Getting stakeholders acquainted with the rationale behind the construct of the English language proficiency test of the University of Costa Rica for the Ministry of Education of Costa Rica

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

Walter Araya Garita
Universidad de Costa Rica,
Facultad de Letras,
Escuela de Lenguas Modernas
walter.arayagarita@ucr.ac.cr

Ana Carolina González Ramírez
Universidad de Costa Rica,
Facultad de Letras,
Escuela de Lenguas Modernas
ana.gonzalezramirez@ucr.ac.cr

José Fabián Elizondo González
Universidad de Costa Rica,
Facultad de Letras,
Escuela de Lenguas Modernas
josefabian.elizondo@ucr.ac.cr

Recepción: 27 de mayo de 2021
Aceptación: 21 de septiembre del 2021
doi: 10.22201/enallt.01852647p.2022.75.1013

Abstract

This paper collects evidence to build a content-validity argument of the language proficiency test of the University of Costa Rica for high school students by analyzing the theoretical foundations supporting the construction and administration of a custom-made language test and its localized context, including a description of the test and the input of external stakeholders. Additionally, this paper provides suggestions and recommendations for future testing experiences of this type, while paving the way for researchers who intend to follow this area of interest.

Keywords: language proficiency; standardized testing; content validity; localization; foreign language assessment

Resumen

Este artículo tiene como objetivo recopilar evidencia para construir un argumento de validez de contenido acerca de la prueba de dominio lingüístico de la Universidad de Costa Rica para estudiantes de secundaria, a través de un análisis de los fundamentos teóricos que sustentan la construcción y administración de una prueba de idioma hecha a medida para un contexto localizado. El documento incluye una descripción de la prueba y las aportaciones de las principales partes interesadas. Al mismo tiempo, este artículo proporciona sugerencias y recomendaciones para futuras experiencias en pruebas de este tipo, mientras que allana el camino para futuros investigadores que pretendan continuar con estos esfuerzos académicos.

Palabras clave: dominio lingüístico; evaluación estandarizada; validez de contenido; localización; evaluación de lengua extranjera

1. Introduction

In 2016, the Costa Rican Ministry of Education (Ministerio de Educación Pública; MEP, for its abbreviation in Spanish) modified the English curriculum nationwide. In a national effort coined as Alliance for Bilingualism, the MEP implemented multiple changes in the English programs to meet cultural, societal, and financial demands as a strategy to transform Costa Rica into a bilingual country, which in turn would attract investors, generate jobs, revitalize the economy, and foster study opportunities abroad (Azofeifa, 2019). Additionally, in 2019 the MEP decided to eliminate their traditional national English tests, which were pass-or-fail, reading comprehension multiple-choice tests. Before 2019, senior high school students were required to obtain a mark of at least 70 out of 100 to be eligible for graduation. Students failing to meet this score would require taking this test as many times as needed to achieve the minimum pass score, impeding them from starting college or getting a job. However, instead of administering the traditional test, the MEP opted for administering a language proficiency test, which is diagnostic in nature as defined by Brown and Abeywickrama (2019: 10). This new test will not evaluate content but the English performance of students based on the Common European Framework of Reference for Languages (CEFR descriptors) (Cordero, 2019: November 29).

Despite the multiple language certifications currently available, none seems to meet MEP's specific requirements and needs. First, the test administration and publication of its results must consider MEP's annual schedule, which requires test administrators to carry out all related activities within a very tight time window. Second, the uneven distribution and availability of resources have introduced challenges for public education institutions. Finally, MEP's economic restrictions should also be considered when choosing between the different options for English certification.

In an attempt to address these needs, the School of Modern Languages (Escuela de Lenguas Modernas; ELM, for its abbrevia-

tion in Spanish) of the University of Costa Rica (UCR) decided to create a more locally-sensitive option called Language Proficiency Test (prueba de dominio lingüístico; PDL). Over the past 30 years, the ELM has accumulated vast experience in the design and administration of a variety of reliable language tests providing valid evidence supporting the interpretation of their scores according to the intended uses. The language proficiency test designed for the English for Specific Purposes program of the National Council of University Presidents — a test for faculty members, and the official certification of translators and interpreters, among others — attest to ELM's expertise in the field, broadened by the guidance of international language testing professionals. The School not only administers its own language tests but is also an internationally recognized center for some renowned language certifications. This acknowledgment has required many faculty members to receive extensive training and gain valuable experience as authorized certified examiners. In addition, the ELM possesses the know-how for administering large-scale, high-stakes tests nationwide, such as the Entrance Examination designed by the Psychological Research Institute and the School of Statistics at UCR. More recently, the ELM has started digitizing some of its tests to facilitate their application and recording of results. This test automatization has widened the options towards considering three possible formats for test administration: online, offline, and hybrid (a combination of offline and online options that requires minimum bandwidth) suitable for the digital infrastructure of Costa Rica's public high schools, which may or may not have a strong Internet connection.

2. Literature review

Standardized language testing may seem overwhelming and intimidating to many stakeholders,[1] especially educators, whose work

[1] The term *stakeholder* is understood as "[any] person whose interests should be taken into consideration" (Coombe, 2018: 38). Furthermore, the four types

effectiveness is at stake, and their students. As Shohamy (2001, as cited in Fulcher, 2010: 8) argued, "one reason why test takers and teachers dislike tests so much is that they are a means of control". Students conceive this type of testing as a punishment focused on identifying their weaknesses and areas for improvement; teachers, on the other hand, believe it to a demonstration of lack of trust in their expertise by supervisors. Brown and Abeywickrama (2019: 1) summarized this view when they stated that students and teachers, both stakeholders of the assessment process, "are not likely to view a test as positive, pleasant, or affirming". Despite these negative perspectives, different scholars have viewed standardized language assessments as a means through which distributive justice, as Messick (1989, as cited in Fulcher, 2010: 4) proposed, could be achieved. Fulcher (2010: 4) further acknowledged the importance of testing as necessary and acceptable worldwide norms on which high-stakes decisions can be made, not only for those in charge of designing the instruments but also for test users, who need to see the test as designed, administered, scored, and reported fairly and equitably.

The language competence assessment concept has been continuously redefined through time, according to the needs of users and the evolution of language teaching and learning theories. Authors such as Oller (1979, as cited in Brown & Abeywickrama, 2019: 13) stated that during the decades of 1970 and 1980, language competence was viewed as "a unified set of interacting activities that could not be tested separately". Cloze and dictation exercises, where several skills were assessed simultaneously, represented the language competence concept. However, during the mid-1980s, Canale and Swain (1980, as cited in Brown & Abeywickrama, 2019: 14) recommended a shift from this struc-

of stakeholders described by Hooge and Helderman (2008, as mentioned in Hooge, Burns & Wilkoszewski, 2012: 13) are also used in this paper, to wit: primary stakeholders, internal stakeholders, vertical stakeholders, and horizontal stakeholders.

ture-centered approach to assessment towards a more communicative one, which involved more real-life tasks that language learners may eventually face. Accordingly, Savignon (1985: 131) agreed that "communicative competence certainly requires more than knowledge of surface features of sentence-level grammar". What is more, when it comes to authenticity in assessment, Bachman (1990) and Weir (1990: 86, as cited in Brown & Abeywickrama, 2019: 16) highlighted the importance of asking questions such as "where, when, how, with whom, and why language is to be used, and on what topics, and with what effect" in order to measure language competence. Supporting this view, Jamieson, Eignor, Grabe and Kunnan (2008: 57) asserted that communicative competence "accounts for language performance across a wide range of contexts, includes complex abilities responsible for a particular range of goals and takes into account relevant contexts". More recently, Bachman and Palmer (2010, as cited in Brown & Abeywickrama, 2019: 15) included "the need for a correspondence between language test performance and language use" among the fundamental principles of language testing. This more realistic communicative view of language assessment permeates some of the most renowned language tests currently on the market.

Today, assessing communicative language competence is performed more holistically. Since proficiency in a given language goes beyond knowing its grammar, other equally — if not more — important features should also be accounted for when testing language proficiency. In fact, as Badger and Yan (2012: 7), stated, "the main feature of the pedagogic orientation of a CLT [Communicative Language Teaching] course is students' ability to use the second language (L2), rather than knowledge about language, with a balance between the four skills". Along these same lines, the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2002) provides a framework that lists the necessary communicative language activities and strategies, as well as the communicative language competences (linguistic, sociolin-

guistic, and pragmatic) that should be considered when designing language assessment instruments. Likewise, the American Council on the Teaching of Foreign Languages (ACTFL) shares CEFR's emphasis on communication, expanding it to an intercultural communication approach. More recently, ACTFL coined the term *intercultural communicative competence,* defined as "using language skills, and cultural knowledge and understanding, in authentic contexts to effectively interact with people. It is not simply knowing about the language and about the products and practices of a culture" (Van & Shelton, 2018, January: 35). Hence, it is evident that the concept of mastering a language as a second or foreign language by a speaker keeps changing as new theories continue to evolve.

One may think that the analysis and construction of standardized tests may have reached a stagnation point; however, the validation of language assessment is an ongoing process (Chapelle, 2012; Brown & Abeywickrama, 2019). A key step in the test validation process entails defining validity: "an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment" (Messick, 1989, as cited in Messick, 1995: 741). Therefore, this concept "is not a property of the test or assessment as such, but rather of the meaning of the test scores" (Messick, 1996: 245). Recently, this view has been supported and expanded in the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], National Council on Measurement in Education [NCME], and Joint Committee on Standards for Educational and Psychological Testing [US], 2014: 11), which further highlights the importance of "accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretations". The second step entails collecting evidence to build a validity argument, which analyzes it "to make a case justifying the score-based

inferences and the intended uses of the test" (Carr, 2015: 331). To operationalize this validity argument, the guide *Standards for Educational and Psychological Testing* outlines five sources of validity evidence: evidence based on response processes, evidence based on internal structure, evidence of relations to other variables, evidence for validity and consequences of testing, and content-based evidence (American Educational Research Association [AERA] *et al.*, 2014).

In light of the vast scope of validation processes, the numerous types of evidence available to demonstrate it, and the multiple research approaches that can be adopted, this paper will attempt to gather 'content-based evidence' as its primary focus aiming to meet some of the validity standards stated above. First, a test can claim to have content validity "if [it] actually samples the subject matter about which conclusions are to be drawn, and if it requires the test-taker to perform the behavior measured" (Brown & Abeywickrama, 2019: 32). To add further detail to the elements to be analyzed and demonstrate content validity, the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA] *et al.*, 2014: 14) state that "test content refers to the themes, wording, and format of the items, tasks, or questions on a test". Content validity is also linked to being ecologically sensitive: "serving the local needs of teachers and learners. What this means in practice is that the outcomes of testing — whether these are traditional 'scores' or more complex profiles of performance — are interpreted in relation to a specific learning environment" (Fulcher, 2010: 2). More recently, other authors have also studied this concept; some of them, such as Coombe (2018: 28), even renamed it *localization* and have defined it as:

> A test that is designed to cater to 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.

Based on the concepts and context provided above, it is possible to affirm that UCR's language proficiency test supports this first claim: it serves the local needs of teachers and learners in Costa Rican high schools as test users can interpret students' scores and profiles of performance as sensitive to this specific learning environment.

Last, the final block towards building a solid validity argument for a standardized test consists of aligning a test with language proficiency descriptors such as those provided by the CEFR the characteristics of test takers and administrators (O'Sullivan, 2016: 148). Similarly, the Association of Language Testers in Europe (2020: 26) further emphasizes that the test should be linked to a theoretical construct as a minimum standard in test construction. Hence, a second claim to be made about UCR's language proficiency test is that this test is properly aligned with the CEFR language proficiency descriptors.

3. Description of the test

3.1. *Validity argument*

For the purpose of the validity argument in this paper, we provide the following descriptions and analyses as suggested in the *Standards for Educational and Psychological Testing* (section 4.1).

3.1.1. Purpose of the test

The purpose of the PDL-MEP test is to assess Costa Rican high school students regarding their understanding and production of non-technical English related to both regional and global contexts pertaining to formal and informal socio-interpersonal, transac-

tional, and academic domains while using the CEFR descriptors as reference. This test is merely diagnostic. All senior high school students in Costa Rica who take this test will be able to know their proficiency level in terms of reading and listening comprehension skills, according to the CEFR. This will not be considered a language certification test; hence, it should not be used for college admissions, visa applications, or job applications.

3.1.2. Score interpretation

As indicated by MEP's authorities, the purpose of this new test is to determine the language proficiency of students as a means to diagnose the efficacy of the language programs recently adopted in Costa Rica, as well as the language development stage of students. Hence, testees should interpret the results as a reflection of progress in their foreign language education, which will evidence the areas where they perform strongly and those that need improvement. Although students are not required to obtain a particular score to graduate from high school, they should take this test as a requirement for graduation.

Based on the scores obtained nationwide, the MEP might make the necessary adjustments to better achieve its goal: getting students to perform at the B1 level by the end of their high school years. To illustrate such adjustments, if the test results show a clear lack of command of B2-level tasks, MEP teachers will be able to use this information and address the deficiencies identified by reinforcing the unsatisfactory tasks in class. Internal stakeholders, as outlined by Hooge, Burns, and Wilkoszewski (2012: 13) might use the test results to determine, for example, where to recruit new bilingual personnel or whether to invest in additional language programs for underprivileged populations.

3.2. *The constructs of the test*

3.2.1. Reading comprehension

Reading comprehension proficiency is defined as demonstrating an understanding of non-technical texts in English related to both regional and global contexts that pertain to formal and informal socio-interpersonal, transactional, and academic domains, taking CEFR's descriptors as reference. The contents to be included are determined following the MEP guidelines. Furthermore, the skills assessed range from recognizing "familiar words accompanied by pictures, such as a fast-food restaurant menu illustrated with photos or a picture book using familiar vocabulary" to understanding "in detail lengthy, complex texts, whether or not they relate to [examinees'] own area of speciality" (North, Piccardo & Goodier, 2018: 60). Finally, some of the strategies to be demonstrated by testees are included in the CEFR descriptors, such as skimming, scanning, understanding a writer's tone and humor, and identifying attitudes and implied opinions (CEFR, 2018).

3.2.2. Listening comprehension

Listening comprehension proficiency is defined as demonstrating an understanding of non-technical English aural texts related to both regional and global contexts that pertain to formal and informal socio-interpersonal, transactional, and academic domains, using the CERF descriptors as reference. The contents to be included are determined following the MEP guidelines. Some of the skills to be assessed range from recognizing "numbers, prices, dates, and days of the week, provided they are delivered slowly and clearly in a defined, familiar, everyday context" to following "extended speech even when it is not clearly structured and when relationships are only implied and not signaled explicitly" (CEFR, 2018: 55). Last, some of the strategies to be demonstrated by testees are encompassed in the CEFR descriptors, such as understanding

the main ideas and specific details, making inferences, and discerning speakers' attitudes (CEFR, 2018).

3.3. *Claims*

3.3.1. Claim 1 (MEP)

The UCR language test gives the MEP valid and reliable information about the English performance of students regarding nationwide language standards and CEFR proficiency bands, including communicative activities, strategies, and language competences. Based on this information, the MEP can report students' language performance by classroom, school, district, and region. With this in mind, the Ministry will be able to design strategies to focus on those areas most in need of support regarding language proficiency.

3.3.2. Claim 2 (teachers)

The UCR language test gives the MEP valid and reliable information about the English performance of students regarding nationwide standards and CEFR proficiency bands, including communicative activities, strategies, and language competences. Based on this information, the Ministry of Education can adjust classroom activities — formative and summative — to meet the standards established by the MEP.

3.3.3. Claim 3 (parents and students)

The UCR language test gives parents and students valid and reliable information about the English performance of students regarding nationwide language standards and CEFR proficiency bands, including communicative activities, strategies and language competences. Based on this information, these stakeholders can determine students' progress across the entire education system.

3.4. *Rationale*

Given the scenario described above, the Ministry of Education and the University of Costa Rica agreed to build, administer, and deliver the results of an English language proficiency test that would represent a customized and convenient option for both institutions. Accordingly, the MEP would have an instrument that meets their specific needs and the UCR would have another opportunity to give back to society all the knowledge it has acquired through research and its socially-oriented education programs over the years. Moreover, as part of a well-documented and reliable process, "language testers shall endeavor to communicate the information they produce to all relevant stakeholders in as meaningful a way as possible" (International Language Testing Association, 2018: 2). This transparency is particularly important in documenting such a pioneering initiative whereby a Ministry of Education in a Latin American country enforces a national policy of bilingualism in conjunction with a higher education public institution through a large-scale language test.

This article aims to gather evidence to build a PDL-MEP content validity argument through an analysis of the theoretical foundations supporting the construction and administration of a customized standardized language test and its localized context, including a description of the test and the input of the internal stakeholders. Additionally, this article provides suggestions and recommendations for future testing experiences of this type while setting the grounds for future researchers who intend to follow this academic approach.

The following literature review has been included as a reference to support both the claims established in the validity arguments proposed above and the claims to define the construct of the UCR Language Proficiency Test (content domain).

3.5. *Localized test context*

3.5.1. MEP diagnostic test for high school students

To meet the localization principles mentioned above, the national English language proficiency test will assess the reading and listening comprehension skills of high school students — as per MEP's request — based on the themes, domains, and scenarios set by the Ministry of Education in its *Programas de Estudio de Inglés* (English Language Programs), which have been aligned with the CEFR guidelines (Ministerio de Educación Pública, 2016). The topics addressed in this document include, but are not limited to, conflict resolution, democracy and democratic principles, economic development, environmental sustainability, blurring of national borders, and human rights defense and protection (Ministerio de Educación Pública, 2016: 13). Three axes encompass all these topics: a global citizenship with local belonging, education for sustainable development, and new digital citizenship (Ministerio de Educación Pública, 2016: 13, 55). The contexts or domains where the target language is to be used and selected for this test include socio-interpersonal, transactional, and academic domains (Ministerio de Educación Pública, 2016: 38). Among the multiple scenarios provided by the MEP for all secondary education levels, the test may include, for example, "Enjoying Life" (7th grade), "Going Shopping" (8th grade), "Lights, Camera, Action" (9th grade), "Stories Come in All Shapes and Sizes" (10th grade), and "The Earth – Our Gift and Our Responsibility" (11th grade). Therefore, a third claim is that UCR language proficiency test items address the topics and themes identified as important by the Ministry of Education, including, but not limited to, conflict resolution, democracy and democratic principles, economic development and environmental sustainability, blurring of national borders, and defense and protection of human rights. A fourth claim is that the test items address the three axes identified by the Ministry of Education: a global citizenship with local belonging, education for sustainable development, and

new digital citizenship. Finally, a fifth claim is that the test items reflect three domains in which students must demonstrate English language proficiency: socio-interpersonal, transactional, and academic domains.

The contextual needs addressed above will be further operationalized by simple and clear wording items and instructions that not only meet the expected language proficiency levels of testees but also comply with the design requirements of trained language specialists. Brown and Abeywickrama (2019: 74) warned against wordiness, redundancy, and unnecessarily complex lexical items that might confuse the testee. Consequently, to ensure that test takers can understand what is expected of them, the language complexity in the instrument should correspond to the one described in the respective CEFR band. The "Text Inspector" tool (Cambridge University Press, 2015) will be used to guarantee this match. Finally, as mandated by the *Standards for Educational and Psychological Testing,* items will be designed and proofread by trained specialists, some of whom are native English speakers and whose teaching expertise is also valuable (American Educational Research Association [AERA] *et al.*, 2014: 75).

The format of the items, tasks, and questions will follow the guidelines in the document *Programas de Estudio de Inglés* (Ministerio de Educación Pública, 2016). For example, the MEP (2016: 44) suggests the following items to assess reading comprehension proficiency in the classroom: "reading aloud, multiple choice, and picture-cued items. Selective reading performances are gap filling, matching tasks, and editing". The test will prioritize those tasks that lend themselves to be used in large-scale, standardized, computer-based scenarios. The MEP also provides a list of possible tasks to be used to assess the listening skill, such as summarizing, note taking, and identifying specific information (Ministerio de Educación Pública, 2016: 42). These tasks should all mimic real-life scenarios similar to those that students encounter in the classroom over their learning process.

3.5.2. Stakeholders' expectations

To further contextualize the test and accurately place it within the given ecosystem, stakeholders' views should also be carefully considered and analyzed. This analysis includes the information provided by two of the most relevant decision-makers: the national coordinator of the Alliance for Bilingualism at the MEP and the director of the School of Modern Languages at the UCR. The former acknowledged that the transition from a traditional reading comprehension test to a skills-based one took the MEP approximately 11 years, led by the *Dirección de Gestión y Evaluación de la Calidad* [Quality Management and Evaluation Department]. The first administration of the test aims to diagnose the English proficiency level of high school students; comparing these results against those to be obtained in the future will demonstrate the impact of the recently introduced *Programas de Estudio de Inglés*. Said impact could then be further analyzed using predictive validity evidence studies. In the words of the national coordinator of the Alliance for Bilingualism at the MEP (personal communication, November 5, 2020), this first test administration will also determine whether or not the MEP has the adequate physical and digital infrastructure to administer the test to more than 60 000 students. The coordinator also stated that this test will help other interested organizations to evaluate their adaptation capacity to the potential scenarios that may arise when monitoring their tests at the MEP. For example, some students might need to take the test at facilities other than their school due to poor or no Internet connectivity. Considering the multiple circumstances that regions or institutions may face (e.g., insufficient number of working computers, lack of personnel to supervise the test administration, or the variety of school types), several administration schedules should be arranged, for example, three or four different agreed-upon times according to the particular requirements of each institution. Additionally, this stakeholder emphasized some of the requirements to be met by any

candidate testing organization when working with the MEP: availability of immediate technical support, verifiable language testing experience, standardized administration protocols, and provision and modification of physical resources according to the particular needs of testees.

The director of the School of Modern Languages at the UCR — the second stakeholder interviewed in this study — highlighted several points (personal communication, December 15, 2020). The School of Modern Languages presents itself as a valuable part of the process because of its significant technical and human capacity and previous experience in standardized language assessment. The director underlined the fact that the School has the necessary capacity, in terms of technology and human resources, to administer such a high-stakes test successfully. However, he also recognized that further support and investment from UCR authorities would be advisable to improve computer laboratories, security protocols, and data collection and analysis instruments. He also acknowledged the added value that collaborating with other University departments could provide in the future. In terms of staff, the institution has properly trained personnel for the successful administration of the test in any of the formats requested by the MEP, although additional support for the continuous training of these professionals would be advisable.

In terms of the School capacity, the director affirmed that the University can administer this test three times a year at large-scale, specifically examining 5000 MEP students per day, with results delivered within three weeks. Moreover, the technological know-how and experience in testing facilitate any specific adaptations and modifications that the MEP may require. Finally, the director highlighted the importance of familiarizing the target population with the test by facilitating a customized digital mock test to bridge the gap between their knowledge of computerized testing and classroom testing.

The director expressed his confidence in the reliability of the listening and reading online tests based on the evidence gathered

but also considers that using printed instruments could be feasible, and equally reliable, depending on the needs of the target population. Both types of formats can be equally reliable regarding data collection, provided they fulfill the safety and procedural protocols designed by the UCR.

Production skills may be evaluated in the future, although further training and studies are necessary to ensure the reliability and validity of the results assumptions based on student performance. As a novel feature, he elaborated on the future possibility of using artificial intelligence as a tool in the development and scoring of UCR language tests.

This stakeholder believes the test should accurately measure the language level of testees by evaluating their understanding of everyday and academic English, which is the core of the construct approved by the MEP for this test. This goal will be achieved by using an instrument that will include 40 to 60 items per skill and which would take approximately 60 to 70 minutes to complete per macro skill. The items may include multiple-choice, sequencing, matching, drag-and-drop, and short answers; the specific items selected will depend upon how these perform in pilot testing. Finally, the director added that the text reading and listening sources will be authentic and ecologically sensitive material that will match the CEFR levels intended to assess. The operationalization of this test and its construct have been approved by the MEP.

4. Next steps

Given the large-scale nature of this assessment and its implications, and as a pioneering enterprise, institutions should work together in organizing the test logistics. The expertise of the involved parties (in this case, MEP and UCR) is key in successfully attaining the goals set by localizing the test. This would require the University of Costa Rica to conduct the following:

1) Arrange meetings with MEP representatives and decision-makers to agree on the operationalization of the test features
2) Carry out analyses of needs:
   a. Survey teachers of English in Costa Rican high schools regarding, among other topics, their technological literacy, type of instruction, expectations from the test, and attitude towards standardized testing
   b. Gather information on the views and experience of students with standardized and computer-based testing, familiarity with online testing and item types, and preferred topics for English language proficiency assessment, among others
   c. Interview regional and national English advisors to collect information regarding the availability of human resources and infrastructure required for reliable test monitoring
   d. Ensure that all students are treated fairly throughout the assessment process, having an unobstructed opportunity to demonstrate their level of English language proficiency
   e. Conduct additional analyses of MEP's English curricula to determine additional topics and domains that could be tested
3) Organize and hold massive training programs for all stakeholders in this ecosystem
4) Design the test around the *language proficiency* concept and its two main pillars according to the CEFR: communicative language activities and strategies and communicative language competences
5) Create a draft test blueprint to share with stakeholders to obtain their feedback before starting the item development phase

6) Share the draft blueprint with key stakeholders and have them complete a survey with questions about it to gather their opinion about its adequacy
7) Pilot-test items in the real population and conduct statistical analyses to assess their usefulness and reliability before the official test administration. This step would ensure these items are fair for various subgroups (e.g., male/female, urban/rural/suburban, different racial/ethnic groups, low SES/high SES) by conducting differential item functioning (DIF) analyses.

As Brown and Abeywickrama (2019), Fulcher (2010), and Coombe (2018) argued, building a localized validity argument for a national standardized test from scratch requires multiple steps and studies that would involve massive amounts of fieldwork to meet the particular needs and characteristics of the context and population assessed. By localizing the English proficiency test to meet the specific needs of Costa Rican high school students, the latter should consider the test fair after realizing it was not developed lightly but resulted from careful consideration and design, as Fulcher (2010: 4) recommended. In turn, this may contribute to neutralizing the generalized negative perceptions of standardized testing. Since this is a customized test, it will need to address our country's needs, lacks, and wants in foreign language standardized testing by basing its tests on MEP's unit contents, theoretical constructs, and item familiarity.

The locality-sensitive assessment will be produced in parallel with the new national policy of bilingualism (Ministerio de Educación Pública, 2016) where, in agreement with Badger and Yan (2012) as well as Brown and Abeywickrama (2019), students should learn to *use* the language. This rationale underlies UCR's choice of the skill-based assessment provided by the CEFR, which emphasizes and evaluates testees' competences. The customized nature of the test would not only reduce the anxiety and fear of those involved (managers and students), but would also help us

obtain more precise evidence of the testees' performance in language receptive skills. This is indeed the current communicative concept of language assessment and advocated by Canale and Swain (1980), Brown and Abeywickrama (2019), and Jamieson *et al.* (2008).

The results from this test will be diagnostic (see Brown and Abeywickrama, 2019: 10), which would, in turn, provide authorities with a clearer perspective of the system's strengths and weaknesses in so far as such interpretation is aligned with the theoretical construct of the test (Carr, 2015; Fulcher, 2010).

Since validation is a never-ending process (Chapelle, 2008; Brown & Abeywickrama, 2019), this pioneering nationwide standardized testing exercise is an ongoing project that has just started with this first step in language standardized testing in Costa Rica and Latin America.

5. Recommendations

The following recommendations are suggestions for those researchers who are developing localized and standardized language tests.

1) Researchers should review international guidelines on developing standardized tests. Some of these guidelines have been issued by institutions such as ILTE, ALTE, and APA. Guides such as the *Standards for Educational and Psychological Testing* are user-friendly starting points for researchers in the field.
2) Localizing a standardized language test requires more than designing an assessment instrument for a specific population. As outlined above, this continuous process should be done from the beginning with hand in hand with stakeholders, especially students. Given the short- and long-term impact of these tests and since multiple actors will be invol-

ved in the process, researchers are advised to consider the opinion of all stakeholders before making any decisions.

3) Researchers should take the input of some stakeholders with caution. For example, some may over- or underrepresent their particular context needs, lacks, or preferences. Consequently, it is imperative to corroborate the information with real-time observations and multiple sources to confirm the test requirements.
4) Investigators are advised to seek the assistance of language-testing specialists during the development of their own standardized tests. These specialists should help investigators to solve any issues since the former may have already dealt with these previously. There is nothing wrong with asking for help when it comes to such high-stakes tests.
5) Institutions aiming to develop standardized language tests may consider the possibility of certifying their language professionals in the different areas they intend to test. For instance, the ACTFL offers international certification for professionals who want to become official certified testers of English (for oral and written production). Having certified testers as part of the team constructing the test would provide valuable support to the process of developing, pilot-testing, and assessing the performance of items designed to measure those skills.
6) If an institution is planning to develop a standardized language test, it should consider the available human resources. Since this is a continuous process, it is convenient to have team members in charge of the different tasks related to the test, so as not to burden them with excessive workloads. For example, one group of language specialists could be dedicated to item writing; another, to item analysis; and yet another, to collecting evidence for the multiple claims. Assigning all of these tasks to the same team may induce a "burnout" feeling in the team members.

7. References

American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME), & Joint Committee on Standards for Educational and Psychological Testing (US) (2014). *Standards for educational and psychological testing*. Washington: American Psychological Association.

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

Azofeifa, Mauricio (2019, February 11). Más de 5.300 estudian inglés gracias a Alianza para el Bilingüismo (ABi). *Ministerio de Educación Pública. Gobierno de Costa Rica.* https://www.mep.go.cr/noticias/mas-5300-estudian-ingles-gracias-alianza-bilingueismo-abi

Bachman, Lyle (1990). *Fundamental considerations in language testing.* New York: Oxford University Press.

Badger, Richard, & Yan, Xiaobiao (2012). To what extent is communicative language teaching a feature of IELTS classes in China? In Jenny Osborne & IDP: IELTS Australia (Eds.), *IELTS research reports 2012* (Vol. 13, pp. 1–44). Australia, United Kingdom: IDP: IELTS Australia Pty Limited, British Council.

Brown, H. Douglas, & Abeywickrama, Priyanvada (2019). *Language assessment: Principles and classroom practices* (3rd. ed.). Hoboken: Pearson Education.

Cambridge University Press (2015). *Text Inspector.* https://languageresearch.cambridge.org/wordlists/text-inspector

Carr, Nathan T. (2015). *Designing and analyzing language tests*. Oxford: Oxford University Press.

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

Chapelle, Carol A. (2008). The TOEFL validity argument. In Carol A. Chapelle, Mary K. Enright & Joan M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 319–352). New York: Routledge.

Council of Europe (2002). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Press Syndicate of the University of Cambridge.

Coombe, Christine (2018). *An A to Z of second language assessment: How language teachers understand assessment concepts*. London: British Council.

Cordero, Monserrat (2019, November 29). MEP: colegios públicos tienen nivel básico en dominio del inglés. *Semanario Universidad.* https://semanariouniversidad.com/ultima-hora/mep-colegios-publicos-tienen-nivel-basico-en-dominio-del-ingles/

Fulcher, Glenn (2010). *Practical language testing*. London: Hodder Education.

Hooge, Edith; Burns, Tracy, & Wilkoszewski, Harald (2012). Looking beyond the numbers: Stakeholders and multiple school accountability. *OECD Education Working Papers*, No. 85. Paris: OECD. http://dx.doi.org/10.1787/5k91dl7ct6q6-en

International Language Testing Association (2018). *ILTA code of ethics*. https://www.iltaonline.com/page/CodeofEthics

Jamieson, Joan M.; Eignor, Daniel; Grabe, William, & Kunnan, Antony John (2008). Frameworks for a new TOEFL. In Carol A. Chapelle, Mary K. Enright & Joan M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 55–95). New York: Routledge.

Messick, Samuel (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Messick, Samuel (1996). Validity and washback in language testing. *Language Testing*, *13*(3), 241–256. doi:10.1177/026553229601300302

Ministerio de Educación Pública (2016). *Programas de estudio de inglés: tercer ciclo y educación diversificada*. Costa Rica: Imprenta Nacional.

North, Brian; Piccardo, Enrica, & Goodier, Tim (2018). *Common European Framework of Reference for Languages: Learning, teaching, assessment. Companion volume with new descriptors*. Strasbourg: Council of Europe.

O'Sullivan, Barry (2016). Adapting tests to the local context. *British Council new directions in language assessment: JASELE journal special edi-*

tion (pp.145–158). Tokyo: Japan Society of English Language Education, British Council.

Van Houten, Jacque, & Shelton, Kathleen (2018, January). Leading with culture. *The Language Educator*. https://www.actfl.org/sites/default/files/tle/TLE_JanFeb18_Article.pdf

Savignon, Sandra J. (1985). Evaluation of communicative competence: The ACTFL provisional proficiency guidelines. *The Modern Language Journal*, *69*(2), 129–134. doi:10.1111/j.1540-4781.1985.tb01928.x