//UCIENCIA/23
CONVENCIÓN CIENTÍFICA INTERNACIONAL

Topic: Curricular design and contextualization of references in the teaching-learning process and the certification of language proficiency.

Evaluación Estandarizada en la Universidad de Costa Rica: PELEx y la Integración de IA

Standardized Evaluation Experience at Universidad de Costa Rica: PELEx and AI integration

Mag. Jennifer Céspedes Araya [1*], Mag. Walter Araya Garita[2]

[1] Universidad de Costa Rica. 2060. jennifer.cespedesaraya@ucr.ac.cr
[2] Universidad de Costa Rica. 2060. walter.arayagarita@ucr.ac.cr

[*] Autor para correspondencia: jennifer.cespedesaraya@ucr.ac.cr

---

Resumen
El Programa de Evaluación en Lengua Extranjera (PELEx) de la Universidad de Costa Rica ha evaluado a más de 80.000 personas desde el 2019 para diagnosticar, ubicar o certificar el dominio lingüístico en idiomas como inglés, francés, alemán, italiano y portugués, entre otros. Así mismo, estas evaluaciones son realizadas por medios tecnológicos que permiten obtener resultados en tiempo real y que también tienen capas de seguridad que garantizan aplicaciones consistentes y con diferentes evidencias de validez. Las poblaciones que se evalúan van desde niños en primera infancia (no saben leer ni escribir) hasta personas cuyo resultado tiene altas consecuencias para hacer un proyecto de vida en traducción e interpretación oficial, carrera diplomática y/o otros ámbitos académicos o laborales. PELEx utiliza inteligencia artificial para desarrollar exámenes adaptativos a través de los cuales cada persona toma una evaluación diferente de acuerdo a su nivel. Además, se utiliza la inteligencia artificial (IA) para evaluar la destreza oral a nivel diagnóstico a diferentes poblaciones. Sin esta herramienta que complementa estos ejercicios, las evaluaciones diagnósticas a gran escala serían imposibles debido al costo que representa tener talento humano calificado para realizarlas. Finalmente, el programa ha desarrollado diferentes formatos que permiten realizar las evaluaciones en una modalidad en línea, híbrida (para aquellas instituciones en las que la conexión a internet no es óptima), y sin conexión a internet. Para estas tres modalidades la interfaz es idéntica, lo que contribuye con la consistencia de los resultados.

Palabras clave: Programa de Evaluación de Lenguas Extranjeras, Universidad de Costa Rica, inteligencia artificial, medios tecnológicos, evaluar

Abstract

The Program for the Evaluation of Foreign Languages (PELEx) at Universidad de Costa Rica has assessed over 80 000 test takers since 2019 in order to diagnose, place or certify their proficiency in languages such as English, French, German, Italian, and Portuguese, among others. In addition, these tests are carried out using technological means that allow for real-time results, and various safety layers that guarantee more consistent applications with various forms of evidence of validity. Among the assessed populations are young children who do not write or read, and adults (such as official translators and interpreters) from various backgrounds and academic and professional fields who take high-stake tests to continue their life projects. PELEx uses artificial intelligence to diagnose the oral skills of different populations. Without this complementary tool, such mass assessments would be impossible due to the elevated cost that the required qualified human talent might have. Finally, the program has developed different test modalities that have allowed an offline, online, and hybrid application (ideal for institutions that have very little Internet connection). The interface is the same for the three formats, adding to the level of consistency of the results.

Keywords: *Program for the Evaluation of Foreign Languages, University of Costa Rica, artificial intelligence, technological means, standardized testing*

Introduction

The need for certifying the proficiency level of different Costa Rican populations in a variety of languages can be traced back to 1990 when the Consejo Nacional de Rectores (CONARE), the national entity in charge of overseeing the appropriate planning and development of the public tertiary education institutions of the country (CONARE, 2023), requested the School of Modern Languages from Universidad de Costa Rica the design of an English certification test for teachers. The need for these kinds of tests continued to evolve until Programa de Evaluación de Lenguas Extranjeras (Program for the Evaluation of Foreign Languages; PELEx) was solidified in 2018. As a result, the program has assessed over 80 000 local and international test takers since 2019 in order to diagnose, place or certify their proficiency in various languages.

Considering the international agreements and commitments, such as Incheon 2030 (UNESCO, 2016), and the national calls to attend to specific educational challenges, such as the creation for a national system of evaluation (CONARE, 2019), PELEx has become an active pioneering contributor to not only addressing the country's needs for certifying its population's language use but also to setting a base for the development of a localized assessment system involving standardized testing that can currently and later provide valuable information to the country's policy makers. With the purpose of expanding on the implications of PELEx's work, this article intends to expand the context in which the program has evolved historically, its impact on the Costa Rican assessment scenario, and its inclusion of technology and artificial intelligence in various steps of the assessment processes.

Methodology

In this article, the impact of PELEx has been analyzed in order to accomplish the following objectives: 1) identify the program's characteristics through the compilation, collection, and discussion of information and data provided regarding its tests and technical programs, 2) determine the importance of such a program and its social impact through the referral and reflection of the journalistic coverage on the tests administered, and 3) reflect on the applications of AI through the exemplifications and discussion of technological resources incorporated into the processes of various assessments.

Results and Discussion

Historical Overview

With more than 25 professionals in language assessment, 16 international certifiers, and several international collaborators, PELEx has evolved significantly in the last 20 years. As displayed in Figure 1, a variety of purposes have been directing the tasks and tests the program carries out, which have allowed it to attend to a spectrum of populations and needs as demanded. From the paper-based test constructions of the 1990s' resources that have transformed into the current design of more complex tests and international applications, the program has been able to

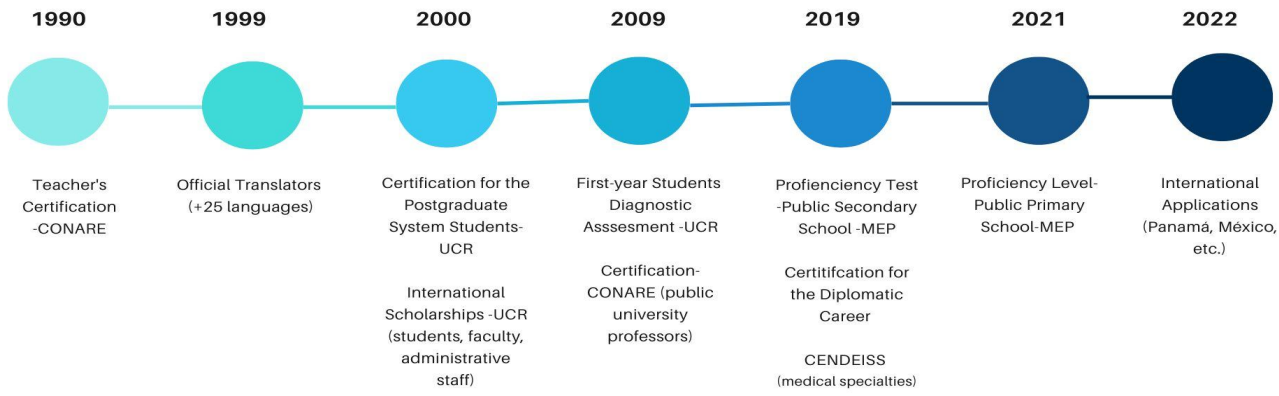

Figure 1. Examples of the general historical transformation of PELEx and its current milestones based on Araya & Quesada (2023)

refine the protocols of test information, design, administration, and result use, enabling a significant growth and an evident improvement in the rigorousness of the tests. As an example, according to Araya and Quesada (2023), the first

certification for CONARE was administered in 2010; in 2011, the first application of that test was carried out, and, in 2013, the first certifications for Oral Proficiency Interview (OPI) were obtained; in 2018, the test was aligned to the CEFR's (Common European Framework of Reference) descriptors; in 2019, computer-based applications started, and, in 2022, artificial intelligence started being used in different stages of the process of the test administration.

As of now, PELEx is working on the design, application, and analysis of different tests targeting Costa Rican and international test takers, responding to national and international requests as seen in Table 1. For this purpose, the target populations are usually selected by stakeholders (the Ministry of Public Education -MEP, or other tertiary institutions, for example); the skills and targeted languages are to be assessed according to the primary needs of the population and national/international interests; the frequency of the test applications depends on the available resources, requirements from stakeholders, and urgency to collect information on the population's proficiency level; and the use of artificial intelligence will depend on all of the above and the resources available.

Test and Target Population	**Skills**	**Target Language(s)**	**First application**	**Frequency**	**Use of Artificial Intelligence**
Exámenes de Régimen Académico UCR teaching faculty	Reading Listening Oral Production Written Production	English French Italian Portuguese German	1990	Every semester	No
Exámenes de Traductores e intérpretes oficiales Ministerio de Relaciones Exteriores y Culto, Costa Rica	Translation	+20 languages	1999	On demand	No
Exámenes de Inglés para Posgrado Postgraduate students entering the Postgraduate Studies System-UCR	Reading	English	2002(+/-)	Every four months	No

Test and Target Population	Skills	Target Language(s)	First application	Frequency	Use of Artificial Intelligence
Exámenes de Dominio Lingüístico General Public	Reading Listening Oral Production Written Production	English	2010	On demand	No
Exámenes de certificación -PAI Students ending the English for Specific Purposes Program -UCR	Reading Listening Oral Production Written Production	English	2010	Every four months	No
Exámenes de Diagnóstico de Primer Ingreso First-year students-UCR	Reading Reading Listening Oral Production	English	2010 2024	Yearly	Not at the beginning The entire test is adaptive now.
UCR-English Placement Test (EPT) International Populations (e.g. Panama, Honduras, Mexico, El Salvador, among others)	Reading Listening Oral Production Written Production	English	2019	On Demand	Yes Adaptive Oral production
Exámenes de Dominio Lingüístico -Secundaria-MEP Senior students from secondary school, Costa Rica	Listening Reading	English	2019	Yearly	Linear (not adaptive) Format: hybrid or offline
Exámenes de Monitoreo Primaria-MEP (monitoring purposes) Primary school students	Reading Listening Oral Production Written Production	English	2021	Yearly	Yes (productive skills)

Test and Target Population	Skills	Target Language(s)	First application	Frequency	Use of Artificial Intelligence
Exámenes de Monitoreo Secundaria (monitoring purposes)	Reading Listening Oral Production	English French Italian	2023	Yearly	Yes (the entire test; adaptive)
Exámenes de Dominio Lingüístico-Primaria-MEP Primary School Students	Reading Listening Oral Production Written Production	English	Only monitoring purposes	—	—
Exámen de Dominio Lingüístico de Francés-Secundaria, MEP 9° and 11° grade students	Listening Reading	French	2023	Yearly	No
Exámen de Dominio Lingüístico de Docentes de Inglés English Teachers	Reading Listening Oral Production Written Production	English	2023	On Demand	Yes (format)

Note. Based on PELEx's website and Araya (personal communication, July 9th, 2023)

Table 1. Tests designed, applied, analyzed, and monitored by PELEx

Between 2019 and 2022, PELEx had more than 1100 enrolled test takers, 1100 locations nationwide in mass test applications, 8000 assessed test takers in a day (2019), and 1000 designed items. Additionally, during the same period, PELEx was able to work with an interface with greater security protocols, continue the three application-modality systems (online, hybrid, and offline), obtain real-time results, assess three skills in some locations (reading, listening, and speaking), and avoid test loss and safety violations (Araya & Quesada, 2023), all of which translate into certain clear benefits for the evaluation process and its users. First, the variety of test takers and locations has allowed the assessment of individuals that might not have the necessary resources to pay for high-quality monitoring or certifying proficiency tests, contributing to not only the provision of information on their language skills, but also to the collection of data under a broader scope. Second, the expedited process of application, result handling, modality versatility, and safety assurance have become core characteristics of PELEx's exams, enabling test takers and test users with prompt and relevant results.

PELEx has also become a certified center for the PTE General Authorised Test Center (Pearson) since 2019, and for the Celpe Bras (Portuguese) and the Marlin Test (Seafarers) in 2023. Furthermore, in 2023, the program will start developing 3 different technical programs (technical diploma) in didactic and assessment of foreign languages, interpretation, and translation.

The Use of Artificial Intelligence at PELEx

When AI is incorporated into the assessment process, three are the main purposes: to detect the modality (format) of a test, to adapt to each test taker's proficiency level in order to present specific questions, or to mediate the assessment of oral production. In the first case, because of the differences between the resources available in each of the test locations (e.g. unstable vs. stable Internet connection), the format or modality of the test needs to be flexible. For example, in the period 2021-2022, the first monitoring test for primary school students was carried out. Around 10 000 fifth-grade students were assessed and, without sacrificing the interface ("the feel and look") of the test, the AI was able to work out the offline or hybrid formats and send the results in real time to the servers of the program. PELEx's tests can also be administered offline (with the same interface as well), guaranteeing its access to many different populations without restrictions related to the internet accessibility .

In the second case, artificial intelligence is used in adaptive tests to assess reading and listening. In order to do so, each test taker is presented with a series of questions. As the assessee completes the test items, the AI starts aligning the level of proficiency of the student with the level of the questions presented until a convergent point is reached. This process and types of tests have certainly different benefits specifically centered on their flexibility, usability and practicality; however, at PELEx, the drawbacks regarding their use are treated with care, which frames the actions taken to continue the improvement of the AI training and result use (see Rezaie and Golshan, 2015 for an expanded discussion on the advantages and disadvantages on CATs). This cautionary use explains why AI-based tests fulfill only diagnostic or informal placement purposes in the program, and certificates of proficiency when a high-level of reliability and test-result validity is reached.

In the third case, artificial intelligence mediates the process of the Automatic Standardized Oral Test of English (POA-IE) from UCR. The purposes of this test are assessing the personal, academic, and linguistic competence of the target populations, and it has been specifically designed for non-native speakers of English. In addition, the test seeks to place the test takers on a specific CEFR band in a practical and reliable manner. As Figure 2 shows, the POA-IE

has been designed through a process of AI incorporation, calibration, and analysis to guarantee a balance between practicality and thoroughness. However, as in the second case, because the human calibrators are still fundamental in AI-language assessment, the POA-IE can only be used for purposes related to monitoring, informal placement, or diagnosis of one's oral proficiency level, and it is not used to certify language proficiency yet.

According to PELEx (n.d), the POA-EI intends to measure the oral expression and production skills on national, regional, and global topics within the CEFR interpersonal domain through descriptions, presentations, narrations and argumentations. The oral production is assessed in one-way exchanges (monologues) with an avatar by audio recording the test taker's contributions in a computer. The topics and themes to orally express about are selected according to the CEFR and the target population's needs. Furthermore, coherence is reached through the use of the CEFR's descriptors in order to interpret the test takers' speech punctuations, which correspond to a specific level of proficiency (i.e. A1, A1+, B1, B1+, B2, B2+, and C1).

Figure 2. Features, General Components and Principles of the POA-EI at PELEx

To collect the necessary information to place a test taker into a specific CEFR band, three language tasks are carried out in a 12-15 minute test. In the first task, a description of a photo, drawing, or life situation is requested. In the

second one, a narration on one's everyday life must be provided, and, in the third task, the assessee must record a specific opinion, point of view, or position regarding a given topic. The level of complexity of the grammar, pronunciation features, content, cohesion, vocabulary, and task performance (relevance) in the test takers' utterances is expected to increase according to each of the tasks. Before recording each of the answers, the assessees have 30 seconds to prepare after reading the task prompt and will have up to 60 seconds to record their response (depending on the task).

Each recording is punctuated using an AI model, which is trained using the input from human language assessors to reflect the CEFR guides that PELEx adapted. The POA-EI punctuations in relation to another standardized test are displayed in Table 2.

IELTS	**CEFR**	**POA-EI UCR (powered by Speechace)**
9.0	C2	100
8.5	C2	94
8.0	C1+	89
7.5	C1	83
7.0	B2	78
6.5	B1+	72
6.0	B1	67
5.5	A2+	61
5.0	A2	56
4.5	A1+	50
4-0	A1	44
0 to 3.5	A0	0-43

Table 2. POA-EI's Punctuation Interpretation in comparison to the CEFR and the IELTS

PELEx's Impact on the Costa Rican Testing Scenario

Taking into account the imminent need for transforming and responding to the needs of the Costa Rican education system (see Programa Estado de la Nación, 2021), PELEx's work represents a significant and primary contribution to building up the national system of evaluation that is so dire. In the past few years, the relevance of standardized testing and PELEx's work has gained relevance, and, in this sense, O'neal (2018), Montero (2021), Cordero (2023), La República (2023), and Martínez (2023) exemplify the journalistic coverage PELEx's tests have at a national level.

From a general perspective, the tests benefit several populations, contribute to the appropriate management of infrastructure, human and financial resources (i.e. virtual vs. in-person applications, and linear vs. adaptive nature of tests), and present a more-user friendly interface (see Figure 3). Additionally, by training AI and incorporating it into diagnostic assessment processes, PELEx is also pioneering in the field of AI in Latin America, collecting relevant information for improving test design, administration, interpretation, analysis and use in the region and positively adding to the understanding of the Costa Rican's students levels of proficiency in different languages.

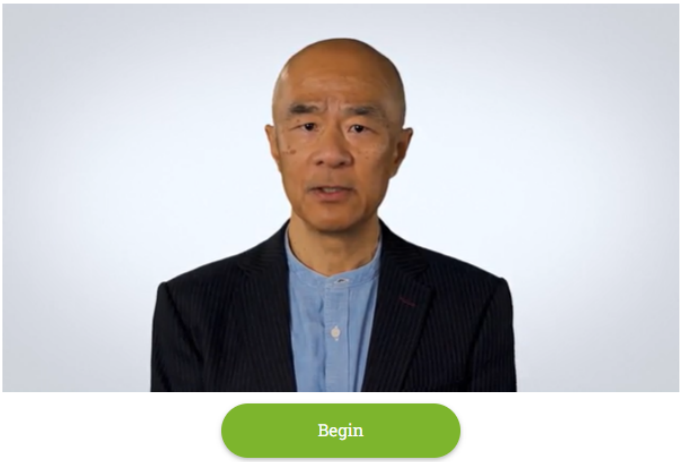

Figure 3. Avatar interface for the POA-EI

Conclusion

By working on standardized testing, improving the measurement protocols, transparently addressing the needs of diverse populations in the public and private sectors, aiding the transformation of the assessment culture, and contributing with assessment training, PELEx continues to work hard on attending to Costa Rica's and international evaluation needs while incorporating artificial intelligence and finding solutions to the issues that may arise as part of the assessment process.

References

Araya, W. & Quesada, A. (2023). *Programa de Evaluación en Lenguas Extranjeras (PELEx)* [Slide Presentation]. (Unpublished institutional document). Escuela de Lenguas Modernas, Universidad de Costa Rica

CONARE. (2023). *La institución*. Retrieved from https://www.conare.ac.cr/conare/la-institucion/

CONARE. (2019). Desafíos de la educación en Costa Rica y aportes de las Universidades Públicas. Retrieved from https://repositorio.conare.ac.cr/handle/20.500.12337/7953

Cordero, M. (March 17, 2023). UCR dona al MEP 25 mil pruebas estandarizadas de lenguas extranjeras. Retrieved from https://tinyurl.com/yecebjtu

Green, A. (2021). *Exploring Language Assessment and Testing: Language in Action (2nd ed.)*. New York: Routledge

La República. (June 23, 2023). *Costa Rica será primero en Latinoamérica en evaluar inglés de estudiantes de primaria y secundaria con inteligencia artificial.* Retrieved from *https://www.larepublica.net/noticia/costa-rica-sera-primero-en-latinoamerica-en-evaluar-ingles-de-estudiantes-de-primaria-y-secundaria-con-inteligencia-artificial*

O'Sullivan, B. (2016). *Adapting Tests to the Local Context. New Directions in Language Assessment, special edition of the JASELE Journal*. Tokyo: Coombe, Christine; Folse, Keith and Hubbley, Nancy. (2007). A practical guide to assessing English language learners. Michigan: The University of Michigan Press. Japan Society of English Language Education & the British Council, pp.145-158

O'neal, K. (July 27, 2018). *Evaluación permite seleccionar los mejores traductores e intérpretes oficiales.* https://www.ucr.ac.cr/noticias/2018/07/27/evaluacion-permite-seleccionar-a-los-mejores-traductores-e-interpretes-oficiales.html

PELEx. (2023). *Programa de Evaluación de Lenguas Extranjeras.* Retrieved from http://www.pelex.ucr.ac.cr/

PELEx. (n.d.). *Prueba Oral Automática de Inglés Estandarizado (POA-EI) de la UCR* (Unpublished institutional document). Escuela de Lenguas Modernas, Universidad de Costa Rica

Programa Estado de la Nación. (2021). *Octavo Informe Estado de la Educación.* Retrieved from https://estadonacion.or.cr/wpcontent/uploads/2021/09/Educacion_WEB.pdf

Martínez, V. (June 23, 2023). *MEP elimina prueba de dominio lingüístico para este año*. Retrieved from https://www.larepublica.net/noticia/costa-rica-sera-primero-en-latinoamerica-en-evaluar-ingles-de-estudiantes-de-primaria-y-secundaria-con-inteligencia-artificial

Montero, F. (April 23, 2021). *La UCR realizó el primer examen virtual de diagnóstico de inglés a gran escala en el país*. https://tinyurl.com/5c8asrk4

Rezai, M., & Golshan, M. (2015). Computer Adaptive Test (CAT): Advantages and limitations. *International Journal of Educational Investigations*, 2(5), 128-137. Retrieved from http://www.ijeionline.com/attachments/article/42/IJEI_Vol.2_No.5_2015-5-11.pdf

UNESCO. (2016). *Educación 2030: Declaración de Incheon y Marco de Acción para la realización del Objetivo de Desarrollo Sostenible 4: Garantizar una educación inclusiva y equitativa de calidad y promover oportunidades de aprendizaje permanente para todos*. Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000245656_spa