A Diagnostic study on teachers' beliefs and practices in foreign language assessment

Estudio diagnóstico sobre creencias y prácticas de docentes en evaluación en lenguas extranjeras

Étude diagnostique sur les croyances et pratiques des enseignants dans l'évaluation en langue étrangère

Frank Giraldo Aristizábal
Master of Arts in Teaching English as a Second Language, University of Illinois at Urbana-Champaign.
M. Ed. Didáctica del Inglés, Universidad de Caldas (Unicaldas).
Academic consultant to Instituto de Lenguas Extranjeras (ILEX), Universidad Tecnológica de Pereira.
Professor Bachelor's Degree on Bilingüismo con Énfasis en Inglés, Universidad Tecnológica de Pereira.
Mailing address: Carrera 8, N.° 2E77, Alfonso López, Pereira, Colombia.
E-mail: icaros@utp.edu.co

Abstract

This paper reports the findings of the qualitative diagnostic stage within an action research study whose purpose is to improve the language assessment literacy (LAL) of English teachers at a Colombian language institute. A questionnaire, interviews, and document analyses were used to inquire into beliefs and practices in the design of an achievement test. Findings suggest that these teachers believe tests should have core qualities that are partially mirrored in their practices. The research also highlights that beliefs and practices in test design exhibit a dynamic relationship. Conclusions are based on findings and provide information useful for professional development experiences in LAL.

Keywords: beliefs, communicative competence, language assessment literacy, testing practices, test qualities

Resumen

Este artículo presenta los resultados de la etapa de diagnóstico cualitativo de una investigación acción cuyo propósito es mejorar la literacidad en evaluación de lenguas (LEL) de los docentes de inglés de un instituto colombiano de lenguas. Un cuestionario, entrevistas y análisis de documentos fueron los instrumentos usados para indagar sobre las creencias y prácticas en el diseño de un examen final de lengua. Los resultados indican que estos profesores creen que las pruebas de lengua deben tener cualidades centrales, lo cual se refleja parcialmente en sus prácticas. La investigación resalta que las creencias y prácticas tienen una relación dinámica. Las conclusiones se basan los resultados y generan información para experiencias de desarrollo profesional en LEL.

Palabras clave: competencia comunicativa, creencias, cualidades de las pruebas, competencia en evaluación en lenguas, prácticas evaluativas

Résumé

Cet article présente les résultats de l'étape diagnostique d'une recherche-action dont l'objectif est l'amélioration de la connaissance et des habilités dans l'évaluation de la langue parmi des professeurs d'anglais dans un institut colombien. Un questionnaire, des entretiens et des analyses de documents ont

Received: 2016-07-07/Accepted: 2017-07-12
DOI: 10.17533/udea.ikala.v23n01a04

été utilisés pour enquêter sur les croyances et les pratiques sur la conception d'un examen final. Les résultats suggèrent que ces professeurs croient que les examens doivent avoir des qualités centrales, lesquelles sont partiellement mises en évidence dans leurs pratiques. La recherche souligne que les croyances et les pratiques dans la conception des examens ont une relation dynamique. Les conclusions sont basées sur la nature des résultats comme instruments informatifs pour le développement d'expériences professionnelles en évaluation de la langue.

Mots-clés : compétence communicative, croyances, qualités des examens, compétence et habilités d'évaluation en langues étrangères, pratiques évaluatives

Introduction

English language teachers are expected to make decisions based on information about students' language ability (e.g. promote them to a higher level). Given this responsibility and the impact it can have on students, teachers, schools, and society, scholars (see Popham, 2009; Brookhart, 2011) argue that teachers need to be knowledgeable of what assessment entails. Also, teachers need to use assessment results to document, report, and improve learning (McNamara & Hill, 2011; Rea-Dickins, 2001). For language teachers, language assessment literacy (LAL) encompasses large-scale and classroom-based assessment knowledge, skills, and practices, including design, implementation, and evaluation of assessment instruments. Finally, LAL includes the appropriate, ethical, and fair use of assessment to improve teaching and learning (Davies, 2008; Fulcher, 2012; Inbar-Lourie, 2008, 2012). In Colombia, some researchers (Herrera & Macías, 2015; López & Bernal, 2009) have called for the methodological and theoretical preparation of language teachers for language assessment. These authors argue that there is a need to help pre- and in- service language teachers to improve their language assessment theory and practice.

Furthermore, Scarino (2013) argues that teachers' contexts and beliefs play a crucial role in the meaning of LAL, as prior knowledge helps teachers shape this ability. Therefore, in order to help teachers to develop LAL, their contexts should be considered. This is, in fact, a call for language teachers' professional development (Giraldo, 2014; González, 2007), of which language assessment is already a component. For fostering LAL, Brindley (2001) proposes that programs start off from teachers' contexts and build on their previous experiences. In turn, Scarino (2013) states that language teachers should be able to improve their LAL practices while they understand the intricacies of their own context.

This study explored the beliefs and practices of a group of Colombian English teachers regarding the design of an achievement test. Specifically, the study used Bachman & Palmer's framework of test usefulness (Bachman & Palmer, 1996) and validity argument framework (Bachman & Palmer, 2010; Kane, 2006) to analyze test qualities emerging from the beliefs and practices of the participating teachers. This article first gives an overview of the notions of LAL, teacher beliefs and practices, as well as test qualities and achievement tests. It then reviews research on beliefs and practices in language assessment. Later, the method and findings are explained to bring forth discussion and conclusions for the action stage in the ongoing study.

Literature review

LAL includes knowledge, skills, and principles for assessment processes and instruments. In general education as well as language teaching, scholars have argued that design of classroom assessments (e.g. tests, portfolios, peer and self-assessment) is key for a teacher's LAL (Brookhart, 2011; Fulcher, 2012; Popham, 2009; Taylor, 2009). Thus, for language assessment teachers can —among other tasks— design closed- and open-ended tasks, provide clear rubrics for speaking and writing, and use assessment feedback for learning and teaching (Coombe, Troudi, & Al-Hamly, 2012; Fulcher, 2012). In addition to technical and theoretical dimensions, Scarino (2013) has placed attention on a teacher's philosophies for LAL; this cognitive dimension includes teacher beliefs, the focus of the next section.

Teacher beliefs

Johnson (1992) classified teacher beliefs as explicit, those that teachers easily verbalize, and implicit, those that need be inferred from actions. Encompassing more than beliefs, Borg (2003) used the term *teacher cognition* to refer to knowledge, thoughts, actions and beliefs that language teachers have. According to Borg, teachers have cognitions about teaching, learning, assessment, and others. Additionally, Johnson (1994) argued that beliefs are aligned with teachers'

judgments; this, in turn, influences their classroom practice. What is more, as Borg stated, practices can also influence beliefs, so beliefs and practices seem to be complementary rather than mutually exclusive. This is why both Johnson (1994) and Borg (2003) have agreed that understanding teacher beliefs is a step forward in their education.

One trend in the literature is that teacher beliefs are complex. For example, Gabillon (2012) stated that beliefs are personal-social, practical-theoretical, dynamic-resistant, and complex-systematic. The interplay among schooling, experience, and professional challenges contributes to a teacher's beliefs about instructional decision-making (Borg, 2003). In discussing the complexity of teacher beliefs, Borg (in Birello, 2012) remarked on the difference between core and peripheral beliefs; he explained that fixed ideas (core beliefs), such as speaking English all the time in class, can interact with secondary ideas (peripheral beliefs), such as the use of L1 for explaining grammar. This is a specific example of how beliefs are complex in nature. Because of this complexity in beliefs, Fang (1996) proposed the use of interviews to investigate teacher thinking and argued that interviews add data to the widely used paper-and-pencil approach to research into teacher thinking.

Research studies have shown that teachers believe assessment should be used to improve learning and teaching, and provide reports of student progress (Brown, 2004). In the same article, Brown highlighted that teachers consider assessment relevant when valid and informative of student learning. Conversely, they see it as irrelevant when it has a negative impact and is used only for accountability purposes (e.g. to evaluate a teacher or school). In language teaching, beliefs about assessment also reflect some of the trends Brown has discussed. For example, Muñoz, Palacio & Escobar (2012) found that teachers think assessment can be used to improve teaching and learning, but, unlike Brown's study, participants in Muñoz et al. did not see assessment as irrelevant. In this study the teachers believed that language assessment should be formative, even though their practices tend to be summative. The use of summative and formative assessments has been, in fact, a major focus of discussion in language assessment practices.

Teacher practices in language assessment

There are two lenses through which classroom language assessment can be viewed. On the one hand, summative assessments record, evaluate, and document student progress and learning or lack thereof (Brown & Abeywickrama, 2010). The other paradigm is formative assessment, also called alternative. This second type also evaluates students' progress in language. Hamidi (2010) has labeled the former approach as *product-oriented*, focused on the *what* or knowledge students have. Product-oriented assessment reflects norm-referenced testing (Fulcher & Davidson, 2007), whereby students are treated under standard procedures and compared to one another. On the other hand, Hamidi (2010) used the term *process-oriented* for assessment focusing on the analysis of information to strengthen learning. This approach is more focused on *how* students learn language.

Rea-Dickins (2001) and Hill & McNamara (2011) have proposed a description of language assessment practices identified in four major stages. During planning, language teachers decide *what*, *why*, and *how* to assess. The second stage focuses on setting the assessment in motion, whereby teachers introduce tasks for students and explain what tasks involve. A third stage is emergent and occurs during instruction while teachers observe and give feedback to students in class. Rea-Dickins (2004) argued that this observational stance is also part of a language teacher's assessment toolbox. The fourth and last stage in language assessment practices refers to how teachers record and report their summative and formative assessment observations.

In addition to these stages, previous research into assessment practices has focused on the different types of instruments—whether formative or

summative—teachers use. One of the latter is the achievement test.

Achievement tests

Achievement tests rely on a close connection between test tasks/items and curriculum objectives (Hughes, 2003). A crucial characteristic is that these tests must display content validity. That means they must collect information about what is stated in a syllabus in sufficient and direct ways (Fulcher & Davidson, 2007; Brown, 2000). A second characteristic is that achievement tests are primarily product-oriented; because of this characteristic, they are considered instruments for summative purposes (Brown & Abeywickrama, 2010).

Since achievement tests are used for summative purposes, and therefore accountability assessment (Popham, 2009), their importance should not be underestimated. Language testing experts (Bachman & Palmer, 2010; Hughes, 2003) have contended that for a test to be useful, it must have qualities that ensure its scores lead to valid interpretations of what test-takers can do in the language. Thus, the next section highlights the major features of test qualities.

Language test qualities

After Messick's (1989) seminal work on the meaning of validity in language assessment, scholars seem to agree that test qualities provide information to determine the validity of decisions made based on assessment scores (Bachman & Palmer, 2010; Kane, 2006). The next section overviews these qualities and how they are connected to score interpretation.

Reliability refers to the consistency of test results across different conditions. Test-takers should have similar results when they take the same assessment within reasonable time differences between administrations (Bachman & Palmer, 2010; Hughes, 2003) or when scoring involves several evaluators. To strengthen reliability, teachers should include clear instructions in their assessments and design clear rubrics (Brown & Abeywickrama, 2010). If reliability is present, it becomes a piece of evidence to argue for the validity of score interpretations. In other words, a reliable assessment gives clear information about the language ability of students (Fulcher & Davidson, 2007).

Traditionally, validity has referred to the extent to which a test measures what it should measure. However, in 1989, Messick shifted the attention from the assessments themselves to score interpretation, arguing that inferences and decisions made from scores must be valid, not the assessment itself. The meaning of a score, then, is pooled together from evidence to ascertain that one can trust test results —for example, in the case of achievement tests— to argue whether students have or have not met curriculum objectives. Central to the meaning of validity is the concept of construct: the particular test-taker attribute or skill that the test is assessing (Brown, 2000; Fulcher & Davidson, 2007). An achievement test can be considered valid if it assesses the language attributes or skills in a syllabus, such as listening, speaking, reading, writing, grammar, or vocabulary.

Since the construct of language ability in language assessment is crucial, descriptions have been offered to support assessment design. In frameworks such as Bachman & Palmer (1996; 2010), Canale & Swain (1980), and Council of Europe (2001), four major foci can be identified. First, knowledge and use of language involves discourse; this refers to how language is constructed and understood by an individual, as well as knowledge and action regarding how it is co-constructed with others. The ability to understand and produce stretches of discourse, whether oral or written, is central to language ability. Second, sociolinguistic competence refers to understanding and using social conventions, such as register, politeness, idioms, and others. Third, linguistic competence entails the mechanical aspect of language, and how words, meanings, sounds, and symbols are put together to use language. Lastly, these scholars

have proposed strategic competence as part of a person's language ability; this competence refers to the activation of strategies to sustain communication or repair breakdowns in it (Canale & Swain, 1980), or the deliberate actions to project, monitor, and evaluate performance during an assessment (Bachman & Palmer, 2010).

Authenticity in language assessment is the correspondence between assessments and their target language use (TLU) domain. This relationship is a basic component for developing communication-based generalizations from test scores (Brown & Abeywickrama 2010; Coombe et al., 2007). In essence, within a validity argument framework, the level of authenticity of an assessment helps with making interpretations about what test-takers can do with language ability in real contexts. For example, a grammar-only multiple-choice test may not give much information about how a student can use language ability in a real-life situation; however, a performance-based assessment (e.g. a roleplay) may be more useful to assess language use.

Bachman & Palmer (2010) explained that interactive tests foster students' use of language competences, cognitive skills, knowledge of general topics, use of strategies, and affective dimension. The authors explain that interactiveness "must be considered essential to language tests if these are to reflect current views about the nature of language use, language learning, and language teaching" (p. 29). Since an interactive assessment triggers language ability (i.e. the construct), it should lead to valid interpretations and uses of test scores.

Once an assessment has been used, it can impact stakeholders such as students and teachers. For instruction, this impact is expected to be mostly positive (Coombe et al., 2007; Hughes, 2003), although assessment can also have a negative impact (Shohamy, 2001). Positive washback benefits the *what* and *how* of teaching and learning, helps students to be ready for a test, provides feedback so students can improve language ability, and is formative in nature (Brown & Abeywickrama, 2010).

Finally, language assessments should be practical so that resources are used effectively and are not overly costly (Bachman & Palmer, 2010; Coombe et al. 2007). For practicality, Brown & Abeywickrama (2010) have suggested that time, physical and financial resources, and even test marking should be considered for the effective administration of language assessments. Practicality needs to be evaluated because it influences language ability. For example, a multiple-choice test about writing conventions may be practical but not valid to estimate how much writing a person can have.

Related research

Research studies exploring assessment have looked at test qualities, how and what teachers assess, and the nature of instruments. Frodden, Restrepo, & Maturana (2004) have reported the preliminary findings of a study in which the participants were twelve English teachers and five French teachers. The researchers used assessment instruments, interviews, and workshops as ways to collect data for the study. The results came from the analysis of summative and formative assessment instruments. Summative instruments were quizzes, exams, and written drills to assess knowledge of grammar and vocabulary. The rubrics in these instruments included institution and level, but lacked general instructions and time for sitting the test. Students usually interacted with visual —not oral— language, which was contrived, neutral, and had few cultural references.

The researchers found that teachers used communicative tasks, and were starting to use self- and peer- assessment. Frodden et al. (2004) concluded that teachers used summative assessment much more than they did formative, and they seemed to take reliability and practicality into account. Concerning constructs, the researchers concluded

that the instruments mostly assessed grammar and vocabulary.

Since there were no scoring criteria and procedures, the researchers argued summative instruments in this study were unreliable. The tests did not fully embody communicative competence as a construct, given their focus on grammar and vocabulary. Also, authenticity was deemed low as test situations were not contextualized; interactiveness in tests involved using language for personal matters, but there were sections about Miami, Florida, which may not have been relevant to some students. Because teachers designed their own tests, the washback they expected was positive. Finally, regarding practicality, the researchers concluded it was influential in test design, due to challenges such as time, other teaching workloads, and the number of students; because of these, participants used more selected-response items in their tests.

Similar results to those in Frodden et al. (2004) can be found in López & Bernal (2009), who found that teachers rely on summative assessment to assess language ability. Similarly, because of a standards-based influence in China, Cheng, Rogers, & Hu (2004) reported that language teachers tended to overuse traditional, summative assessment.

In Díaz, Alarcón, & Ortiz's (2012) study, the English teachers focused their assessment practices on achievement tests for grammar, vocabulary, and pronunciation. However, the primary school teachers in this study claimed to adhere to a communicative approach to language teaching, whereby all language skills are targeted; nevertheless, this was not evident in their assessments, which can be considered a mismatch between implicit and explicit beliefs (see Johnson, 1992). Contrarily, testing practices of teachers in higher education have shown an alignment with the communicative approach. Even though they still used written closed-ended tests, they included speaking in assessment. Similar results in Díaz et al. (2012) can be found in Arias & Maturana (2005), who claimed there was a limited view (mostly linguistic) of communicative competence in their research with university teachers.

Muñoz et al. (2012) conducted a study with sixty-two teachers from a language institute in a Colombian university. The findings indicated that teachers found assessment central to language education, since it can improve teaching and learning. The teachers conceived assessment as a means to academic improvement, not a route towards accountability and certification. The teachers expressed that their practices were formative rather than summative, as they sought to improve instruction and learning. Nevertheless, the researchers identified a discrepancy between this belief and the teachers' practice because they employed mostly summative assessment. This discrepancy further supports Johnson's (1992) statements about implicit and explicit beliefs.

Arias, Maturana & Restrepo (2012) reported the findings of a study involving five teachers from two universities in Medellín, Colombia. The study engaged the teachers in critical decision-making regarding assessment by employing an assessment framework. The results in this study suggest that test design should be systematic and rigorous to benefit learners, teachers, and institutions. The authors conclude that justice and democracy were key findings in this study, because these two principles helped the teachers to improve their assessment practices. The teachers' assessment improved thanks to a common, etic assessment language and the use of varied and well-designed assessment instruments. The teachers improved as individuals because their practices were fair and democratic, thanks to their concern about students' well-being and their belief in assessing ethically. Another finding Arias et al. highlighted was the relationship between the assessment system and the definition of communicative competence; this led to coherence in the framework. In conclusion, this study provides evidence that training in language assessment can

have a positive impact on teachers' assessment practices, which Colombian scholars have urged.

The problem

The previous research studies have focused on what teachers do and think in terms of their assessment practices in general. The present study, however, has focused on one summative instrument: the achievement test. Thus, the diagnostic stage of the present study sought to "take a picture" of an important testing practice at the institute where it was conducted. This became the problematizing core (Burns, 2005) of the diagnostic stage of an ongoing action research study, as it examined beliefs and practices for the design of this instrument. With this "picture," the study has served as a needs assessment to propose paths to LAL based on teachers' practices, skills, beliefs, and contexts (Brindley, 2001; Scarino, 2013). The study was framed by this question: What are the beliefs and practices of English teachers at a language institute in relation to the design of an achievement test?

In synthesis, the purpose of the study was to identify needs in the area of language assessment and propose paths for LAL improvement. Specifically, as will be apparent in the findings, trends in the beliefs and practices for an achievement test may shed light on what knowledge, skills, and principles of language assessment (Davies, 2008) the teachers can reflect upon and improve.

Context

This study was conducted at the language institute of a Colombian state university. The institute teaches English to university students and has an eight-course program, from elementary to upper-intermediate levels —A1 to B2 in the Common European Framework of Reference for Languages (Council of Europe, 2001).

Teachers at this institute assess students based on a curriculum aligned with a communicative approach, operationalized through task-based and content-based lessons. Assessment in each course is divided into two parts: 60% of the course is assessed through both formative and summative instruments for listening, speaking, reading, writing, and grammar and vocabulary in context. The final 40% is assessed through an achievement test. The teachers must follow guidelines in a document that has been designed by the institute's academic advisors. The following are major issues teachers must consider in test design:

- Inclusion of listening, reading, speaking, writing, and use of the English language (grammar and vocabulary)
- 100 items, each skill weighing 20 points
- Validity: test language constructs and course competences/contents
- Authenticity: employ tasks students are likely to do in the real world
- Administration: similar circumstances for all students in a course
- Evaluation: evaluated by academic advisors
- Washback: provide students with feedback for learning

Method

The diagnostic stage of this study consisted of descriptive qualitative research design, which is a naturalistic, anti-positivist, and idiographic (Cohen, Manion & Morrison, 1998) paradigm because it describes people's thinking and actions. As Cohen et al. (1998) have claimed, researchers "search for meaningful relationships and the discovery of their consequences for action" (p. 10).

The participants were sixty English teachers who anonymously completed a Likert-scale questionnaire about beliefs and practices. Furthermore, the achievement test analysis included fifty artifacts, and fifteen teachers were interviewed as a follow-up to the questionnaire. Thus, two approaches to sampling were used: *convenience* and *purposive* (Mackey & Gass, 2005). Convenience reflects ease in access to informants, in this case for the questionnaire and the interview. The questionnaire

was administered during an institutional meeting. When the sixty teachers completed the questionnaire, they were familiar with the guidelines for test design (see "Context" section) and had already designed at least one achievement test. For the interview, the researcher met with fifteen teachers, one at a time, and asked them about their practices and beliefs in achievement test design in general.

Additionally, purposive sampling occurred across data collection instruments. The purpose was to describe something generalizable for the entire sample of teachers (Mackey & Gass, 2005). The open questions in the questionnaire (see Appendix A), for example, were similar to those in the interviews, which aided in bringing together generalities for testing beliefs and practices at the institute.

Data Analysis

The study used a priori coding for data analysis taken from the test qualities proposed by Bachman & Palmer (1996; 2010), Brown & Abeywickrama (2010) and others. This was done in order to sift through questionnaire and interview answers, as well as document analysis; the codings were *construct validity*, *reliability*, *authenticity*, *washback*, and *practicality*. As Mackey & Gass (2005) argue, a priori categories are welcomed in qualitative research, provided that there is room for emergent categories in findings, as will be shown in the results below.

Beliefs and practices —the overarching categories for data analysis— were operationalized in the questionnaire about test qualities. In order to examine it, an external evaluator familiar with the achievement test at the institute read the items for clarity based on the questions below and provided applicable comments.

Is the statement clear? Yes__ No__
Does the statement ask about beliefs and/or practices related to language assessment? Yes__ No__
Comment on this item (if needed)

After this analysis, a reliability item was modified. Cronbach's alpha coefficient for the questionnaire was 0.69, which is an acceptable coefficient for Likert scales (Dörnyei, 2003).

Next, to examine the language constructs in the tests and teachers' answers, the model of communicative competence presented by the Common European Framework (CEF) was chosen. The CEF has guided English language education at the institute for more than a decade, and teaching at this institute is driven by linguistic, pragmatic, and sociolinguistic competences. Table 1 delineates the data collection instruments and their foci; the number of respondents is displayed next to each instrument. Finally, answers to the interview were classified under the categories and codings for this study but were open to welcome emerging categories. The interview was semi-structured and did not seek to bias teachers towards answers but rather prompt beliefs and practices as they would naturally emerge in an interview (see Appendix B).

Triangulation was used to combine information from the questionnaire, sample tests, and interview answers, based on the aforementioned codings and grouped under beliefs and practices. Analysis was iterative, as the researcher grouped data from these three instruments. For example, the results about practices in construct validity in the questionnaire were grouped with sample items or tasks from the tests, and excerpts from interviews.

Table 1 Data collection instruments and their foci

Instruments/Foci	Beliefs	Practices
Questionnaire (60 respondents)	X	X
Documents: Tests (50 tests)		X
Interviews (15 respondents)	X	X

Findings

The findings are presented in two categories. The information from the questionnaire, interviews, and document analysis sheds light on teachers' beliefs on test design. Additionally, data from

instruments and interviews provides information regarding testing practices based on the a priori categories for data analysis.

Achievement Test Design: Beliefs

Questionnaire and interview results suggested that the majority of the teachers believed tests should be content valid. They stated that the achievement test should assess what students do during the course and be based on its objectives, give clear information about communicative competence, and replicate course tasks. The data are taken from the aforementioned instruments; the Likert scale goes from 1 (strongly disagree), to 5 (strongly agree).

Table 2 Beliefs in achievement test design

Characteristic	Agree	Disagree
The progress test should…		
assess what students studied during the course	59	1
be based on the objectives of the course	58	2
include tasks that are similar to those during the course	51	9
test linguistic knowledge	54	6
test pragmatic knowledge	57	3
test sociolinguistic knowledge	55	5

The following comment comes from an interview and focuses on alignment with course content. It relates to what this teacher thinks about content validity in the achievement test. For coding, *I* equals interview and *T#* equals teacher number.

> I think… take a look at the syllabus and then I make a draft of the test based on the competences we, we have there… as the course goes I um, I make a lot of changes, I make a lot of changes, I change things, I include things that um, so by the end of the course I think, I could have a test that is valid that tests overall language ummm achievement but also that is consistent to the content of the course. (IT5, Pereira, Colombia. 03/27/2015)

Another belief among teachers is that tests should be authentic in the tasks and language they contain. Below, evidence from the questionnaire and interviews that support this claim are shown.

Table 3 Beliefs about authenticity in test design

Characteristic	Agree	Disagree
The Progress Test should…		
have tasks that resemble real-life use of English	57	3
contain language that is natural "sounding"	51	9

In the excerpt below, the teacher explains what she thinks about authentic test tasks and authentic language use.

> I try to make it as real as possible… a situation; something that they can encounter in real life. For example, a job application….They will live when they finish their programs, or applying to a university giving scholarships…not, write a message telling about your last vacation. What for? I always give them the purpose…
>
> It's knowing what they're going to do with that in their lives, in their speaking. (IT13 Pereira, Colombia. 03/27/2015)

Regarding washback, the teachers agreed it should be positive. In general, the answers below suggest that this type of impact characterizes these teachers' thinking in the administration of the progress test.

Table 4 Beliefs about washback in test design

Characteristic	Agree	Disagree
The Progress Test should…		
help you improve your teaching (based on test results)	57	3
help students improve their language learning	55	5

The data from the interview below show this teacher's ideas about positive washback on learning and teaching.

> I always have a self-reflection questionnaire…for them and for me. Uh, what aspects did you find easy or difficult? What do you think it is necessary to be improved? I think you always you need to reflect on

> what you're doing. And that is going to help them to reflect on what they did, with the test and also, it helps me to reflect on some issues regarding to my teaching practice. (IT4 Pereira, Colombia. 03/27/2015)

The next finding is divided into two parts: reliability and practicality. Overall, teachers felt that the achievement test they design should be reliable and practical. The information in table 5 was taken from the questionnaire.

Table 5 Beliefs about reliability and practicality in test design

Characteristic	Agree	Disagree
The Progress Test should…		
Reliability		
give consistent results if students took them twice	54	6
give consistent results if another teacher scored them	53	7
Practicality		
be completed by students within appropriate time constraints.	51	9
designed and scored within appropriate time constraints.	53	7

In the interviews, it was a general trend that teachers included rubrics for speaking and writing tasks. They did this for clarity in assessment—an issue related to reliability and washback. The datum below comes from a teacher's practices. Another use for rubrics is to show students what they are assessed on, which this teacher considers valid.

> I use rubrics for writing and speaking. I design the rubrics based on one I saw on PET and FCE. I always change it too. Writing has, uh, it has 6 traits, and speaking has 4....I think this is good for clarity when I assess their speaking and writing. Also, I show them the rubrics because if I don't, it wouldn't be valid. (IT11 Pereira, Colombia. 03/27/2015)

Achievement test design: practices

The second set of findings illustrates practices in test design and summarizes the analysis of the 50 tests that were scrutinized. Data from the questionnaire and interviews are triangulated to support results. Test qualities described in depth are reliability, validity, authenticity, interactiveness, and washback. The last focus of this section is on the construct of communicative competence and the ways it is embedded in assessment.

After analysis, consistency in the results may be present in tests requiring learners to choose the correct answer in a task. This occurred in reading and listening, where most tests included tasks such as matching, multiple choice, and "true-false-does not say." All tests included an answer key for close-ended test items. However, actual reliability is ascertained more clearly with a statistical analysis of test responses, a research procedure that would need consent from students. This was beyond the scope of this research study.

In the rubrics for speaking and writing, there was no uniformity as teachers used different approaches for their design. Table 6 shows two examples of rubrics for writing, and Table 7 presents the rubrics for speaking in the same course. Also evident in the samples below is a variety of approaches to weight in the rubrics. Test #1 in writing assigns 5 maximum points to the two assessed components, while test #23 does not assign any points. For speaking, test #3 assigns points to level descriptors, while test #22 assigns 4 points for each criterion.

Table 6 Sample Rubrics for Writing

Writing (Course Four)
Test #1
Content: Should address both parts of the task. Write a text message card and include adjectives. (___/5)
Language accuracy: Should not contain major errors that lead to misunderstandings or that irritated the reader (___/5)
Test # 23 (no score given)
Spelling / Grammar / Structure / Punctuation / Vocabulary

In terms of content validity, most teachers argued they base their tests on the contents of the course syllabus. They explained that test topics and tasks resembled those that students performed during their course. The excerpts below describe two teachers' practices as they relate to validity. In the

Table 7 Sample Rubrics for Speaking

Speaking (Course Seven)				
Test # 3				
F/A: Fairly Accomplished = 1pt		*O/S: On the Standard = 2pts*		
A/S: Above Standards = 3 pts		*H/S: Highly Over Standards = 4 pts*		
Descriptor				*F/A O/S A/S H/S*
1. Pronunciation and intonation is appropriate and understandable according to the level				
2. The oral discourse is coherent and clear.				
3. The participant incorporated all the info required for the description with sufficient range of vocabulary.				
4. Hesitation is presented; it doesn't affect communication though.				
5. Accuracy is evidently well incorporated. Learner respects mainly the subject-verb agreement and basic grammar rules.				
Test # 22 (4 points each)				
Preparation: The student investigated about the country and was prepared to speak.				
Vocabulary: Appropriate use of vocabulary was identified, it made the message understood.				
Coherence: The ideas presented were very clear. The presentation had coherent sequence.				
Fluency: Certain fluency was identified according to the learner's level.				
General achievement: The student has done fairly a good job, paying attention to pronunciation, appropriate vocabulary and grammar.				

questionnaire (where *Q* means questionnaire, *T#* teacher number, and *OQ#* open question number), one of the teachers stated that "The tests I design are totally connected to the topics and language items studied during the course. They also are based on the syllabus and updated topics. (QT15-OQ2)." The following datum comes from an interview in which the teacher explains her approach to validity in test design.

> The first thing that I take into account is all the topics and language issues that I work during the course. And the kind of activities that I designed with them, so I start to structure my exam based on that....I also take a look at the syllabus again so I am not testing something that is not appropriate. (IT7 Pereira, Colombia. 03/27/2015)

Test analysis revealed that the specific language areas teachers assessed were different across tests. In table 8 below, Test #8 included register, a component of sociolinguistic competence that was not assessed in test #34. In test #34, students had to understand instructions, a skill related to reading comprehension, and they were asked to check their spelling and punctuation, areas of writing which were not addressed in test #8. The samples below were taken from two different tests designed by different teachers teaching the same course.

Authenticity occurred in most tests as teachers included authentic content and tasks. In reading and listening, the general trend was to use topics of interest to a general university audience: Twitter for business, job applications, NGOs, and everyday conversations. In writing, tests generally asked learners to reply to letters and emails, write short articles, statements of purpose, and others. In speaking, tasks included spontaneous conversations with teacher and/or classmates, oral presentations, and debates.

This interview excerpt shows how the teacher addressed authenticity in test design: "One of them was replying an email, your concern about certain situation in your country. Emails, short articles....

Table 8 Differences in Language Skills Being Assessed

Test #8 writing (Course One)
You went to the cinema last weekend to watch a very nice film. One of your friends asked you about the film, so that he/she can go and watch it with his/her friends.
What's the film story?
How long did it last?
Did you enjoy it?
Content (2 Points)
It should describe main activities.
Language Accuracy (2 Points)
It should not contain major errors that lead to misunderstanding or which irritate the reader
Range (2 Points)
It should have vocabulary about daily routines.
Register (2 Points)
It could range from fairly formal to semi-formal but should be the same throughout the description.
Target reader (2 Points)
The reader should be informed about the subject and his/her occupation
Test #34 writing (Course One)
Your friend wants to visit Pereira, he wants you to express your opinions about things that you like or dislike about the city (food, people, weather, transportation, etc) (25-30 words). Include a greeting and an ending.
Point Values (1, 2, 3, 4, or 5 for each descriptor)
Task understood (The students understood the instruction and they are sequentially followed to achieve the aim).
Vocabulary mostly appropriate (The vocabulary used is according to their level and effective words are used)
Mistakes do not affect meaning (The paper is neat, legible, and presented in an appropriate format.)
Punctuation (Sentences are punctuated correctly, and the piece is free of fragments and run-ons.)
Spelling (The writing is free of misspellings, and words are capitalized correctly, comas and periods are also used)

Also, they are supposed to write short motivational letters, curriculum vitas, um, reply comments. (IT1 Pereira, Colombia. 03/27/2015)." The questionnaire data in table 9 shows how frequently teachers employ authenticity in test design (1. Never (N)...4. Frequently (F), 5. Always (A)). Lastly, an excerpt from test # 5 shows what could be considered authentic for university students.

Teachers leaned towards using materials students would relate to; this aligns with interactiveness as it deals with students' topic knowledge, communicative competence, and emotional relatedness. The tests included topics such as technological issues, the lives of people and everyday issues, Colombian and world tourist destinations, applying for jobs and scholarships, and others. These topics are related to the majors at the university where the study was conducted. This interview excerpt shows the content the teacher uses in test design: "I like to use technology because it makes them feel less, ahhhh, stressed out and they feel confident when they see those types of topics like technology or sciences." (IT3 Pereira, Colombia. 03/27/2015). The instructions in table 10 below were taken from a test and show the content for the reading section.

Most teachers use test results for positive washback. Teachers ask learners to revise the test to see what they can improve, and teachers in turn see what they need to improve themselves. The interview extract below shows a teacher's approach to engage students in analyzing test results for positive feedback on learning; likewise, the data show

Table 9 Authenticity in Test Design

Questionnaire		
The progress tests YOU design	# of teachers who do it	
have tasks that resemble real-life use of English	F: 23, A: 28	Total: 51
contain language that is natural "sounding"	F: 20, A: 29	Total: 49
Test # 05 Instructions for a writing task		
You just saw on the NYU website a scholarship that fits your interests. They offer students the possibility to travel to New York and spend a whole semester in their campus so that they improve their professional skills before graduating. Write a statement of purpose to this university. Tell them about yourself, your background, your professional goals and why NYU is perfect for you (80 words).		

Table 10 Sample Topic in a Reading Test

Test #30 (Course Seven)
Your friends Alli, Frank, Jhon and Peter are looking for a technological device with special characteristics. They are asking you to give them advice about the perfect device for them. Read the characteristics of the devices, then read the text and choose the appropriate one for each one of your friends. Write the number (#) of the device on the lines provided.

how the teacher discusses what to improve in his instruction.

> I really like to check the exams in class, with the students.... They can see the results.... I sit with each student and check their writing. And they know what, like, the mistakes that they have.... Of course, after that, I talk to some friends [teachers] about the results and the things we need to improve for the other, I mean, uhm, the next exam. (IT9 Pereira, Colombia. 03/27/2015)

To confirm the washback trend, Table 11 below shows questionnaire answers with frequency of washback in testing practices.

There was relative consistency between what the tests assessed and communicative competence, the institute's overall learning goal. First, discourse competences are included in the reading and listening sections, where students are asked to understand general and specific information, vocabulary in context, connections among ideas, and whether the texts contain certain information or not. Concerning functional competence, teachers assessed functions in speaking or writing tasks. Second, linguistic knowledge and skills were directly addressed in the grammar and vocabulary section. The analysis of test tasks revealed that teachers asked students to use correct forms of grammar or vocabulary items where they would logically fit in conversations.

Lastly, tests and interview answers imply that sociolinguistic competence is tested superficially. The only assessed area in this competence was formal and/or informal register as a criterion for writing production in several tests; for example, see Test #8 Writing in table 8 above. This practice, however, is not consistent across instruments, as can be observed in the samples in the same table. Sociolinguistic issues such as linguistic markers of social relations, politeness conventions and others are not explicitly assessed in the instruments. When asked about specific skills they assess in writing and speaking, two teachers mentioned sociolinguistic skills (formality) superficially for writing but not for speaking, as shown in Test # 45 below. Table 12 shows sample tasks that relate to specific aspects of communicative competence.

The sample below shows how this teacher addresses sociolinguistic skills in writing, specifically register as a subcomponent of this competence. The sample confirms how sociolinguistic is included consistently in writing tests, but not so much in speaking tests.

Table 11 Washback Based on Test Results

Questionnaire	
The progress tests YOU design	# of teachers who do it
help you improve your teaching (based on test results)	1(1) 2(0) 3(7) 4(17) 5(35)
include feedback for students	1(0) 2(4) 3(5) 4(18) 5(34)
Key: 1=never, 2=rarely, 3=sometimes, 4=frequently, 5=always. Result in parentheses.	

Table 12 Communicative Competence Skills in Test Design

Discourse competence (understanding related ideas)
Test #49
You are going to read a newspaper article from the New York Times about FARC in Colombia. Seven sentences have been removed from the article. Choose from the sentences A-H the one which fits each gap (1-7). Be careful, there is one extra sentence which you don't need to use.
Discourse competence (understanding specific information)
Test #10
Two friends are talking about vacation and clothes. What does the man plan to wear during the summer months?
Functional competence (proposing ideas, promising, expressing obligation)
Test #6
You are a candidate for the coming Pereira's mayor elections. Write an 80 words proposal for the citizens telling them the situations you would change or do if you won the elections. Also, state some of the things people will be able to do and what things they will have to do.
Linguistic competence
Test #12
You are going to read and complete the following interview using the past tense or the present perfect of the verbs given in brackets.
Sociolinguistic competence
Test #45
You recently met someone on-line who lives abroad. You have decided to send him/her an E-mail to tell a bit more about yourself and your English lessons. Do not forget to include the following points in your e-mail:. . .
. . . Remember that this is an informal letter. An informal letter contains an introduction, body and conclusion.

> They need to show us if they know how to use English in a formal register or an in informal register. We are expecting them to have the two possibilities. They need to know formal expressions and academic writing. The first part is informal, a letter, comments... the second part formal, like an essay. That way, they are showing me if they can use the formal and informal registers. (IT10 Pereira, Colombia. 03/27/2015)

Achievement test design: Challenges

Three challenges emerged from data analysis in the study. Firstly, the teachers found it demanding to obtain the right material and gauge task difficulty for listening. This questionnaire answer shows the problem with the listening section: "Listening skill. It's always the last skill to be evaluated because I'd like to find listenings that are related to the topic studied during the class and sometimes doing that is very challenging. (QT29-OQ2)." Secondly, the grammar section (called *Use of English* by the teacher) was another area for teachers to improve, as this questionnaire answer shows: "I think it should be great to learn more about the implementation of use of English tests. (QT10-OQ2)." Similarly, in the interviews, teachers highlighted the need to improve the design of the listening and grammar tasks in tests. For listening, this teacher stated that "[f]inding the specific content for the listening

is not easy; sometimes the recordings are very advanced. You know, should I use authentic adapted or materials?" (IT2 Pereira, Colombia. 03/27/2015). In terms of the grammar section, this teacher argues that it "is very challenging because it is hard to, to make, or design things in, in context". (IT15 Pereira, Colombia. 03/27/2015).

A final challenge had to do with lack of opportunities to provide positive washback. Even though the teachers wanted to give formative feedback for students based on the results of the achievement test, the dynamics of the institute have been a constraint because teachers do not generally continue with the same students during two courses. When grades are reported, teachers and students do not have follow-up conferences where formative feedback can be provided. This interview excerpt shows why the teacher did not have a chance for positive washback on language learning: "In this semester I was not able to do that because we changed, eh, students. When you have two courses in a row with them, you finish the course and you start the course doing that formative assessment." (IT12 Pereira, Colombia. 03/27/2015).

Discussion

Findings in this study suggest that, as Borg (2003), Johnson (1994) and others have argued, beliefs and practices coexist in teachers' cognition in intricate ways. This study shows evidence that the teachers beliefs tend to align with their practices, but this is not always the case. Specifically, these teachers believed language tests should be reliable; however, the lack of a unitary approach to rubric design may impede consistent decisions when teachers assess students' performance in speaking and writing. The teachers also believed achievement tests should assess sociolinguistic competence, but their practices were found to be rather limited for this language construct. The discrepancy between beliefs and practices has also been documented in other studies (Cheng et al., 2004; Díaz et al., 2012; López & Bernal, 2009), where teachers believed they should assess communicatively but their practices showed otherwise.

Another similarity between the present study and others (for example Brown, 2004; Díaz et al., 2012) is that teachers' believed assessment should be used for learning. Even though there are limitations to this practice, the questionnaire responses and interview data show positive washback on teaching and learning; this, therefore, makes this achievement test both summative and formative.

The findings in this study, on the other hand, differ substantially from those conducted by Arias & Maturana (2005), Díaz et al. (2012), and Frodden et al. (2004). While in those studies, assessment focused mostly on linguistic competence (grammar and vocabulary), the present study's findings suggest the teachers have a more comprehensive approach to language ability, albeit with the discrepancy regarding sociolinguistic competence. Also, the contrived use of language in tasks in Frodden et al.'s (2004) study contrasts with the authentic tasks teachers in the present study designed.

Limitations

One limitation in this study must be addressed. The answers to the questionnaire may have been influenced by what is called social desirability bias, whereby participants respond with what they believe is right or seems to be right (King & Bruner, 2000). One way to tackle this issue was to ask the teachers to answer anonymously and honestly. This limitation may have been the reason why Cronbach's alpha for 50 items was 0.69—an acceptable but not ideal result that may have had an impact on the overall validity of the questionnaire.

Conclusions

The present research study describes the beliefs and practices teachers at a language institute have in designing an achievement test. The teachers believe tests should comply with four fundamental

principles: validity, reliability, authenticity, and positive washback on learning and teaching.

Reliability may be present in test design when it comes to receptive skills and grammar and vocabulary in use, given the design of close-ended tasks. As stated earlier, further statistical analyses would be needed to determine reliability, but this would require access to students' test scores. The findings also show a lack of reliability in inconsistent rubrics for speaking and writing, even across groups of the same course and level. Therefore, not having a unified approach to designing rubrics for speaking and writing could become problematic for the validity of inferences from scores. Concerning validity, tests align with communicative competence in the sense that they do assess pragmatic, linguistic, and sociolinguistic skills. Nevertheless, the superficial assessment of sociolinguistic competence leads to a relative lack of construct validity in test design, and it contradicts the belief that language tests should assess this construct. Additionally, tests tend to be authentic as they possess tasks from the real world. Teachers also strive to provide learners with a testing experience that has face validity and engages them emotionally; coupled with relevant areas of language ability, tests in this study tend to be interactive. Finally, the efforts in giving feedback to learners make test washback positive, but there are administrative constraints at the language institute that impede this practice.

The purpose of this diagnostic study was to identify needs in language assessment of English language teachers at a Colombian institute and use this information for furthering LAL. Based on beliefs and practices specifically for an achievement test, three foci for an LAL program can be suggested. First, teachers in this study may benefit from a program that examines reliability and validity for classroom achievement tests, and how these qualities nurture one another for appropriate inferences from test scores. Specifically, workshops about unitary design for reliable and valid rubrics vis-à-vis a communicative competence model could prove enriching. Second, the test design practices that the teachers already follow can be used as a jumping off point for further professional development in language assessment (Scarino, 2013). These ideas should prove useful for the future action stage in the action research cycle of this study. Finally, a program which includes the design of listening and grammar sections for tests, as well as alternatives for positive washback, may be welcomed by the group of teachers.

References

Arias, C. I., & Maturana, L. (2005). Evaluación en lenguas extranjeras: discursos y prácticas. *Íkala, Revista de Lenguaje y Cultura, 10*(1), 63-91.

Arias, C. I., Maturana, L. & Restrepo, M. I. (2012). Evaluación de los aprendizajes en lenguas extranjeras: hacia prácticas justas y democráticas. *Lenguaje, 40*(1), 99-126.

Bachman, L. & Palmer, A. (1996). *Language testing in practice: Developing useful language tests*. Oxford: Oxford University Press.

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

Birello, M. (2012). Teacher cognition and language teacher education: Beliefs and practice. A conversation with Simon Borg. *Bellaterra, Journal of Teaching & Learning Language & Literature, 5*(2), 88-94.

Borg, S. (2003). Teacher cognition in language teaching: A review of research on what language teachers think, know, believe, and do. *Language Teaching, 36*(2), 81-109.

Brindley, G. (2001). Language assessment and professional development. In C. Elder, A. Brown, K. Hill, N. Iwashita, T. Lumley, T. McNamara, & K. O'Loughlin (Eds.), *Experimenting with Uncertainty: Essays in Honour of Alan Davies* (pp. 126-136). Cambridge: Cambridge University Press.

Brookhart, S. M. (2011). Educational assessment knowledge and skills for teachers. *Educational Measurement: Issues and Practice, 30*, 3-12.

Brown, J. D. (2000). Questions and answers about language testing statistics. *Japanese Association of Language Testing*, (2), 8-12.

Brown, H. D. & Abeywickrama, P. (2010). *Language Assessment: Principles and Classroom Practice*. New York: Pearson Longman.

Brown, G. (2004) Teachers' conceptions of assessment: implications for policy and professional development. *Assessment in Education: Principles, Policy & Practice, 11*(3), 301-318.

Burns, A. (2005). Action research. In E. Hinkel (Ed.), *Handbook of Research in Second Language Teaching and Learning* (pp. 241-262). London: Lawrence Erlbaum Associates, Inc. Cambridge: Cambridge University Press.

Canale, M. & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics* (1), 1-47.

Cheng, L., Rogers, T., & Hu, H. (2004). ESL=EFL instructors' classroom assessment practices: Purposes, methods, and procedures. *Language Testing, 21*(3), 360-389.

Cohen, L., Manion, L. & Morrison, K. (1998). *Research Methods in Education*. London: Routledge.

Coombe, C., Folse, K. & Hubley, N. (2007). *A Practical Guide to Assessing English Language Learners*. Michigan (US): The University of Michigan Press.

Coombe, C., Troudi, S. & Al-Hamly, M. (2012). Foreign/second language assessment literacy. In C. Coombe, P. Davidson, B. O'Sullivan, & S. Stoynoff (Eds.), *The Cambridge guide to second language assessment*, (pp. 20-29). Cambridge: Cambridge University Press.

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Davies, A. (2008). Textbook trends in teaching language testing. *Language Testing, 25*(3), 327-347.

Díaz, C., Alarcón, P. & Ortiz, M. (2012). The English teacher: His beliefs about English language assessment at primary, secondary and tertiary levels. *Íkala, Revista de Lenguaje y Cultura, 17*(1), 15-26.

Dörnyei, Z. (2003). *Questionnaires in second language research: construction, administration, and processing*. London: Lawrence Erlbaum Associates, Inc.

Fang, Z. (1996). A review of research on teacher beliefs and practices. *Educational Research, 38*(1), 47-65.

Frodden, M. C., Restrepo, M. I., & Maturana, L. (2004). Analysis of assessment instruments used in foreign language teaching. *Íkala, Revista de Lengua y Cultura, 9*(15), 171-201.

Fulcher, G. & Davidson, F. (2007). *Language Testing and Assessment: An Advanced Resource Book*. New York: Routledge.

Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly, 9*(2), 113-132.

Gabillon, Z. (2012). Revisiting Foreign Language Teacher Beliefs. IOKSP. *International Online Language Conference 2012, November 3* (pp. 190-203). Online: Universal Publishers. halshs-00799937.

Giraldo, F. (2014). The impact of a professional development program on English language teachers' classroom performance. *Profile. Issues in Teachers' Professional Development 16*(1), 63-76.

González, A. (2007). Professional development of EFL teachers in Colombia. *Íkala, Revista de Lenguaje y Cultura, 12*(18), 309-332.

Hamidi, E. (2010). Fundamental issues in L2 classroom assessment practices. *Academic Leadership: The Online Journal, 8*(2).

Herrera, L. & Macías, D. (2015). A call for language assessment literacy in the education and development of Teachers of English as a foreign language. *Colombian Applied Linguistics Journal, 17*(2), 302-312.

Hughes, A. (2003). *Testing for language teachers*. Cambridge: Cambridge University Press.

Inbar-Lourie, O. (2008). Constructing a language assessment knowledge base: A focus on language assessment courses. *Language Testing, 25*(3), 385-402.

Inbar-Lourie, O. (2012). Language assessment literacy. In C. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics* (pp. 1-9). Oxford: John Wiley & Sons, Inc.

Johnson, K. E. (1992). The relationship between teachers' beliefs and practices during literacy instruction for non-native speakers of English. *Journal of Reading Behavior, 24*(1), 83-108.

Johnson, K. E. (1994). The emerging beliefs and instructional practices of preservice English-as-a-second-language teachers. *Teaching and Teacher Education, 10*(4), 439-452.

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational Measurement* (4th Ed.) (pp. 17-64). Westport, CT: American Council on Education and Praeger.

King, M. & Bruner, G. (2000). Social desirability bias: A neglected aspect of validity testing. *Psychology & Marketing, 17*(2), 79-103.

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.

Mackey, A. & Gass, S. (2005). *Second Language Research: Methodology and Design*. London: Lawrence Erlbaum Associates, Inc.

McNamara, T. & Hill, K. (2011). Developing a comprehensive, empirically based research framework for classroom-based assessment. *Language Testing, 29*(3), 395-420.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.), 13-103. New York: American Council on Education and Macmillan.

Muñoz, A., Palacio, M. & Escobar, L. (2012). Teachers' beliefs about assessment in an EFL context in Colombia. *Profile. Issues in teachers' professional development, 14*(1), 143-158.

Popham, W. J. (2009). Assessment literacy for teachers: Faddish or fundamental? *Theory Into Practice, 48*, 4-11.

Rea-Dickins, P. (2001). Mirror, mirror on the wall: identifying processes of classroom assessment. *Language Testing, 18*(4), 429-462.

Rea-Dickins, P. (2004). Understanding teachers as agents of assessment. *Language Testing, 21*(3), 249-258.

Scarino, A. (2013). Language assessment literacy as self-awareness: Understanding the role of interpretation in assessment and in teacher learning. *Language Testing, 30*(3), 309-327.

Shohamy, E. (2001). Democratic assessment as an alternative. *Language testing, 18*(4), 373-391.

Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics,* (29), 21-36.

Appendix A: Questionnaire for beliefs and practices

Dear teacher:

I would like you to complete this survey about your assessment beliefs and practices in the English class. It is anonymous, so please do not write your name anywhere on this document. The information will be used for research and academic purposes only.

Part one: Please, rate statements *A* to *Y* based on this scale:

1: Strongly disagree | 2: Disagree | 3: Undecided | 4: Agree | 5: Strongly agree

The progress tests should

A. assess what students studied during the course 1__ 2__ 3__ 4__ 5__
B. be based on the objectives of the course
C. give clear information about students' language competence
D. include tasks that are similar to those during the course
E. give clear information about what students can/can't do
F. test linguistic knowledge
G. test pragmatic knowledge
H. test sociolinguistic knowledge
I. have tasks that resemble real-life use of English
J. include contextualized items (exercises)
K. include topics that are meaningful/relevant to students
L. contain language that is natural "sounding"
M. help you improve your teaching (based on test results)
N. help students improve their language learning
O. be designed for students to do their best
P. tell students clearly what they can/can't do
Q. stay within budgetary limits

R. be completed by students within appropriate time constraints
S. have clear instructions for performance
T. be designed and scored within appropriate time constraints
U. give consistent results if students took them twice
V. give consistent results if another teacher scored them
W. have clear instructions for scoring
X. have clear and uniform rubrics for assessment
Y. contain tasks (exercises) that are clear for students

Part two: Please, rate the frequency with which you do the following:

1: Never | 2: Rarely | 3: Sometimes | 4: Frequently | 5: Always
The progress tests YOU design
(Items are the same as part one)

Part three: Please, answer the following questions based on your experience.

What other tasks do you use to assess students (60% of the course)?
What are your strengths in the design of progress tests and other assessment instruments?
What are your challenges? What do you feel you need to improve?

Appendix B: Questions for semi-structured interview

Dear teacher:

I would like you to answer the questions in this interview, which are about your testing beliefs and practices at ___. The information from this interview will be used for research and academic purposes only. Your name will not be made public.

Semi-structured interview:

1. How do you design the final progress test?
   1.1 What resources do you use?
   1.2 What steps do you take to design it?
2. What do you assess in this test? Why?
3. What do you do with the results of the test? Why?
4. What are your strengths in the design of the test? Why?
5. What are your challenges? Why?
6. Is there anything you need to improve? What do you feel you need to improve?

How to reference this article: Giraldo Aristizábal, F. (2018). A Diagnostic Study on Teachers' Beliefs and Practices in Foreign Language Assessment. *Íkala, Revista de Lenguaje y Cultura, 23*(1), 25-44. DOI: 10.17533/udea.ikala.v23n01a04