Gist Education and Learning Research Journal. ISSN 1692-5777.
No. 21 (July-December, 2020). pp. 191-208.

Challenges in the Design and Implementation of an English Placement Test for a Colombian Public University[1]

Desafíos en el Diseño e Implementación de una Prueba de Clasificación de Inglés para una Universidad Pública Colombiana

Alexander Ramírez[2*]
Universidad del Valle, Colombia

[1] Received: November 12th 2019/ Accepted: September 7th 2020

[2] alexander.ramirez.e@correounivalle.edu.co

Abstract

This reflection paper disseminates the design and implementation of an English placement test for first semester students from different majors. This design was carried out within the framework of a research project in a Colombian public university. The article has a twofold purpose: to scrutinize the stages leading to a test, and to highlight the complexity that underlies the process of conceiving evaluation, so that language teachers who embark for the first time on the task of designing placement exams have a model and can anticipate vicissitudes of academic and administrative nature that may arise. The conclusions derived from the process of construction and implementation of the test highlight the lack of training and literacy in the area of language testing, as well as the existing tensions between academic expectations and the administration within a university institution.

Key words: language testing; placement test; assessment; evaluation; test design

Resumen

Este artículo de reflexión disemina el diseño e implementación de una prueba de clasificación en inglés para estudiantes de primer semestre en diversos programas académicos. Tal diseño se llevó a cabo en el marco de un proyecto de investigación en una universidad pública colombiana. El artículo tiene un doble propósito: diseccionar las etapas conducentes a una prueba y resaltar la complejidad que subyace a la concepción de procesos evaluativos, de manera que los docentes de lengua que se embarcan por primera vez en la tarea de diseñar exámenes de clasificación tengan un modelo y puedan prever las diversas vicisitudes de índole académica y administrativa que se pueden presentar. Las conclusiones derivadas del proceso de construcción e implementación de la prueba remarcan la falta de formación y literacidad en el área de diseño de pruebas, así como las tensiones existentes entre las expectativas académicas y la administración en el seno de una institución universitaria.

Palabras clave: evaluación de lenguas; examen de clasificación; evaluación; diseño de pruebas.

Resumo

Este artigo de reflexão dissemina o desenho e implementação de uma prova de classificação em inglês para estudantes de primeiro semestre em diversos programas acadêmicos. Tal desenho foi realizado no âmbito de um projeto de pesquisa em uma universidade pública colombiana. O artigo tem um duplo propósito: analisar as etapas que conduzem a uma prova, e ressaltar a complexidade que subjaz à concepção de processos avaliativos, de maneira que os docentes de língua que se embarcam pela primeira vez na tarefa de desenhar provas de classificação, tenham um modelo e possam prever as diversas vicissitudes de índole acadêmica e administrativa que se possam apresentar. As conclusões derivadas do processo de construção e implementação da prova remarcam a falta de formação e alfabetização na área de desenho de provas, bem como as tensões existentes entre as expectativas acadêmicas e a administração dentro de uma instituição universitária.

Palavras chave: avaliação de línguas; prova de classificação; avaliação; desenho de provas.

Introduction

The use of tests to keep track of progress and measure users' language proficiency constitutes an ongoing and essential activity in the field of foreign language assessment, evaluation, and testing processes (Douglas, 2010). However, the literature on language testing in Colombia suggests that language teachers lack literacy and training in this field, which means that they are in need of strong skills for the design, implementation, and research of language tests for various purposes (Giraldo, 2018a, 2018b; López & Bernal, 2009). Using tests for evaluation, however, is an activity that underlies the teaching profession; this allows to infer that gaps in terms of testing literacy and training might be leading to commit several irregularities in their evaluation processes. In the same vein, Giraldo (2019) remarks that poor quality in test design "may disorient language teaching and learning" (p. 124) as decisions that derive from test results may lack solid grounds.

The design and development of language tests within an institution is desirable, not only because it can help minimize the purchase of commercial tests—which are not tailored to the needs of a particular institution or context—but it also empowers teachers and contributes to their continuous training in terms of Language Assessment Literacy (LAL) (López & Bernal, 2009). The latter translates into greater expertise and diversification of evaluation techniques, as well as the development of constant reflection and critical thinking towards assessment and evaluation.

With this in mind, and in view of the need for an English placement test in a Colombian public university, a research project was designed with a twofold objective: on the one hand, to train a group of teachers in the techniques for the design of tests and their subsequent statistical analysis; and on the other hand, to design a placement test, with high quality standards, that fulfilled the institutional needs. This paper presents the different stages that made up the design of the placement tests, interwoven with the literature review, the theoretical considerations, and the reflection on the various challenges experienced at each stage by me as the person who conducted the training, and by my colleagues as trainees. It is my hope that other teachers who embark for the first time on this ambitious task have a model and can anticipate the various vicissitudes of academic and administrative nature that may arise along the process.

Context

The research project from where this paper derives took place at Universidad del Valle, a public University in the Southwest Colombia. Initially, this institution offered a program of Reading in English for Specific Purposes (ESP) to all the majors on campus, composed by four courses. However, the current demands of globalization have pushed the University into academic dynamics that require moving from merely consulting information in a foreign language, to a more complex need of communicating in general and academic English. Accordingly, the University shifted from the ESP program to an English for General and Academic Purposes (EGAP) program in order for all undergraduate program students to reach a B1 proficiency level in a foreign language (Resolution No. 136 by the Academic Council of the University of Valle, December 22nd 2017).

With a new English program in execution, the University administration commissioned its English teachers to take on the design of the test, although such a task would be much more complicated than expected. First of all, most English teachers lacked from *Language Assessment Literacy* (LAL), as stated by the teachers themselves. LAL is defined as the "knowledge, skills, and principles in language testing" (p. 180) by different stakeholders. Such lack had been previously documented by López and Bernal (2009), who drew attention to the low presence of language assessment training in university undergraduate and graduate programs in Colombia.

With this in mind, a team of 7 teachers was gathered and trained in the basics of language testing for one semester. The trained teachers (hereinafter referred to as the team) were part of the English for General and Academic Purposes (EGAP) Section at Universidad del Valle, which offers its services to all the undergraduate programs on campus. After the training, a research project[3] was conducted, resulting in the design and piloting of the Univalle's English Placement Test UVEPLAT. The test comprises 52 different items and places test takers in levels A2 to B2 according to the Common European Framework of Reference for Languages (CEFR). UVEPLAT, which requires one hour for its completion, evaluates test takers in reading comprehension, listening comprehension, grammar structures and vocabulary. The next section of this paper gathers the guidelines, recommendations and reflections that resulted from designing team's experience.

Guidelines for the Design of an English Placement Test

Diagnosing language testing practices in the institution

The starting point for the design of a test that will fulfill institutional purposes must be the meticulous analysis of the context in which the test will be applied, in terms of four fundamental elements: the evaluation practices established in the institution, the specific situations in which a language test is needed, the institutional language policies

[3] This research project aimed at piloting the test and measuring its item difficulty and discrimination indexes. The results, as well as a deeper technical characterization of the test, can be found in Ramírez (2020).

(or lack thereof), and the different stakeholders involved in the testing process. I now move on to exploring each one of these elements.

First of all, the recognition of pre-established institutional practices regarding the use of language tests ensures a clear landscape of what to keep and what to avoid in the design of a new test. For instance, in the particular case that feeds this article, a first glimpse into the University allowed for the identification of two practices that were common and traditionally passed on among language teachers: 'frankensteining' tests out of different materials, and solving any testing need through the use of commercial testing platforms. 'Frankensteining' or making exams out of several pieces coming from test preparation books, previously discarded tests, or even the Internet, revealed a deep lack of awareness towards testing design on the part of the teachers. Similarly, the analysis of the institutional context revealed the unsystematic use of commercial tests for multiple purposes within the institution; for example, a commercial quick-placement test had been used to verify different levels of proficiency, to place undergraduate students into levels of the curriculum, and to determine the entrance of graduate students into certain programs. Not only did both of these practices exposed the Administration's misinformation towards fundamental principles of testing, such as validity and reliability, but it also evinced that, when it comes to language policies at a university, there might be institutional decisions made by administrative authorities that do not take into account the participation or scrutiny of language teachers or testing experts. The excessive use of commercial tests and the practice of 'frankensteining' respond to the Aministration's lack of awareness about the complexity of language testing, and about the intricacy of several aspects (time, resources, budget, training, etc.) behind the design, piloting, and implementation of a test; such a lack of awareness might explain the Administration's constant and urgent demand for language test scores for various purposes, as well as the seeming institutional idea that one single test might solve a plethora of needs.

It is necessary to establish the different needs that the institution has, as well as the nature of the tests required for each need. In the Colombian context, for instance, several universities usually make use of placement tests and proficiency tests for different types of test takers including aspiring students, enrolled students, faculty and staff. Not all of these audiences pursue the same language programs or linguistic skills, so chances are that the nature of the test might vary, going from language for general communicative purposes, or a particular skill, such as reading or writing, for specific or academic purposes.

Third, the recognition of language policies (or their inexistence) determines a legal framework for language evaluation, assessment and testing in the institution. For example, in the university where this study was carried out, language policies were still in the making, and no light had been shed on the case of indigenous people, who were required to demonstrate proficiency in an L2 through an English test, completely ignoring that they are users of two or more languages already. If language policies exist, then the design and implementation of new tests have to be planned around the former. Otherwise, the issuing of new language policies must encompass the type of evaluation and tests that will be needed.

Finally, recognizing the different stakeholders involved in the evaluation activity is paramount for the quality of the assessment processes: language teachers, test designers, test takers, and the university administration shape an intricate gear where the role of each one must be clearly established and held accountable for their participation in the testing matters.

Setting up a test design team

As stated before, a team of 7 teachers was gathered and trained in the basics of language testing for one semester, in weekly workshop sessions of three hours. The rationale behind this training program is the fact that test design is a daunting task that "requires attention to a considerable number of theoretical and technical details" (Giraldo, 2019, p. 124); thus, the training program encompassed six elements, as summarized in the following graph:

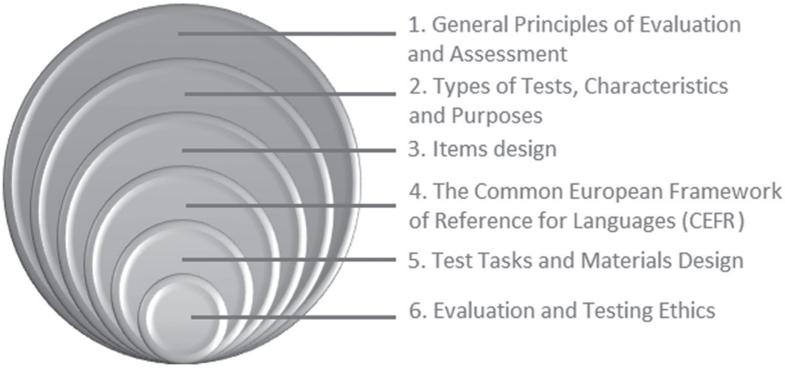
1. General Principles of Evaluation and Assessment
2. Types of Tests, Characteristics and Purposes
3. Items design
4. The Common European Framework of Reference for Languages (CEFR)
5. Test Tasks and Materials Design
6. Evaluation and Testing Ethics

Figure 1: Test Design Training Components

For the purposes of this paper, I have intentionally chosen the term training, and not professional development, mainly because the workshops imparted to the trainees were conducted with a specific and terminal purpose, seeking to respond, in the short term, to an urgent request from the University. However, this type of training could and should evolve into professional development, understood as a constant process, a continuous activity focused on the teacher, and not merely consisting of 'one-shot' workshops or lectures focused on a specific product (Armbruster & Osborn, 2001; Joyce & Showers, 2002; Cooper, 2009).

The first three components were covered through the study of different postulates by Bachman and Palmer (1996), Carr (2011), Douglas (2010), and Shohamy (2001). Through several study-group sessions and workshops, this component started with the discussion of how abstract and challenging the idea of measuring linguistic performance can be and went on to deepen into the definitions of fundamental concepts. It is worth mentioning that at the beginning of the training, most participants seemed overconfident regarding the training program components as these topics are expected to be mastered by any seasoned teacher. Furthermore, according to participants, testing and assessing languages are tasks that they perform on a regular basis, so they trusted they had a decent mastery of assessment design. The first training sessions, however, showed otherwise; not only did the teachers ignore, or had forgotten, basic concepts such as construct, reliability, validity, washback effect, and the very difference between evaluation and assessment, but they also manifested some ideas that unveiled lack of fairness; for instance, all of them explicitly stated that a high quality test was a synonym of a difficult-to-pass test, rendering evaluation a tool of power and control (Shohamy, 2001).

One of the very first challenges, then, was to establish a group philosophy that a test—and evaluation in general, for that matter—has to do with offering learners as many opportunities as possible for them to show what they know, rather than making them hesitate with a riddle type of exam, because "the more opportunities we give test takers to show what they know, the more accurate and fair the measurement is likely to be" (Douglas, 2010, p. 4).

This was a great opportunity to openly discuss fairness and the ethics of evaluation and testing, which are topics that usually stay under the individual domain of every teacher inside his or her classroom. After this common ground was reached, the second and the third component were covered by focusing on the study of taxonomies of language tests, their purposes, as well as different types of items (Bachman & Palmer, 1996; Carr, 2011).

The study of the fourth element, the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001) was the second major challenge we faced in this training program. The CEFR comprises "a set of common reference levels for language teaching, learning and assessment stretching from basic to mastery" (Taylor & Geranpayeh, 2011, p. 91); as indicated by its name, the CEFR is expected to serve as a reference, a lighthouse to look at for guidance. However, a recurring phenomenon among the trainees (also detected to a greater or lesser extent in many teachers formally and informally surveyed in the framework of this study) is the fact that there is a subjective and mostly intuitively-derived construct with respect to the CEFR levels. Although the CEFR is precisely a reference document, most English teachers agreed upon the fact that they do not consult the reference, but trust 'their gut' when it comes to evaluating and placing language users. It seems that some teachers have formed an idea of what it means to be an A1, a B2, or a C1 kind of language user, with very diffuse boundaries between the levels that vary from one teacher to the other; these boundaries are usually rooted in a long tradition of comparing students who have been previously evaluated by the teacher, in a kind of rationale as follows: 'if John is a B2 and Jane's level is a little lower than John's, then Jane must be an A2'. Thus, individual and subjective criteria are created, with rough borderlines that have nothing to do with the descriptors established in the CEFR for each skill. This type of mindset implies quite transcendental incidents for assessment, as teachers will be much more challenging or lax in the type of demands they make for each level, and to a greater extent, these individual mindsets will affect pedagogical relationships within the classroom, as pointed out by Giraldo (2019), when he argues that poor assessment design "may disorient language teaching and learning" (p. 124).

The training also encompassed a component of test tasks and material design, as well as a component of testing ethics. The former was included as the team members were found to lack formal training in the design of language materials, which translated into ambiguity in instruction writing, as well as poor skills for adapting or designing written and oral documents; in other words, material design and test design are expected to be done by teachers on a regular basis, yet the trainees manifested their lack of theoretical foundations to effectively perform both of them. Finally, the training concluded with the study and discussion of the International Language Testing Association's (ILTA) Code of Ethics[3] for language testing.

Defining Test Specifications and Establishing Test Constructs and Audience

After the training, the next step was the creation of a blueprint containing the test specifications, which have been defined as "a detailed set of documentation normally drawn up during the process of designing a new test or revising an existing one" (ALTE, 1998, p. 166). The purpose for the creation of this document was twofold: on the one hand, it was meant to provide a framework of reference that would contextualize future teachers (and other stakeholders) within the institution, to understand the rationale that guided the design of the test, so that future revisions, corrections and improvements would be more easily addressed; and on the other hand, the writing

[4] This document can be consulted at: https://www.iltaonline.com/page/CodeofEthics

of the document was intended to help teachers establish the test construct and make informed decisions about test length, "(…) design, content, level, task and item types used, target population, use of the test, etc." (ALTE, 1998, p. 166).

For the definition of the test construct, and building on Taylor and Geranpayeh's (2011) proposal, three aspects were taken into consideration: the cognitive processing in the skills to be tested, the characteristics of the target population of test takers, and contextual factors for the implementation of the test. First, understanding the nature of every language skill, the features comprised in each one of them, and the cognitive processes involved in them, allows test developers to sharpen the demands they make from test takers in each of the levels that the test is expected to discriminate. It is necessary to foster discussion among test developers around fundamental questions such as *what does reading comprehension mean?* Or *what does listening comprehension imply?* With this in mind, the teachers were invited to revise the nature of the skills to be tested, in this case listening comprehension, reading comprehension, and language use (grammar and vocabulary). Later, by contrasting the CEFR descriptors and the list of contents in the syllabus of each level course, the teachers came up with what Taylor and Geranpayeh (2011) call a set of 'can-do statements'; the fact that the team in this study had a copious expertise teaching the different language levels made it easier for them to determine the contents they considered more relevant to be included in the test tasks, as well as the proficiency descriptors they deemed mandatory from test takers at each level.

Regarding the characteristics of the target population, the teachers reflected on the characteristics of the prototypical freshman at this University, as audience characteristics are paramount for designing a test. In this regard, Douglas (2010) remarks the importance "that the test tasks reflect to the degree possible the ways test takers have been learning and using the language" (p. 44); test tasks have to be designed taking into account the strategic competence and background knowledge that the potential audience possesses. In our experience, after the study of items design, the team came up with varied and creative tasks and items that they wanted to include in the test construction. They hadn't taken into account, however, that the prototypical first semester student in our contexts is a 16–18 year-old youngster who has recently graduated from a public high school, mainly familiarized with multiple-choice tests; in this case, test performance might be negatively influenced by the lack of experience with the intricate design in tasks and items, and not necessarily by the language knowledge that test takers have.

Similarly, a small portion of the freshmen population in our context often include blind students, which imply other versions of the test that cater properly to their needs. Finally, our target population also includes a considerable percentage of students who come from indigenous communities, or from the Colombian deaf community, who are usually bilingual in a minority language and Spanish as a second language. This portion of the population raised the discussion around several questions: What are the entry requirements in terms of second language and bilingualism? Who should be exempted from these requirements and, therefore, the placement test? Who meets already the bilingualism requirement by being speakers of Spanish as a second language? The emerging reflections testified for the need of a clear linguistic policy within the institution regarding bilingualism and showed the importance of the participation of administrative representatives, as the development of tests is best done collaboratively between different stakeholders (Douglas, 2011; Giraldo, 2018a).

Designing Items

Although many high-stake language tests are designed by specialists in the field of testing who do not necessarily teach the language, we believe that language teachers "make good item writers as they have developed a deep understanding of language learners and of the language" (ALTE, 2011, p. 26); therefore, university language teachers should be offered more participation in test-design teams, as well as in language policy decision-making. After the item design workshops, three challenges were identified and manifested by the team: variety in the nature of items, instructions writing, and distractors design.

The first challenge in this stage was fostering variety within the test design. In our training experience, the workshops allowed the teachers to discover several forms of items they did not know, or that they had never included in a test before. Thus, once the teachers started designing the first test blueprint, they felt a creative rush to incorporate as many different items as possible, in sake of rich variety in the test. The result was an intricate test that could mislead test takers, as the proposed items and the tasks did not match the target audience's background experience in solving language tests. Once again, the first designs produced by the team would propose items and tasks that were very complex for the target audience. The trainees in our team would draw heavily upon a framework of previous experiences as test takers, and so they would try to mimic tests they had solved, tasks they had found particularly complex and demanding, or tasks that they deemed interesting.

At this point, the team had to reflect upon how paramount it is to carefully choose items that test takers have been previously exposed to, since both success or failure in a test also depends on the level of familiarity that the test taker has with the item type; for instance, it might happen that a test taker with a high level of language proficiency performs poorly in a test: he might be linguistically good enough but he might get confused with the format of the item stem, resulting in poor task completion. This desire to include several types of items opened the discussion towards what it means to foster variety within the test. Although a good test should offer tasks that are varied in nature, the real challenge is to offer as many varied opportunities for language learners to show their mastery of different linguistic functions. In that sense, variety in tests can be achieved through the incorporation of different communicative functions in each of the assessed skills; for instance, there may be tests with a vast array of item types that revolve around the same topic or function, rendering the test tricky in the solving but monotonous in the content; on the other hand, there may be an exam containing only multiple choice and cloze items, but whose content is rich in the type of communicative functions it evaluates.

The second major challenge was instruction writing. During the design of items and peer evaluation within the training sessions, it became clear that, in general, there was a lot of difficulty writing clear and concise instructions on item stems. To this respect, the language teachers stated that perhaps the reason for such difficulty was the fact that they were more familiar with giving instructions and explanations orally, and not so much in written form. The following common errors could be identified in the first item proposals:

- The instruction was ambiguous, giving rise to more than one interpretation
- The instruction did not clearly state what the test taker was expected to do
- The instruction was too long, with redundant explanations of what the test taker was expected to do.

After finally learning how to write clear instructions in item stems, our rookie design team had to face the third challenge at this stage: designing distractors. Initially the team would design an item with the expected correct answer, and then would randomly include scattered words to function as distractors; the result was an item that could easily be solved without actual knowledge of the topic, as the correct answer was very obvious amidst poorly selected distractors. With this in mind, the team was led to understand that distractors in multiple-choice items are actually designed, based on the target function, structure or vocabulary that the item intends to assess.

This stage of the process revealed that the teachers in the designing team felt much more comfortable while designing tasks for grammar and vocabulary assessment. At the opposite side, they did not show to be so at ease while defining assessment for listening and reading. This suggests the need to constantly revise the concept of construct in relation to all the language skills, and the necessity for teachers to reflect upon what it means to 'measure' listening and reading comprehension, as well as oral production.

Making Decisions

One of the hardest lessons we had to learn as rookie test designers was the fact that there is no such thing as a step-by-step recipe to craft a new test. In that sense, Taylor and Geranpayeh (2011) were very assertive when they mention that "all language testing, including large-scale assessment, is 'the art of the possible'" (p. 94); this means that for every case of test design there are many decisions to be made by the designing team according to their particular setting, and that many of the questions and concerns you may have while crafting a new test do not have a unique valid answer. In our case, for example, the teachers felt very unconfident and anxious about questions such as what the best scale to grade the test was, whether text in the reading comprehension section should be utterly authentic, or if a computer-based version of the exam would be better than a paper-based version. In this regard, Taylor and Geranpayeh (2011) explain that "tests remain to some degree provisional, work-in-progress, even experimental, hopefully serving a positive and practical function in the real world of here and now" (p. 94); informed decisions have to be made taking into account the academic, social, and economic context where the test will be designed and implemented, as well as the possibilities of the human resources involved in the making; all of these aspects have to be attentively taken care of, as "language testing occurs in an educational and social setting, and the uses of language tests are determined largely by political needs." (Bachman 1990, p.291).

In our case at Universidad del Valle, for instance, we had to make decisions on the number of items that would eventually compose the test, about the source and length of written and oral texts that would be used for the tasks, and about the weighing assigned to every section of the test; similarly, discussions among the designing team allowed for decisions on the best platform to host the digital version of the test, the braille and special versions for disabled people, and the security protocols for the exam implementation. Every decision, every discussion, every stage of the process brought a deeper understanding of language evaluation and assessment, as "the main purpose of language testing is to provide opportunities for learning, both for the students who are being tested, and for the professionals who are administering the tests" (Tomlinson, 2005, p. 39). We agree profoundly with Tomlinson (2005) when he asserts that "while it is obviously important that tests should be fair, valid, and reliable, the most important of all is that tests should provide useful opportunities for learning" (p. 40). All in all, through this process we could attest that the design of any test, whether for placement, proficiency or achievement purposes, bears "An intrinsic potential as [a] research tool[s] whose outcomes will help enrich our understanding of the nature of language proficiency so we can develop better tests in the future" (Taylor & Geranpayeh, 2011, p. 94).

Limitations and Further Constraints

The reflections and advice given here are far from being the ultimate complete guide for the construction of a placement test, as there are still many more elements at each stage of the process that deserve deep discussion and analysis, such as adaptation of oral and written texts, the care with copyright norms, the study of statistical variables in the piloting of the test and the creation of parallel versions, or mirrors, just to mention some relevant topics. However, it is expected that this paper has been able to highlight the complexity that underlies the process of conceiving evaluation, as assessment practices and the construction of evaluation devices are bound to face multiple challenges, some of which have been discussed here.

More reflection is needed, however, in order to establish a frank dialogue between academia and administration around issues such as financial support, time constraints, and the joint construction of clear language policies. Finally, it is worth mentioning that some of the challenges faced by the team in terms of test design such as difficulties for writing item stems, designing answer distractors, or designing tasks might be related to their experiences with material design; thus, further studies in this particular context could focus on the correlation between the team's skills and formal knowledge (or lack thereof) in materials design and test design.

Conclusion

I agree with López and Bernal (2009) and with Giraldo (2018) that it is an imperative need to train language teachers in the design, analysis and research on language exams. Beyond merely being aware of the different types of tests and their uses, language teachers require "knowledge on how to write, administer and analyze tests" (Inbar-Lourie, 2013, p. 32). So far, it seems that many language teachers are mainly users and consumers of tests but not their designers or producers. In this regard, it is worth remembering what Giraldo (2018a) asserts the teachers who have received language assessment training "use[d] assessment to improve teaching and learning, whereas those with no training used it as a way to solely obtain grades" (p. 181).

Administrative stakeholders within universities usually demand from language departments to implement exams and report results, for them to make decisions. The process, however, needs fixing: administrative stakeholders should work hand-in-hand with language teachers (especially the ones involved in test design) so that clear and fair evaluation policies are established in the institution. Administrative stakeholders need to understand how demanding the process of designing a high-quality test is, so that more support for teachers and researchers may be granted; similarly, they need to know how crucial their participation is, if tests are to become fair instruments that enlighten long-term decisions, they will absolutely have an impact in the test taker's life, as well as in the institution itself.

A strong focus on LAL needs to be fostered, starting with pre-service teachers in undergraduate language teaching programs so that they understand, from the very beginning, the far-reaching responsibility that comes with evaluating. On these grounds, it is recommended that undergraduate and graduate programs offer more subjects with a practical component on test design, so that both pre-service and in-service teachers may sharpen their skills in defining constructs, designing items, and piloting tests on grounds on fairness and social justice.

Finally, the long-term success of a test design resides in team effort: a team composed by the test designers, the language teachers and the representatives from the institutional administration. In this sense, it is paramount to devote time for group training if the endeavor of test design is to succeed, but it is also equally important to foster constant dialogue between academia and administration, so that time and financial resources may be granted for research, as well as good conditions for the planning, designing and piloting of tests. Similarly, in the long term, institutions must think up language evaluation processes in relation to a linguistic and evaluative policy, one that should be consistent with the academic objectives of the program, in accordance with institutional purposes, and fair to the population evaluated.

This article derives from the research study "Pilotaje de una Prueba de Clasificación de Inglés para Estudiantes de Primer Semestre en la Universidad del Valle: Análisis de Ítems," carried out at Escuela de Ciencias del Lenguaje, Universidad del Valle, and sponsored by the Research Vice-rectory from the Humanities Faculty (registration code CI-4383).

References

ALTE, Association of Language Testers in Europe (1998) *Multilingual glossary of language testing terms*. Cambridge: United Kingdom. Available at: https://www.alte.org/Materials

ALTE, Association of Language Testers in Europe (2011) *Manual for Language Test Development and Examining*. Cambridge: United Kingdom. Available at: https://www.alte.org/Materials

Armbruster, B. B., and Osborn, J. (2001). *Put reading first: The research building blocks for teaching children to read (K–3).* Ann Arbor, MI: Center for the Improvement of Early Reading

Achievement, University of Michigan. Distributed by National Institute for Literacy (NIL), National Institute for Child Health and Human Development (NICHD), and the U.S. Department of Education.

Bachman, L. (2007) What is the Construct? The Dialectic of Abilities and Contexts in Defining Constructs in Language Assessment. In Janna D. Fox, Mari Wesche, Doreen Bayliss, Liying Cheng, Carolyn E. Turner & Christine Doe (Eds.), *Language Testing Reconsidered* (pp. 41–71). Ottawa: University of Ottawa Press/Les Presses de l'Université d'Ottawa.

Bachman, L. and Palmer, A. (1996) *Language Testing in Practice*. Oxford: Oxford University Press

Carr, N. (2011) Designing and Analyzing Language Tests. Oxford: Oxford University Press Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning teaching, assessment*. Cambridge: Cambridge University Press.

Cooper, J. D. (2009). *Professional development: An effective research-based model.* Houghton Mifflin Harcourt. Retrieved from http://dataworks-ed.com/wp-content/uploads/2016/05/JoyceShowers.pdf

Douglas, D. (2010). *Understanding Language Testing*. London: Hodder-Arnold

Giraldo, F. (2018a) Language assessment literacy: Implications for language teachers. *Profile: Issues in Teachers' Professional Development, 20* (1), 179–195. Available at http://doi.org/10.15446/profile.v20n1.62089

Giraldo, F. (2018b). A Diagnostic Study on Teachers' Beliefs and Practices in Foreign Language Assessment. *Íkala, Revista de Lenguaje y Cultura*, *23*(1), 25–44. Available at: http://doi.org/10.17533/udea.ikala.v23n01a04

Giraldo, F. (2019). Designing Language Assessments in Context: Theoretical, Technical, and Institutional Considerations. *HOW Journal*, *26*(2), 123–143. https://doi.org/10.19183/how.26.2.512

Inbar-Lourie, O. (2013). Guest Editorial to the special issue on language assessment literacy. *Language Testing, 30*(3) 301–307. Available at: https://doi.org/10.1177/0265532213480126.

Joyce, B., and Showers, B. (2002). *Student achievement through staff development* (3rd ed.). Alexandria, VA: Association for Supervision and Curriculum Development

López, A., & Bernal, R. (2009). Language testing in Colombia: A call for more teacher education and teacher training in language assessment. *Profile: Issues in Teachers' Professional Development,* 11(2), 55–70.

Ramírez, A. (2020). Análisis de ítems para prueba de clasificación en inglés en una Universidad colombiana. *Cultura, Educación y Sociedad*, *11*(2). 177-190
DOI: http://dx.doi.org/10.17981/cultedusoc.11.2.2020.11

Shohamy, E. (2001). *The power of tests: A critical perspective of the uses of language tests.* Harlow: Longman.

Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics*, *29*, 21–36. https://doi.org/10.1017/S0267190509090035.

Taylor, L. & Geranpayeh, A. (2011) Assessing listening for academic purposes: Defining and peratoinalizing the test construct. Journal of English for Academic Purposes, 10, 89–101

Tomlinson, B. (2005). Testing to learn: a personal view of language testing. *ELT Journal Volume 59*, *1*, 39-46. https://doi.org/10.1093/elt/cci005

Universidad del Valle (2017) Resolución No. 136 del Consejo Académico de diciembre 22 de 2017. Cali: Colombia.

Author

Alexander Ramírez holds a B.A in Foreign Languages English-French and an M.A in Linguistics. He is a first-year PhD student at Universidad del Valle, where he also works as an associate professor of English and Linguistics. His research interests include language testing, intercultural communication, and Queer linguistics.

ORCID: https://orcid.org/0000-0002-7122-9537

How to reference this article: Ramirez, A. (2020). Challenges in the Design and Implementation of an English Placement Test for a Colombian Public University. *GIST – Education and Learning Research Journal, 21*, 191-208. https://doi.org/10.26817/16925777.825