument

The 40–item 5–point Likert–scale questionnaire used in the present study was an adapted version of the Language Assessment Literacy Survey, created by Kremmel and Harding (2020). The initial survey was intended to be easily understood, using plain English to ensure comprehensibility for respondents who may not be familiar with assessment–related terminology. To enhance clarity and comprehension, the first version of the questionnaire was tested through think-aloud protocols with three colleagues. Based on their feedback, numerous modifications were implemented to improve the wording and structure of the items. Additionally, the questionnaire was piloted using FACETS analyses, which helped identify inconsistencies in item performance. As a result, two items were removed to increase reliability and ensure the clarity of the instrument

(Linacre, 2018). The finalized version of the adapted questionnaire exclusively focused on assessing language assessment literacy and underwent refinements tailored to the specific requirements of the research. The questionnaire items were structured as closed-ended questions, employing a Likert scale to provide response alternatives. Participants assessed their level of self-perceived knowledge and agreement using a 5-point scale, with 1 representing "no knowledge" and 5 representing "high knowledge" for knowledge items, and with 1 representing "strong disagreement" and 5 representing "strong agreement" for agreement items.

The finalized questionnaire explores two main constructs: (a) knowledge about the principles, theory, and practice of language assessment (constructs 1-5) and (b) understanding the role of examiner training (constructs 6-8). The following list describes various aspects of language assessment expertise. Each category is accompanied by a short description, the number of items, and the reliability coefficient (Cronbach's alpha) to ensure that the questionnaire measures the intended constructs and provides consistent results across different items within each category. The questionnaire integrates various theoretical viewpoints that Fulcher (2013) analyzed, particularly the distinctions between realism and anti-realism. In this context, realism assumes that language constructs, such as competence, represent real and measurable phenomena. In contrast, anti-realism regards these constructs as theoretical tools that facilitate the organization and prediction of observations without asserting a direct correspondence to reality. Fulcher (2013) differentiated between constructivist perspectives, which prioritize subjective interpretations of language constructs, and instrumentalist (or operationalist) approaches that focus on the practical utility of tests. He proposed four criteria for evaluating theories in language testing: testability, simplicity, coherence, and comprehensiveness. He emphasized the necessity of evidence-based validation to ensure that language assessments possess a robust theoretical foundation and maintain their relevance over time. The final version of the questionnaire, as presented in Appendix A, reflects these refinements and ensures that the assessment is both theoretically and empirically grounded.

(1) *Language assessment development and use* (11 items, α = 0.95): refers to the processes involved in developing and evaluating language assessment, with the aim of ensuring reliability, validity, and fairness in language assessment (example: *How knowledgeable do you consider yourself about selecting appropriate tasks for a particular assessment purpose?)*.

(2) *Assessment and scoring for learning* (7 items, α = 0.92): measures how assessment is used to guide learning, diagnose strengths and weaknesses, and score written and spoken performances (example: *How knowledgeable do you consider yourself about how to use assessment to guide learning goals?*).

(3) Learner preparation for assessment (5 items, α = 0.86): explores preparing learners for assessment, the impact of assessment on exam design and teaching materials (example: *How knowledgeable do you consider yourself about preparing learners to take language assessment?*).

(4) Understanding language–learning processes (4 items, α = 0.79): examines language skill development, and understanding the levels of the Common European Framework of Reference (example: *How knowledgeable do you consider yourself about how language skills develop?*).

(5) Assessment and administration (3 items, α = 0.69): refers to designing rating scales for exams and conducting assessment before their official administration (example: *How knowledgeable do you consider yourself about trying out assessment before their administration?)*.

(6) Understanding the role of examiner training (4 items, α = 0.85): measures the role of examiner training in ensuring consistency, understanding scoring rubrics and their essential role in high stakes written and speaking exams (example: *To what extent do you agree that examiner training helps to maintain consistency in giving marks to students?*).

(7) Benefits of examiner training (4 items, α = 0.74): explores the usefulness of group discussions in examiner training, and in revising candidate scores (example: *To what extent do you agree that examiner training helps with revising scores given to candidates?*).

(8) Evaluating the importance of training (2 items, α = 0.90): measures the effectiveness of examiner training in assessing oral and written performances (example: *To what extent do you agree that examiner training is essential in high stakes speaking exams?*).

3.4 Data collection and data analysis

After thoroughly designing it and conducting two think-aloud protocols to ensure empirical validity (Dörnyei, 2007), the 5–point Likert–scale questionnaire was shared across different Facebook groups specifically for EFL examiners, and then it was sent to EFL teacher email lists and professional contacts, from September 2022 until December 2022. Some participants were requested to share the information with their co–workers and other professional connections to involve them in "snowball sampling" (Dörnyei, 2007, p. 85). Since participant self–selection is a concern with online questionnaire studies, it is essential to address it now. It is possible that the sample predominantly reflects the perspectives of the respondents who choose to answer the questionnaire. This is because the participants filled out the questionnaire voluntarily, and there might be substantial differences in attitudes between the volunteers and those who do not volunteer. This can put the representativeness and validity of the study at risk (Dörnyei, 2007). To address this problem and balance the number of

motivated and unmotivated respondents, various English department heads in different secondary school teaching environments, such as secondary grammar schools, technical schools, and vocational schools, were contacted by e–mail and phone. They were requested to ask their EFL teacher colleagues involved in the evaluation of A–level Matura exams to complete the questionnaire as a component of their meetings.

The participants were introduced to the study, given information about the researcher, and informed about the focus of the study at the beginning of the questionnaire. Anonymity was guaranteed to ensure ethical principles, and the questionnaire's estimated completion time was also given. Participants were asked to consent to the use of their data for research purposes. In addition, participants were requested to provide biographical data, including demographic details, years of expertise, type of present educational institution, and their positions.

The quantitative data collected through the questionnaire was exported to SPSS Version 29 for analysis. Before conducting statistical tests, a data preparation process was carried out, including coding, input, screening, and cleaning to ensure accuracy and reliability (Dörnyei, 2007). Firstly, coding was applied to categorical variables, where qualitative responses were assigned numerical values (e.g., "strongly agree" = 5, "agree" = 4, etc.). Any open-ended responses that required categorization were systematically coded based on identified themes. The data was entered into SPSS, with each row representing an individual participant and each column corresponding to a questionnaire variable. The dataset was subsequently examined for missing values and inconsistencies. The research questions determined the data analysis process. The reliability of the scales was assessed using Cronbach's alpha values, which were reported in the description of the scales in the methods section. To address the first research question, independent samples t–tests were employed and to answer the second and third research questions regression analyses were run.

4 Results

4.1 Knowledge of examiners about the principles, theory, and practice of language assessment

RQ1 How much self-perceived knowledge do the different groups of examiners in the sample have about the principles, theory, and practice of language assessment?

The first research question can be answered based on Table 1, Table 2, and Table 3, containing the descriptive statistics of the scales and independent samples t–tests to compare the mean scores of the different groups of examiners.

	Assessment experience					
Scales	**1 –14 years n=31**		**15–27 years n=33**			
	M	**SD**	**M**	**SD**	***t***	***p***
Language assessment development and use	3.41	9.58	3.81	8.77	–1.873	.033
Assessment and scoring for learning	3.87	4.98	4.18	4.35	–1.856	.034
Learner preparation for assessment	3.82	3.53	4.08	3.42	–1.522	.067
Understanding language learning processes	4.07	2.13	4.27	2.27	–1.328	.095
Assessment and administration	3.53	2.45	3.86	2.36	–1.590	.058

Table 1. Grouping factor: Assessment experience

Table 1 compares examiners with 1–14 years of experience to those with 15–27 years based on their self-perceived knowledge in different aspects of language assessment. The two groups were formed according to the eligibility of secondary school teachers to apply for the Master Level within the Pedagogical Career Model (Oktatási Hivatal, 2019). According to the findings examiners with more experience tended to rate themselves higher in most areas. Only two of the five measured scales showed statistically significant differences between the two groups. In particular, participants with 15–27 years of experience reported higher self-perceived competence in language assessment development and use (M = 3.81 SD =8.77) compared to less experienced examiners (M = 3.41, SD = 9.58) (t = –1.873, sig. <.05) and assessment and scoring for learning (M = 4.18, SD = 4.35) compared to the less experienced group (M = 3.87, SD = 4.98) (t = –1.856, sig. <.05). On the other hand, for the other three areas—learner preparation for assessment, understanding language learning processes, and assessment and administration—no significant differences were found (sig. <.05). It is possible that self–perceived knowledge in these broader areas is not heavily influenced by experience and could be influenced by general teaching skills or personal confidence levels rather than assessment-specific expertise, as the mean scores were relatively close across both groups. The results highlight that examiners' self–perceptions of their knowledge grow with experience, particularly in technical and applied areas of language assessment. Nevertheless, the lack of significant differences in more general domains suggests that some aspects of assessment knowledge may be shaped by factors beyond years of experience alone and also cast some doubt whether the forming of the two groups (1–14 years and 15–27 years of experience) is appropriate.

	School–type					
Scales	**grammar school n=32**		**technical school n=32**			
	M	**SD**	**M**	**SD**	***t***	***p***
Language assessment development and use	3.74	9.06	3.48	9.01	1.248	.108
Assessment and scoring for learning	4.18	3.63	3.88	5.52	1.817	.037
Learner preparation for assessment	4.08	3.23	3.83	3.71	1.507	.068
Understanding language learning processes	4.25	1.86	4.12	2.53	.841	.202
Assessment and Administration	3.83	2.68	3.61	2.15	1.077	.143

Table 2. Grouping factor: School–type

The grouping variable in this analysis (Table 2) was whether the participants worked in grammar or technical schools. As for language assessment development and use, although Group 1 had a higher mean score (3.74) compared to Group 2 (3.48), the t–value of 1.248 and p–value of .217 indicated no statistically significant difference between the groups. Regarding assessment and scoring for learning the t–value of 1.817 and the p–value of .037 showed a significant difference, suggesting a potential distinction between the groups. However, for learner preparation for assessment, understanding language learning processes, and assessment and administration, differences between the groups were not statistically significant, as indicated by t–values of 1.507, 0.841, and 1.077, and with (sig. > .05). The results indicate that the type of school has little or no impact on most self–perceived assessment–related skills, except for the assessment and scoring for learning scale, where a significant difference is noted between participants from grammar and technical schools.

	Role					
Scales	**EFL teacher n=39**		**professional examiner n=25**			
	M	**SD**	**M**	**SD**	***t***	***p***
Language assessment development and use	3.49	9.14	3.80	9.48	–1.438	.078
Assessment and scoring for learning	4.04	3.92	4.02	5.93	.150	.441
Learner preparation for assessment	3.86	3.31	4.12	3.74	–1.446	.047
Understanding language learning processes	4.15	2.14	4.22	2.37	–.442	.330
Assessment and Administration	3.60	2.34	3.90	2.52	–1.451	.076

Table 3. Grouping factor: Role

4.2 Understanding the role of examiner training

RQ2 How does self-perceived understanding the role and benefits of examiner training impact the assessment outcomes of productive skills in A–level Matura exams?

In order to explore the impact of understanding the role of examiner training, multiple linear regression was conducted. This model was chosen, because it quantifies the impact of training by measuring its strength and direction, helping to determine whether increased training leads to more consistent or improved assessment. Multiple regression facilitates the use of additional variables such as examiner experience and candidate proficiency, thereby enhancing the accuracy of the analysis. Moreover, regression offers predictive insights that are valuable for policymaking and the preparation of examiners. Statistical measures, including p-values and confidence intervals, are essential for determining the significance and reliability of the findings.

	B	SE	*β*	*t*	*p*
Understanding the role of examiner training	–0.171	0.236	–0.114	–0.727	.470
Benefits of examiner training	1.140	0.324	0.552	3.517	<.001
R^2	0.230				
F	9.100	(p < .001)			

Table 4. Linear regression with assessment and scoring for learning as a dependent scale

The linear regression model in Table 4 makes predictions about the assessment and scoring for learning variables using two scales: (i) understanding the role of examiner training and (ii) benefits of examiner training. The coefficients (B) suggest that understanding the role of examiner training has a minimal effect (B = –0.171, p = .470), whereas the benefits of examiner training have a considerable impact on assessment and scoring for learning (B = 1.140, p < .001). The R^2 value of 0.230 suggests that around 23.0% of the variation in assessment and scoring for learning can be accounted for by the predictors. The analysis reveals that the regression model is statistically significant (F = 9.100, p < .001), indicating that the variables understanding the role of examiner training and the benefits of examiner training together predict the outcome of assessment and scoring for learning.

4.3 Effects of assessment literacy

RQ3 How does assessment literacy impact understanding the role of examiner training?

To address the third research question and explore how background theoretical knowledge impacts understanding the role of examiner training, linear regression

was run. This method helped determine the strength and direction of the relationship, quantify the impact of assessment literacy on examiner training, and make predictions about how improvements in literacy might enhance understanding. Also, regression analysis provided statistical validation through measures like p-values and R^2, ensuring the findings are reliable.

	B	SE	*β*	t	p
Language assessment development and use	–0.077	0.099	–0.228	–0.777	.440
Assessment and scoring for learning	0.290	0.203	0.435	1.430	.158
Learner preparation for assessment	–0.019	0.241	–0.021	–0.078	.938
Understanding language learning processes	–0.064	0.273	–0.045	–0.232	.817
Assessment and administration	0.115	0.352	0.088	0.326	.746
R^2	0.088				
F	1.117	(p = .362)			

Table 5. Linear regression with understanding the role of examiner training as a dependent scale

Table 5 shows the results of a linear regression analysis that assessed the relationship between various predictors and understanding the role of examiner training. Language assessment development and use showed a coefficient (B) of –0.077, indicating a negligible impact on understanding the role of examiner training, as its associated *t*–value of –0.777 was not statistically significant (*p* = .440). Similarly, assessment and scoring for learning, learner preparation for assessment, understanding language learning processes, and assessment and administration demonstrated coefficients (B) of 0.290, –0.019, –0.064, and 0.115. None of the predictors yielded statistically significant relationships with understanding the role of examiner training. The R^2 value of 0.088 suggests that these predictors explain around 8.8% of the variance in understanding the role of examiner training. Nevertheless, the F–value of 1.117, with a corresponding *p*–value of .362, indicates that the overall regression model lacks statistical significance, implying that the predictor scales do not significantly predict understanding the role of examiner training.

5 Discussion

The analysis conducted for the first research question, which involved comparing the mean scores of different groups of examiners using various scales, produced informative results. Significant differences were identified between the groups in their ratings for the language assessment development and use scale as well as the assessment and scoring for learning scale, indicating that examiners with

differing levels of assessment experience evaluate these aspects distinctively. In contrast, no statistically significant differences were observed between the groups for the learner preparation for assessment scale, the understanding language learning processes scale, and the assessment and administration scale. This lack of statistical significance suggests that any variation in these areas might be attributed to random chance rather than systematic differences in the assessment practices of examiners.

Linear regression analysis provided insights into the second research question, which explored how understanding the role and advantages of examiner training influence assessment outcomes. The analysis revealed that perceived understanding of the role of training had minimal impact on assessment and scoring practices, suggesting that simply recognizing the need for training may not directly enhance performance. However, the tangible advantages gained through training, such as improved consistency in scoring or enhanced understanding of assessment criteria, demonstrated a significant positive effect on assessment outcomes. This indicates that while theoretical acknowledgment of training is important, it is the practical benefits and applied knowledge from training sessions that play an important role in improving assessment practices. The findings suggest that examiners who actively integrate the advantages of training into their practices are better equipped to deliver more reliable and effective assessment. Moreover, the statistical significance of the model underscores its predictive validity, suggesting that training–related benefits are a meaningful factor in determining assessment quality. These results highlight the necessity of designing training programs that emphasize practical applications and measurable outcomes, directly aligning with the needs and practices of examiners.

Concerning the third research question, which investigated the influence of assessment literacy on the comprehension of the significance of examiner training, linear regression analysis produced unexpected findings. Although factors such as language assessment development and use, assessment and scoring for learning, learner preparation for assessment, understanding language learning processes, and assessment and administration accounted for approximately 8.8% of the differences in understanding the role of examiner training, the regression model lacked statistical significance. This suggests that assessment literacy does not predict the awareness of the significance of examiner training. Therefore, it is necessary to conduct additional research to explore other potential aspects that may have an impact.

Concerning the third research question, which investigated the influence of assessment literacy on the comprehension of the significance of examiner training, linear regression analysis produced unexpected findings. While factors such as language assessment development and use, assessment and scoring for learning, learner preparation for assessment, understanding language learning processes, and assessment and administration together explained approximately 8.8% of the variance in understanding the role of examiner training, the regression model itself

was not statistically significant. This suggests that assessment literacy, as defined by these dimensions, does not predict awareness of the importance of examiner training.

These findings imply that a direct relationship between assessment literacy and the recognition of training importance may not exist or may be mediated by other unexamined factors. For instance, it is possible that personal attitudes toward professional development, institutional support, or prior training experiences play a more significant role in shaping how examiners perceive the role of training. This highlights the complexity of the relationship between knowledge of assessment principles and the practical value placed on training.

6 Conclusions

The research investigated language assessment, specifically the impact of examiner training on the assessment process of the A–level Matura exam in English. After providing an insight into the complexities surrounding the assessment of productive skills, the need for standardised scoring criteria, grading scales, and examiner training programs has been addressed. The results reveal differences in certain assessment scales based on examiner experience, reflecting the challenges highlighted in the literature regarding the assessment of speaking proficiency. While understanding the role of examiner training had minimal direct impact on assessment outcomes, the benefits of examiner training influenced results, highlighting the importance of actual training benefits in enhancing assessment reliability and validity. Interestingly, while neither theoretical nor professional knowledge predicted awareness of the significance of examiner training, the substantial benefits of training programs were evident. The findings suggest that examiner training should be approached as a practical process, directly addressing the real–world needs of examiners and aligning training objectives with measurable outcomes in assessment performance. The multifaceted nature of assessment is evident in the interplay of multiple factors influencing scoring consistency, including examiner experience, rating scale interpretation, and training effectiveness. The findings indicate that assessment is not merely a mechanical process but is shaped by subjective judgment, cognitive biases, and varying levels of examiner expertise.

Despite certain limitations, including a small sample size and uneven geographical distribution, the research contributes perspectives on language assessment practices and examiner training among EFL examiners. The insights gained reveal the need for future research to address sampling biases and to explore additional contextual factors influencing training efficacy, such as institutional policies and examiners' attitudes toward professional development. By addressing these limitations, future studies can enhance the validity, reliability, and generalizability of findings, further strengthening the knowledge base in this critical area.

In conclusion, this study provides actionable insights into improving examiner training and assessment practices, advocating for a shift from theoretical frameworks to experiential learning and practical application. These findings offer essential guidance for educators, policymakers, and institutions aiming to develop robust and equitable assessment systems. Future research should continue to explore innovative methodologies and examine long–term impacts to ensure sustained improvements in language assessment practices and training programs.

Proofread for the use of English by: Francis J. Prescott-Pickup, Department of English Language Pedagogy, Eötvös Loránd University, Budapest.

References

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford University Press.

Bachman, L. F. (2000). Modern language testing at the turn of the century: assuring that what we count counts. *Language Testing, 17*(1), 1–42. https://doi.org/10.1191/026553200675041464.

Baker, B. A. (2012). Individual differences in rater decision–making style: An exploratory mixed–methods study. *Language Assessment Quarterly, 9*(3), 225–248. https://doi.org/10.1080/15434303.2011.637262

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

Council of Europe (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press. https://rm.coe.int/1680459f97

Council of Europe (2020). *Common European framework of reference for languages: Learning, teaching, assessment. Companion volume with new descriptors*. Council of Europe. https://rm.coe.int/16809ea0d4

Csépes, I. (2014). Language assessment literacy in English teacher training programmes in Hungary. In Horváth, J., & Medgyes, P., (Eds.), *Studies in honour of Marianne Nikolov*. 399–411.

Davies, A. (2006). *Dictionary of language testing*. Cambridge University Press.

Dávid, G. (2012). A szintleírások nyelvének szerepe a Közös Európai Referenciakeret magyar, angol és német nyelvű kiadásában. [The role of the language of level descriptors in the Hungarian, English and German versions of the Common European Framework of Reference for Languages]. *Magyar Pedagógia*, *112*(1), 19–39.

Dimova, S. (2022). Performance–based speaking tests: Possibilities in local language testing. *Language Teaching Research Quarterly*, *29*, 120–133. https://doi.org/10.32038/ltrq.2022.29.08

Doosti, M., & Ahmadi Safa, M. (2021). Fairness in oral language assessment: Training raters and considering examinees' expectations. *International Journal of Language Testing*, *11*(2), 64–90.

Dörnyei, Z. (2007). *Research methods in applied linguistics*. Oxford University Press.

Együd, G., Kiszely, Z., & Szabó, G. (2012). Comparing the Hungarian school–leaving examination with international language examinations. *Language Testing and Evaluation*, 109–126.

Einhorn, Á. (2015). *A pedagógiai modernizáció és az idegennyelv–tanítás*. [Pedagogical modernisation and foreign language teaching]. Miskolci Egyetemi Kiadó.

Fulcher, G. (2003). *Testing second language speaking*. Routledge.

Fulcher, G. (2013). Philosophy and language testing. *The Companion to Language Assessment*, 1431–1451.https://doi.org/10.1002/9781118411360.wbcla032

Fulcher, G., & Davidson, F. (Eds.). (2013). *The Routledge handbook of language testing*. Routledge.

Ginther, A. (2012). Assessment of speaking. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics*. Blackwell Publishing. https://doi.org/10.1002/9781405198431.wbeal0052.pub2

ITM rendelet 64/2021. (XII. 29.) *A tanári felkészítés közös követelményeiről és az egyes tanárszakok képzési és kimeneti követelményeiről szóló 8/2013. (I. 30.) EMMI rendelet, valamint egyes kapcsolódó miniszteri rendeletek módosításáról (2021)*. [ITM Decree 64/2021. (XII. 29). Amendment of EMMI Decree No. 8/2013 (I. 30.) on common requirements for teacher preparation and on training and outcome requirements for certain teacher specialisations, and certain related ministerial decrees (2021)]. https://net.jogtar.hu/jogszabaly?docid=A2100064.ITM&txtreferer=00000001.txt

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

Kremmel, B., & Harding, L. (2020). Towards a comprehensive, empirical model of language assessment literacy across stakeholder groups: Developing the language assessment literacy survey. *Language Assessment Quarterly*, *17*(1). 100–120. https://doi.org/10.1080/15434303.2019.1674855

Linacre, J. M. (2018): *A user's guide to FACETS: Rasch-model computer programs. Program manual*. Chicago, IL: Winsteps.com. https://www.winsteps.com/facets.htm

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, *12*(1), 54–71. https://doi.org/10.1177/026553229501200104

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

Mead, N. A. (1980). Assessing speaking skills: Issues of feasibility, reliability, validity, and bias. *Paper presented at the Annual Meeting of the Speech Communication Association New York*.

Oktatási Hivatal (2019). Útmutató a pedagógusok minősítési rendszerében a Pedagógus I. és Pedagógus II. fokozatba lépéshez.[Guide to the Teacher I and Teacher II levels in the teacher career model]. https://www.oktatas.hu/kiadvanyok

Oktatási Hivatal (2024). Útmutató a tantárgyi bizottságok munkájához. [Guide to the work of subject committees]. https://www.oktatas.hu/kozneveles/erettsegi/erettsegi_vizsgakkal_kapcsolatos_informaciok

Oktatási Hivatal (2025*): Akkreditációs Kézikönyv*. [Accreditation handbook] https://nyak.oh.gov.hu/nyat/doc/ak2025/word/Akkreditacios_Kezikonyv_2025.pdf

Sahib, F. H., & Stapa, M. (2021). The impact of implementing the Common European Framework of Reference on language education: A critical review. *International Journal of Academic Research in Business and Social Sciences*, *11*(11), 644–660. https://doi.org/10.6007/ijarbss/v11-i11/11160

Stiggins, R. J. (1991). Relevant classroom assessment training for teachers. *Educational Measurement: Issues and Practice, 10* (1), 7–12. https://doi.org/10.1111/j.1745-3992.1991.tb00171.x

Sundqvist, P., Wikström, P., Sandlund, E., & Nyroos, L. (2018). The teacher as examiner of L2 oral tests: A challenge to standardization. *Language Testing, 35*(2), 217–238. https://doi.org/10.1177/0265532217690782

Tankó, Gy., & Andréka, Zs. (2021). Probing the advanced level EFL school–leaving examination: The use of English paper. In Tankó, Gy. & Csizér, K. (Eds.), *DEAL 2021: Current explorations in English Applied Linguistics* 65–105. Eötvös Loránd University – Faculty of Humanities.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263–287.

Yan, X. (2014). An examination of rater performance on a local oral English proficiency test: A mixed–methods approach. *Language Testing, 31*(4), 501–527. https://doi.org/10.1177/0265532214536171

APPENDIX A

Perceptions of examiners on the role of examiner training in language exam assessment
(Adapted from *Language Assessment Literacy Survey* by Kremmel and Harding, 2020)

Dear Colleague,

My name is _____ and I am conducting this questionnaire study for my research related to my PhD studies in English applied linguistics and language pedagogy. In this research, I would like to look into how much knowledge the different groups involved in language assessment have about the principles, theory and practice of language assessment and to what extent examiner training influences assessment in high–stakes exams.
Please provide your answers in the following survey, which will take about 8–10 minutes. Participation in this survey is anonymous. By clicking submit to the survey results you consent to your answers being used in this research.

Thank you in advance for your participation!

First, I would like to ask for some basic information.

1) How old are you?

2) Please select where you currently work:
Budapest
county seat
town
other

3) How long have you been teaching English?

4) How long have you been taking part in the assessment of high–stakes exams?

5) Please select the educational institution that you currently work at:

secondary grammar school (gimnázium)
technical school (technikum) or vocational school (szakgimnázium, szakközépiskola, szakiskola)
higher education
other

6) Which of these positions are you regularly involved in?
language teacher
professional rater
test developer
language assessment researcher
assessment policy maker

I. How knowledgeable do you consider yourself about each aspect of language assessment below?

Please respond according to the following scale:

1 = not knowledgeable at all
2 = slightly knowledgeable
3 = moderately knowledgeable
4 = very knowledgeable
5 = extremely knowledgeable

1) developing rating criteria for assessing speaking performance
2) developing rating criteria for assessing writing performance
3) selecting appropriate rating scales
4) selecting appropriate tasks for a particular assessment purpose
5) training others to use rating scales appropriately
6) training others to write good–quality items for language assessment
7) writing good quality items for language assessment
8) the levels of the Common European Framework of Reference
9) determining pass–fail marks
10) determining cut–scores
11) identifying bias in assessing speaking exams
12) designing rating scales for speaking exams
13) designing rating scales for written exams
14) making decisions about what aspects of language to assess
15) trying out assessment before their administration
16) how to use assessment to guide learning goals
17) how to use assessment to diagnose learners' strengths and weaknesses
18) how to use assessment to encourage student learning
19) how to give useful feedback based on an assessment
20) how to prepare learners to take language assessment
21) how assessment can influence the design of a language exam
22) how assessment can influence teaching materials
23) how assessment can influence teaching in the classroom
24) how language skills develop
25) how foreign languages are learned
26) using rating scales to score speaking performances
27) using rating scales to score written performances
28) scoring open–ended questions in speaking
29) scoring discursive speaking tasks
30) scoring monologic speaking tasks

II. To what extent do you agree with the following?

1 = do not agree at all
2 = slightly agree
3 = moderately agree
4 = very much agree
5 = completely agree

1) examiner training helps to maintain consistency in giving marks to students
2) examiner training helps to understand the speaking task(s)
3) examiner training helps to understand the writing task(s)
4) examiner training helps to understand the scoring rubric(s)
5) examiner training helps with the understanding of the allocation of marks
6) group discussions are useful during examiner training
7) examiner training helps with revising scores given to candidates
8) during examiner training, it is possible to consult with other professionals/ examiners to 9) clarify doubts about assessment made
10) examiner training is essential in high stakes written exams
11) examiner training is essential in high stakes speaking exams