Revista de Lenguas Modernas, N.º 40, 2025 / 01-18
ISSN electrónico: 2215-5643
ISSN impreso: 1659-1933
DOI: 10.15517/7EXXMD35

Assessing Validity of the UCR Standardized English Automatic Oral Test in the MEP Monitoring Test

Evaluación de la validez de la Prueba Oral Automática Estandarizada de Inglés de la UCR en la Prueba de Seguimiento del MEP

Mag. Mónica Bradley Harvey
Universidad de Costa Rica, San José, Costa Rica
monica.bradley@ucr.ac.cr
ORCID: 0000-0003-0755-6389

Mag. Héctor Alvarado Picado
Universidad de Costa Rica, San José, Costa Rica
hector.alvaradopicado@ucr.ac.cr
ORCID: 0009-0004-8248-7904

Mag. Alejandro Fallas Godinez
Université Laval, Québec, Canada
jose-alejandro.fallas-godinez.1@ulaval.ca
ORCID: 0000-0002-9767-3126

Abstract

The Program of the Evaluation of Foreign Languages (PELEx) from the University of Costa Rica is at the forefront of the application of automatic speech recognition (ASR) and artificial intelligence (AI) in the assessment of foreign languages within Latin America (LATAM), creating a need for continual research, data collection, and reliability assessment to collect further evidence of validity for these tests. This paper compares the results obtained during the application of Costa Rica's Ministry of Public Education (MEP) Monitoring Test in 2023 by PELEx of the University of Costa Rica (UCR) with those of 7,454 Costa Rican public high school students through a statistical sampling of the tests, compared with the grading of a certified human tester from PELEx, to produce evidence of the test's concurrent validity—specifically, the degree of agreement between

Recepción: 07-03-2024 Aceptación: 29-07-25

human and AI scores. It offers an analysis of the ratings as well as recommendations for further validation research and continued collection of evidence of construct validity and scoring reliability for the AI-based oral exam.

Keywords: Artificial Intelligence, Foreign Language Evaluation, Test Validity, Speaking Assessment

Resumen

El Programa de Evaluación de Lenguas Extranjeras (PELEx) de la Universidad de Costa Rica está a la vanguardia en la aplicación del reconocimiento automático de voz (ASR, por sus siglas en inglés) y la inteligencia artificial (IA) en la evaluación de lenguas extranjeras dentro de América Latina (LATAM), lo cual genera la necesidad de una investigación continua, recopilación de datos y evaluación de la confiabilidad para recopilar más evidencias de validez de estas pruebas. Este trabajo compara los resultados obtenidos durante la aplicación de la Prueba de Seguimiento del Ministerio de Educación Pública (MEP) de Costa Rica en 2023 por el PELEx de la Universidad de Costa Rica (UCR) a 7454 estudiantes de secundaria pública costarricenses, mediante un muestreo estadístico de las pruebas, en comparación con la calificación de un evaluador humano certificado del PELEx, con el fin de aportar evidencia de validez concurrente, es decir, el grado de coincidencia entre la calificación humana y la de la IA. Se ofrece un análisis de los puntajes y recomendaciones para investigaciones futuras orientadas a la validación continua y a la recopilación de evidencias de validez de constructo y confiabilidad de calificación del examen oral automatizado.

Palabras clave: Inteligencia Artificial, Evaluación de Lenguas Extranjeras, Validez de Pruebas, Evaluación del Habla

Introduction

Since the creation of the Turing Test in 1950, artificial intelligence (AI) has fascinated artists, scientists, and educators alike, prompting a proliferation of research and development in areas such as machine learning, deep learning, and generative AI (Agrawal et al., 2023). In the last decade, from Amazon's Alexa to Apple's Siri to applications such as ChatGPT, AI and Automatic Speech Recognition (ASR) systems have expanded global capitalist technological markets. ASR refers to technologies that allow for human interaction with computational interfaces; although these systems still require more development to fully "pass" the Turing Test—where humans can no longer distinguish between human and artificial intelligence—research into ASR and its applications has continually progressed since the 1980s (Das et al., 2022; Shabtai, 2010; Waibel, 1990; Johnstone & Altmann, 1984).

Within Latin America (LATAM), investment in and application of AI—of which ASR is a central component—carries significant political and socioeconomic implications. A report by Economist Impact (2022), supported by Google, reveals that global AI investment has increased over 9,000%, from less than one billion dollars in 2010 to $78 billion in 2021. In LATAM specifically, AI is projected to raise regional GDP by over 5% by 2030 (p. 5). The region's top priorities in AI include "cultivating local talent, strengthening technological infrastructure, and ensuring that AI is deployed in a responsible manner" (Economist Impact, 2022, p. 5). This aligns with an OECD regional report identifying Costa Rica as one of the few LATAM countries to adopt national AI strategies and officially endorse the OECD AI Principles (OECD, 2022). As with other emerging technologies, educational institutions have increasingly integrated AI within foreign language education (Mkrttchian, 2021; Rizvi et al., 2023; Zhai & Wibowo, 2023).

In this context, Costa Rican public institutions—particularly universities, language testing programs, and the Ministry of Public Education (MEP)—must invest in AI, foster local innovation, and build the necessary infrastructure to ensure ethical and effective implementation. Within education, this includes developing and validating AI-based tools for language assessment as part of broader efforts to strengthen public education policy and practice. In Costa Rica, this task holds particular urgency. Research reveals that most public secondary students do not meet expected levels of English proficiency, especially in productive skills such as speaking (Araya Garita, 2021). Additionally, significant structural gaps persist between urban and rural schools, underscoring the need for equitable and context-sensitive evaluation tools (Quesada et al., 2023). AI-based oral assessments offer a promising yet complex response to these challenges, requiring rigorous, localized empirical validation to ensure fairness and reliability.

Yet despite advances in ASR and AI, language testing utilizing AI has become both a productive and contested site of assessment, as questions of validity, bias, and accessibility continue to emerge (Hao, 2020; Swiecki et al., 2022). The growing use of AI in language assessment has intensified

debates about the need for rigorous validation processes to ensure testing equity (Kunnan, 2000; McNamara & Roever, 2006). Major international exams such as the Pearson Test of English (PTE) and TOEFL iBT claim that AI reduces bias, while others argue it may increase it (Hao, 2020; Nakatsuhara & Berry, 2021). Anecdotally, some native speakers report passing all other components of the PTE but failing the computerized speaking section (The Guardian, 2017). While testing with human judges also reveals inconsistencies (Witzigmann & Sachse, 2021; Kang et al., 2019), major validity issues—such as native speakers scoring abnormally low—occur more frequently with AI-based systems due to issues like accent variation, audio quality, or breathing interference, which human raters can typically interpret with greater nuance.

Nevertheless, the application of AI in education continues to expand, as demonstrated by works such as *Robot-Proof: Higher Education in the Age of Artificial Intelligence* (Aoun, 2017), proceedings from Artificial Intelligence in Education (AIED) conferences (Wang et al., 2023; Rodrigo et al., 2022; Roll et al., 2021), and early explorations like *Artificial Intelligence in Second Language Learning* (Dodigovic, 2005). As with any standardized evaluative instrument, testing institutions must regularly examine the reliability and validity of speaking constructs to ensure accurate results and make necessary improvements (Cizek et al., 2008; Weir, 2005; Schilling, 2004). As such, this paper situates the application of the POA-IE-UCR in the context of Latin America and Costa Rica and provides an analysis of the accuracy of the results obtained during the largest application of this test to date: the 2023 MEP Monitoring Test, which evaluated 7,454 Costa Rican public secondary students. The study offers recommendations for improving the reliability and validity of the test and contributing to the broader conversation on AI-based assessment in the Global South.

Costa Rican National Context

The Costa Rican educational system includes a wide range of both public and private institutions from preschool through university. This study focuses specifically on the evaluation of English language proficiency in secondary schools within the Ministry of Public Education (MEP) system. Research consistently shows that Costa Rican secondary students struggle to reach the expected B1 level of English proficiency, particularly in oral production, underscoring the need for valid and large-scale assessment tools that can measure productive skills (Araya Garita, 2021; Quesada et al., 2023).

Prior to 2023, standardized language assessments conducted by PELEx in 2019, 2021, and 2022 evaluated only the receptive skills of listening and reading for more than 60,000 students through the Linguistic Proficiency Exam (Prueba de Dominio Lingüístico, or PDL), but did not assess the productive skills of writing or speaking (PELEx, n.d.; W. Araya, personal communication, October 2023). These evaluations reached nearly every public secondary school in the country, across both rural and urban areas, with considerable investment to ensure access to the technology needed

for administration. However, applying large-scale oral assessments based on human scoring was logistically and financially unfeasible. The number of certified oral raters needed to evaluate such a vast student population exceeded both the available personnel and the testing budget of a low-to-middle-income country.

In response to this limitation, and after several smaller pilot applications, PELEx and MEP implemented the first national pilot of the UCR Standardized English Automatic Oral Test (POA-IE-UCR) in the 2023 Monitoring Test. This test assessed all four language competencies—writing, speaking, reading, and listening—and marked the first time a standardized speaking test using AI was applied at scale in Costa Rican public secondary education (PELEx, n.d.). The use of AI-based oral assessment allowed the program to reach a broader population with significantly reduced costs while maintaining a focus on alignment with national curriculum goals. By reducing dependence on costly international testing systems and introducing scalable, locally developed evaluation tools, this innovation strengthens Costa Rica's capacity for context-sensitive language assessment. The current study examines the validity of the speaking component of this test, with a particular focus on evidence based on internal structure and testing consequences, as defined in educational assessment frameworks (Brown & Abeywickrama, 2019).

*The Standardized Eng*lish Automatic Oral Test (POA-IE-UC*R)*

The Program for the Evaluation of Foreign Languages (PELEx), housed within the School of Modern Languages (ELM) at the University of Costa Rica (UCR), developed the POA-IE-UCR, a standardized, adaptable speaking test that incorporates artificial intelligence (AI) technology in collaboration with the company Speechace. Speechace provides what it describes as "the first and only speech recognition API designed for evaluating and giving feedback on pronunciation and fluency" within educational settings (Speechace, 2023). The test integrates realistic human avatars that serve as examiners, posing localized, culturally appropriate questions specifically tailored for learners in Costa Rica and Latin America. PELEx oral assessment specialists design the test prompts with careful attention to the targeted population. These prompts typically include items that elicit description, narration, and argumentation, depending on the expected language proficiency level. The test content and scoring criteria are grounded in internationally recognized frameworks such as the American Council on the Teaching of Foreign Languages (ACTFL) Proficiency Guidelines (ACTFL, 2012) and the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2020), while also aligning with the Costa Rican Ministry of Public Education (MEP) national English curriculum. This ensures the test remains both globally benchmarked and locally relevant.

MEP Monitoring Test 2023

In 2023, the POA-IE-UCR was administered as part of Costa Rica's MEP Monitoring Test to a sample of 7,454 secondary students drawn from diverse schools within the public education

system (PELEx, n.d.; W. Araya, personal communication, October 2023). The sample was selected based on statistical projections developed by MEP's planning team to ensure national representativeness. The test measured students' proficiency in the four main language competencies: speaking, reading, writing, and listening. The speaking section applied four parallel versions of the POA-IE-UCR, each developed specifically for this population in accordance with the MEP curriculum. To administer the test, MEP and PELEx coordinated extensive logistical support to ensure participating schools—many of which lacked adequate technological infrastructure—were equipped with the necessary devices and internet connectivity. Each version of the speaking test consisted of three questions, designed to prompt responses in description, narration, and opinion. Students had 30 seconds to prepare and up to one minute to respond to each prompt, for a total speaking time of approximately three minutes. The questions were delivered by an avatar that both read the question aloud and displayed it on screen. AI-powered scoring through Speechace assessed each student's responses according to four main criteria: pronunciation, fluency, vocabulary, and grammar, along with a holistic CEFR-aligned score. These features allow for scalable, standardized assessment across thousands of test-takers while still reflecting key communicative competencies emphasized in national and international curricula.

Justification

Globally, the use of Artificial Intelligence (AI) and Automatic Speech Recognition (ASR) has opened new possibilities for the evaluation of linguistic competencies and spontaneous speech production. These technologies allow for more affordable, scalable, and accessible testing, especially for large populations of learners. However, despite their potential, AI-based assessment tools remain under development and present known risks related to bias, misrecognition, and limitations in adaptive interaction, particularly in high-stakes contexts. In Latin America, the University of Costa Rica's PELEx program stands at the forefront of AI and ASR integration in foreign language assessment. By developing the POA-IE-UCR oral proficiency test, PELEx has taken a pioneering step toward context-sensitive, cost-effective evaluation in public education. However, to ensure the responsible and equitable use of these technologies, ongoing validation studies are needed to examine how well automated scoring systems align with human evaluation standards and reflect students' true communicative competence. This paper presents a validity study based on evidence from internal structure and testing consequences (Brown & Abeywickrama, 2019). Specifically, it analyzes the results from the 2023 application of the POA-IE-UCR test, which assessed 7,454 Costa Rican public high school students as part of the Ministry of Public Education (MEP) Monitoring Test. Through a statistical sampling of test recordings scored both by the AI system and by certified human raters from PELEx, this study compares

scores to evaluate the reliability and validity of the AI-generated results and identify areas for improvement in test design and implementation.

Literature Review

The integration of Artificial Intelligence (AI) into education has expanded steadily over the past two decades, with increasingly sophisticated tools emerging across applications, websites, and mobile platforms (Dodigovic, 2005; Aoun, 2017). The COVID-19 pandemic significantly accelerated the adoption of AI technologies in education, as students, teachers, and institutions were forced to rapidly adapt to remote learning environments. Tools such as ChatGPT have transformed how students interact with educational content, including the ability to generate written compositions instantaneously.

The area of oral language assessment has seen especially notable advances. Mobile-based speaking tests now use real-time scoring engines to evaluate pronunciation and fluency with high correlations to human raters (Cheng, 2018). For example, the Duolingo English Test incorporates automated scoring for tasks such as "Speak about the Photo," with psychometric validation demonstrating strong reliability (Duolingo, 2023). Similarly, the TOEFL iBT incorporates ETS's SpeechRater®, which automatically scores pronunciation, fluency, and other linguistic features; research shows that combining machine and human ratings improves scoring reliability (Educational Testing Service, 2020; Evanini et al., 2013). Pearson's Versant and PTE Academic tests also employ AI-based scoring systems with documented validity in large-scale testing contexts (Pearson, 2024). However, these systems still require careful review and adjustment, particularly in high-stakes, cross-cultural contexts where validity and fairness are essential (Edmett et al., 2024).

The growing literature on AI-based language assessment examines issues of comprehension, adaptability, bias, and transparency (Mercer & Cannon, 2022). While machine scoring shows efficiency, it lacks the nuanced comprehension and adaptive interaction of human interlocutors. For example, when evaluating open-ended, dialogic speech, AI often struggles to determine the appropriateness of student responses (Yong, 2020). Emotional tone and pragmatic coherence—critical aspects of spoken communication—remain difficult for machines to assess accurately. Therefore, AI cannot fully replace human judgment in all assessment situations. In addition to cognitive limitations, scholars also critique the standardized and inflexible nature of AI assessments. Pirrone (2023) questions the assumption that uniform question sets offer equity, suggesting that teacher-adapted oral exams allow for more authentic assessment. Similarly, Young (2023) notes that AI systems like Sherpa lack interactional adaptability, reducing their effectiveness in dynamic speaking evaluations. Although AI tools can supplement instruction (e.g., tracking student progress or enabling practice outside the classroom), their use in formal assessment still faces challenges of adaptability and fairness (Yu et al., 2022).

Ethical concerns also arise around transparency, cultural bias, and

algorithmic design. Swiecki et al. (2022) argue that automated decision-making often lacks transparency and may marginalize professional expertise. While removing teacher bias is one benefit of machine scoring, it can introduce new risks, especially when assessments are designed and administered by third-party engineers with no knowledge of the students or context. Similarly, Zahi and Wibowo (2023) warns that algorithms often reproduce racial, gendered, and cultural biases embedded in training data, raising serious legal and ethical questions. Ethical frameworks have been proposed to address these risks, but enforceable safeguards remain limited. Narasimhan (2023) adds that GenAI tools trained on biased data can reinforce systemic inequalities through both test design and student use. These challenges also influence test-taker perceptions. Studies by Pearson (Fan et al., 2021; Richardson et al., 2023) reveal polarized opinions: while some students appreciate the lower anxiety associated with machine raters, others find the format inauthentic, overly short, or alienating. Factors such as equipment quality, access to preparation, and socioeconomic background also affect outcomes, highlighting persistent equity concerns even in AI-mediated assessments.

Despite these concerns, AI-based oral assessment is rapidly expanding, including in Latin American contexts. However, empirical research on its use in public secondary education in Latin America remains scarce. In Costa Rica, recent studies underscore persistent disparities in English language teaching and low proficiency outcomes in oral communication (Araya Garita, 2021; Quesada et al., 2023). These findings demonstrate the urgency of evaluating whether AI-based speaking assessments—such as the POA-IE-UCR—are valid, equitable, and contextually appropriate. This study addresses this gap by providing empirical validity evidence—particularly related to internal structure and test consequences (Brown & Abeywickrama, 2019)—from the largest application of an AI-based oral test in Costa Rican public education. By comparing scores assigned by the POA-IE-UCR system with those of certified human raters, the study contributes to both the global debate on AI in assessment and the local development of scalable, just, and pedagogically meaningful evaluation tools.

Methodology

This study uses a quantitative research design based on statistical correlation analysis to examine the reliability and validity of the speaking component of the MEP Monitoring Test (2023). A representative sample of 445 students was selected from the total population of 7,454 Costa Rican public high school students who took the exam. The sample was stratified by CEFR band, as determined by Speechace's AI-generated ratings, to ensure 95% confidence in proportional representation. It includes 33 tests from Pre-A1 (7.33%), 103 from A1 (23.18%), 179 from A2 (40.21%), 93 from B1 (20.84%), 30 from B2 (6.76%), and 7 from C1 (1.68%). A single rater was employed to evaluate the 445 oral proficiency samples from the stratified sample. The selected rater is a certified ACTFL Oral Proficiency Interview (OPI) tester, affiliated with the PELEx

program, and has over ten years of experience in evaluating English oral performance. The purpose of this evaluation was not to produce official proficiency certifications, but to serve as a consistent human benchmark against which to compare the AI-generated scores from Speechace. Using one highly experienced rater ensures consistency across all evaluations, minimizing the variability commonly found in multi-rater studies. Since the primary goal of the study is to analyze agreement between human and machine-generated scores, rather than interrater reliability, this methodological choice strengthens the internal validity of the comparison. Similar approaches have been adopted in prior research on human–machine scoring alignment. This study analyzes concurrent validity by comparing the CEFR levels assigned by the AI system to those assigned by the human rater. Additionally, it examines convergent validity by analyzing the relationships between students' speaking scores and their performance in the other three skills: listening, reading, and writing. Together, these analyses contribute to the test's broader construct validity, and help inform recommendations for the continued development of reliable, fair, and contextually appropriate AI-based assessment tools in Costa Rican public education.

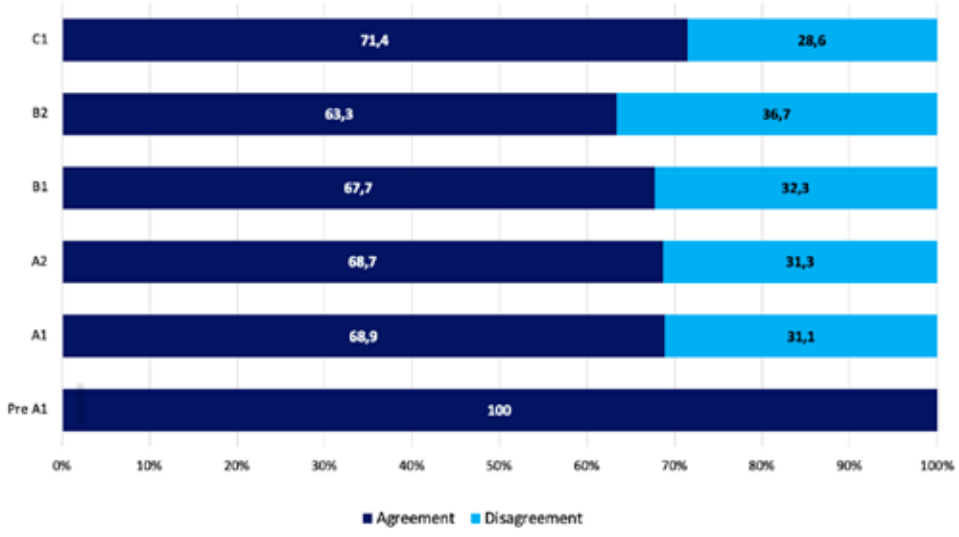
Results and Discussion

Analysis of AI and Human Judge

The comparison between the certified human rater and Speechace's AI technology revealed a 73% agreement across the six stratified CEFR bands, and a 71% overall agreement based on total matching scores. The highest agreement occurred at the Pre-A1 level, where there was a 100% match between the human and AI-assigned ratings. In contrast, the lowest agreement was found at the B2 level, with a disagreement rate of 36.7% (Figure 1). The A1 and A2 levels showed similar disagreement rates of 31.1% and 31.3%, respectively, while B1 exhibited a slightly higher rate of 32.3%. These three bands (A1, A2, and B1) represented the largest portion of the student sample.

Although some disagreement was observed, it is important to recognize that variability among human raters is also common, influenced by factors such as background, training, native speaker status, and unconscious social bias (Kang et al., 2019). As such, a certain degree of scoring discrepancy is expected—whether between human raters or between human and machine assessments. Despite the observed variation, the findings demonstrate a strong overall correlation between human and AI-generated scores, particularly in the lower CEFR bands. These results provide initial support for the concurrent validity of the POA-IE-UCR test and suggest that, with further calibration and refinement, AI-based scoring can offer a reliable and scalable option for oral language assessment in the Costa Rican public education system.

Figure 1.
Percentage Distributions of Judgement Agreement with the Artificial Intelligence Marking by CERF Band



Moreover, when examining the overall distribution of CEFR band scores across all 445 students—as an indicator of the speaking proficiency of the sampled MEP student body rather than of the functionality of individual items—the results show even greater similarity between the human and AI-generated ratings. The largest number of discrepancies appeared in the A2 band, which also had the highest number of test-takers, followed by Pre-A1. This discrepancy can be explained by the 12 tests in which the human rater assigned a Pre-A1 level due to the test-takers either responding in Spanish or merely reading the prompt aloud without producing meaningful language. In such cases, the human rater judged the task as incomplete or invalid. In contrast, the AI system frequently assigned A2-level scores to these same tests because it interpreted the reading of the cue card as a completed response. Similar mismatches occurred in other bands where students read the cue card rather than answering the question, resulting in artificially inflated AI scores.

To address this issue, Speechace is currently developing a "task completion" category that will enable the system to detect when students fail to provide an original spoken response. This feature aims to filter out non-responses or misinterpreted readings from being scored as successful task completions. Based on the current data, the authors anticipate that future iterations of the POA-IE-UCR test—once integrated with this task completion feature—will demonstrate an even higher level of accuracy in assigning CEFR bands at the population level.

Figure 2.
Overall Comparison Between AI Results and PELEx Human Tester

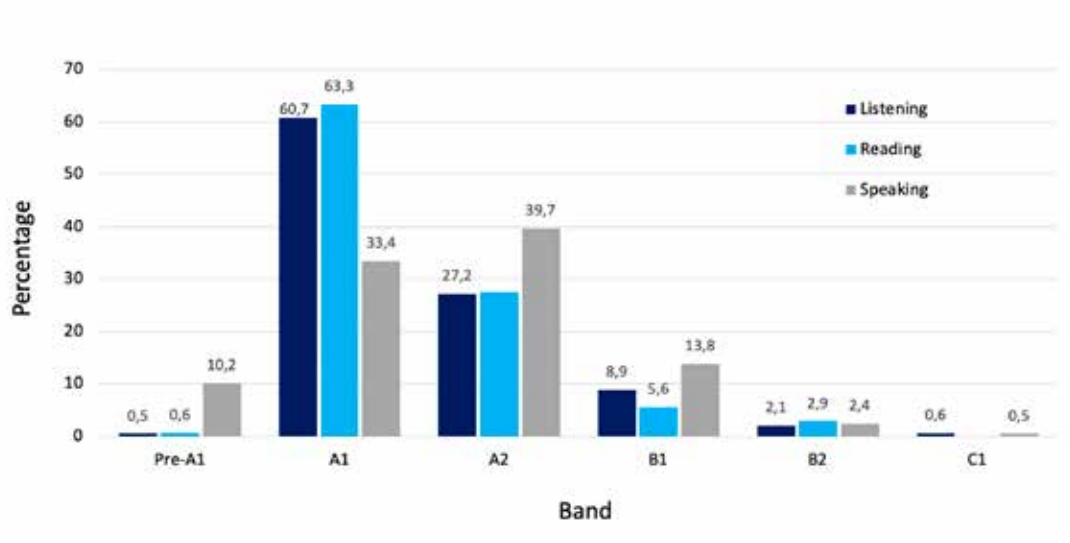
		Human Rater						
		Pre-A1	A1	A2	B1	B2	C1	Invalid
Speechace AI	Pre-A1	13						17
	A1	28	71	3				1
	A2		37	123	7			12
	B1			23	37	7		
	B2				11	19		
	C1					2	5	

Overall, the human rater assigned lower CEFR ratings than the AI in every band except Pre-A1, largely due to the presence of unratable or invalid responses that were automatically classified as A1 or A2 by the system. As further research is conducted to correlate human and AI scores, algorithmic refinements will continue to address these misclassifications and improve overall rating accuracy.

Figure 3 presents a detailed breakdown of AI-generated CEFR ratings compared with those assigned by the human rater. In the Pre-A1 category, the AI identified 30 tests as Pre-A1. However, the human rater confirmed only 13 of these as valid Pre-A1 responses and classified the remaining 17 as invalid (e.g., students reading the prompt or responding in Spanish). For the A1 band, the AI scored 103 tests at this level. The human rater agreed with 73 of these, placed 3 tests in a higher band, and assigned the rest to Pre-A1 or invalid categories.

In the A2 band, Speechace assigned 179 tests, and the human rater agreed with 123. Of the remaining 56 tests, 12 were deemed invalid, 7 were rated one band higher, and 37 were rated one band lower. For B1, the AI scored 93 tests, and the human rater concurred with approximately two-thirds of them; the rest were either downgraded one band or marked invalid. At the B2 and C1 levels, no tests were considered invalid by the human rater, and all disagreement occurred within a single adjacent band—i.e., B2 scores downgraded to B1, and C1 scores downgraded to B2. Importantly, in no case—aside from invalid tests—did the AI and the human rater differ by more than one CEFR level. This consistency marks a notable improvement over previous test iterations and reflects the ongoing collaboration between PELEx human testers and Speechace to refine algorithmic performance and scoring accuracy.

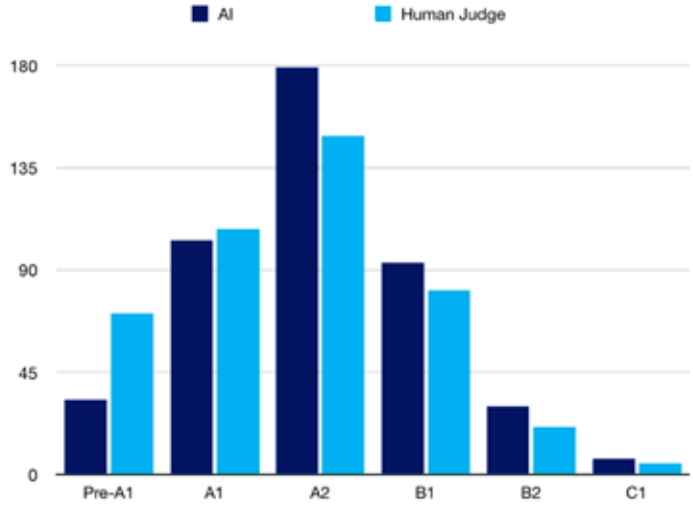
Figure 3.
Confusion Matrix for Itemized Comparison between AI Results and Human Tester



Initial Results Compared to Other Skill Areas

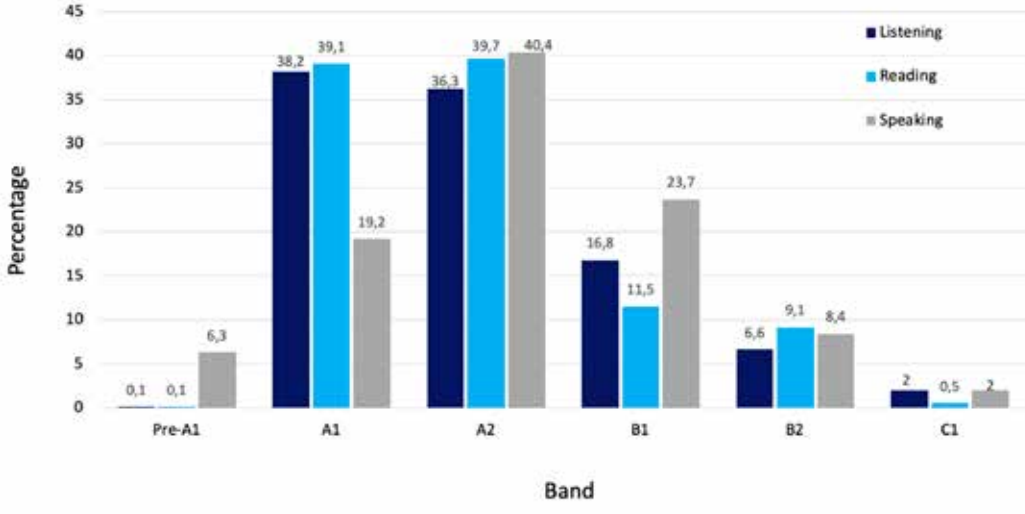
In addition to the comparison between AI and the human tester, the paper also briefly compares these results to the results of all of the oral exams compared to the other skills of reading and listening broken down in 9th grade (Figure 4) and 11th and 12th grades (Figure 5).

Figure 4
Percentage Distributions of the Bands Obtained by Skill: 9th Grade, 2023



Note. *nL* = 2195, *nR* = 2005, *nS* = 2127. Adapted from 2023 *Pruebas de Monitoreo, Ministerio de Educación Pública 2023*, by PELEx, 2023 (https://sites.google.com/view/mep2023/home). In the public domain.

Figure 5
Percentage Distribution of the Bands Obtained by Skill: 11th and 12th Grades, 2023



Note. *nL* = 2195, *nR* = 2005, *nS* = 2127. Adapted from 2023 *Pruebas de Monitoreo, Ministerio de Educación Pública 2023*, by PELEx, 2023 (https://sites.google.com/view/mep2023/home). In the public domain.

While there is not always a direct correlation between receptive and productive language skills, comparing them can offer a general sense of students' linguistic development. This is particularly relevant in Costa Rica, where instructional emphasis often falls more heavily on receptive skills such as reading and listening. In the current data, reading and speaking levels show a notable discrepancy, with virtually no students placed at the Pre-A1 level in reading, while Speechace reported approximately 10% of 9th graders and 6% of 11th and 12th graders at the Pre-A1 level in speaking. Given that productive skills tend to lag behind, especially in public education contexts, these results are not unexpected. If we rely on the human rater's scores—who generally rated lower than the AI—this percentage would likely be even higher. However, it is promising to note that the proportion of students rated at Pre-A1 decreases in higher grades, indicating progress over time.

Interestingly, in the A1 band, the gap between receptive and productive scores becomes even more pronounced. In several cases, students performed better in speaking than in reading, particularly in both 9th and 11th/12th grades. This may reflect improved confidence or fluency due to curriculum changes or test format familiarity. These trends support the interpretation that the human rater's results—lower than the AI's—may more accurately represent students' actual oral proficiency, reinforcing the need for continued human oversight in AI-based assessment systems.

Conclusions and Recommendations

AI-based language assessment tools do not operate independently of the human systems that design, implement, and interpret them. Technologies like Speechace reflect the values, decisions, and biases of their programmers and institutional partners, as well as the diverse experiences of test-takers themselves. This validity study underscores that collaboration between human experts and AI developers is essential to improving the fairness, accuracy, and pedagogical usefulness of these tools. The findings suggest that the POA-IE-UCR oral production test offers a reliable general measure of population-level speaking ability, but should not yet be used in high-stakes assessments, such as placement or certification exams. The risk of misclassification—particularly for invalid responses or lower-proficiency students—requires careful attention to ensure fairness and avoid misuse.

Instead, the technology could be more effectively employed in classroom settings and practice assessments, where students gain familiarity with the test format and receive formative feedback to support their learning. The ongoing development of a "task completion" feature within the Speechace algorithm is a promising step toward filtering out invalid responses and improving rating accuracy. Additionally, adjusting the AI rating thresholds to more closely align with human evaluators will further enhance scoring consistency. Moving forward, the partnership between PELEx and Speechace should continue refining the tool, using input from trained human raters to fine-tune the system. While neither

human nor AI evaluation can guarantee 100% accuracy, each benefits from continual training, calibration, and shared learning. This study offers a first step in the validation of an AI-based oral proficiency test, but future research should expand to include multiple human raters to assess inter-rater reliability and further compare those results with AI-generated scores. As this technology continues to evolve, so too must our commitment to ethical, rigorous, and contextually grounded assessment practices.

References

ACTFL. (2012). Oral proficiency interview. ACTFL. https://www.actfl.org/assessments/postsecondary-assessments/opi

Agrawal, A., Gans, J. S., & Goldfarb, A. (2023, June 12). *The Turing transformation: Artificial intelligence, intelligence augmentation, and skill premiums*. Brookings Institution. https://www.brookings.edu/articles/the-turing-transformation-artificial-intelligence-intelligence-augmentation-and-skill-premiums/

Aoun, J. (2017). *Robot-proof: Higher education in the age of artificial intelligence*. The MIT Press.

Araya Garita, W. (2021). Dominio lingüístico en inglés en estudiantes de secundaria para el año 2019 en Costa Rica. *Revista de Lenguas Modernas*, (34), 1–21. https://doi.org/10.15517/rlm.v0i34.43364

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

Cheng, J. (2018). Realtime scoring of an oral reading assessment on mobile devices. In *Proceedings of Interspeech 2018*(pp. 1621–1625). International Speech Communication Association. https://doi.org/10.21437/Interspeech.2018-34

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement*, *68* (3), 397-412.

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

Das, A. K., Nayak, J., Naik, B., Dutta, S., & Pelusi, D. (2022). *Computational intelligence in pattern recognition proceedings of CIPR 2021*. Springer.

Dodigovic, P. M. (2005). *Artificial intelligence in second language learning: Raising error awareness*. Multilingual Matters.

Duolingo, Inc. (2023, June 12). *Assessing speaking on the Duolingo English Test* (Duolingo Research Report DRR 23 03). https://web.archive.org/web/20240324111501/https://duolingo-testcenter.s3.amazonaws.com/media/resources/speaking-whitepaper.pdf

Economist Impact. (2022). *Seizing the opportunity: The future of AI in Latin America*. https://impact.economist.com/technology-innovation/seizing-opportunity-future-ai-latin-america

Edmett, A., Ichaporia, N., Crompton, H., & Crichton, R. (2024). *Artificial*

intelligence and English language teaching: Preparing for the future (2nd ed.). British Council. https://www.teachingenglish.org.uk/sites/teacheng/files/2024-08/AI_and_ELT_Jul_2024.pdf

Educational Testing Service. (2020). *Reliability and comparability of TOEFL iBT® scores* (TOEFL Research Insight Series, Vol. 3). https://www.ets.org/content/dam/ets-india/pdfs/toefl/toefl-ibt-insight-s1v3.pdf

Evanini, K., Xie, S., & Zechner, K. (2013). Prompt-based content scoring for automated spoken language assessment. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 157–162). Association for Computational Linguistics. https://aclweb.org/anthology/W/W13/W13-1721.pdf

Fan, J., Jin, Y., Kong, X., & Zhang, X. (2021). *How do test takers prepare for computerized speaking tests? A comparative study of the PTE Academic and the CET-SET*. University of Melbourne.

The Guardian. (2017, August 8). Computer says no: Irish vet fails oral English test needed to stay in Australia. *The Guardian*. https://www.theguardian.com/australia-news/2017/aug/08/computer-says-no-irish-vet-fails-oral-english-test-needed-to-stay-in-australia

Hao, K. (2020, April 2). This is how AI bias really happens-and why it's so hard to fix. *MIT Technology Review*. https://www.technologyreview.com/2019/02/04/137602/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/

Johnstone, A., & Altmann, G. (1984). *Automated speech recognition: A framework for research*. Department of Artificial Intelligence, University of Edinburgh.

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

Kunnan, A. J. (2000). Fairness and justice for all. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 1-14). Cambridge University Press.

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Blackwell Publishing.

Mercer, S. H., & Cannon, J. E. (2022). Validity of automated learning progress assessment in English written expression for students with learning difficulties. *Journal for Educational Research Online / Journal für Bildungsforschung Online, 14*(1), 39–60. https://doi.org/10.31244/jero.2022.01.03

Mkrttchian, V. (2021). Impact of Artificial and Natural Intelligence Technologies With Avatar-Based Teaching, Learning, and Research in Russian Modern Universities. In S. Verma & P. Tomar (Eds.), *Impact of AI Technologies on Teaching, Learning, and Research in Higher Education* (pp. 204-221). IGI Global Scientific Publishing. https://doi.org/10.4018/978-1-7998-4763-2.ch013

Nakatsuhara, F., & Berry, V. (2021). Use of innovative technology in oral language assessment. *Assessment in Education: Principles, Policy & Practice*, *28*(4), 343–349.

https://doi.org/10.1080/0969594x.2021.2004530

Narasimhan, S. (2023, October 26). Challenges and opportunities of generative ai in assessments. *Hurix-Digital*. https://web.archive.org/web/20250420171101/https://www.hurix.com/blogs/challenges-and-opportunities-of-generative-ai-in-assessments/

OECD. (2022). *The strategic and responsible use of artificial intelligence in the public sector of Latin America and the Caribbean* (OECD Public Governance Reviews). OECD Publishing. https://doi.org/10.1787/1f334543-en

Pearson. (2024). *PTE Academic score guide*. https://web.archive.org/web/20250219130145/https://www.pearsonpte.com/ctf-assets/yqwtwibiobs4/5Sz9Ur4qbus8AEOQdetkAj/69f6c1f2e2870980740b10a2ea9b467f/pte-academic-test-taker-score-guide-nov-2024-v4.pdf

Pirrone, A. (2023, March 23). *Resist AI by rethinking assessment*. LSE Higher Education Blog. https://blogs.lse.ac.uk/highereducation/2023/03/23/resist-ai-by-rethinking-assessment/

PELEx. (n.d.). *Prueba de Dominio Lingüístico (PDL) para secundaria – MEP*. Universidad de Costa Rica. https://www.pelex.ucr.ac.cr/pruebasMEPDominioSecundaria.html

Quesada, A., Araya, W., & Fallas, J. A. (2023). *La enseñanza y aprendizaje del inglés en secundaria pública costarricense del siglo XXI: Innovaciones, brechas y desafíos*. Informe final para el *Noveno Informe Estado de la Educación*. Programa Estado de la Nación – CONARE.

Richardson, M., Leaton Gray, S., Popov, J., & Maddox, B. (2023). *Computer-based tests and machine marking: Candidates' perceptions and beliefs about the test taking experience*. UCL Discovery. https://discovery.ucl.ac.uk/id/eprint/10177154/

Rizvi, S., Waite, J., & Sentance, S. (2023). Artificial Intelligence Teaching and learning in K-12 from 2019 to 2022: A systematic literature review. *Computers and Education: Artificial Intelligence*, *4*, 100145. https://doi.org/10.1016/j.caeai.2023.100145

Rodrigo, M. M., Matsuda, N., Cristea, A. I., & Dimitrova, V. (Eds.). (2022). *Artificial intelligence in education: 23rd international conference, AIED 2022, proceedings*. Springer.

Roll, I., McNamara, D., Sosnovsky, S., Luckin, R., & Dimitrova, V. (Eds.). (2021). *Artificial intelligence in education: 22nd international conference, AIED 2021, proceedings*. Springer.

Shabtai, N. R. (Ed.) (2010). *Advances in speech recognition*. Sciyo.

Schilling, S. G. (2004). Conceptualizing the validity argument: An alternative approach. *Measurement*, *2*, 178-182.

Speechace. (2023). *Speechace speaking test*. https://www.speechace.com/speaking-test/

Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldonado, R., Lodge, J. M., Milligan, S., Selwyn, N., & Gašević, D. (2022). Assessment in the age of Artificial Intelligence. *Computers and Education: Artificial Intelligence*, *3*, 100075. https://doi.org/10.1016/j.caeai.2022.100075

Waibel, A., & Lee, K.-F. (1990). *Readings in speech recognition*. Morgan Kaufmann.

Wang, N., Rebolledo-Mendez, G., Matsuda, N., Santos, O. C., & Dimitrova, V. (Eds.). (2023). *Artificial intelligence in education: 24th International Conference, AIED 2023, Tokyo, Japan, July 3–7, 2023, proceedings* (Vol. 13916). Springer. https://doi.org/10.1007/978-3-031-36272-9

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Palgrave Macmillan.

Wet, F. de, Walt, C. van, & Niesler, T. (2007). Automatic large-scale oral language proficiency assessment. *Interspeech 2007*. https://doi.org/10.21437/interspeech.2007-90

Witzigmann, S. & Sachse, S. (2021). Diagnostic competencies of prospective teachers of French as a foreign language: judgement of oral language samples. Research in Subject-matter Teaching and Learning (RISTAL), 4(1), 71-87. https://doi.org/10.23770/rt1847

Yong, Q. (2020). Application analysis of artificial intelligence in oral English assessment. *Journal of Physics: Conference Series*, *1533*(3), 032028. https://doi.org/10.1088/1742-6596/1533/3/032028

Young, J. R. (2023, October 5). *As AI chatbots rise, more educators look to oral exams—With high-tech twist*. EdSurge. https://www.edsurge.com/news/2023-10-05-as-ai-chatbots-rise-more-educators-look-to-oral-exams-with-high-tech-twist

Yu, Y., Han, L., Du, X., & Yu, J. (2022). An oral English evaluation model using artificial intelligence method. *Mobile Information Systems*, *2022*, Article 3998886. https://doi.org/10.1155/2022/3998886

Zhai, C., & Wibowo, S. (2023). A systematic review on artificial intelligence dialogue systems for enhancing English as foreign language students' interactional competence in the University. *Computers and Education: Artificial Intelligence*, *4*, 100134. https://doi.org/10.1016/j.caeai.2023.100134