UCR-ELTA: Constructing a Standardized Proficiency Test for English Teachers in Costa Rica*
UCR-ELTA: La Construcción de un Argumento de Evaluación para una Prueba Estandarizada de Dominio Lingüístico de los Docentes de Inglés en Costa Rica

Walter Araya García [1]
Universidad de Costa Rica

Jennifer Céspedes Araya [2]
Universidad de Costa Rica

Resumen

En este artículo se analizan los principios para llevar a cabo procesos de evaluación más relevantes con el fin de iniciar la construcción y sistematización de un argumento de validez con base en los procesos de toma de decisiones para el diseño y sustentación del examen llamado Universidad de Costa Rica - English Language Teacher Assessment (UCR-ELTA). El análisis teórico respalda la necesidad (1) de la contextualización de una prueba estandarizada para medir el dominio lingüístico de las personas docentes de inglés en Costa Rica, (2) de la reflexión de los requisitos que una prueba de esta naturaleza debe cumplir, (3) del reconocimiento de los argumentos a favor y en contra de este tipo de evaluación, y (4) de la descripción de los procesos de diseño, implementación, validación, y mejora. Además, se busca fundamentar de mejor manera los parámetros de validez de la interpretación de los resultados de esta prueba. Esto lleva a la conclusión y reconocimiento no solo de que (a) una prueba de esta naturaleza tiene beneficios considerables para comprender mejor las características del dominio lingüístico de la población meta sino también (b) que hay una necesidad fundamental de crear un sistema de evaluación explícito y transparente que permita contemplar y atender todos aquellos aspectos que pueden perjudicar a la población meta.

Palabras clave: *contextualización de una prueba estandarizada, Costa Rica, dominio lingüístico de las personas docentes de inglés, prueba estandarizada, UCR-ELTA*

Abstract

In this article, the most relevant principles for creating a sound assessment cycle to start building and systematizing a validity assessment argument based on the decision making processes for the design and foundation of the test called Universidad de Costa Rica - English Language Teacher Assessment (UCR-ELTA) are presented. The theoretical analysis supports the need for (1) the contextualization of a standardized test for English teachers' language proficiency in Costa Rica, (2) the reflection on the requirements that this test must comply with, (3) the acknowledgement of the arguments in favor and against this type of testing, and (4) the description of the processes of test design, implementation, validation, and improvement; the UCR-ELTA parameters and validity of interpretation are better founded as a result. This leads to the conclusion and recognition not only of (a) the fact that this test has considerable benefits to better understand the characteristics of the proficiency level of the target population, but also of (b) the fundamental need to create an explicit and transparent assessment system which allows to foresee and attend to all those issues that might negatively affect the target population.

Keywords: *contextualization of a standardized test, Costa Rica, English teachers' language proficiency, standardized test, UCR-ELTA.*

In 1824 (Marín-Arroyo, 2013), the formal teaching of English began, and this practice has been developing and evolving. This long-lasting tradition has probably contributed to the fact that Costa Rica has been one of the countries where English is spoken the best as a foreign language in Latin America according to the various rankings published (see Education First, 2023). Consequently, the different policies implemented in Costa Rica translated into covering more than 25% of the preschool population, 92% of the primary school population, and 92% of the secondary school population (Política Educativa de Promoción de Idiomas, 2019). In spite of this positive outcome, the country continues to seek to improve through evolving policies and strategic plans more systematically addressing English teachers' language performance.

In this regard, the University of Costa Rica has strongly invested in research and design processes that respond to what are considered good qualities and principles of assessment as reliability, localization, practicality, tailoring, and fairness (see O'Sullivan, 2016; Brown & Abeywickrama, 2019). Furthermore, according to Marco Nacional de Cualificaciones para las Carreras de Educación (2021), there has been a historical interest in promoting the training quality of future teachers. This has enriched the processes of hiring and evaluating education professionals (including English teachers), improving the Program for International Student Assessment (PISA) results, attending to educational quality gaps, and addressing the limited number of accredited educational majors.

Based on this scenario, in this document, the processes of design, administration, interpretation, validation of the Universidad de Costa Rica - English Language Teacher Assessment (UCR-ELTA) are defined, described, and explained. Even though there are quite valid arguments against the application of standardized tests to make high-stakes decisions, the fact is that there is an evident need to ensure, in some way or the other, that English teachers can handle the linguistic demands of the target language as part of the pedagogical knowledge that they must develop and acquire (Dadvand & Behzadpoor, 2020) and which are framed in the evident need to correspond to historical transformations and education-related contexts in the present and future in Costa Rica. In this sense, this article aims at presenting and reflecting on the conceptual, constructive, and theoretical support for the UCR-ELTA.

Literature Review

Based on Araya Garita et al. (2022), Coombe et al. (2007), Brown and Abeywickrama (2019), it is evident that standardized testing has been a complex practice in educational settings because it must meet specific criteria framed in language testing principles of validity, authenticity, reliability, practicality, and washback. In addition, localization in standardized testing is a must (O'Sullivan, 2016), especially because the needs and the contexts of specific populations vary from place to place.

On the one hand, standardized testing complexities pose a great challenge for stakeholders, assessment professionals, and test takers. In other words, "standardized

language testing may seem overwhelming and intimidating" (Araya Garita et al., 2022, p. 122). On the other hand, there are also controversies connected to the use of standardized testing as it is considered by some as a tool to control (Shohamy, 2001; Fulcher, 2010). In spite of the evident criticisms associated with standardized language testing, it has become a widely used tool to monitor the current situation of a specific educational context under very particular circumstances to later design and implement better national policies.

In this fashion, there are some challenges that the Costa Rican system of education has been facing: access to it, teacher-focus policies, learning environments and infrastructure, and the management of the educational system (CONARE, 2019). For this reason, the UCR-ELTA can provide meaningful information on the applied, contextualized, and task-based linguistic knowledge English teachers make/have of the L2, contributing, as CONARE (2019) proposes, to the creation of a national system of educational evaluation that allows to place teacher in ability levels that, in some cases, could encourage the improvement of competencies (p. 7), among other benefits.

Taking into account the potential evaluative, professional, and national implications of a test as the UCR-ELTA and considering that every single language assessment has a direct impact in the real world (Bachman & Palmer, 2010), the relevance of clarifying the steps followed to collect information and make decisions based on the interpretation of test results is indispensable. When constructing an argument for assessing one's language ability, it is necessary, in addition, to acknowledge with careful consideration and transparency what the reach of the interpretations should have for the test users and how their impact will be responsive to a well-thought and theoretical ground that may not fully fit into a specific context but that is enriched by it as much as possible. In this sense, the test is

> [Inicio de cita]designed to cater for the local needs of the test population. This may mean choosing appropriate cultural topics and making sure the processes of test design, piloting, administration, and scoring reflect local needs and expectations. In more recent localization movements, this has also involved localization of language use in context to include the spread and changing shape of English in countries that use English as an official language. (Coombe, 2018, p. 28) [Fin de cita]

To cater to those needs, a key participant in this process is the target population to be assessed: The language teacher must be taken into consideration when designing a specific test as UCR-ELTA. Slomp (2005), when reflecting on the assessment of test takers' writing skills, argued that collaboration is fundamental for building up a well-rounded and dialogic assessment process. Historically, it seems that assessment specialists have underestimated the contributions of the various actors involved in the process (e.g., language experts, experienced teachers, test-takers in general, and other stakeholders). Furthermore, Slomp (2005) also reinforced the idea that

> [Inicio de cita]rather than minimizing the expectations for test validity on the basis of construct complexity and the difficulty involved in defining measurable

constructs, assessment specialists should recognize the need to engage in collaborative design… Validity-based research provides both the rationale and the push for collaborative assessment design in language education. The issue is real, the time is now. (p. 153) [Fin de cita]

In addition to the collaboration with stakeholders and the target population, the final block towards building a solid validity argument for a standardized test is aligning a test with language proficiency descriptors as those provided by Council of Europe (2020) and the test takers' and administrators' characteristics (O'Sullivan, 2016, p. 148). In the same manner, the Association of Language Testers in Europe (ALTE, 2020) further emphasized that, as a minimum standard in test construction, the test must be linked to a theoretical construct (p. 26). One may think that the analysis and construction of standardized tests have plateaued; however, validation of language assessment is a never-ending, ongoing process (Chapelle, 2012; Brown & Abeywickrama, 2019) that can shed light on a multiple variety of language-related aspects.

Assessing teachers is a critical action that aims at creating opportunities for reflection and improvement and informing any relevant decision-making processes (Loredo, 2021). Standardized testing, therefore, is one of the possible tools to collect information for making these decisions, and it is highly practical when assessing large target populations. Even though the UCR-ELTA falls into the category of a high-stakes diagnostic test, it must also respond to the ongoing nature of language evaluation (Loredo, 2021) that is discussed throughout the literature. This is a systematic process to collect reliable and valid evidence of language teachers' performance that permits the Ministry of Public Education to hire new teachers or give them tenure based on specific criteria that is technically and scientifically sound based on the Costa Rican context. In other words, assessment is a mean for strengthening the national teacher profiles (Loredo, 2021, p. 7) and policies. As a result, to have sound foundations to construct a test (Bachman & Palmer, 2010) as UCR-ELTA and to contextualize it (Coombe, 2018; O'Sullivan, 2016) for a fairer assessment practice, it is necessary to take into consideration different aspects: national regulations, language teachers' assessment expectations and attitudes, and test-related standards.

Language proficiency tests and qualifications: requirements in Costa Rica

In Costa Rica, the Marco Nacional de Cualificaciones para las Carreras de Educación (2021), based on international trends set by UNICEF, ONU, OCDE, CEFR, among others, outlines the minimum requirements an English Teacher must comply with, and they revolve around communicative and interactive skills, language ability (C1-CEFR based), teaching-related knowledge, use of technology, global citizenship skills, language assessment knowledge, among others. These requirements and competencies can also be found in various types of qualifications-based examinations and frameworks such as the Cambridge Teaching Framework (see Cambridge University Press & Assessment, 2022).

It can be noted, as a result, that the qualifications and competencies language teachers must have and develop are responsive to a multiple set of professional, cultural, and

educational needs. Furthermore, an English teacher must be able to deal with the language in multiple ways to be able to respond to the needs of their populations and the world; in this case, there must be evidence of how well they can work with the language to determine if they have developed a proficient level. Unfortunately, scrutinizing their language abilities is not positively viewed by many language-teaching professionals.

Teachers' Language Assessment and its connection to professional development

There are many benefits for the different actors involved in an assessment process. Not only are these benefits related to a quantitative value but also to a qualitative insight gained through the reflections that this process triggers. At the same time, these reflections work as a basis for improving teachers' language proficiency and for other stakeholders to make appropriate decisions to benefit the language assessment system.

Around the world, there have been some relevant experiences and examples to take into consideration when understanding the impact of teachers' language assessment. For example, Chu and Jaca (2019) have defined professional development that is "primarily concerned about keeping one's skills career fresh and on top of the game… [it] takes into account the skills and knowledge employees acquire to optimize their personal development and job growth" (p. 421). They also add that professional development "refers to the skills, knowledge and ongoing learning opportunities undertaken to enhance an individual's ability to carry out their jobs and achieve professional growth" (p. 421). In addition, Harding (2009, as cited in Chu & Jaca, 2019, p. 421)) stated that Continual Professional Development (CPD) is key to improving one's performance. This also implies the recognition of the teachers' responsibility for their own improvement and needed fulfillment as educational professionals. As a result, the implications of CPD require evident effort, awareness, and resources, especially for a language teacher.

If carrying out rigorous and consistent assessment is the first step in this cyclical process of improvement in professional development, Davidson et al. (2012, as cited in Chu & Jaca, 2019) highlighted the importance of being aware of other activities and measures to expand teachers' CPD: self-reflection, skills and knowledge expansion, collaborative learning and sharing, and training and workshop engagement. This complex scenario also implies that the authorities and higher-education institutions must ensure the availability of resources to meet the quality standards for language teacher training.

Sarwar et al. (2014) reinforced this idea of continuous assessment considering a study carried out in Pakistan in which the main finding suggests the need of improving "speaking skills in teacher education programs" (p. 7), and, in Costa Rica, the panorama seems to be somewhat similar; in fact, the diagnostic processes (e.g. TOEIC tests applications in 2008 and 2015) showed that training can have a positive impact on teachers' language performance, and that there is indeed an impact on higher education and teacher training (Diálogo Interamericano y Unidos por la Educación, 2018, pp. 33-34).

Expectations about teacher's language proficiency and professional development: a competency-based approach

Language teacher assessment is complex not only because of the cyclical and reflective systems it responds to but because of the external expectations that are pressed on it. These demands come from the diversity and interconnectivity of the world which seem to pose major needs on handling information (Rueda, 2009, p. 3); these demands also have a direct impact on how English teachers are viewed in the world and on what they need to be able to do to be considered professionally prepared. Therefore, understanding and developing the necessary teaching competencies have also become relevant, and having a proficient command of the second language is one of them. Rueda (2009), when discussing the competency-based approach in tertiary education, indicated that a competency involves the ability to face complex demands while supporting oneself and using psycho-social resources, skills, and attitudes in a particular context (p. 3). Therefore, a "competent" English teacher will be the one who can handle the language and the teaching skills and knowledge necessary to responsibly carry out work-related tasks. For this reason, assessment and standardized testing can shed light on the linguistic notion of the competencies of the English-language teachers.

Teacher's attitudes towards standardized testing

 Historically, tests have been perceived as punishment. Brown and Abeywrickrama (2019) mentioned that those who are assessed (in general) "are not likely to view a test as positive, pleasant, or affirming" (p. 1), and it is expected that English teachers may share this view. When discussing teachers' perspectives on examinations administered to their students, Kellaghan et al. (1982) considered that

> [Inicio de cita]if teachers perceive standardized tests and the constructs measured by them to be inaccurate, biased, unstable, or unimportant, they probably will be less likely to utilize test results in a practical way than if they perceive tests in a more favorable light. Even in those cases where teachers do have favorable perceptions, we should consider the weight teachers say they accord to test information relative to other forms of evidence about pupils (e.g., observations, prior teacher recommendations, classroom tests) before we suggest a strong relationship between standardized testing and various classroom practices. (p. 64)[Fin de cita]

Based on a general impression, English teachers seem to have a similar view regarding tests that are administered to them in order to measure their proficiency level. This negative perspective on testing and assessing might be reinforced when the test results, consequences and impact affect the test-takers' lives directly, and this view seems to still be present nowadays. In this way, Bachman and Palmer's (2010) observation on building an argument for assessment provides theoretical support to create a more consistent, transparent, and reliable system of assessment that considers the needs of the test takers.

Difficulties when assessing teachers and the role of standardized testing

Standardized language high stakes testing for pre-serving teachers is a great challenge for any educational system around the world. Therefore, on the one hand, negative views on these types of testing are not surprising as explained previously and in other sources (Brown & Abeywickrama, 2019; Bachman & Palmer, 2010; Green, 2021), and, on the other hand, this also results in more traditional approaches to assess students during their practicums; these are quite relevant, but they do not provide standardized information with a quantitatively sound interpretation. In this sense, Bolitho (2013) highlighted that

> [Inicio de cita]the practicum in pre-service training and developmental observation for serving teachers are acknowledged as crucial planks in maintaining and improving standards of teaching, and yet the trainer's or educator's role as an observer, supervisor or assessor remains largely underexplored, susceptible to subjectivity in its practices and cloaked in silence and handed-down traditions rather than opened up in public debate. (p. 12)[Fin de cita]

Even though these micro-assessing tasks (individual practicums) provide the practitioners and other actors with relevant information, the fact is that, at the macro-level (country-wise) such information is usually undervalued or ignored when making policies and decisions. However, the fact is that systematizing the qualitative information of practicums from various institutions, multiple individuals and scenarios is not practical. In addition, according to Chapelle (2008), the warrant is the observations of test performance that reveal relevant information on knowledge, abilities, and skills within a specific target domain, and the assumptions are those inferences expected when the test takers scores are interpreted as indicative of performance and scores that they would receive in the target setting. For this reason, standardized testing is a more reliable means for collecting data that provides evidence on language proficiency of new and experienced practitioners, allowing a closer approximation to the tested individuals' realities, a more practical administration of the test, and greater opportunities for building a sound validity argument. As a result, the test warrants and assumptions must be responsive to its purpose, population, uses, and impacts.

Slomp (2005) recommended the involvement of different actors when designing a test, for example, the test takers, the test users, and the test designers, among others. This translates into the need for communicating with the target test takers to better understand who they are, what they do with the language, and what kinds of tasks they carry out using English, a process that corresponds with a particular evidence-designed approach focused on the examining and task models by providing relevant data and linguistic performance on the professional domain (Tschirner, 2018). This will add more detailed information that will help build a stronger assessment cycle as shown in Figure 1. In other words, the test designers can fortify the validity of the interpretations if the different actors' voices are taken into consideration when defining the test domain, the constructs, the evaluation

process, the test generalizations, the explanations, the extrapolations, and the utilizations of the results.

Figure 1

Chain of Analysis, Design, Auditing, and Reflection of the UCR-ELTA
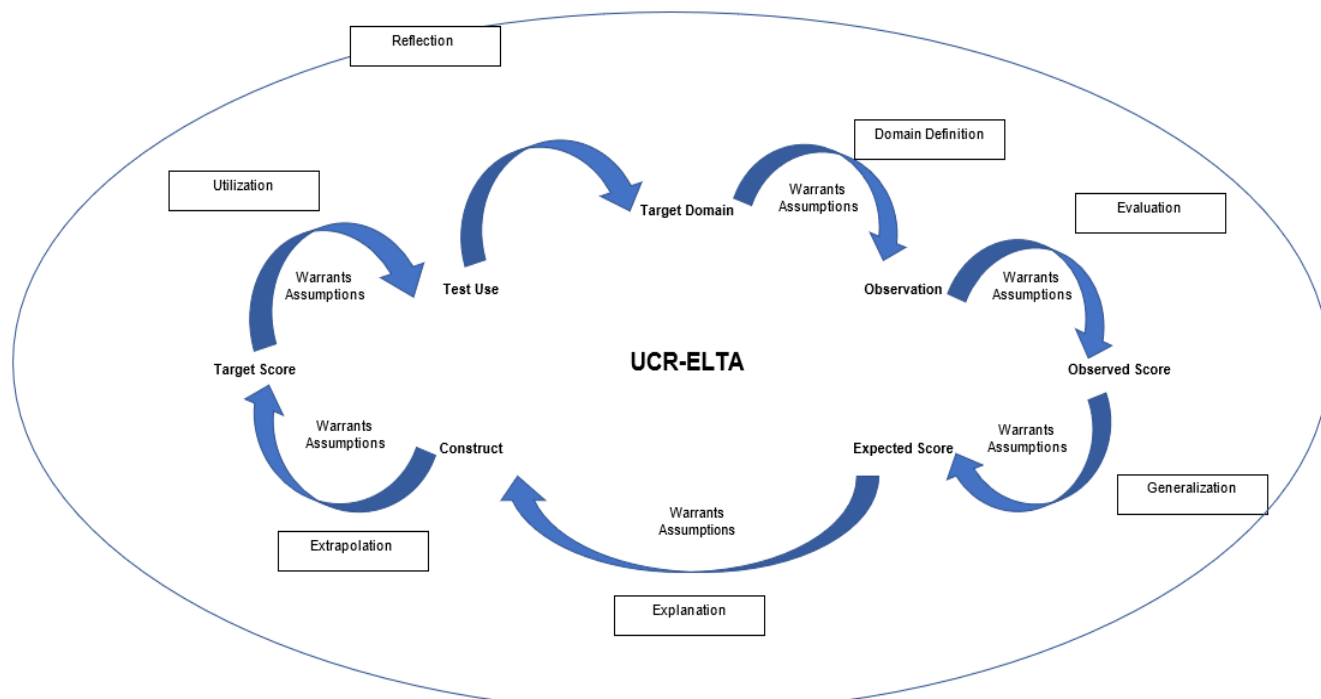
Note. Based on Chapelle (2008) and ALTE (2011)

Chapelle (2008) also recommended that the specific warrants and assumptions of each specific test associated with each of the inferences need to be identified. This, for sure, can only be accomplished through teamwork. Precisely, as Romero (2007) pointed out, from a critical perspective, assessment is a reflexive process in which all of the involved actors are encouraged to participate and whose purpose is to contribute with the growth and development of the test taker. This, of course, means that every single stage of the design of a standardized test has to be addressed and described clearly. For this reason, in the following section the authors describe and discuss the starting and grounding phases for the design of UCR- ELTA.

The UCR-ELTA Test

Considering the benefits that a standardized test has and the needs to be assessed in the Costa Rican English teaching scenario, many actors involved in the process must be included when making decisions to build a healthy assessment cycle. For example, the

coordinator of Costa Rican Alliance for Bilingualism acknowledges that "English teachers' effective mastery of a second language, in the coming years, will assure academic success of students in academic and work domains" (M. Rojas, personal communication, June 8, 2022). He also adds that it is necessary to develop actions that allow improving the quality of the teaching staff; in other words, it is necessary to evaluate language teachers periodically and systematically. The Costa Rican Ministry of Public Education currently has a majority of qualified and certified English teachers in terms of language proficiency.

The Director of the School of Modern Languages of University of Costa Rica mentions that teacher quality is linked with the success of students' learning. Another important aspect mentioned is the need for accountability of data to improve the educational system. This means that it is expected that evaluation enhances teachers' practices and improves their effectiveness in the classroom (A. Quesada, personal communication, June 22, 2022).

English language coordinators also support the project of evaluating English teachers. For instance, a head of an English department from a private institution mentions that

> [Inicio de cita]undoubtedly, assessing an English teacher's language proficiency is a constant need, at least in our school. One of the reasons behind it is that we are expecting our students to achieve a high command of the different language skills, so we must verify and guarantee that our English teachers have the language mastery and competency that is required for their position. We are aware that the result of a standardized test might be influenced by different variables. However, we recognize that the results have a good level of reliability and are a good resource for decision-making when hiring or offering a promotion. The second reason is related to professional development which is also a key element for us. I consider that one of the biggest challenges all teachers in general have is to develop the capacity to reflect on their own experience and abilities, so assessing teachers' language proficiency could be a starting point. This allows teachers and English coordinators to recognize not only the points where improvement is possible, but also the teacher's strengths. (S.Víquez, personal communication, June 15, 2022)[Fin de cita]

In the public sector, there are a variety of opinions involving different attitudes regarding standardized tests in general or the TOEIC (which is the test administered to teachers so far). When asked about the relevance of assessing teachers' proficiency, a head of an English department and English teacher at a public secondary institution expressed the following:

> [Inicio de cita]I believe evaluating language teachers is of great importance in order to get a position and to be able to continue with their tenure. I think a standardized test such as the TOEIC should not only be administered, but there should also be one where methodologies are assessed. It is well-known that a good language teacher must not only be language proficient but also have teaching skills. (C. Retana, personal communication, September 2, 2022)[Fin de cita]

Evidently, individual views on standardized tests administered to teachers should not be generalized to the entire related populations; however, they do reflect the need for better understanding the perceptions of the different groups of stakeholders and users of the results. These micro views help guide the process of warrant and assumption determination and allow to be able to tackle specific questions and arguments in favor and against particular decisions made based on the test interpretations and possible impacts.

Overall profile of the UCR-ELT test takers and K (knowledge), A (abilities), S (skills), and tasks: English Language Teachers in Costa Rica

The UCR-ELTA test is targeting English teachers from pre, primary, and secondary schools in Costa Rica (approximately 6000 professionals). However, to do so, piloting the test first is fundamental. At the piloting stage, a sample of 300 participants will be tested. The sample population should have very similar characteristics to the target test takers; for this reason, the piloting test takers will be English Teaching students about to graduate from university and in-service teachers from the public and private sectors. The piloting process will be carried out through an online platform with a digital test at UCR all over Costa Rica.

As with any other EFL population, Costa Rican English Teachers must be able to handle the language in many different ways and in relation to several diverse topics, especially because of the various kinds of populations and learners they usually work with. In this sense, typifying KAS, tasks, roles and contents related to English poses an evidently complex issue. Firstly, understanding what these concepts imply is a must. In this sense, Green (2021) provides a more manageable insight into what KAS are, and what their connection with language tasks entails:

> [Inicio de cita]**Knowledge** about language may include recognizing a word written in a foreign language and knowing a translation equivalent…, or knowing a grammatical rule…, or knowing pragmatic conventions… A distinction is often made in language education following Hymes (1972) between knowledge of the rules governing language as a system and the **ability** to use language in unrehearsed interaction… Language **skills** involve drawing on language knowledge and language abilities in order to read, listen, write, speak, to interact with others, or to mediate between them. The evidence we have of a person using a language may be very limited – a few telephone conversations and a handful of emails, perhaps – but based on what we observe in these few instances, we often make inferences about their more general knowledge of a language, their ability to use the language and their skill in carrying out language-related **tasks**. (p. 4) [Fin de cita]

The relationship between knowledge, abilities, skills, and tasks is characterized by a seemingly symbiotic reflection of one another although it is quite clear that observing one's performance in a task will never be enough to understand the wholeness of one's language capabilities. In addition, there are other complexities that must be considered when determining and categorizing English teachers' KAS. On the one hand, these professionals

are themselves English Language learners in the sense that their learning process will very likely continue throughout their lives, and, on the other hand, they are, at the same time, English facilitators and language consultants. The former implies that language improvement will always be part of the needs of any English teacher, and the latter that they must consistently be able to help others (their students) develop their language potential as well; this, of course, cannot be "measured" or assessed solely on the bases of the English teachers' language abilities because teaching (and learning) a language cannot be reduced to "being able to speak, write, listen, or read" only. However, the responsibilities of an English teacher include being able to deal with the target language in a way that their use of it works as a "good" example for the learners. For this reason, collecting as much information as possible regarding the teachers' language abilities and proficiency level can aid in the process of identifying linguistic aspects to improve and strengths, which might positively guide other related decision-making processes. These aspects also frame how relevant it is for a test of this nature to be well designed and to be aware of the consequences it has.

Despite the difficulties and complexities that describing KAS might pose, the attempt must be made, especially because decisions the test users will make need as much support as possible. This need for gathering meaningful information and evidence from the teachers' language ability also frames the design and creation of tasks (and items). As a result, test takers should be involved in the assessment cycle.

In June 2022, a preliminary survey designed as part of the process to create the test, involving test takers more in the assessment process (Tschirner, 2018), was administered to 380 English teachers. As shown in Table 1, the survey takers reported a significant use of English in the classroom, especially with other English teachers (50.8% - *almost always* or *always)*, and their students (91.9% - *almost always* or *always*). This sheds light on the kinds of interactions these professionals may be having through English and sets a certain standard for contextualizing items, scenarios, and tasks for the test.

Table 1
Frequency Percentage on the Use of English with other interlocutors as Reported by the Participants

Interlocutors	Frequency Scale					
I… use English with…	Never	1	2	3	4	Always
Other English teachers		16.3	32.9	35.8	15	
The institution's principal		81.1	15	3.7	0.3	
Regional Advisors		28.4	19.7	29.5	22.4	
National Advisors		53.2	10.8	17.4	18.7	
Students' Parents		85.8	11.6	1.6	1.1	
Students		0.5	7.6	45.8	46.1	

Note. N=380. Taken from the preliminary results of the survey "*Usos específicos del Idioma Inglés por parte de los Docentes de Enseñanza del Inglés en Costa Rica*" (PELEx, 2022b).

In addition, the participants of the survey provided information on the variety of English-based tasks they carry out very frequently or frequently. As noticeable in Table 2,

English *is* relevant for these professionals mostly in connection with those aspects revolving around designing and teaching a class and carrying out assessment processes. In this sense, the tasks presented also provided guidance into how certain testing scenarios and input should be chosen and designed (i.e., layout) for them to be more naturally presented to this population, strengthening the support for making decisions regarding the operationalization of the CEFR statements as well. In this sense, in the development stage, feasibility plays a key role in the process for task design. For instance, requesting test takers to write a lesson plan (one of the most frequent tasks) as part of an extended writing task for the UCR-ELTA is indeed relevant for them in their professional setting, but it will not allow test developers and judges to obtain grounded information on their writing performance.

Table 2

Percentage of Very Frequent or Frequent Tasks Carried Out in English as Reported by the Participants

	Task	**Percentage**
1	Development of mediation processes (the class itself)	92.4
2	Writing lesson plans	92.4
3	Design of written assessment instruments (tests)	88.9
4	Administration of oral exams	85.8
5	Reading lesson plans	70.5
6	Reading narrative texts	49.5
7	Listening to podcasts related to education	46.1
8	Participation in meetings of the English department	44.7
9	Creation of observations to register the students' performance	40
10	Participation in training workshops	39.7
11	Reading emails	30
12	Reading administrative documents (e.g., from MEP)	32.9
13	Reading students' reports	26.1
14	Writing progress reports addressed to the students' parents/guardians	25.3
15	Writing emails	23.9
16	Reading scientific articles	23.4
17	Writing reports of curricular accommodations	22.6
18	Completion of relevant administrative documents (e.g., score records)	20
19	Organization of institutional events (e.g., graduation ceremonies)	18.2
20	Organization of school assemblies	13.7
21	Organization of school trips	4.7

Note. N=380. Taken from the preliminary results from the survey "*Usos específicos del Idioma Inglés por parte de los Docentes de Enseñanza del Inglés en Costa Rica*" (PELEx, 2022b).

As mentioned before, the tasks and input to be incorporated in the test must be clearly connected to the test takers' realities to draw inferences from the results based on three main assumptions: assessment instruments will allow data collection on the targeted test takers' language abilities; the assessments tasks will be more responsive towards the test takers' cultural context in relation to their language abilities; and future and more robust statistical analyses will shed light on the tasks, procedures, items, forms, and task judges and designers (in accordance with Chapelle, 2008). For example, choosing appropriate texts

or themes for reading and listening tasks is of extreme importance as well. Table 3, in this regard, displays a selection of key curricular and teaching themes that teachers must be able to handle in the Costa Rican context in a variety of student populations as reported in the survey. As observed, most of the relevant topics revolve around cultural, technological, and personal contexts, which should inform the parameters for the test. Practicality and cultural relevance will also guide the decision-process related to the creation of the UCR-ELTA. The themes/topics below, even though they are informed by potential test takers, will have to be selected and some of them might not be suitable for item design, for instance.

Table 3

Percentage of Very Important or Important Curricular and Teaching Themes for the Costa Rican English Teachers and their Current Populations

	Curricular and Teaching Themes/Topics	Percentage
1	Technological Advances	76.6
2	Personal Identity	65.5
3	Intercultural Communication	63.4
4	Self-care (physical, emotional, among others)	62.9
5	Teaching Practice	56.1
6	Costa Rican Cultural Diversity	53.7
7	National Affairs	50
8	National Identity	46.3
9	National biodiversity reality	42.6
10	Family diversity in Costa Rica	41.6
11	Citizen ethics	37.4
12	International affairs	33.9

Note. N=380. Taken from the preliminary results from the survey "*Usos específicos del Idioma Inglés por parte de los Docentes de Enseñanza del Inglés en Costa Rica*" (PELEx, 2022b).

As a result, English teachers' KAS are indeed connected mostly to the interactions they have among themselves and with their students, reinforcing the idea of choosing topics, themes, and tasks that relate to this reality as much as possible within the professional domain.

Test construct and target language use analysis

As in any other test of this nature, the construct to be assessed must be clarified. For the UCR-ELTA, the assessment construct has been built up considering the following general aspects according to the Common European Framework of Reference for Languages (CEFR) by the Council of Europe (2020):

Oral production. Proficiency in oral production is understood as the ability to maintain a two-way oral exchange with an interlocutor, using the English language to speak about both formal and informal contexts on regional and global issues, at the professional domain based on Council of Europe (2020). Some of the skills to be tested range from "giving clear, detailed descriptions and presentations on complex subjects, integrating sub-

themes, developing particular points and rounding off with an appropriate conclusion" (Council of Europe, 2020).

Oral comprehension. This skill is defined as comprehension in live, face-to-face communication and its remote and/or recorded equivalent. It thus includes visual-gestural and audio-vocal modalities. The aspects of oral comprehension included here under reception are different kinds of one-way comprehension. The user can understand enough to follow extended discourse on abstract and complex topics beyond their own field, though they may need to confirm occasional details, especially if the variety is unfamiliar. The language user can also recognize a wide range of idiomatic expressions and colloquialisms, appreciating register shifts and can follow extended discourse even when it is not clearly structured and when relationships are only implied and not signaled explicitly at the professional level of CEFR (Council of Europe, 2020).

Reading Comprehension. Proficiency in reading comprehension is defined as the ability to understand different texts and images of general English in a professional teaching context at regional and global levels both formal and informal, at the professional domain of CEFR (Council of Europe, 2020). This comprises the users' ability to understand in detail lengthy, complex texts, whether or not these relate to their own area of specialty. They can also understand a wide variety of texts including literary writings, newspaper or magazine articles, and specialized academic or professional publications (Council of Europe, 2020).

Written production. Proficiency in written production is understood as the ability to write academic essays written in non-technical English, in formal contexts at the regional and global levels, in the professional domain as developed by the Council of Europe. Some of the skills to be tested are "can produce clear, smoothly flowing, well-structured text, showing controlled use of organizational patterns" and "can employ the structure and conventions of a variety of genres, varying the tone, style and register according to addressee, text type and theme" (Council of Europe, 2020).

Considering that tasks are the basic units of assessment for the UCR-ELTA, they must also correspond to the sub-specifications that describe them. In this sense, Table 4, presents the main aspects considered for each of the skills being assessed for the B2 and C1 bands considering the topics, themes, and common exchanges presented above:

Table 4

Particular Subskill Specifications for the UCR-ELTA

Reading comprehension
Reading for the main idea, reading for major points, reading for specific details, reading for the gist, inferencing, distinguishing fact from opinion, and identifying author's purpose or tone.
Listening comprehension
Listening for the main idea, listening for points, listening for specific details, listening for the gist, inferencing, distinguishing fact from opinion, and determining speaker's intent or tone.

Speaking Production

Grammar, vocabulary, segmental pronunciation (vowels and consonants), suprasegmental pronunciation (for example, stress, rhythm, intonation, prominence, connected speech phenomena), content, organization, cohesion, task performance, and appropriate use of performance of language functions, sociolinguistic appropriacy.

Written production

Grammar, vocabulary, content, rhetorical, cohesion, task performance, use of appropriate rhetorical mode, and register.

Note. Extracted from Council of Europe (2020)

The UCR-ELTA test provides stakeholders with access to evidence that shows valid and reliable information about language proficiency with respect to CEFR bands, including communicative activities, strategies, and language competences. Based on this information, different decision makers can report prospects' language performance to make relevant decisions.

UCR-ELTA's item design and creation

Considering (a) that test-takers will be carrying out a series of tasks which will be a reflection of the nature of language-connected scenarios they experience in their teaching life and (b) that those tasks will help collect evidence of teachers being highly independent (B2) or proficient users of the language (C1), the need for creating meaningful, construct-responsive, and high-quality items is fundamental. For this reason, items must be designed according to set parameters that will, hopefully, shed light on test takers' KAS. This explains why the process of item creation has to be thoroughly observed and developed and why item writers must be carefully selected, because they are the ones that concretize the minimum units of assessment: the tasks (Bachman and Palmer, 2010, p. 306). In this light, Table 5 displays the distribution and general specifications of the UCR-ELTA's tasks:

Table 5

Distribution and number of items and tasks for the UCR-ELTA

| Reading comprehension: 50 items	
Type	Number of items
B2 descriptors according to CEFR	30
C1 descriptors according to CEFR	20
Listening comprehension: 50 items	
Type	Number of items
B2 (descriptors according to CEFR)	30
C1 (descriptors according to CEFR)	20
Written Production: 2 tasks	
Type	Number of tasks
B2 (descriptors according to CEFR)	1
C1 (descriptors according to CEFR)	1
Oral Production: 1 task	
Type	Number of tasks
One-on-one, adaptive interview (interviewer moves up/down the CEFR proficiency scale as the interview progresses based on testees' answers.	1

The reason for choosing the 20 (C1)-30 (B2) distribution of items responds to the need to ensure that test takers' responses have the opportunity to adjust to the items (from the easiest one to the most difficult ones) and guarantee plenty of opportunities to expose themselves to relevant items that will provide significant information on the test takers' language abilities for listening and reading. In the case of speaking and writing, the number of tasks has been chosen based on the available resources, practicality, and task (input) authenticity.

Reading and listening items. Systematically and as illustrated in Figure 2, the cycle for item construction is carried out through a series of steps that seek to ensure the application of various quality filters to accept, improve, or discard items. Firstly, the designers are chosen based on their teaching, research, or testing experience; their qualifications are taken into consideration as well. This allows the recruitment of professionals that have the necessary expertise for carrying out a responsible and conscientious process of design and creation of items, a fundamental feature of any process of this nature.

Figure 2
Process of item construction for the UCR-ELTA's Listening and Reading Sections
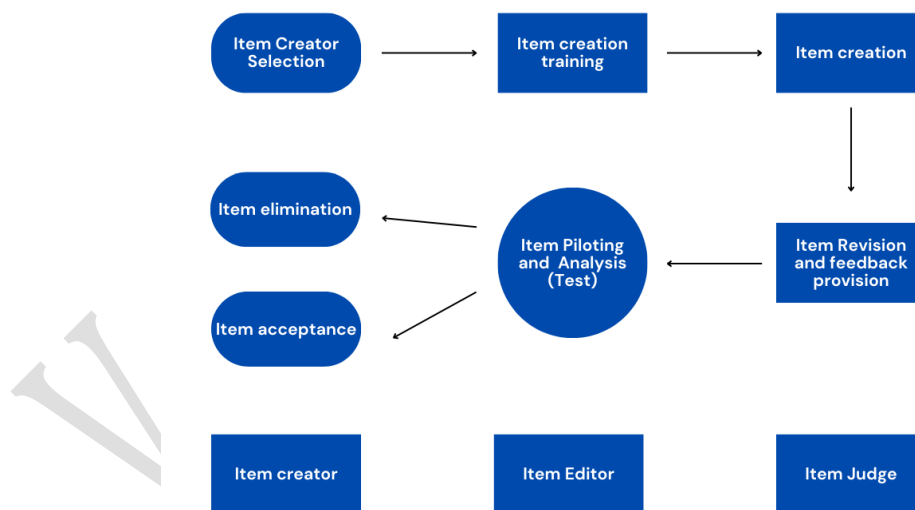
Note. Created by the authors

Secondly, once they have accepted to be part of the task design stage, item designers need to attend an informative session/workshop where the requirements their tasks must comply with are stated and exemplified. In addition, a set of instructions accompanying the training session and other documents for item design (e.g., assigned CEFR can do

statements; number of items to create) are shared with the designers as well. Furthermore, this training session has the overall purpose of informing the creators of the minimum quality standards and the steps to follow when fulfilling their responsibilities, and it also provides further guidance for creating the best items possible since the beginning.

Thirdly, the item creators start the designing stage. In this one, they must apply the necessary recommendations and follow the guidelines provided during the training sessions. To give more support and allow for doubt clarification, the item creators can reach out to the examination coordinator at any time. The designing stage takes around one month, and they will have to come up with a set of 20-50 different items for either reading or listening.

Finally, the item revision, feedback, and improvement processes begin when the designers send their first drafts to the test coordinator. This process provides opportunities for both parties to discuss the changes the items need and for filtering those ones that do not comply with the minimum and primary requirements of the quality standards. Either newer versions of the items can be requested and created, or the items are discarded completely once piloted. In addition, the process of item creation is complemented by a process of item judgment and editing in which they are improved as much as necessary to guarantee their sound construction. For the most part, the items that comply with the requirements of the construct, the parameters of internal consistency (e.g., Cronbach alpha analysis), and the indicators of difficulty and discrimination are the ones to be used.

Speaking and writing tasks. The task specifications for the writing tasks are handed in to the task designers. Based on them, the designers will create the tasks which will be then revised by an experienced team as many times as necessary until they are approved. The prompts and contexts presented to the students reflect the scenarios and themes teachers face frequently in their professional lives.

The adaptive oral interview method has been chosen as it is practical enough in terms of administration and resources available. The questions asked to the test takers will reflect a B2 or C1 nature and will be administered by experienced interviewers. The interviews can either be carried out in person or via virtual platforms (e.g., Zoom).

Test organization and administration

The UCR-ELTA is digital using three different modalities: online through the University of Costa Rica portal called *PELEx*, a hybrid application, through which the test is installed on a computer and the results are sent in real time to the servers of the University of Costa Rica. This procedure requires a minimal connection to the Internet. The third modality does not require Internet connection at all: the test can be installed on the takers' computers and then the results have to be imported to storage devices to later be incorporated into UCR servers.

The guidelines for applying and taking the test are prepared as well. The former involved the test invigilators (if applied in person) and contain the administration and technical steps to follow. The latter contain the general instructions test takers must follow. The responsibility of foreseeing the entire process and ensuring the appropriate conditions for the delivery of the test are both the responsibility of UCR (primarily) and of the institutions where the individuals will be taking the test. The workflows, storage, and transport processes are corroborated, surveilled, guaranteed, and reflected upon by UCR as well before, during, and after the application of the test.

Due to the fact that the UCR-ELTA is sectioned according to the skills being assessed (e.g., oral production, written production, listening, and reading), some aspects connected to its administration must be clarified.

Section of reading and listening. Once the design and filtering of the test has been finalized and as soon as its best version has been approved, the test administration is carried out. For this purpose, the written and listening sections of the examination are tested on an online platform designed and supported by IT experts, when the format of the test is either hybrid or online.

If the test is administered in a face-to-face environment, then a set of steps are followed: (a) trained personnel work as invigilator coordinators or invigilators themselves, (b) a group of other invigilators are recruited (these are usually other professors from the School of Modern Languages) to be able to supervise each of the assessed groups. The higher the number of test takers, the more invigilators are needed, so more people are recruited. All of the invigilators must attend a session where the purpose of the test, the security and safeguarding measures, and other logistics protocols are presented, explained and clarified; the steps to follow before, during, and after the test application are also presented to guarantee standardization and questions are attended to. UCR is in charge of guaranteeing the administration procedures are followed and, when unexpected situations happen before, during, or after the administration of the test, the invigilators must contact their coordinator to attend to any issue.

All of the documents and materials for the face-to-face test administration are kept under UCR's supervision before and immediately after the application. In addition, the invigilators only have access to them a few minutes before the administration, and every single document and materials (e.g. test booklets, answer sheets, attendance lists, among others) must be accounted for once the test ends.

Auditing processes are usually carried out after the application of the test with the purpose of identifying strengths and aspects to improve in future applications. Members of PELEx who were involved in the administration analyze the entire flow of the administration of the test and pinpoint the areas for improvement.

Sections of speaking and writing. The oral production skill is assessed virtually via Zoom or face-to-face. This section consists of a 10 to 20-minute interview in which a

multilevel set of questions are asked to the test takers. The oral interviewers are either certified or trained to carry out this a task. The questions and rubrics reflect the nature of the CEFR. The auditing processes here are related to making sure that the interviewing protocols are as standard as possible as well. However, this might be difficult to do since this auditing process requires a considerable amount of time to analyze the interview protocols, the inter-rater reliability, the interview-protocol application, the reception process, among others. This is, indeed, a work in-progress.

In the case of the writing section, three different tasks are provided to the test takers. The tasks are connected to each of the CEFR bands under assessment and contain context that reflects the scenarios and topics teachers handle on a daily basis. The test takers have around 35 minutes to complete the tasks and provide their answers and complete this section on the same date they complete the listening and reading sections.
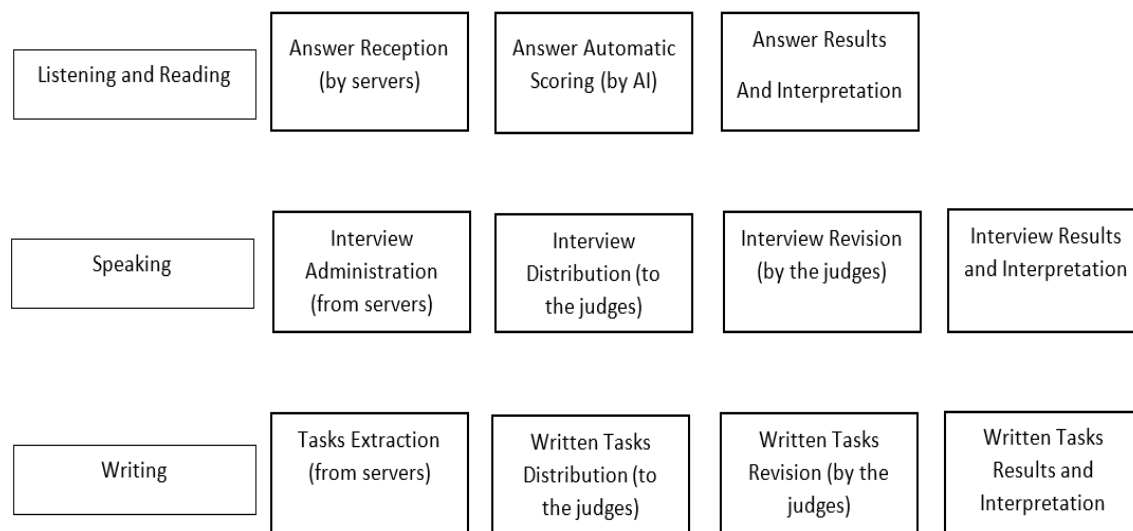
Test scoring

The following are the overall procedures for scoring the different items in the test.

Reading and listening tasks. The general item scoring process for reading and listening also consists of a series of steps illustrated in Figure 3.

Figure 3

General Process of Scoring of the UCR-ELTA Test Tasks and Items

Note. Created by the authors

More specifically, in the case of the reading and listening sections, the test takers' responses are automatically scored using AI, and the results are extracted from the server in charge of storing the data. After this, the results are sent to the administration and

coordination of the test. As noticed, the process requires access to technological tools, servers, and scoring software and programming that must ensure the responses of the students are actually saved and secured appropriately. The later process of result interpretation can then initiate once the answers have been obtained.

Considering the relevance of this test and the standards set for the minimum language use requirements for English teachers, the summative parameters for categorizing the test takers' performance is shown in Table 6.

Table 6

Quantitative Parameters for CEFR Band Categorization

Total Items	100
Minimum parameters to be placed in the band	C1-80% of correct answers
	B2-55% of correct answers

Note. Taken from PELEx (2022a), UCR-ELTA's Table of Specifications

The number of minimum correct responses has been decided based on the fact that the test is a high-stakes one and high-quality performance is expected considering the possible uses it will have (for hiring and retraining).

Speaking and writing tasks. The oral interviews and written tasks are scored manually. In the case of the written tasks, experienced language professors are recruited to assign a score to an anonymous set of tasks. The scorers must have significant experience in teaching English composition and rhetoric. This guarantees that the nature of the skill under assessment is more profoundly understood. However, this does pose a difficulty: the raters must be trained to assess the written tasks considering the criteria stated (e.g. Can Do statements) and not as writing teachers with particular views on writing and classroom writing needs. To attend to this need, the scorers do attend a meeting/workshop in which a series of written tasks samples are used as examples and context to calibrate the understanding of the scoring criteria.

The scorers take around 6 weeks to check the assigned tasks using a rubric prepared for that purpose. After that, they must send the tasks and scores to the "referee" -a person in charge of organizing the information and carrying out the inter-rater agreement process. There are two scorers per band, per task, so, when there is a disagreement in the scores and, after double-checking the task, the "referee" is in charge of making the final decision as to which score must be assigned to a test taker.

In the case of the oral section, the scoring process is quite similar. Nonetheless, all of the interviews are assessed by two judges trained to assess these kinds of tasks. The calibration processes are carried out as well, and the role of the "referee" is the same one as played for the written section.

Test interpretation, validation, and awareness of its impact

As stated previously, this proficiency test revolves around the need to explicitly present a set of parameters for linguistic skills and abilities within the B2 and C1 levels of proficiency, which can serve as a base for employment and retraining opportunities. For this reason, validators of the results of the test will be analyzing the systems developed for designing and administering the test and interpreting the results. Hopefully, these analyses on the coherence, cohesiveness, and accuracy of the test will provide useful information and feedback to incorporate later and better these processes.

Indeed, it is of extreme importance to refer more in-depth to the kinds of decisions that the results of UCR-ELTA can be used as support for. In the first place, there is a clear acknowledgement to the fact that this test is a language proficiency one; this evidently means that no assumptions regarding the teaching abilities of the test takers can be made through this test. In order words, the UCR-ELTA just shows certain evidence of what the test takers can do with the target language under very specific conditions, but this cannot be extrapolated to conclusions regarding their actual language teaching competency in a classroom setting. Of course, it can be noted that, as language consultants and exemplary users of English, teachers should have vast knowledge of the target language to handle it in various scenarios relevant to what the students need to do; therefore, expectations regarding the levels of English they must have are evident: the more proficient users they are, the better. This test can shed light on this point.

Conclusion

The UCR-ELTA test has the potential to offer different benefits. The process of developing, administering, and interpreting the results of such a test can provide valuable insights into the abilities and skills of the test takers. These insights can be used to make informed decisions in various domains, especially employment, professional development, and education.

The development of this test involves a meticulous process that ensures the test is comprehensive, fair, and reliable. It is designed to measure the specific skills and knowledge that are relevant to the test takers. Furthermore, the administration of the test has to be carefully planned and executed to ensure that all test takers are given an equal opportunity to demonstrate their abilities; in other words, fairness should be guaranteed.

Through the appropriate interpretation of results, valuable information about the test takers' language performance in English can be collected. Moreover, it can highlight their strengths and areas for improvement, providing them with valuable feedback that can further guide their language learning and professional development. Furthermore, the results can be used to identify trends and patterns, providing insights that can inform policy and practice.

However, the use of the UCR-ELTA test results extends beyond the individual test takers. There is a growing interest among test users, particularly employers, to utilize results as

parameters for hiring, which might have a meaningful impact: a high score on the test can open up opportunities for employment and professional advancement, while a low score can limit these opportunities.

This potential impact on the professional lives of the test takers is not to be taken lightly; as a consequence, test designers must ensure that the test is developed and validated responsibly, transparently, and consciously through the continuous improvement and systematization of the chain of analysis, design, auditing, and reflection of the test. Transparency, in this sense, further supports (a) the process of evidence collection that ensures the test is reliable and consistent, and (b) the provision of support and resources to help test takers prepare for the test and interpret its results as intended.

References

Araya Garita, W., Elizondo González, J., & González Ramírez, A. (2022). Un acercamiento al constructo de la prueba de dominio lingüístico del idioma inglés desarrollada por la Universidad de Costa Rica para el Ministerio de Educación de Costa Rica. *Estudios de Lingüística Aplicada*, *40*(75), 119-143. https://doi.org/10.22201/enallt.01852647p.2022.75.1013

Association of Language Testers in Europe (ALTE). (2020). *ALTE principles of good practice.* Council of Europe. https://alte.org/resources/Documents/ALTE%20Principles%20of%20Good%20Practice%20Online%20version%20Proof%204.pdf

Association of Language Testers in Europe (ALTE). (2011). *Manual for language test development and examining.* Council of Europe. https://www.alte.org/resources/Documents/ManualLanguageTest-Alte2011_EN.pdf

Bachman, L., & Palmer, A. (2010). *Language assessment in practice*. Oxford University Press.

Bolitho, R. (2013). Dilemmas in observing, supervising and assessing teachers. In P. PD (Ed.), *Assessing and Evaluating English Language Teacher Education, Teaching,*

and Learning (pp. 7-12). British Council.

https://www.britishcouncil.in/sites/default/files/tec12_publication_1.pdf

Brown, D., & Abeywickrama, P. (2019). *Language assessment: Principles and classroom practices.* Pearson Education.

Cambridge University Press & Assessment. (2022). *Cambridge English teaching framework*. https://www.cambridgeenglish.org/teaching-english/professional-development/cambridge-english-teaching-framework/

Chapelle, C. A. (2012). Validity argument for language assessment: The framework is simple… *Language Testing*, *29*(1), 19-27.

Chapelle, C. A. (2008). The TOEFL validity argument. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 319-352). Routledge.

Chu, R. K., & Jaca, C. A. (2019). A study on the factors affecting the professional development levels of ESL Filipino teachers according to the Cambridge English Teaching Framework. *International Journal of English Education*, *8*(1), 420-441. https://ijee.org/assets/docs/34cristie.373846.pdf

Coombe, C., Folse, K., & Hubley, N. (2007). *A practical guide to assessing English language learning*. The University of Michigan Press.

Coombe, C. (2018). *An A to Z of second language assessment: How language teachers understand assessment concepts*. British Council. http://www.britishcouncil.org/exam/aptis/research/assessment-literacy

Council of Europe. (2020). *Common European Framework of Reference for Languages:*

  Learning, teaching, assessment. https://rm.coe.int/common-european-framework-

  of-reference-for-languages-learning-teaching/16809ea0d4

Consejo Nacional de Rectores (CONARE). (2019). *Desafíos de la educación en Costa Rica*

  y aportes de las universidades públicas. CONARE-OPES.

  https://repositorio.conare.ac.cr/handle/20.500.12337/7953

Diálogo Interamericano y Unidos por la Educación. (2018). *Costa Rica: El estado de*

  políticas públicas docentes. Estado de la Educación.

  https://www.thedialogue.org/wp-content/uploads/2018/08/El-estado-de-politicas-

  publicas-abril-15.pdf

Dadvand, B., & Behzadpoor, F. (2020). Pedagogical knowledge in English language

  teaching: A lifelong-learning, complex-system perspective. *London Review of*

  Education, *18*(1), 107–125. https://doi.org/10.18546/LRE.18.1.08

Education First. (2023). *Índice del dominio del inglés de EF: Una clasificación de 113*

  países y regiones en función de su nivel de inglés.

  https://www.ef.com/assetscdn/WIBIwq6RdJvcD9bc8RMd/cefcom-epi-

  site/reports/2023/ef-epi-2023-spanish.pdf

Fulcher, G. (2010). *Practical language testing*. Hodder Education.

Green, A. (2021). *Exploring language assessment and testing: Language in action* (2nd

  ed.). Routledge.

Kellaghan, T., Madaus, G., & Airasian, P. (1982). *The effects of standardized testing*.

  Kluwer- Nijhoff Publishing.

Loredo, J. (2021). Evaluación docente. *Revista Iberoamericana de Evaluación Educativa*, *14*(1), 7-11. https://doi.org/10.15366/riee2021.14.1.001

Marco Nacional de Cualificaciones para las Carreras de Educación (MNC-CE-CR). (2021). *Resultados de aprendizaje de la carrera de enseñanza del inglés*. https://cualificaciones.cr/mnc-ce/images/documentos/carreras/MNCCE_INGLES.pdf

Marín-Arroyo, E. (2013). La enseñanza del inglés en Costa Rica en el siglo XIX: Una respuesta al modelo económico. *Revista Comunicación, 13*(2), 47-55. https://revistas.tec.ac.cr/index.php/comunicacion/article/view/1132

O'Sullivan, B. (2016). Adapting tests to the local context: New directions in language assessment. *JASELE Journal* special edition*, 145-158.

Programa de Evaluación de Lenguas Extranjeras (PELEx). (2022a). *UCR-ELTA: Table of Specifications* [Unpublished document]. Universidad de Costa Rica

Programa de Evaluación de Lenguas Extranjeras (PELEx). (2022b). *Usos específicos del idioma inglés por parte de los docentes de enseñanza del inglés en Costa Rica*. [Unpublished raw data]. Universidad de Costa Rica.

Política Educativa de Promoción de Idiomas. (2019). *Hacia una Costa Rica bilingüe*. Ministerio de Educación Pública. http://cse.go.cr/sites/default/files/acuerdos/politica_educativa_para_la_promocion_de_idiomas.pdf

Romero, J. (2007). Concepciones de evaluación y de evaluación docente. *Cuadernos de Lingüística Hispánica*, *10*, 137-148. https://www.redalyc.org/pdf/3222/322227484009.pdf

Rueda, M. (2009). La evaluación del desempeño docente: Consideraciones desde el enfoque por competencias. *Revista Electrónica de Investigación Educativa*, *11*(2), 1-16. https://www.redalyc.org/pdf/155/15512151004.pdf

Sarwar, M. Alam, M., Hussain, S., Shah, A., & Jabeen, M. (2014). Assessing English speaking skills of prospective teachers at entry and graduation level in teacher education program. *Language Testing in Asia, 4*(5). http://www.languagetestingasia.com/content/4/1/5

Shohamy, E. (2001). *The power of tests: A critical perspective on the use of language tests*. Routledge.

Slomp, D. H. (2005). Teaching and assessing language skills: Defining the knowledge that matters. *English Teaching: Practice and Critique*, 4(3), 141-155. https://files.eric.ed.gov/fulltext/EJ847274.pdf

Tschirner, E. (2018). Language Testing: Current practices and future developments. *Die Unterrichtspraxis/Teaching German*, *51*(2), 105-120. https://www.jstor.org/stable/10.2307/90026419