



Received: February 10, 2024
Accepted: March 18, 2024
Available online: March 25, 2024

Mukaddas Butabaeva

Teacher, Uzbekistan State World
Languages University
Tashkent, Uzbekistan
E-mail: m.butabayeva@uzswlu.uz
ORCID iD: 0009-0000-6502-6584

Muqaddas Bo‘tabayeva

O‘qituvchi,
O‘zbekiston davlat jahon tillari universiteti
Tashkent, Uzbekistan

A SCIENTIFIC ANALYSIS OF THE RELIABILITY OF MULTILEVEL ASSESSMENT OF SPEAKING SKILLS

KO‘P DARAJALI (MULTILEVEL) GAPIRISH KO‘NIKMASINI BAHOLASH ISHONCHLILIGINING ILMIY TAHLILI

ABSTRACT

The article examines the reliability of the Multilevel Speaking Test administrated by the Knowledge and Skills Assessment Agency under the Ministry of Higher Education, Science and Innovation of the Republic of Uzbekistan. Assessing speaking test – oral proficiency is essential in language testing, but it can be challenging due to various factors influencing evaluations. The piece delves into reliability in language testing, emphasizing the importance of consistent assessment outcomes. It explores the role of raters, discussing intra-rater and inter-rater reliability, and how subjective evaluation can impact test scores.

The article is about the Multilevel Speaking Test, an English proficiency examination aiming to evaluate students' language levels in Uzbekistan's educational system. The test is a criterion-referenced proficiency examination and integral to the Common European Framework of Reference for Languages (CEFR). The speaking section of the test is discussed in detail, including its structure, evaluation framework, and factors that affect its reliability. Rater variability and test conditions are two critical aspects that are evaluated to ensure the test's reliability. The test employs two raters to mitigate subjectivity, but deficiencies in rater training and cultural biases may affect reliability. The article also highlights the benefits of computer-delivered tests for reliability, but it also points out issues such as limited sample materials and the singular task format that may impact the test-taker's performance. In conclusion, the Multilevel Speaking Test is reliable because it uses two raters, but the article suggests that rater

ANNOTATSIYA

Maqolada O‘zbekiston Respublikasi Oliy ta’lim, fan va innovatsiyalar vazirligi huzuridagi Bilim va malakalarni baholash agentligi tomonidan o‘tkaziladigan multilevel (ko‘p darajali) sinovi tarkibidagi gapirish ko‘nikmasini baholashning ishonchliligi tahlil qilinadi. Gapirish ko‘nikmasi – og‘zaki nutqni tekshirish tilni bilish darajasini baholashda hal qiluvchi ahamiyatga ega, ammo baholashga ta’sir qilishi mumkin bo‘lgan turli omillar tufayli bu jarayon har doim ham oson kechmaydi. Maqolada mazkur nutq kompetensiyasini imtihon qilishda ishonchlilik konsepsiyasi o‘rganilib, uni izchil baholash natijalarining muhimligi ta’kidlanadi. Bundan tashqari, maqolada baholovchilar rolining ishonchliligi muhokama qilinadi va subyektiv baholash test ballariga qanday ta’sir qilishi mumkinligi ko‘rib chiqiladi.

Ushbu sinov O‘zbekiston ta’lim tizimi doirasida “Chet tilini o‘rganishning umumiyevropa kompetensiyalari: o‘rganish, o‘qitish, baholash” (CEFR) standartlariga integratsiya qilingan va ingliz tilini bilish darajasini (B1, B2 yoki C1) baholashga qaratilgan. Sinov talabalar va o‘qituvchilar uchun muhim bo‘lib, muayyan mezonlarga asoslangan malaka imtihoni vazifasini bajaradi. Unda baholash ikkita muhim jihatni o‘z ichiga oladi: baholovchining o‘zgaruvchanligi va sinov shartlari. Sinov subyektivlikni yumshatish uchun ikkita baholovchini qo‘llaydi, ammo baholovchini tayyorlashdagi kamchiliklar va madaniy tarafdashliklar ishonchlilikka ta’sir qilishi mumkin. Maqolada, shuningdek, kompyuterda topshirilgan testlarning ishonchlilik afzalliklari e’tirof etilgan, biroq cheklangan

training should be improved to reduce biases. The reliability of the test can be further improved by providing more sample materials, adding more tasks, and improving test conditions.

After a comprehensive evaluation, the Multilevel Speaking Test was found to be reliable. However, there is room for improvement in rater training, test conditions, and sample availability to enhance its effectiveness in measuring oral proficiency.

Key words: language testing, oral competence, reliability, multilevel speaking test, language proficiency, rater variability, educational system, cultural bias.

namunaviy materiallar va test topshiruvchilarning tayyorgarligi hamda ishlashiga ta'sir qiluvchi yagona topshiriq formati kabi masalalar tilga olingan. Xulosa shuni ko'rsatadiki, multilevel og'zaki testi yaxshi baholovchilararo ishonchlilikka ega, chunki u ikkita baholovchidan foydalanadi. Shu bilan birga, u noto'g'ri fikrlarni kamaytirish uchun baholovchilarni bilim va malakalarini oshirish maqsadida, qayta tayyorlov kurslarini doimiy tashkil etishni taklif qiladi. Xulosa shuki, test kompyuter tomonidan boshqarilganda ishonchli ekanligini tan oladi, ammo u test shartlarini yaxshilashni, ko'proq namunaviy materiallarni taqdim qilishni va ishonchlilikni yanada oshirish uchun qo'shimcha vazifalarni qo'shishni tavsiya qiladi.

Umuman olganda, maqolada multilevel gapirish ko'nikmasini baholashning ishonchligi har tomonlama baholanadi, uning kuchli tomonlarini e'tirof etiladi, shu bilan birga, baholashning umumiy ishonchligi va og'zaki nutqni o'lchashda samaradorlikni oshirishda baholovchilarni tayyorlash, test sharoitlari hamda namunalar mavjudligini yaxshilash ta'kidlanadi.

Kalit so'zlar: til testi, og'zaki kompetensiya, ishonchlilik, ko'p darajali gapirish sinovi, tilni bilish, baholovchining o'zgaruvchanligi, ta'lim tizimi, madaniy tarafkashlik.

INTRODUCTION

Communicative language instruction, emphasising fostering speaking abilities, has underscored the necessity of incorporating oral competence testing into language assessment practices [Nakamura, 1996]. Nevertheless, evaluating speaking skills poses a considerable challenge due to various factors. These challenges include the inherently subjective nature of the evaluation process, the potential influence of contextual factors, time constraints, the nuanced impact of non-verbal cues, and the inherent linguistic complexity of spoken language [Luoma, 2004]. As articulated by A.Green, the assessment of a language test's quality hinges on specific characteristics, namely practicality, reliability, validity, and the overall beneficial outcomes derived from the assessment [Green, 2014]. Among these characteristics, reliability and validity stand out as essential measurement qualities critical for the efficacy of language tests.

A.Green contends that reliability and validity are paramount in evaluating the effectiveness of language tests [Green, 2014]. The concept of reliability, particularly in the context of a speaking exam, has garnered substantial attention from researchers [Bachman, 1990; Bachman & Palmer, 1996; Fulcher & Davidson, 2007]. It is considered a fundamental issue and a key measuring feature that necessitates comprehensive investigation in language testing. Reliability, in this context, refers to

the consistency and stability of the test results. A reliable speaking test should yield consistent scores for the same individual across multiple testing instances, assuming their language proficiency remains constant.

This article is focused primarily on probing the reliability of the multilevel speaking test administered by the Knowledge and Skills Assessment Agency under the Ministry of Higher Education, Science and Innovation of the Republic of Uzbekistan. The central question guiding this inquiry is, ‘To what extent is the oral examination of the multilevel test offered by the Knowledge and Skills Assessment Agency under the Ministry of Higher Education, Science and Innovation of the Republic of Uzbekistan reliable?’. Addressing this question requires meticulously examining various factors, including the test design, scoring rubrics, inter-rater reliability, and the overall consistency of scores over repeated administrations.

In evaluating the reliability of the multilevel speaking test, it is essential to scrutinise the test’s design and the clarity of its instructions. A well-constructed test with unambiguous guidelines contributes to the stability of results. Additionally, the reliability of the test is contingent on the consistency of scoring, which necessitates well-defined scoring rubrics and adequate training for raters. The issue of inter-rater reliability, referring to the extent to which different raters produce consistent scores for the same oral performance, is also a crucial aspect of the reliability assessment. The reliability of oral language exams, especially in the context of multilevel speaking tests, is a multifaceted and critical aspect of language testing. This article explores and analyses the reliability of the multilevel speaking test conducted by the Knowledge and Skills Assessment Agency under the Ministry of Higher Education, Science and Innovation of the Republic of Uzbekistan. Through a comprehensive examination of test design, scoring rubrics, and inter-rater reliability, this work aims to contribute to the ongoing discourse on effective language testing practices to enhance the reliability of oral language assessments.

The concept of reliability in language testing and oral ability

The reliability of a testing instrument is measured by how consistently it produces the same results [O’Mahony, 2019]. The definition of the term ‘reliability’ is ‘the constancy of assessment’ [Bachman & Palmer, 1996]. To put it another way, a measurement is considered accurate if a person’s results on the same test taken twice are comparable. According to Jacob, reliability is a crucial feature of a successful exam since if a test does not measure consistently (reliably), then the results from a particular administration cannot be relied upon as an accurate estimate of students’ achievement [Chiedu & Omenogor, 2014]. Therefore, reliability facilitates identifying that the results are the most accurate and complete way to determine how skilled a candidate is. A test’s potential to accurately reflect an evaluation’s accuracy and consistency is one of the characteristics defining its reliability [Khan et al., 2022].

The role of raters in identifying speaking competence is essential when examining productive language learning skills. Two main types of rater reliability are discussed in the academic literature: intra-rater reliability and inter-rater reliability [Rohde et al., 2022]. Data collected by the same person in two or more trials have

high intra-rater reliability [Hoffman et al., 2014]. When two or more raters measure the same group of people, and their results are consistent, it can be stated that there is high inter-rater reliability. M.Park suggests monitoring the same events on the same individuals multiple times to avoid the potential confounding influence of actual testing variances [Park, 2020]. However, it is sometimes considered an issue since it includes individual subjectivity, which influences the grades issued to different students [Schoonen, 2020]. Thus, the reliability may decrease due to the subjective nature of the spoken evaluation if some raters are more tolerant than others [Blais et al., 2017].

This shows several factors in the speaking exam scores [Winke & Gass, 2013; Lee, 2017; Zhao, 2017]. The assessment scale, rubrics, and marking criteria used in each oral test will also determine how much the raters' assessments conflict with one another [Khan et al., 2022]. This rating criterion may affect inter-rater reliability due to differences in understanding the grading system. Other factors affecting test performance and score reliability are bad health, a lack of interest, and test comprehension [Bachman & Palmer, 1996]. Also, L.Bachman and A.Palmer point out that differences in testing conditions, fatigue, and anxiety may impact candidates' performance, resulting in inconsistent scores on different occasions [Bachman & Palmer, 2010; 135]. By understanding the impact of these factors, researchers can minimise measurement error and maximise reliability [Bachman & Palmer, 2010].

The purpose of the Multilevel Test

There is always a reason behind why a test is being administered [Fulcher, 2010]. As A.Hughes argues, the purpose of a language test is to see if a candidate knows the language well enough to be called 'fluent' in some context [Hughes, 2003]. Most language proficiency examinations aim to identify whether candidates have acquired a particular level of competence regarding a set of skills [Quaid, 2018]. As a result, the 'Multilevel Test' that the Knowledge and Skills Assessment Agency under the Ministry of Higher Education, Science and Innovation of the Republic of Uzbekistan administers is known as a proficiency test.

Uzbekistan is one of the countries that has adopted the CEFR into its school curriculum. It is quite recent, having only occurred after the implementation of Presidential Decree No. 1875, 'The measures of enhancing the system of studying foreign languages' (2012, December 10). On March 1, 2022, the Knowledge and Skills Assessment Agency under the Ministry of Higher Education, Science and Innovation of the Republic of Uzbekistan created a brand-new testing methodology called the 'Multilevel' test to assess the language proficiency of Uzbek EFL students and teachers. Before this test, CEFR B1, B2, and C1 level examinations were conducted for EFL students, and a test was conducted to determine a monthly salary increase for EFL teachers. However, they are now being grouped into a single 'Multilevel CEFR' exam. In other words, this test assesses the level between B1 and C1 and allows high school pupils to successfully pass university English exams if they obtain B1 or B2. In addition, Uzbek EFL teachers will receive a wage promotion if they get a C1 level. Furthermore, beginning in the academic year 2022/2023, all newly hired

foreign language teachers at state educational organisations must have at least a B2 level, and all foreign language teachers at state higher education institutions must have at least a C1 level or an equivalent international certificate, according to Clause 9 of Resolution No. 312 of the Cabinet of Ministers of the Republic of Uzbekistan ‘On measures for effective organization of promotion of learning foreign languages’ (2021, May 19) [Resolution No. 321 of the CMRU, 2021]. It indicates that the test is a high-stakes test that benefits educators and students. In addition, the test is regarded as a standard proficiency test, which is often criterion-referenced rather than norm-referenced [Fulcher & Davidson, 2007]. In other words, rather than being compared to other test takers, a candidate’s performance on the proficiency test is measured against the rating criterion standard [Fulcher, 2010].

Structure of the Multilevel Test: context of speaking part

According to L.Bachman and A.Palmer, competence assessments demonstrate that the four primary language abilities are available to individuals of all levels willing to take them [Bachman & Palmer, 2010]. Likewise, the multi-level test is divided into four parts – speaking, listening, reading, and writing – to evaluate a candidate’s English language command comprehensively.

The multilevel speaking test is conducted as an oral interview using a computer. Test questions appear on the computer screen, and the test taker is given time to think and answer. The test is a certified examination procedure that awards test takers with one of three English proficiency levels (B1, B2, or C1). The lowest possible level to pass the test is a B1.

Discourse management, grammar range and accuracy, vocabulary range and appropriacy, and pronunciation are the four aspects of candidates’ speaking skills that must be evaluated in the Multilevel Speaking Test. The objective of the oral discourse section is to evaluate a candidate’s ability to communicate correctly and fluently in a foreign language. This part, which is divided into three parts, has been given 15 minutes.

Part 1 (1-6 questions). Unprepared discourse on personal interests or common, everyday topics; the goal is to provide knowledge and express opinions about important topics (This section lasts four minutes).

Part 2 (Question 7). The applicant must compose a two-minute monologue based on the specified difficult situation. During the speech, the candidate must describe his or her approach to the challenge (a solution to the problem). Questions should be based on the candidate’s personal interests. The candidate has one minute to prepare.

Part 3 (Questions 8–12). The candidate will be asked supplementary questions related to the topic of Part 2 (the monologue). The candidate receives no preparation time.

METHODS

In pursuit of a thorough examination of the reliability of the Multilevel Speaking Test conducted by the Knowledge and Skills Assessment Agency under the Ministry of Higher Education, Science and Innovation of the Republic of Uzbekistan, this

study employs a methodological approach centred on the analysis of sample tests, test formats, and scoring rules. The investigation is conducted by meticulously reviewing the official website of the Knowledge and Skills Assessment Agency, accessible at <https://uzbmb.uz/>. By navigating the official online platform, this research aims to gain insights into the inherent characteristics and structures of the Multilevel Speaking Test.

The methodological framework adopted for this study involves systematically exploring the sample tests available on the official website. Due to inherent constraints in space and time, the research will focus on a carefully selected subset of sample tests, test formats, and scoring rules. This deliberate selection ensures a targeted and in-depth analysis while acknowledging the practical limitations of the extensive examination of the entire test corpus. The chosen sample tests will be scrutinised for their representative nature, reflecting the diversity of language proficiency levels and communicative situations that the Multilevel Speaking Test aims to assess. The analysis will delve into the linguistic and contextual elements embedded within these samples, seeking to identify patterns, consistencies, and potential areas of concern related to the reliability of the speaking test.

Simultaneously, the study will investigate the various test formats employed within the Multilevel Speaking Test. Understanding the structure and organisation of the test is crucial for evaluating its reliability. This analysis encompasses exploring task types, time allocations, and the overall coherence of the test design. By dissecting these components, the study aims to uncover insights into how effectively the test measures oral proficiency across different language levels. Moreover, examining scoring rules constitutes a pivotal aspect of this research. The scoring criteria play a crucial role in determining the consistency and fairness of evaluations. This study will assess the clarity and specificity of the scoring rubrics, aiming to ascertain the extent to which they provide guidance to raters and contribute to the overall reliability of the test results.

The appendices will include a comprehensive list of all examined sample tests, test formats, and scoring rules to ensure transparency and facilitate further reference. This supplementary material will be a valuable resource for readers seeking a detailed understanding of the specific aspects investigated during the study.

Data analysis

Relevant information on sample tests, test formats, and scoring rules will be systematically extracted and analyzed. The goal is to gain insights into the structure and design of the Multilevel Speaking Test. Furthermore, the study will draw upon existing literature, particularly the work of L.Bachman and A.Palmer [Bachman & Palmer, 2010], to understand the factors influencing the Multilevel Speaking Test's reliability. Special attention will be given to two key factors: inter-rater reliability and testing conditions. The research will explore the degree of agreement between different raters in scoring the Multilevel Speaking Test. This assessment is crucial for understanding the consistency of scores assigned by different evaluators. Finally, the study will examine the testing conditions outlined in the literature to determine their impact on the reliability of the Multilevel Speaking Test. Factors, such as standardized

procedures, environment, and participant characteristics will be considered.

The article will conclude by synthesizing the findings from the analysis of sample tests, test formats, and scoring rules. Additionally, the evaluation of reliability factors, based on the literature review, will contribute to a comprehensive understanding of the Multilevel Speaking Test.

The concluding section will answer the research question by integrating insights from the website analysis and reliability evaluation. It will also discuss any limitations of the study and suggestions for future research.

By adopting these methods, the study aims to thoroughly examine the Multilevel Speaking Test, offering valuable insights for educators and policymakers involved in language assessment.

RESULTS

Analysis of rater variability

The following analysis will focus on two aspects of test reliability. Due to the subjective nature of the marking process, there will always be some degree of marker inconsistency on the Multilevel Speaking Test. In the parts mentioned above of this article, there are two distinct forms of grading: inter-rater and intra-rater [Karim & Haq, 2014]. There is high intra-rater reliability when the same rater repeatedly evaluates the same performance under the same conditions, and high inter-rater reliability when multiple raters consistently agree on the same results [Hoffman et al., 2014]. Since two raters are in the Multilevel Speaking Test, inter-rater reliability is an essential factor. One rater conducts the interview via computer; another listens to the recording later that day and assigns a score depending on what they hear. The administration typically takes the average of the two ratings when deciding. From this perspective, it is evident that the test can be reliable because it is less subjective. In other words, most research on oral testing recommends employing at least two raters to avoid the potential for bias created by a single rater. In other words, most research on oral testing advises using at least two raters to prevent any potential bias that a single rater may introduce [De Charruf, 1984].

However, the behavior of the raters themselves can occasionally have a significant impact on the final scores given [Xi & Mollaun, 2009]. According to A.Brown and K.Hill, 'difficult interviewers' and 'easy interviewers' are two distinct categories of test raters. Challenging interviewers emphasise the sophisticated skills of speculating and defending viewpoints. They may, for instance, ask candidates another crucial question before they have finished the current one. As opposed to this, easy interviewers ask candidates straightforward, cost-effective questions and rarely waste candidates' time with contentious ones. They may also engage in scaffolding activities, such as asking follow-up questions [Brown & Hill, 1998].

Although there are two raters in the multilevel test, most test takers still feel they have reasons to dispute their scores and the testing environment. Therefore, training raters is a crucial step in improving the reliability of the examination. Wilkinson claimed that training raters was necessary for consistent grading [Fulcher & Davidson, 2007].

There is no way to eliminate differences in severity between raters; however, rater training studies show that training does reduce random error in rating. Yet, from my own experience, in the Knowledge and Skills Assessment Agency, rater training is not always well organized; some raters are instructed online and begin their careers with limited knowledge and experience. I have completed one of the rater training courses offered by the agency in Uzbekistan in 2021 and have acquired some knowledge of evaluating the speaking ability of test participants. However, I was unsatisfied with my experience and did not feel prepared to evaluate test takers' knowledge due to a lack of training. Due to these issues in terms of training courses, it may be evident that the 'multilevel test' is not very reliable. For higher inter-rater reliability coefficients, the administration should organize additional training courses and periodically train its raters [Fulcher, 2010], as acceptable levels of inter-rater correlations require extensive training and close monitoring of examiners [Galaczi, 2010].

Moreover, cultural expectations under test conditions contribute significantly to the reliability of the evaluation. G.Fulcher noted that culture-specific examiners may emphasise specific linguistic processes [Fulcher, 2010]. Most test-takers who were questioned personally about their speaking exam hold a negative view of the examiners because they believe the Uzbek raters are biased when assigning scores. G.Wigglesworth claimed that raters were more consistent with their ratings when they received feedback on their evaluations between rating sessions [Wigglesworth, 1994]. Thus, assessment program administrators should uncover rater biases to build rater training programs that address and minimise recognised issues.

Analysis of test conditions

The testing environment might also impact the test's reliability and quality. L.Bachman says that the test setting, the test participants' appropriateness, and the testing method are the most critical factors in a test's reliability [Bachman & Palmer, 2010]. The administration of a multi-level computerised speaking test may have positive and negative consequences. According to E.Galaczi, computer-delivered (CD) speaking tests are highly reliable, one of their positive qualities. She argues that the questions and how they are asked are always the same in CD-speaking tests, and there is no room for error. The two primary sources of variation that threaten testing reliability are interviewer variability's influence on the testing process and scoring. Thus, CD oral assessment reduces variability and eliminates causes of measuring error in both contexts [Galaczi, 2010]. However, when candidates have trouble with complex tasks or do not hear a question on the first try, the computer test environment and time constraints might cause stress and anxiety for the candidate [Nakatsuhara et al., 2017]. From this perspective, determining whether the multilevel test is reliable is difficult. Still, we are more likely to assert that the speaking exam, administered by computer, is more reliable than a face-to-face examination.

In addition, using a single-task format, such as an interview in the Multilevel Test, to evaluate an applicant's communicative ability may not be ideal, especially if the candidate is nervous or uncomfortable in such an official and controlled setting [Li, 2019]. A test-taker may do exceptionally well on role-playing assignments

but poorly on the stationary question mode. According to A. Hughes, incorporating additional items will improve the reliability of a test [Hughes, 2003; 44]. He also proposes that the second thing should be distinct from the first to acquire further insight. Furthermore, the exam website does not provide enough sample tests, making it difficult for test-takers with limited opportunities for preparation. If candidates are unfamiliar with any aspect of a test, they are likely to perform less well than they would otherwise [Hughes, 2003].

DISCUSSION

The article discusses the Multilevel Speaking Test and evaluates its reliability concerning inter-rater reliability, testing conditions, and potential areas for improvement. The presence of two raters in the test contributes to its reliability, aligning with the standard recommendation in oral testing research to involve multiple raters to minimise subjectivity. However, the article highlights a significant concern – the shortage of rater training. This raises the importance of addressing potential biases and implementing structured rater training programs to enhance the overall reliability of the assessment.

The testing conditions, specifically the administration through a computer, are noted to reduce test-taker anxiety, a positive aspect that contributes to the reliability of the test results. However, a notable drawback is identified – the insufficient availability of sample questions on the exam website. This limitation may hinder test-takers' preparation, potentially impacting their performance and the overall fairness of the assessment. As the article suggests, future developments should provide more examples for candidates to practice, ensuring a more comprehensive and equitable testing experience.

Moreover, the article recommends incorporating additional tasks, such as role-playing or photo description, to assess candidates' language proficiency beyond a single assignment or interview. This recommendation aims to diversify the assessment methods, potentially enhancing the overall reliability of the Multilevel Speaking Test.

Recommendations:

1. *Rater training programs.* Administrators of the Multilevel Speaking Test should prioritise developing and implementing rater training programs. These programs should explicitly target and minimise biases, ensuring that raters are adequately prepared to assess test-takers consistently and fairly.

2. *Increased availability of sample questions.* To improve test-taker preparation and the overall reliability of the test, the examination program should work on providing a more extensive set of sample questions on the exam website. This will empower candidates to familiarize themselves with the test format and adequately prepare for the assessment.

3. *Diversification of assessment tasks.* Consider incorporating additional tasks, such as role-playing or photo description, into the assessment process. This diversification can provide a more comprehensive evaluation of candidates' language proficiency, contributing to increased reliability.

4. *Continuous improvement.* Recognizing that the Multilevel Speaking Test is relatively new, administrators should commit to ongoing improvement efforts. Regular updates to testing materials, ongoing rater training, and soliciting feedback from test-takers can contribute to refining and enhancing the test's overall reliability over time.

By addressing these recommendations, the administrators of the Multilevel Speaking Test can make significant strides in improving the test's reliability, validity, and overall effectiveness as an assessment tool.

CONCLUSION

In conclusion, the Multilevel Speaking Test stands out for its commendable reliability, particularly in inter-rater consistency, thanks to including two raters in the evaluation process. This dual-rater approach adds a layer of objectivity to the assessment, contributing to the overall credibility of the results. However, promptly addressing the observed deficiency in rater training is crucial. The lack of comprehensive training for raters may inadvertently introduce biases, compromising the fairness and accuracy of the evaluation. Urgent attention to rater training is necessary to ensure a standardised and unbiased assessment process.

Furthermore, introducing a computer-administered format for the Multilevel Speaking Test marks a positive shift by alleviating test-taker anxiety. This technological adaptation enhances the testing experience and contributes to a more controlled and consistent evaluation environment. Despite this advancement, it is imperative to acknowledge and rectify the existing challenges, particularly concerning the scarcity of sample questions available on the exam website. The limited pool of practice materials poses a significant obstacle to adequate test preparation, hindering candidates from familiarising themselves with the specific test format and requirements.

Future improvements should prioritize expanding sample materials to enhance the overall effectiveness of the Multilevel Speaking Test. This expansion would serve the dual purpose of providing candidates with ample opportunities to practice and allowing them to gain a comprehensive understanding of the test structure. Access to more extensive sample questions would empower test-takers to develop their skills and build confidence, ultimately contributing to a more equitable and inclusive testing experience.

In navigating the future developments of the Multilevel Speaking Test, stakeholders must collaborate to address the identified challenges and build upon the test's strengths. Investing in robust rater training programs is essential to maintain the reliability of the assessment process, while concurrently expanding the pool of sample materials is crucial for fostering effective test preparation. By addressing these key areas, the Multilevel Speaking Test can continue to evolve as a reliable and accessible tool for evaluating language proficiency, ensuring fairness, and promoting confidence among test-takers.

REFERENCES

1. Resolution No. 321 of the Cabinet of Ministers of the Republic of Uzbekistan. (2021, May 19). *On measures for effective organization of promotion of learning foreign languages*. <https://lex.uz/docs/-5431845>.

2. Bachman, L.F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.
3. Bachman, L.F., & Palmer, A.S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.
4. Bachman, L.F., & Palmer, A.S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.
5. Blais, J., Forth, A.E., & Hare, R.D. (2017). Examining the interrater reliability of the Hare Psychopathy Checklist-Revised across a large sample of trained raters. *Psychological Assessment*, 29(6), 762–775. <https://doi.org/10.1037/pas0000455>.
6. Brown, A., & Hill, K. (1998). Interviewer style and candidate performance in the IELTS oral interview. *IELTS Research Reports*, 1, 1–19.
7. Chiedu, R.E., & Omenogor, H.D. (2014). The concept of reliability in language testing: issues and solutions. *Journal of Resourcefulness and Distinction*, 8(1), 1–9.
8. De Charruf, L.F. (1984). Oral testing. *MEXTESOL Journal*, 8(2), 63–79.
9. Fulcher, G. (2010). *Practical language testing*. London: Routledge.
10. Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. London: Routledge.
11. Galaczi, E.D. (2010). Face-to-face and computer-based assessment of speaking: Challenges and opportunities. In L.Araújo (Ed.), *Computer-Based Assessment of Foreign Language Speaking Skills* (pp. 29–51). Luxembourg, LU: Publications Office of the European Union.
12. Green, A. (2014). *Exploring language assessment and testing: Language in action*. London: Routledge.
13. Hoffman, L., Wilson, L., Copley, A. (2014). The reliability of a severity rating scale to measure stuttering in an unfamiliar language. *International Journal of Speech-Language Pathology*, 16(3), 317–326. <https://doi.org/10.3109/17549507.2014.898097>.
14. Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge: Cambridge University Press.
15. Karim, S., & Haq, N. (2014). An assessment of IELTS speaking test. *International Journal of Evaluation and Research in Education*, 3(3), 152–157.
16. Khan, R.M.I., Kumar, T., Benyo, A. (2022). The reliability analysis of speaking test in computer-assisted language learning (CALL) environment. *Education Research International*, 2022, 1–10. <https://doi.org/10.1155/2022/8984330>.
17. Lee, H. (2017). The effects of rater's familiarity with test taker's L1 in assessing accentedness and comprehensibility of independent speaking tasks. *SNU Working Papers in English Linguistics and Language*, 15, 93–111.
18. Li, J. (2019). An evaluation of IELTS speaking test. *Open Access Library Journal*, 6(12), 1–17. <https://doi.org/10.4236/oalib.1105935>.
19. Luoma, S. (2004). *Assessing speaking*. Cambridge: Cambridge University Press.
20. Nakamura, Y. (1996). Assessment of English speaking ability. *Journal of Humanities and Natural Sciences*, 102, 25–53.
21. Nakatsuhara, F., Inoue, C., Berry, V., & Galaczi, E.D. (2017). Exploring the use of video-conferencing technology in the assessment of spoken language: A mixed-methods study. *Language Assessment Quarterly*, 14(1), 1–18.
22. O'Mahony, C. (2019). Reliability and validity in language testing—a real conflict?. *Center for English Language Education*, 27, 133–140.
23. Park, M.S. (2020). Rater effects on L2 oral assessment: focusing on accent familiarity of L2 teachers. *Language Assessment Quarterly*, 17(3), 231–243.
24. Quaid, E.D. (2018). Reviewing the IELTS speaking test in East Asia: Theoretical and practice-based insights. *Language Testing in Asia*, 8(1), 1–9. <https://doi.org/10.1186/s40468-018-0056-5>.
25. Rohde, A., McCracken, M., Worrall, L. (2022). Inter-rater reliability, intra-rater reliability and internal consistency of the Brisbane Evidence-Based Language Test. *Disability and*

- Rehabilitation*, 44(4), 637–645. <https://doi.org/10.1080/09638288.2020.1776774>.
26. Schoonen, R. (2020). Measurement generalizability: Considerations on reliability and validity. In P. Winke & T. Brunfaut (Eds.), *The Routledge Handbook of Second Language Acquisition and Language Testing* (pp. 483–494). New York: Routledge.
 27. Wigglesworth, G. (1994). Patterns of rater behaviour in the assessment of an oral interaction test. *Australian Review of Applied Linguistics*, 17(2), 77–103. <https://doi.org/10.1075/aral.17.2.04wig>.
 28. Winke, P., & Gass, S. (2013). The influence of second language experience and accent familiarity on oral proficiency rating: A qualitative investigation. *Tesol Quarterly*, 47(4), 762–789.
 29. Xi, X., & Mollaun, P. (2009). How do raters from India perform in scoring the TOEFL iBT™ speaking section and what kind of training helps?. *ETS Research Report Series*, 2009(2), i-37. <https://doi.org/10.1002/j.2333-8504.2009.tb02188.x>.
 30. Zhao, K. (2017). Investigating the effects of rater's second language learning background and familiarity with test-taker's first language on speaking test scores. *MA thes.* USA: Brigham Young University.