Familiarizing users and stakeholders with the reasoning behind UCR's language proficiency test construct for MEP

Abstract

This paper aims at collecting evidence to build a content validity argument of UCR's Language Proficiency Test for high school students in Costa Rica through an analysis of the theoretical foundations that support the construction and administration of a custom-made language test and its localized context, which embeds a description of the test and the input of primary stakeholders. At the same time, this paper provides suggestions and recommendations for future testing experiences of this type, while paving the way for future researchers who intend to follow this line of academic endeavors.

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 de la Prueba de Dominio Lingüístico de la UCR para estudiantes de secundaria en Costa Rica, 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. Background

In 2016, the Costa Rican Ministry of Education (MEP) modified the English curriculum nationwide. In a national effort they called "Alliance for Bilingualism," they implemented a myriad of changes in the English programs to meet cultural, societal, and financial demands as a strategy to turn Costa Rica into a bilingual country, which in turn would attract investors, generate jobs, revitalize the economy, and foster study abroad opportunities (Azofeifa, 2019). As a token of the multiple changes being made, in 2019 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 70 out of 100 to be eligible for graduation. Were they to fail meeting this score, they would have needed to take this test as many times as necessary until achieving the cut-off score, keeping them from starting college or obtaining a job. However, instead of administering this traditional test, MEP decided to administer 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 students' English performance by using as a reference the descriptors of the Common European Framework of Reference for Languages (CEFR) (Cordero, 2019).

In spite of the many existing language certifications available, none of them seems to meet MEP's specific requirements and needs. First, the administration of the test and publication of its results must consider the MEP's yearly programming, which requires test administrators to conduct all related activities within a very tight time window. Second, uneven distribution and availability of resources have proven a challenge for public education institutions themselves. Finally, MEP's monetary restrictions must also be considered when choosing among the different options.

In an attempt to address these needs, the School of Modern Languages (ELM for its acronym in Spanish) of the University of Costa Rica (UCR) decided to create a more locally-sensitive option called Language Proficiency Test (PDL for its acronym in Spanish). Over the last 30 years, ELM has accumulated vast experience in the design and administration of a variety of language tests that have reliability and validity evidence to support the interpretations of their scores according to their intended uses. The language proficiency test designed for Consejo Nacional de Rectores's English for Specific Purposes program, a language proficiency test for faculty members, and the official translators' and interpreters' certification, among others, attest to ELM's expertise in the field, which has been broadened by the guidance of international language testing professionals. Not only does the School administer its own language tests, but it is also an internationally recognized center for some renowned language certifications; this guarantees the input and expertise of authorized certified raters amongst the School's faculty. In addition, ELM possesses the know-how for administering large scale, high-stakes tests nationwide, such as the Entrance Examination designed by UCR's Psychological Research Institute and the university's Statistics School. More recently, given the experience gained through time, ELM has started experimenting with digitizing some of its tests to ease their application and result gathering. This venture into test automatization has widened the options towards considering three possible formats of administration of the test: online, offline, and hybrid (a mix of both that requires a minimum bandwidth).

2. Validity argument

For the sake of this paper's validity argument, the following descriptions and analyses are provided as suggested in the *Standards for Educational and Psychological Testing* (section 4.1).

2.1. *Purpose of the test*

The purpose of the PDL-MEP test (Prueba de Dominio Lingüístico-Ministerio de Educación Pública) is to assess Costa Rican high school students' understanding and production of non-technical English related to both regional and global contexts that pertain to the socio-interpersonal, transactional, and academic domains, formally and informally while using as reference the descriptors of CEFR. This test is merely diagnostic. All senior high school students in Costa Rica who take this test will be able to know what their proficiency level is in terms of their reading and listening comprehension skills, according to CEFR. This test will not be considered a language certification test; hence, it may not be used for college admissions, visa applications, nor job applications.

2.2. *Score interpretation*

As indicated by MEP's authorities, the purpose of this new test is to determine students' language proficiency as a means to diagnose the efficacy of their recently adopted language programs, as well as students' language developmental stage. Hence, testees are to interpret the results as a reflection of their progress in their foreign language education, which will show the areas where they perform strongly as well as those that need improvement. Even though students are not to obtain any specific score to graduate from high school, they are to take this test as a requisite for graduation.

Based on the scores obtained at the nationwide scale, MEP might then 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. As an illustration of said adjustments, if the results show a clear lack of command of B2-level tasks on the test, MEP teachers might take this information and address the lacks identified by reinforcing said tasks in class.

Other stakeholders might make use of the results obtained to determine, for example, where to recruit new bilingual personnel or where to invest in more language programs for underprivileged populations.

2.3. *The constructs of the test*

2.3.1. Reading comprehension

Reading comprehension proficiency is defined as demonstrating understanding of non-technical English written texts related to both regional and global contexts that pertain to the socio-interpersonal, transactional, and academic domains, formally and informally, taking as a reference CEFR's descriptors. The contents to be included are determined as mandated by the guidelines provided by MEP. 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" (CEFR, 2018: 60). Finally, some of the strategies testees are to demonstrate are included in CEFR's descriptors, such as skimming, scanning, understanding a writer's tone and humor, and identifying attitudes and implied opinions (CEFR, 2018).

2.3.2. Listening comprehension

Listening comprehension proficiency is defined as showing understanding of non-technical English aural texts related to both regional and global contexts that pertain to the socio-interpersonal, transactional, and academic domains, formally and informally, using as a reference the descriptors of CERF. The contents to be included are determined as mandated by the guidelines provided by MEP. Some of the skills to be tested 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). Lastly, some of the strategies testees are to demonstrate are encompassed in CEFR's descriptors, such as understanding main ideas and specific details, making inferences, and discriminating speakers' attitudes (CEFR, 2018).

3. Claims

3.1. *Claim 1 (MEP)*

The UCR language test provides MEP with access to evidence that shows valid and reliable information about students' English performance with respect to nationwide language standards and CEFR proficiency bands, including communicative activities, strategies, and language competences. Based on this information, MEP can report students' language performance by classroom, school, district, and region. Having this in mind, the Ministry can then design a strategy to focus on those areas in more need of support in terms of language proficiency.

3.2. *Claim 2 (teachers)*

The UCR language test provides MEP instructors with access to evidence that shows valid and reliable information about students' English performance with respect to nationwide standards and CEFR proficiency bands, including communicative activities, strategies, and language competences. Based on this information, the Ministry of Education can adjust their classroom activities – formative and summative – to meet the standards established by MEP.

3.3 *Claim 3 (parents and students)*

The UCR language test provides parents and students with access to evidence that shows valid and reliable information about students' English performance with respect to nationwide language standards and CEFR proficiency bands, including communicative activities, strategies and language competences. Based on this information, these stakeholders can determine students' progress throughout the entire educational system.

4. Justification

Given the scenario previously described, the Ministry of Education and the University of Costa Rica came to an agreement in order to build, administer, and deliver the results of an English language proficiency test that would represent a tailor-made and more convenient option for both institutions. In this manner, MEP would have in their hands an instrument that meets their specific needs and 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" (ILTA, 2018: 2). This transparency is particularly important in documenting such a pioneering endeavor where one Ministry of Education in Latin America enforces a national policy of bilingualism in conjunction with a higher education public institution through a large-scale language test.

This paper aims at collecting evidence to build a content validity argument of PDL-MEP for students through an analysis of the theoretical foundations that support the construction and administration of a custom-made standardized language test and its localized context, which includes a description of the test and the input of primary stakeholders. At the same time, this paper provides suggestions and recommendations for future testing experiences of this type while paving the way for future researchers who intend to follow this line of academic endeavors.

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

5. Review of the literature

Standardized language testing may seem overwhelming and scary to many stakeholders, professionals in assessment, and 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 think of this kind of testing as a punishment that focuses on identifying their weaknesses and areas in need of improvement; teachers, on the other hand, believe it to be a demonstration of their supervisors' lack of trust in their work. Brown and Abeywickrama (2019: 1) summarized this view when they stated that stakeholders "are not likely to view a test as positive, pleasant, or affirming". In spite of all of these negative perspectives, different scholars have viewed standardized language assessment as a means through which distributive justice, as Messick (1989, as cited in Fulcher, 2010, p. 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 in a fair and equitable manner.

The concept of language competence assessment has been continuously redefined through time, according to users' needs 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 70s and 80s 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 concept of linguistic competence. However, during the mid 80s, Canale and Swain (1980, as cited in Brown & Abeywickrama, 2019: 14) advised a shift from this structure-centered approach to assessment towards a more communicative one, which dealt with more authentic tasks that language learners could eventually face. In the same manner, 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, [...] why language is 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, p.15) included among the fundamental principles of language testing “the need for a correspondence between language test performance and language use”. This more realistic communicative view of language assessment permeates some of the most renowned language tests in the market currently.

Assessing communicative language competence today is approached in a more holistic way. Since proficiency in a language goes beyond knowing its grammar, other equally - if not more - important features when testing an individual’s language ability must also be accounted for. In fact, as Badger and Yan (2008: 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”. In this same manner, the Common European Framework of Reference for Languages (CEFR) (2002) provides a framework that lists the necessary communicative language activities and strategies, as well as the communicative language competences (linguistic, sociolinguistic, and pragmatic) which 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: 35). Hence, it is readily visible that the concept of mastering a language as a second or foreign language 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, validation of language assessment is a never-ending, ongoing process (Chapelle, 2012; Brown & Abeywickrama, 2019). A key step in the process of test validation entails defining what validity is: “an overall evaluative judgement 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], & Joint Committee on Standards for Educational and Psychological Testing (U.S.), 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 evidence based on content (AERA, APA, & NCME, 2014).

In light of the vast scope of validation processes, the numerous types of evidence available to demonstrate it, as well as the multiple paths of research that can be followed, this paper shall attempt to gather ’evidence based on content’ as its primary focus in order to meet some of the validity standards stated above. To start, 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 further detail the elements to be analyzed and demonstrate content validity, the *Standards for Educational and Psychological Testing* (2014, p. 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.

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

5.1. *Description of test context*

5.1.1. MEP-high schoolers' diagnostic test

To meet the previously mentioned localization principles, the national English language proficiency test will assess high school students' reading and listening comprehension skills - as per MEP's request - based on the themes, domains, and scenarios stated by the Ministry of Education in their Programas de Estudio de Inglés (English Language Programs), which have been aligned with CEFR guidelines (MEP, 2016). The topics or themes addressed in said document include but are 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 (MEP, 2016: 13). Three axes encompass all these themes: a global citizenship with local belonging, education for sustainable development, and new digital citizenship (MEP, 2016: 13 & 55). The domains or contexts where the target language is to be used and which have been selected for this test include the socio-interpersonal, transactional, and academic domains (MEP: 38). Amongst the numerous scenarios provided by MEP for all levels of secondary education, the test might make use of some of these, such as "Enjoying Life" (7th grade), "Going Shopping" (8th grade), "Lights, Camera and 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's items address the topics and themes that the Ministry of Education has identified as important to focus on, 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 that can be made is that the test's items address three axes that the Ministry of Education has identified: a global citizenship with local belonging, education for sustainable development, and new digital citizenship. Finally, a fifth claim encompasses that the test's items reflect three domains in which students will need to demonstrate English language proficiency: socio-interpersonal, transactional, and academic domains.

The contextual needs addressed above will be further operationalized by wording items and instructions in a simple, clear manner that not only meets testees' expected language proficiency levels but also complies with the design requirements of trained language specialists. In fact, 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 actually understand what is expected of them, the language complexity shown in the instrument should mirror that which is described in the CEFR band they belong to. The tool "Text Inspector" (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, whose teaching expertise could also prove valuable (AERA, APA, & NCME, 2014: 75).

The format of the items, tasks, and questions will be based on the guidelines shared in the document *Programas de Estudio de Inglés* (MEP, 2016). For example, MEP (2016: 44) suggests the following types of 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. In a similar fashion, 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 (MEP, 2016: 42). These tasks should all mimic real-life scenarios similar to those students have been exposed to in the classroom throughout their learning process.

5.1.2 Stakeholders' expectations

To further contextualize the test and accurately place it within the given ecosystem, stakeholders' views must 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 MEP and the Director of the School of Modern Languages of UCR. The former acknowledges that the transition from a traditional reading comprehension test to a skills-based one took MEP approximately 11 years, a process led by the Office for the Management and Evaluation of the Quality. This first administration of the test aims at diagnosing high schoolers' levels of English proficiency and then comparing these results against those that will be obtained in the future, which will show evidence of the impact caused by the recently introduced *Programas de Estudio de Inglés*; said impact could be further analyzed using predictive validity evidence studies. In the words of the National Coordinator of the Alliance for Bilingualism at MEP (personal communication, November 5, 2020), this first administration of the test will also serve as a means to determine whether or not MEP has the adequate physical and digital infrastructure to give a test to more than 60,000 students. At the same time, he says that it would help other interested organizations to evaluate their capacity to successfully adapt to the various possible scenarios they might encounter when proctoring their exams at MEP. For example, some students might need to take the test at different facilities from those where they normally attend their classes, due to poor - if not completely absent - Internet connectivity. Considering the diverse circumstances that every region or institution might face (e.g., insufficient number of working computers, lack of personnel to oversee the test administration, or variety of school types), several schedules of administration should be arranged, for example three or four different agreed-upon times according to each institution's requirements. Additionally, this stakeholder emphasizes some of the requirements that any candidate testing organization must meet when working with MEP:

availability of immediate technical support, verifiable language testing experience, standardized administration protocols, and provision accommodations and modifications according to testees' special needs.

The Director of the School of Modern Languages of UCR, the second stakeholder interviewed for this paper, acknowledges several points worth mentioning (personal communication, December 15, 2020). The School of Modern Languages presents itself as a valuable part of the process because of assets such as its significant capacity, both technical and human, and its previous experience in standardized language assessment. The Director highlights the fact that the School does have the necessary capacity, in terms of technology and human resources, to successfully administer such a high stakes test; however, he also recognizes that further support and investment from UCR authorities would be advisable to improve the computer laboratories, security protocols, and data collection and analysis instruments. He also acknowledges the added value that collaboration with other University's departments could provide in the future. In terms of staff, the institution has trained personnel that can tackle the challenge of successfully administering the test in any of the formats requested by MEP. Although additional support for continuous training of these professionals would be advisable.

In terms of the School's capacity, the Director affirms that the University can administer this test three times a year at a large scale, specifically testing 5000 MEP students a day and delivering results within three weeks. Moreover, the technological know-how and experience gathered in testing facilitate any accommodations and specific modifications that might be required by MEP. Finally, the Director highlights the importance of familiarizing the target population with the test by facilitating a customized digital mock test that would attempt to bridge the gap between their knowledge of computerized testing and classroom testing.

The Director shows confidence in the reliability of the listening and reading online tests based on the evidence gathered but also considers that using paper-based instruments might be feasible, and equally reliable, depending on the needs of the target population. Both types of format, given the fulfillment of the specific safety and procedural protocols designed by UCR, can prove to be equally secure in regard to data collection.

Productive skills might be evaluated in the future, although further training and studies are necessary to ensure the reliability and validity of the assumptions of the results based on students' performance. As a novelty feature, he elaborates on the future possibility of using artificial intelligence as a tool in the creation and grading of UCR language tests.

This stakeholder believes that the test must accurately measure the language level of the testees by evaluating their understanding of both every day and academic English, which is the core of the construct approved by MEP for this test. This will be done by using an instrument that will consist of 40 to 60 items per skill and which would take approximately 60 to 70 minutes to complete per macro skill. The items might include multiple choice, sequencing, matching, drag-and-drop and short answer; the specific items selected will depend upon how items perform in pilot testing. Finally, the Director adds that the sources of reading and listening texts will be authentic and ecologically-sensitive material that will adequately match the CEFR levels intended to assess. Fortunately, the operationalization of this test and well as its construct have been approved by MEP.

6. Next Steps

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

1. arrange meetings with MEP representatives and decision makers in order to agree on the operationalization of the features of the test
2. carry out needs analyses
    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 their attitude towards standardized testing
    b. question students' views and experience with standardized and computer-based testing, familiarity with online testing and item types, preferred topics for English language proficiency assessment, among others
    c. interview regional and national English advisors to collect information regarding availability of human resources and infrastructure to reliably proctor the test
    d. try to ensure that all students are treated fairly in the assessment process, having an unobstructed opportunity to demonstrate their level of English language proficiency
    e. further analyze 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 exam around the concept of language proficiency and its two main pillars according to CEFR: communicative language activities and strategies, and communicative language competences
5. create a draft test blueprint to share with stakeholders to obtain their input before starting with item development
6. share said draft blueprint with key stakeholders and have them complete a survey in which they are asked questions about the blueprint to gather their judgments about its adequacy
7. pilot testing of items with the real population and provide statistical analyses that prove their usefulness and reliability prior to the official administration to make certain that they 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 field work to meet the specific 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, these might consider the test fair since they will have evidence that it was not done haphazardly, as Fulcher (2010: 4) suggested, which in turn may help contribute to neutralize generalized negative perceptions of standardized testing. Such research would in turn decrease the negative aura that surrounds testing, for the resulting test would be the product of careful considerations and design. Because this is a tailor-made test, it will need to address the needs, lacks, and wants of our country in terms of foreign language standardized testing by basing its tests on MEP's unit contents, theoretical constructs, and item familiarity.

The locally-sensitive assessment produced would go hand in hand with the new national policy of bilingualism (MEP, 2016) where, in agreement with Badger and Yan (2008) as well as Brown and Abeywickrama (2019), students should learn to *use* the language. This reason underlies UCR's choice of the skills-based assessment provided by CEFR, where testees' competences are emphasized and tested. The custom-made nature of the test would not only reduce the anxiety and fear held by those involved (administration and students), but it would also help us obtain more precise evidence of the testees' performance of language receptive skills. This is indeed the communicative concept of language assessment that is currently in vogue and which Canale and Swain (1980), Brown and Abeywickrama (2019), and Jamieson *et al* (2008) advocated for.

The results obtained from this test will be diagnostic (see Brown and Abeywickrama, 2019: 10) in nature, 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 and Abeywickrama, 2019), this pioneering nationwide standardized testing exercise is an ongoing endeavor that has just begun with this first step in the long path of language standardized testing in our country and in Latin America.

7. Recommendations

The following recommendations are written as suggestions for those researchers who are in the process of developing localized and standardized language tests.

- Researchers must consult international guidelines on developing standardized tests. Some of these are provided 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.
- Localizing a standardized language test requires more than designing an assessment instrument for a specific population. As shown above, the process is never ending, and it must be done from the beginning in conjunction with the stakeholders and students. Due to the impact these tests will have in the short and long term and because multiple actors will be involved in the process, researchers are advised to take their time to listen to what all stakeholders have to say before making any decisions.
- Researchers should take the information they receive from some stakeholders with a grain of salt. For example, some might over/underrepresent their context's needs, lacks, or wants. Consequently, it is imperative to triangulate the information with real-time observations and multiple sources to confirm what is required of the test.
- Investigators are advised to seek the help of language testing specialists while in the process of developing their own standardized tests. These specialists might help investigators to overcome obstacles encountered along the way since the former may have already dealt with these issues in previous occasions. There is nothing wrong in asking for help when it comes to such high-stakes tests.

- In the same vein, institutions whose intentions are developing standardized language tests could look into the possibility of certifying their own language professionals in the different areas they plan to test. As an illustration, ACTFL offers international certifications for testers who want to become official certified raters of English (for oral and written production). Having certified raters as part of the team who is constructing the test would help significantly in the process of developing, piloting, and analyzing the performance of the items that were designed to measure those skills.
- If an institution is planning to develop a standardized language test, it is paramount to consider the human resources available. Since this is a never-ending process, it is advisable to have team members in charge of different tasks related to the test, as not to burden them with extreme workloads. To illustrate, there could be one group of language specialists dedicated to item writing, another focused on item analysis, and another one dealing with collecting evidence for the multiple claims. Assigning all of these tasks to the same team might induce a state of “burnout” in any institution’s team.

6. References

ALTE. (2020). *ALTE Principles of Good Practice*. Retrieved from: https://pt.alte.org/resources/Documents/ALTE%20Principles%20of%20Good%20Practice%20Online%20version%20Proof%204.pdf

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 (U.S.) (2014). *Standards for Educational and Psychological Testing*

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

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

Badger, R. and Yan, X. (2008). To what extent is communicative language teaching a feature of IELTS classes in China? In Osborne, J., & Taylor, L. B. (Eds.), *IELTS research reports 2012.* (p. 8). IELTS Australia.

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

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

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

Chapelle, C. (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). New York, NY: Routledge.

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

Coombe, C. (2018). *An A to Z of Second Language Assessment: How Language Teachers Understand Assessment Concepts*. London, UK: British Council.

Cordero, M. (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, G., & Davidson, F. (2017). *The Routledge handbook of language testing*. Milton Park, Abingdon, Oxon, NY: Routledge.

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

ILTA. (2018). *ILTA Code of Ethics*. Retrieved from https://www.iltaonline.com/page/CodeofEthics

Jamieson, J. M., Eignor, D., Grabe, W., & Kunnan, A. J. (2008). Frameworks for a new TOEFL. 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. 55–95). New York: Routledge.

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

O'Sullivan, B. (2016). Adapting Tests to the Local Context. New Directions in Language Assessment, special edition of the JASELE Journal. Tokyo: Japan Society of English Language Education & the British Council, pp.145-158.

Van Houten, J., & Shelton, K. (2018, January). Leading with Culture. Retrieved from https://www.actfl.org/sites/default/files/tle/TLE_JanFeb18_Article.pdf

República de Costa Rica. Ministerio de Educación Pública. (2016). *Programas de Estudio de Inglés: Tercer Ciclo y Educación Diversificada*. Imprenta Nacional, Costa Rica.

Savignon, S. 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