ESTIMATING INTRA-RATER RELIABILITY ON AN ORAL ENGLISH PROFICIENCY TEST FROM A BILINGUAL EDUCATION PROGRAM

VIVIAN ALEJANDRA AGUIRRE ISAZIGA
DANIELA MONTOYA RUIZ

UNIVERSIDAD TECNOLÓGICA DE PEREIRA
FACULTAD DE BELLAS ARTES Y HUMANIDADES
LICENCIATURA EN BILINGÜISMO CON ÉNFASIS EN INGLÉS
PEREIRA
2019

ESTIMATING INTRA-RATER RELIABILITY ON AN ORAL ENGLISH PROFICIENCY TEST FROM A BILINGUAL EDUCATION PROGRAM

TRABAJO DE GRADO COMO REQUISITO PARA OPTAR AL TÍTULO DE LICENCIADO EN BILINGÜISMO CON ÉNFASIS EN INGLÉS

ASESOR
DANIEL MURCIA QUINTERO

UNIVERSIDAD TECNOLÓGICA DE PEREIRA
FACULTAD DE BELLAS ARTES Y HUMANIDADES
LICENCIATURA EN BILINGÜISMO CON ÉNFASIS EN INGLÉS
PEREIRA

Acknowledgments

First of all, I would like to express my gratitude to my friend Vivian Aguirre for being the most perfect research partner I have ever met. Without her commitment and motivation, this project would not have been possible. I owe a lot of gratitude to her for always being there for me and for always managing to make me feel special. I feel privileged to be her friend. She is a great and charming person.

My special words of thanks also goes to our research assessor, professor Daniel Murcia, for trusting our potential and for encouraging us to give our best and write this project. His guidance, cooperation and support were of invaluable help in the development of this investigation.

Finally, I would like to thank my family and friends who have always been a major source of support when things would get a bit discouraging. I am thankful to them for all the emotional support, entertainment, and caring they provided.

Daniela Montoya R.

First and foremost, I thank God because He has never failed in keeping his promises. Throughout "this journey", He has been faithful, He has been merciful, and I can only be grateful about it.

I would also like to express my deepest gratitude to my parents who always supported me and never ceased to push me to the finish line. A very special gratitude also goes to my family and friends who helped me survive all the stress and motivated me to continue working on this project.

I owe special thanks to professor Daniel Murcia, our research advisor, who not only trusted us and gave us the opportunity to conduct this study but also provided us with thoughtful guidance and encouragement in the whole process. Thanks are extended to professor Frank Giraldo whose knowledge and powerful ideas laid the foundations of this project; your inspiring lessons sparked my love for language assessment.

Finally, I am fully indebted to my dear friend and research partner Daniela Montoya for all her patience, positivism, understanding and wisdom. I cannot thank you enough for all you have invested in this study and my life; this accomplishment would have not been possible without you.

Vivian Alejandra Aguirre

Table of Content

Table of Figures

List of Tables

Abstract

This research project aims at presenting the results of an investigation which sought to estimate the level of intra-rater reliability on an oral English proficiency test and to determine the different external and internal factors that affect rater's consistency. The participants involved in the development of this study were two professors in charge of rating the speaking section of a proficiency test administered at the Licenciatura en Bilingüismo con énfasis en inglés program. A correlation coefficient was calculated in order to establish the raters self-agreement, whereas a retrospective verbal report was carried out as an instrument to gather information about the factors influencing rater reliability. The findings suggest that there is a high level of intra-rater reliability on the proficiency test as the correlation coefficient yielded values above .80. However, aspects related to the lack of adherence to rubric's criteria, rater-student relationship, physical conditions and rater's pressure and responsibility to give an accurate score were identified as factors affecting rater's self-consistency. Lastly, some significant implications drawn from this investigation are provided.

Keywords: intra-rater reliability, language assessment, factors influencing reliability, speaking skills

Resumen

Este estudio tiene como objetivo presentar los resultados de una investigación la cual pretendía estimar el nivel de confiabilidad intra-evaluador en un examen de suficiencia oral en inglés, y determinar los diferentes factores internos y externos que afectan la consistencia del evaluador. Los participantes involucrados en el desarrollo de este estudio fueron dos profesores encargados de evaluar la sección de habla de un examen de suficiencia administrado en la Licenciatura en Bilingüismo con énfasis en inglés. Se calculó un coeficiente de correlación con el fin de establecer la consistencia de los evaluadores mientras que un protocolo verbal retrospectivo fue llevado a cabo para recopilar información acerca de los factores que influyen en la confiabilidad del evaluador. Los resultados sugieren que hay un alto nivel de confiabilidad intra-evaluador en el examen de suficiencia en cuanto el coeficiente de correlación arrojó valores superiores a .80. No obstante, aspectos relacionados con la falta de adhesión a los criterios de la rúbrica, la relación evaluador-estudiante, las condiciones físicas, y la presión y responsabilidad del evaluador para dar una nota precisa fueron identificados como factores que afectan la consistencia del evaluador. Finalmente, se proporcionaron algunas implicaciones procedentes de esta investigación.

Palabras clave: confiabilidad intra-evaluador, evaluación de lengua, factores que influencian confiabilidad, habilidad de habla

1. Presentation

Speaking in a second language is a difficult and challenging task since language users are required to have a certain level of mastery in grammar, pronunciation and vocabulary knowledge, pragmatic competence and communicative strategies. What is more, as the spoken language occurs in real time, learners need to combine multiple skills in just seconds. Speaking does not only constitutes a concern to the second language acquisition field but also to language assessment. On the one hand, assessing this skill is complex due to the fact that human raters are required to judge students' performances, and subjectivity may affect the scoring process. On the other hand, teachers and evaluators have to make instantaneous decisions about students' oral proficiency given the fast-fading nature of speech. Considering all of the above-mentioned aspects, it is concluded that speaking is a difficult skill to assess reliably.

The purpose of the present study was to analyze the rater reliability on an oral English proficiency test implemented in a language undergraduate program as well as to explore the factors that influence rater's self-consistency. This research was conducted with two of the raters assessing the speaking test. In order to carry out the study, each rater was asked to score 5 students in two different moments. As a result, two set of scores for each of the students were yielded and collected. The scores were correlated through a statistical method in order to estimate each rater's reliability level. After that, a verbal report was held with the raters with the aim of gathering qualitative information regarding the factors that could impact their reliability.

Overall, it was found that the level of reliability of both raters was satisfactory in spite of the fact that the majority of the factors identified could have negatively influence their consistency. The perspective of the raters was crucial for establishing the factors explored in this study. In regards to the implications derived from the findings, new inquiries arose in terms of the raters' cognitive processes, reliability on the assessment of different skills and additional aspects that can prevent raters from being reliable. Moreover, additional recommendations to foster teachers and raters' language assessment practices and knowledge were provided.

2. Statement of the Problem

Throughout history, the Colombian government has launched several policies and programs aiming at strengthening bilingualism all around the country. The bilingual program currently in force is "The National English Program: Colombia Very Well", which purpose is to position Colombia as the Latin American country with the best English level by the year 2025. As for higher education, one of its aspirations is to have at least 85% of language undergraduate students with a C1 English proficiency level based on the Common European Framework of Reference for Languages (CEFR), whereas students from other bachelor programs are expected to achieve a B2 proficiency level on the language.

Considering the requirements prescribed by the national English program and the responsibility that comes with it, the Universidad Tecnológica de Pereira, located in Risaralda-Colombia, formulated the "Acuerdo 15 del 2015", which established that students from the Licenciatura en Bilingüismo con énfasis en inglés program (LBI) must take a C1 English proficiency test as a requirement for graduation. The resolution also stated that students from the program, by means of a test, have to demonstrate progress in their language proficiency at two specific moments within the curriculum: (1) in the fourth, and (2) in the seventh semester. The idea of implementing these tests, which received the name of "Pruebas de Semaforización", arose from the necessity of tracking students' progress in the language so that the program could guarantee high quality teachers. In order to do so, the test was developed by the professors of the program following the design of the First Certificate in English test (FCE), and including four different skills: Listening, Reading, Writing and Speaking. At first, the "Pruebas de Semaforización" were linked to the Upper-intermediate English course, which belongs to the fourth semester of the program, and the Professional Development course from the seventh semester. However, two years later, the Acuerdo 15 was updated resulting in the Acuerdo 18. The latter modificated the subject in which the seventh semester test was implemented and so the Initiation to the Teaching Practicum was adopted as the new course.

This test was implemented for the first time in the second semester of 2016 and from that time, the test has been administered twice a year. In total, the LBI program has compiled information regarding the results of 4 administered tests for fourth semester, and 5 test administrations for seventh semester. Despite the numerous times students had taken the test, the results have been superficially analyzed, and no in-depth evaluation about the reliability of the test has ever been conducted. For this reason, the scores from previous test implementations were statistically analyzed and tabulated, and it was found that in 7 out of 9 administrations, the speaking skill yielded the highest scores compared to the other language skills. This phenomenon can be the result of a reliability matter. Reliability, as a quality of test scores, is crucial in determining the validity of an assessment. Hughes (2010) asserts that a valid test is one that provides accurate scores that reflect test-takers' performance. In other words, for a test to be valid, it has to be reliable. Considering this assumption, it becomes of paramount importance to

explore the extent in which the "Pruebas de semaforización" are reliable. In this way, test-designers and test-administrators could obtain accurate information about students language ability that would allow them to identify possible shortcomings and take action in this regard.

Bearing in mind that "human error, subjectivity, and bias may enter into the scoring process" (Brown, 2004, p. 28), it can be concluded that a factor of variation on the reliability of a test derives from the scoring process carried out by the raters. From the skills assessed in language tests, speaking and writing are the ones that require human raters to carry out the scoring procedures. Speaking, in this regard, is the productive skill that is more difficult to assess, not only because of the subjectivity of the process, but also because it is fast-fading. A handful of studies have conducted research on this issue. For instance, Shohamy (1983), Sak (2008), Kang and Rubin (2012) and Bolaños, Cerdas and Ramirez (2014) investigated the role of raters on the reliability of second language oral performance tests. The four studies focused on estimating intra-rater consistency because it is the type of reliability in which less research has been done. The findings obtained through different coefficient correlation methods varied from low to high values of reliability. In addition, the authors agreed that intra-rater reliability can be improved by integrating rubrics and teacher training. The identified problem of having raters with low values of reliability is a matter of concern in the field of second language proficiency testing given the implications and consequences of the test itself.

In spite of the necessity of evaluating intra-rater reliability in language assessments, no studies addressing this issue have been found at the LBI program from the Universidad Tecnológica de Pereira. Therefore, this research project intends to analyze raters' self-consistency on the speaking section from the "Pruebas de Semaforización" and the factors that influence such consistency. This decision emerges from the aforementioned issue where speaking assessment may be prone to human error. Intra-rater reliability will, then, be estimated and evaluated through a coefficient correlation, whereas a retrospective verbal protocol with the raters will be conducted to explore the possible factors. The questions underlying this study are:

- What is the level of intra-rater reliability at the proficiency speaking test?
- What factors influence the level of intra-rater reliability?

3. Theoretical Framework

3.1 Conceptual Framework

The main concepts used in the development of this research project are addressed in this section with the aim of providing a clear understanding of them. The terms are classified in two main categories that are *language assessment* and *reliability*. Additionally, each main concept is divided into some minor concepts. For the first main concept, which is *language assessment*, *proficiency test* is addressed along with the explanation of the *Semaforización test* considering that it is the source of data to be analyzed in this research. Moreover, the constructs of *speaking assessment* and *rubrics* are explored . The second main concept, *reliability*, is further explained by addressing *rater reliability* and the *correlation coefficient* used to estimate it.

3.1.1 Language assessment.

Before stepping into the concept of language assessment, it is necessary to understand that assessment in general refers to the process of continuously identifying the extent to which students have acquired certain knowledge and topics from a teaching scenario, and are able to replicate it in different situations (Areiza, 2013). In language teaching and learning contexts, the concept of language assessment derives from the Applied Linguistics discipline, and it is defined by Chiedu and Omenogor (2014) as the practice of evaluating one person's proficiency in a given language. According to the authors, it is better when the assessments include specific language skills to be measured such as listening, reading, speaking or writing. Chapelle and Brindley (2010) complement the previous idea by explaining that this practice comprises not only evaluating, but also collecting and interpreting information about students' language ability. Furthermore, Tsushima (2015) asserts that in order to collect and evaluate learners' use of the language, different instruments or tools can be employed. Examples of these include tests, rubrics, portfolios, self- and peer-assessments, among others.

All assessment should lead teachers and professors to make decisions related to their teaching practices, the contents of the course, and the assessing instruments, all with the primary aim of helping students improve their language ability. Although this is a common goal, the practice of assessing can vary according to the purpose of its implementation. The most prevalent purposes described by Brown (2004) include: (1) checking students' knowledge about the constructs covered in a course, (2) diagnosing aspects of language that represent strengths and weaknesses for students, (3) placing students within different courses depending on their language proficiency level, and (4) measuring the overall language proficiency of a learner. There is a specific type of test for each of the aforementioned purposes, which are addressed as achievement tests, diagnostic tests, placement tests, and proficiency tests. Along this line, the test under which this study will be carried out can be classified as a proficiency test.

3.1.1.1 Proficiency test.

A test is an instrument of assessment that is administered under formal conditions in order to measure and evaluate students' general or specific knowledge and performance within a particular domain (Brown, 2004). When a test aims at gathering information about the general competence of an individual, it is regarded as a proficiency test. Likewise, in the field of second language assessment, proficiency tests seek to measure general language ability without being limited to the contents of a specific language course (Hughes, 2010). Brown (2004) conveys that these types of tests are usually designed and structured to be summative and norm-referenced, which means that students obtain a grade or score that will be analyzed in relation to the norm. In other words, each test-taker score will be compared to the scores of other participants and they will be ranked depending on their performance.

The key question that guides the widespread use of language proficiency tests is whether students are equipped with the necessary language competences to deal with future academic and professional demands such as having the abilities for studying in a university abroad, or getting a job (Chiedu & Omenogor, 2014). People requiring to demonstrate English language competence can take international English proficiency exams like the International English Language Testing System (IELTS), the Test of English as a Foreign Language (TOEFL) and the Cambridge Exams. Traditionally, these tests are designed so that reading comprehension, listening, writing and speaking are assessed separately and include diverse items and tasks.

3.1.1.1.1 Prueba de semaforización.

In the same line, the Semaforización test is a proficiency test carried out at the LBI program of the Universidad Tecnológica de Pereira with the specific purpose of tracking students' ability in the language through different stages of their professional development. This test emerges from the work done by the curriculum committee of the program which identified the necessity of collecting information about students' progress in their English language learning. Based on the subjects offered by the program and the hours of instruction, the members of the curriculum committee proposed two moments in which students had to achieve certain level of proficiency in the language. The first moment was at the end of the fourth semester and the expected level was B1.2, whereas B2 level was expected to be achieved at the end of the seventh semester. As a result, the "Acuerdo 15 del 5 de Mayo de 2015" was formulated, and the proficiency test was officially established within the LBI program, receiving the name of "Prueba de Semaforización" in Spanish.

In 2016, the Semaforización test was administered for the first time, and it was decided to implement it at the end of each semester so that learners could prepare themselves. From that time, students from fourth semester have received special training on the test during the Upper-intermediate English course since the fourth semester test is linked to this subject, meaning that the score they get from the test will represent a percentage of the course's final score. On the

other hand, the seventh semester test is linked to the Initiation to the Teaching Practicum course (Acuerdo 18, 2017), but students enrolled in this subject do not receive any kind of preparation.

The design, piloting, logistics and administration of the test is in charge of the LBI program and coordinated by the English-Spanish area. A set of professors are assigned to the designing of the test; for instance, two Upper-intermediate English course professors develop the test that students from fourth semester take. After the designing, the coordinators of the area revise and conduct a piloting session with a group of students who meet a number of special features. Once the piloting has been conducted, the test is modified if needed, and then, it is uploaded to the learning management system "Schoology". This online test is revised once again by the coordinators to check that everything is working well before the administration. The day of the test implementation, students are required to attend two sessions, one in the morning and the other in the afternoon. Students usually take the listening, reading and writing parts of the test during the morning session on a computer lab while some professors from the LBI program are monitoring the test administration. In the afternoon session, students take the speaking test. To carry out this part, students are randomly divided in groups of two or three people, and each group has around 15 minutes to perform the speaking tasks in front of the interlocutor and the rater.

The design of the test follows the structure of the FCE. It is worth highlighting that the FCE is targeted at a B2 level. So, in the case of the fourth semester test, the degree of complexity of tasks is simplified to meet the desired level. The Semaforización test measures language skills separately through different tasks. For listening, reading and language use, the following tasks are included: multiple choice, gap filling, multiple matching, open-ended questions, and word sentence transformation. As for speaking, the tasks proposed are answering personal questions, describing, comparing and contrasting pictures, and participating in a conversation. Writing has a special characteristic, here, students have to write according to the course linked to the test. Students from fourth semester write an argumentative essay following the structure taught throughout the English course. On the contrary, seventh semester students are asked to write a reflective essay in which they talk about their experience in the Teaching Practicum. Scores from each of the skills are calculated either by the platform of by human raters. Schoology automatically generates the results for the listening and reading parts of the test, whereas professors have to score speaking and writing based on a rubric due to their subjective nature. Lastly, the total score of the test is systematized on Excel by adding the final score from each of the language skills.

3.1.1.3.1 Grading system.

In education, a grading system is used to assess the performance of students. The two most common types of grading systems are norm-referenced and criterion-referenced. Aviles (2001) explains that norm-referenced grading aims at assessing students in relationship to one another. In this case, the "learner performances are compared against each other" (Luoma, 2004,

p. 81) in order to assign a grade. As for criterion-referenced grading, its purpose is to compare students performances to an established standard or criterion instead of comparing them to other students (Aviles, 2001). In other words, an objective is set and students are assessed based on their level of accomplishment of the objective. Wiggins (1994) points out that objectives and criteria are strong points of reference. Therefore, rubrics are an effective way of implementing criterion-referenced grading given the fact that they include the criteria to be assessed. In the Semaforización test, the grading system implemented for measuring students' language proficiency is criterion-referenced. The grading system used for the measurement is numerical, from 0.0 to 5.0, being 0.0 the lowest score, 3.5 the passing score, and 5.0 the highest score.

3.1.1.2 Speaking assessment.

Following the definition of assessment addressed before, speaking assessment could be considered as the process of collecting, evaluating and analyzing information about language learners' oral performance. Nevertheless, this process represents a challenge at the moment of giving a reliable score since learners' speaking ability is, most of the times, being judged in real time by an assessor, who has to pay careful attention to a variety of features. Nunan (as cited in Rahmawati, 2014) argues that speaking is a complex skill that demands language users to be competent in pronunciation, grammar, fluency and vocabulary; thus, such features are usually included in the dimensions of speaking assessment's criteria.

As claim by Qian (2009), there are three types of formats for oral assessments: indirect, direct and semi-direct testing. For indirect testing, speaking ability is assessed in a way that learners are not required to speak, instead, they have to show competence through different items. For example, writing the phonetic transcription of words to assess pronunciation. However, the author states that this type of tests were used during initial stages of language testing but, nowadays, they have become obsolete for speaking assessment since its results cannot be considered reliable and valid given that the test-takers are not even using spoken language. Direct testing, on the other hand, is developed in a face-to-face interaction that involves learners in performing different tasks which allow them to provide evidence of their true language proficiency level. This format is more used and accepted than other types of testing because its purpose of measuring the exact language skill that is supposed to measure makes it more reliable and valid. The last type of testing format proposed by Qian (2009) is semi-direct testing, and it refers to the kind of assessment where learners interact with different types of aural or visual input, and, then, are required to produce an oral response that is taped and later scored by a rater.

The ideal format for a speaking assessment is direct testing inasmuch as it provides enough evidence of spoken language and allows interaction between the participants. As can be noted in the definition provided by Qian (2009), this test is underlined by tasks, which are "activities that involve individuals in using language for the purpose of achieving a particular goal or objective in a particular situation" (Bachman and Palmer, 1996, p. 44), and give test-

takers the opportunity to use language in a more authentic and natural way. Considering that tasks for speaking assessment are designed for students to provide open-ended responses, test designers have to carefully set the rubrics for these tasks in order to be more reliable and ensure that the assessment measures the right features of speaking and task conditions.

3.1.1.3 Rubrics.

In general words, a rubric is an assessment instrument that contains the scoring criteria to assess the performance of a learner in a specific skill (Mertler, 2001; Mansoor and Grant, 2002; Pineda, 2013). Rubrics permit the scoring procedures to be more objective since they state, from the very beginning, specific descriptors of the language features to be measured. Brown (2011) in agreement with Picon (2013) observes that there are two types of rubrics: holistic and analytic. The former presents, in a general scale, the description for the different levels of performance expected by the rater. Therefore, the features of language are not separated but they are implicitly stated in the descriptions. On the contrary, the latter presents separately the different features of language that will be assessed along with clear descriptors and a scoring scale for each of them. Analytical scales, within classroom scenarios, are much more advantageous for learners since teachers can give detailed feedback to help them improve their performance (Gipps as cited in Picon, 2013).

In order to assess speaking performance, there are some features that need to be considered by the raters. In this sense, Torres (1997) proposes to observe the following components: pronunciation, vocabulary and grammar to test accuracy, and mechanical skills (pauses, length, speed), language use (coherence) and judgment skills (create and develop thoughts) to assess fluency. It has been found that when these types of features are assessed, some of them are considered as “concrete” and others as “less discrete” (Kang & Rubin, 2012, p.56). Kang and Rubin (2012) explain that grammar, and vocabulary mistakes are easier to identify and easier to check; thus, they are considered as concrete speech features. On the other hand, pronunciation is less discrete since it is more abstract and difficult to check. Although the researchers do not mention fluency, based on their claims, it can be concluded that fluency is a less discrete aspect.

The rubrics, designed to measure students’ speaking skill during the Semaforización test, were divided into four criteria that represent some features of oral production, these are discourse management, pronunciation, interactive communication, and grammar and vocabulary. Discourse management refers to the students’ fluency, organization of ideas, use of cohesive devices and discourse markers. Pronunciation stands for intelligibility, intonation, sentence and word stress. Interactive communication is understood as the appropriate participation in a conversation by responding, contributing, maintaining, and negotiating ideas. Finally, grammar and vocabulary deals with the use of a range of simple and complex grammatical forms, and appropriate vocabulary. To assess the quality of performance for each criterion, raters are also

provided with a rating scale which values are 0.5, 0.75 and 1.25. Each of these values contains a descriptor providing indicators of performance (see Appendix A).

3.1.2 Reliability.

One of the most critical qualities of test scores is reliability, which has been referred to as the "consistency of test results across different conditions" (Giraldo, 2018, p.29). When scores from a test are reliable, they should remain the same regardless of the time, the assessors and the number of administrations (Bachman, 1990). However, time intervals between administrations should never exceed six months since test scores are likely to be less consistent as time span becomes longer (Anastasi as cited in Weir, 2005). From the foregoing, it can be assumed that raters can experience changes in their scoring capacity, or that learners' performance can improve or decrease over time.

In academic settings, tests' scores may determine many decisions made by teachers, professors, and learners; therefore, reliability in this kind of assessments is crucial because it represents how trustworthy the score is, and the extent to which the score indicates learners' true language ability. Luoma (2004) explains that "unreliable scores (...) can lead to wrong placements, unjustified promotions, or undeservedly low grades on report cards" (p.176). For these reasons, academic communities should make efforts to constantly evaluate the reliability of their assessments, instruments and raters. Nevertheless, it is important to note that rater-reliability is mainly affected when the scoring procedures are subjective, such as in oral or written tests.

3.1.2.1 Rater reliability.

Rater reliability, in simple words, refers to the ability of the rater to yield consistent scores. In the field of language assessment, raters are a necessity for scoring samples of oral and written production; nonetheless, they also represent a problem since their scoring is essentially based on subjective judgements. During the 50's and 60's, the community of language assessment sought to achieve the highest possible level of test reliability. As a consequence, raters' participation on scoring procedures posed a threat to the pursuit of their goal, leading to the indirect assessment of speaking and writing skills (McNamara, 2000). Efforts to increase rater reliability of direct testing in productive skills began to gain ground later, in the 80's (Weir, 2005). Such efforts resulted in the development of a variety of statistical methods that, nowadays, help to estimate not only raters reliability but also task difficulty and rater's severity. These methods will be addressed later in this section, for now, the two types of rater-reliability will be explored.

3.1.2.1.1 Inter-rater reliability.

In the first place, reliability is concerned with inter-rater consistency, or the level of score agreement that two or more raters have on a given performance (Shohamy, 1983). When a test is

administered, the results obtained from this should be similar across the different raters who were in charge of scoring it. If this consistency is present, test-takers' concerns about who is going to score a test should not exist because their score will be the same no matter the rater (Fulcher, 2003).

3.1.2.1.2 Intra-rater reliability.

Intra-rater reliability refers to the level of consistency that raters have with themselves after giving a score (Luoma, 2004). Different from inter-rater reliability, this reliability is concerned with the consistency or variation of scores given by one single rater (Bolaños, Cerdas, & Ramirez, 2014). Shohamy (1983) and Fulcher (2003) state that intra-rater reliability is the extent to which a rater is able to consistently provide the same test score to the same student across different times and conditions. In order to estimate intra-rater reliability two set of scores assigned by the same rater for the same individuals have to be collected and, then, calculated through a correlation coefficient method (Brown, 2011; Sak, 2008). The result from the correlation coefficient is a number that estimates the level of consistency of the rater in question. Brown (2004) explains that low levels of inter- and intra-rater reliability can be the result of internal and external factors such as "lack of adherence to scoring criteria, inexperience, inattention, preconceived biases" (p.28), tiredness and obscure rubrics.

3.1.2.2 Factors affecting rater reliability.

The rater's behavior in scoring procedures can be influenced to some extent by different factors. Two of the most prevalent include raters' experience and fatigue. Experience is defined by Lim (2011) as the number of times that a rater has had the opportunity to take part in scoring procedures as well as the amount of ratings done. There is a notion that rater's reliability is directly related to the rater's experience. Hence, the more experienced a rater is, the more reliable the scores will be. Bolaños et al. (2014) suggest that it occurs due to the fact that experienced raters observe and analyze speaking performance in a holistic way, being able to assess different aspects of the presentation at the same time.

On the other hand, fatigue is understood as mental and physical tiredness, which results in the decline of performance (Ling, Mollaun, & Xi, 2014). In the case of raters, they have higher possibilities to suffer from fatigue given the fact that they are required to assess large amounts of written and oral productions during long periods of time. Taking this into account, the order in which the language samplings are arranged may lead to mistakes in scoring, affect the severity of the raters and compromise the consistency of the grades since tiredness usually occurs at the end of the rating shift. In order to find out if these factors are adversely affecting the reliability of raters, it is necessary to constantly estimate rater's consistency through a correlation coefficient method.

3.1.2.3 Correlation coefficient.

Intra-rater reliability can be estimated through a calculation of the correlation coefficient, which is a statistical method used to measure the relationship between two variables (Luoma, 2004). For this purpose, the variables, which in this case are scores from a test administration along with scores resulting from a second rating session, are collected. Both sets of scores must come from the same test, test-takers and raters. The correlation coefficient ranges from -1 to +1, being +1 the representation of a perfect agreement between the two variables (Fulcher, 2003). This means that the rater, in both rating sessions, provided exactly the same scores to each of the students. However, Luoma (2004) argues that achieving a value of +1 is almost impossible due to the fact that human errors are constantly present in the scoring process. Correlation values above .8 show a good level of self-agreement, whereas values below .5 indicate a low degree of self-agreement (Cronbach as cited in Luoma, 2004). If the correlation yields a value close to 0, no agreement between the two scores is found. There are some cases in which the variables are in complete opposition to each other, meaning that whereas the score in one of the ratings is high, the score in the other rating is low. In such cases, the correlation is represented with a -1 value (Luoma, 2004).

3.2 Literature Review

Over the past few decades, the practice of assessing speaking has gained relevance in the field of second language (L2) assessment. With the arrival of the Communicative Language approach, the focus of language teaching shifted from the mastery of grammatical structures to the communicative uses of the language which "created a demand for oral proficiency in foreign languages" (Richards, 2001, p. 7). Before that shift, speaking tests were avoided because of the reliability issues they entailed; instead, language teachers mainly focused on the assessment of phonetic transcriptions where learners were required to write rather than to speak (Fulcher, 2003). Although, nowadays, the speaking competence is assessed in the majority of second language courses and programs, it still represents a challenge for language assessment due to its scoring procedures.

Reliability of oral assessments remains a concern as spoken language requires human raters to make fast judgments and decisions about the language proficiency of learners. This special condition makes scoring speaking a subjective process that can cause raters' inconsistencies, which negatively affect the reliability of test results and lead to wrong judgments about the learner's ability; besides, "one cannot be confident of the validity of oral assessments (…) when a rater cannot agree with her own rating from occasion to occasion" (Kang & Rubin, 2012, p. 44). On this basis, it is important to determine the intra-rater reliability on oral proficiency tests to enhance the teaching and learning processes, and to ensure students' accurate scores. Therefore, this literature review introduces and describes four sources that examined intra-rater reliability from statistical and qualitative points of view.

One of the first studies analyzing rater reliability on L2 oral assessments is the one carried out by Shohamy (1983), where inter- and intra-rater agreement on the Oral Interview (OI) were examined. The OI is a speaking proficiency test widely used in the US by different government and educational institutions. This test assesses students' ability to use the target language through a casual interview that usually lasts from 15 to 30 minutes, and that measures proficiency based on grammar, pronunciation, fluency and vocabulary. Shohamy's study took place at the University of Minnesota with 106 learners, who had different Hebrew proficiency levels, and with three Hebrew teachers, who were highly proficient in the language. The interviews conducted with the 106 students were recorded and later scored independently by the three raters.

In order to assess students' oral skills, a pre-made rating scale was adapted focusing on the specific characteristics of the target language in regard to the same four features of the interview: grammar, pronunciation, fluency and vocabulary. Furthermore, raters received extensive training on how to use the speaking rating scale. They could practice using the rubric with sample tapes, and then comparing and discussing scores with the other teachers. After the training session, raters were asked to rate the 106 taped interviews using the OI Hebrew scale; then, they had a second rating four weeks later with a quarter of randomly selected recordings. For the purpose of measuring inter-rater reliability of the OI, Cronbach alpha was computed for

each of the speaking components while for intra-rater reliability, Pearson correlations were calculated. Results from Cronbach formula ranged from .94 on pronunciation to .99 on the total ratings of the test. Similarly, Pearson correlations between the two ratings of the speaking test by each rater were high for grammar (.94 and above) and vocabulary (.93 and above); in the case of fluency, the correlations were slightly low (.88 and above), and even lower for pronunciation (.63 and above). Nonetheless, the total score for intra-rater reliability ranged from .95 to .99.

The results yielded by this study indicate that there is a high level of inter- and intra-rater reliability on the OI proficiency test. Such findings are encouraging since it is demonstrated that reliable information can be gathered in spite of the subjective nature of oral assessments. Shohamy (1983) concluded that, in fact, these findings "provide strong evidence to further support and encourage the use of qualitative measuring instruments (...) in testing foreign/second languages" (p. 222). Likewise, the researcher suggests that the high levels of reliability obtained in the test could be attributed to factors such as the experience of the raters, who participated in the study, the training sessions that were conducted, and the rubrics used throughout the development of the research. Thus, it is important to consider and investigate the influence of factors such as the aforementioned to increase the levels of reliability in language assessments.

Aiming to address the same issue of intra-raters reliability in judgments of learners' oral performance, Kang and Rubin (2012) carried out a research study with tape samples taken from an English standardized test. As the researchers considered that long time span between rating sessions might cause a maturation effect on raters' performance, and therefore, affect the true intra-rater reliability of assessments, they decided to evaluate self-rater consistency at a single rating session. In this study, the raters were 187 undergraduate students from a US University, and who were enrolled in a speech communication course. Participants were asked to rate 26 sample responses taped from the sixth TOEFL iBT task, which included male voices samples with different levels of language proficiency. Considering that one of the purposes of the study was to analyze the characteristics of novice raters, the participants received minimal training on how to use the rating scale. Then, they were required to score the speech samples online, and within a short period of time (about an hour).

Drawing from the conclusions by Shohamy's (1983) research, Kang and Rubin also attempted to examine some of the factors affecting rater behavior on oral proficiency assessments. The researchers claimed that one of the external factors influencing rating performance is order effect. This phenomenon occurs when a single rater scores the same language sample twice, but with different levels of severity depending on its order of appearance. For that reason, and with the aim of evaluating intra-rater reliability, they chose two of the 26 TOEFL-iBT speech samples and placed them in the rating queue twice. These two samples were placed within the first 10 slots and repeated within the last 20 slots. As a means to score the speech samples, raters were given two assessment criteria, one designed for assessing comprehensibility, and the other one for measuring oral proficiency. Similar to the speech rating

scale used by Shohamy (1983), this speaking analytic scale included, among others, pronunciation, grammatical accuracy, vocabulary and fluency features.

Rater-reliability was estimated for each of the two assessment criteria using alternative techniques such as Intraclass Correlation Coefficients (ICC) and Many-facet Rasch Measurement (MFRM). The researchers conducted comparisons of the initial and subsequent scores within each rater for the two samples, and the findings confirmed that there exists order effect among raters, with a tendency of leniency towards later ratings. Additionally, a post hoc analysis demonstrated that the mean score for the second half of the ratings was significantly higher than those of the first half for both the comprehensibility, and the oral proficiency ratings. More importantly, the coefficients of intra-rater reliability for each dimension of the oral proficiency scale showed that the consistency of pronunciation was significantly lower than the intra-rater reliabilities for vocabulary, grammatical accuracy, and fluency. Overall, results from the coefficient calculations indicated that the reliability for each rater ranged from .68 and above for comprehensibility and from .72 and above for oral proficiency ratings.

To sum up, this research study proved low to moderate intra-rater reliability even though the two ratings took place within the same session, and therefore, the raters did not undergo any type of maturation effect. In addition, the drift in rater's severity found in this study is associated by the researchers to a fatigue effect that causes raters to be less attentive to the rating process; however, they called for further research on this matter. The findings related to the dimensions of the oral proficiency scale, once again, aligned with Shohamy's (1983) earlier findings on the same topic. Kang and Rubin observed that the low reliability coefficients for pronunciation are attributable to the less "discrete and enumerable" (p. 56) nature of this feature of speaking. On the other hand, more "concrete and enumerated" (p. 56) aspects of speech such as grammatical accuracy and vocabulary are more prone to yield higher reliability coefficients. All of the above-mentioned insights lead to the conclusion that rater training is a key element for achieving intra-rater agreement, and that special attention is needed in the discrete dimensions of oral proficiency, more specifically pronunciation.

The findings by Kang and Rubin (2012) are similar to the ones presented by Sak (2008) whose study aimed at investigating the validity and reliability of a speaking exam conducted at the TBO University of Economics and Technology (TOBB ETU) located in Turkey. In this University, newly admitted students have to take an English exam in order to prove proficiency in the language and those who pass the exam have to take the TOEFL-ITP. The students who do not manage to pass the first exam, or the TOEFL-ITP are required to take classes at the Department of Foreign Languages, which has three different programs for students according to their English levels. One of them is the C program that was designed for students who passed the first exam but who had difficulties with the TOEFL-ITP; these programs have a strong emphasis on speaking and writing. However, the study focuses on the speaking assessment since students will take this kind of exams during different semesters at the university, and are also required to get a score major than 94 at the TOEFL IBT as requirement for graduation.

The participants of the study were 70 students from the C programme that enrolled in the University during the 2007 - 2008 academic year. In addition, six English instructors from the University, who rated a previous speaking exam, volunteered to participate in the study. Video recordings of the speaking exam where given to the six volunteers who were asked to give a new score to the students they had already graded in order to test intra-rater reliability. The scores from both the exam administration and the video recording were then analyzed by calculating the correlation coefficients.

The results of the calculation showed that four out of six raters were reliable since their ratings were 0.776, 0.933, 0.727, and 0.796. However, the correlation between the first and the second scores of two raters were 0.582 and 0.560 respectively, which represent low agreement. To conclude, the researcher states that although the ratings of most of the instructors are "statistically significant, it can be claimed that they are not so high when the minimum desirable level for oral tests (.70) is considered." (Sak, 2008, p. 82). The author suggested that intra-rater reliability could had been adversely affected by the different conditions of the two rating sessions considering that during the first session students were present at the moment of grading and in the second session, raters had to base the scores on video-recordings of the students. However, the author also brought up the study by Shohamy (1983) to point out that although speaking performance of the learners in Shohamy's study was rerated based on recordings, the intra-rater reliability was high, leaving the door open for new possible reasons that could have interfered with intra-rater reliability.

Another study that seeks to evaluate intra-rater reliability is the research done by Bolaños et al. (2014); nevertheless, this study also aims to investigate how experience as well as gender and rubric development affect rater reliability. In order to do so, 20 professors were classified as experienced and novice. In this study, professors teaching for less than 5 years were regarded as novice, and the others as experienced. The first step for the implementation of this study was to interview the participants about their teaching experience and their rubric development; in addition, the professors were given recordings of some students' oral presentation which they were required to grade. The second stage was performed one month after the first, this time, the participants had to score the same recording once more.

The scores collected from both rating sessions were organized in a chart containing the grades, the mean, the standard deviation and the range; information that would help to obtain the Pearson product-moment correlation coefficient, whose purpose is to determine raters reliability. In this calculation, the results can range from -1 up to +1; however, only the results that range from 0.80 to +1 are considered as representing high consistency among the grades, whereas 0.40 or less are indicators of poor consistency. In addition to calculating the Pearson product-moment correlation coefficient, the researchers also compared all the professors coefficients with the aim of analyzing whether or not experience had played an important role in their reliability.

The authors concluded that experience is an important variable in boosting raters reliability. This statement was supported by the results of the study that show that 63% of the

participants that had obtained a high correlation coefficient were also experienced professors. Moreover, it was discovered after the implementation of this study that experienced professors tend to be more severe and strict than novice professors; probably because they "are more likely to pay attention to a broader set of aspects when grading students' performance" (Bolaños et al., 2014, p. 306) such as body language, fluency, and communication. Another factor that gave professors the opportunity to be more reliable was the designing of their own rubrics for assessing speaking. For instance, 67% of the raters who developed their criteria obtained a good correlation coefficient level, whereas only 38% of the participants who used a pre-made rubric were classified as reliable. The researchers suggested that the reliability of the raters who used another rubric is the result of training before implementing the rubric in the assessment. Finally, the results demonstrated that although gender bias exists, it did not affect students' scores considering that male and females raters gave similar grades.

The analysis of these studies helped to shape the development of the present research taking into account the significant findings on the influence of factors such as rubrics and experience on high levels of rater's reliability. The interest of analyzing the role of rubrics derives from the study conducted by Shohamy (1983) who concluded that using analytic rubrics may contribute to high levels of rater consistency on oral assessments. On the other hand, Kang and Rubin (2012) and Bolaños et al. (2014) pointed out that raters' experience is a characteristic that can affect raters reliability; thus, this research also intends to examine if consistency is influenced by experience. In general, one of the purposes of this study is to investigate the factors that may be positively or negatively influencing reliability of the speaking part of the Semaforización test administered at the Universidad Tecnológica de Pereira. The review of the literature also shed light on the appropriate procedures to estimate the level of intra-rater reliability. The studies suggested a series of correlations, from which Pearson-product correlation coefficient was the most mentioned and used (Bolaños et al., 2014; Sak, 2008; Shohamy; 1983). Bearing this in mind, this study attempts to contribute to the limited research on intra-rater reliability on oral proficiency tests that exists in the Colombian context.

3.2.1 Historical background of Semaforización test scores.

Considering that the scores from the Semaforización test have not been thoroughly explored, this research project aimed at collecting, organizing, and tabulating the students' scores obtained from the very first test implementation in the LBI program, which started in the second half of 2016 with students from seventh semester and continued through the years 2017 and 2018 with both fourth and seventh semester students. All the scores were carefully systematized for the purpose of carrying out a statistical analysis of the mean, median, mode and standard deviation of each skill. To guarantee accurate results, the students who had a zero (0.0) in one of the four skills assessed were left apart from the analysis since it was considered that a zero did not represent the student's true language ability.

In the following bar charts, the means of the four skills from each test implementation with fourth semester students are displayed. Axis X represents the language skills assessed in the test while axis Y indicates the average score from zero (0.0) to five (5.0) obtained in each skill.

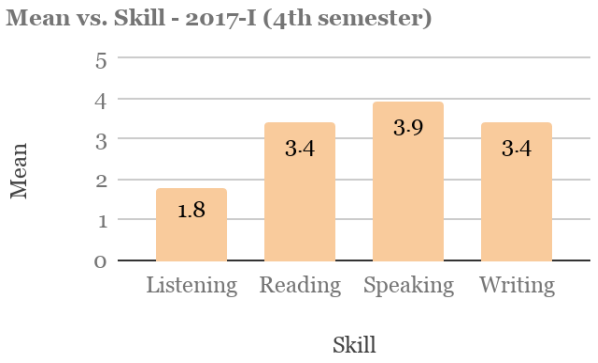

Chart 1. Comparison between language skills' means from 4th semester test (2017-I)

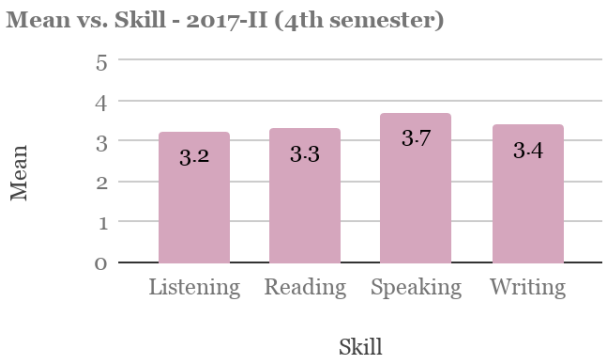

Chart 2. Comparison between language skills' means from 4th semester test (2017-II)

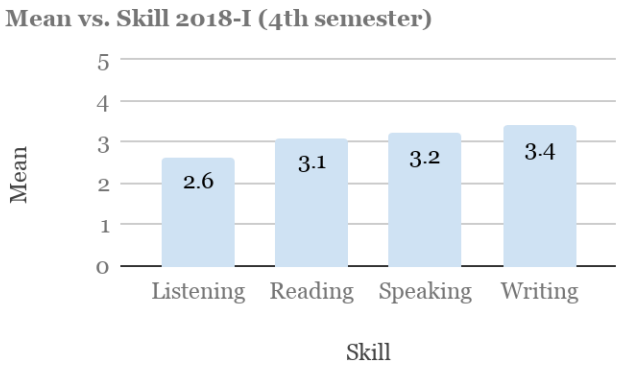

Chart 3. Comparison between language skills' means from 4th semester test (2018-I)

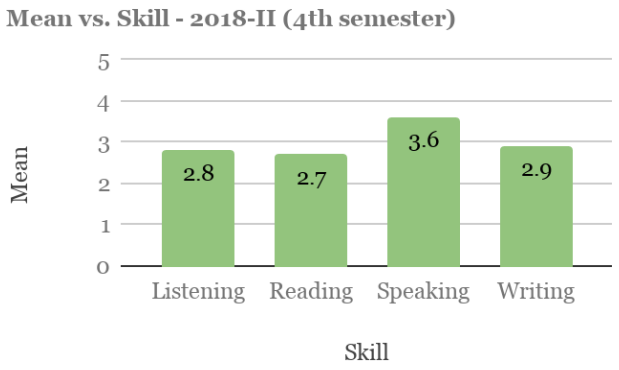

Chart 4. Comparison between language skills' means from 4th semester test (2018-II)

The graphics show that speaking was the skill with the best results except for one implementation (Chart 3) in which it was surpassed by writing. On the other hand, listening and reading obtained the lowest scores from all. Finally, it was found that all of the means were below 4.0.

The charts presented below show the mean scores of the four skills from each test implemented with seventh semester students. Similar to the previous graphs, axis X represents the language skills assessed in the test while Y indicates the average score from zero (0.0) to five (5.0).

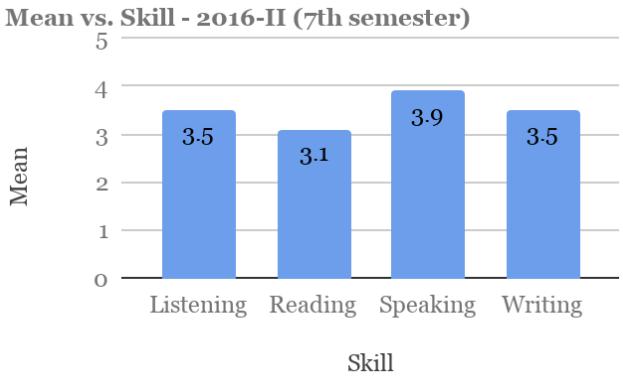

Chart 5. Comparison between language skills' means from 7th semester test (2016-II)

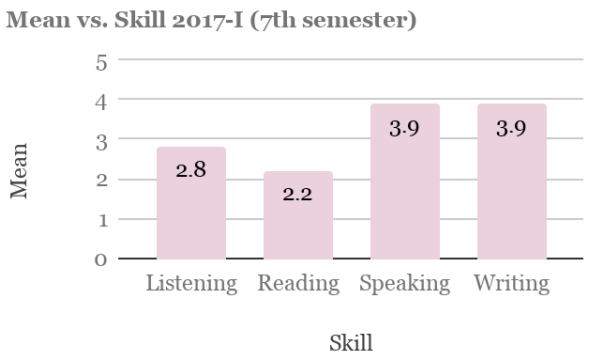

Chart 6. Comparison between language skills' means from 7th semester test (2017-I)

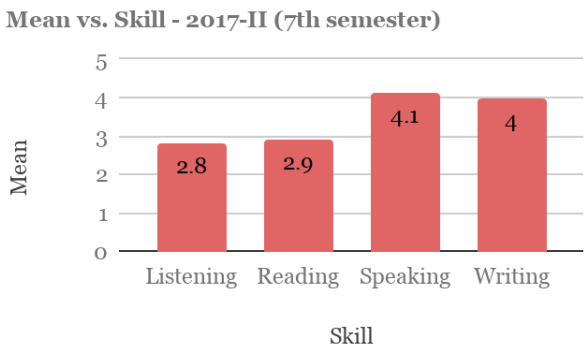

Chart 7. Comparison between language skills' means from 7th semester test (2017-II)

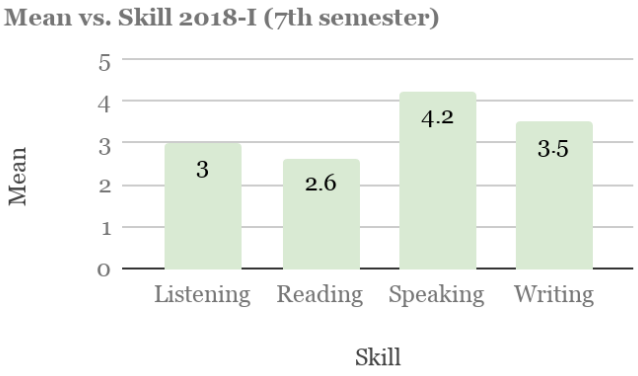

Chart 8. Comparison between language skills' means from 7th semester test (2018-I)

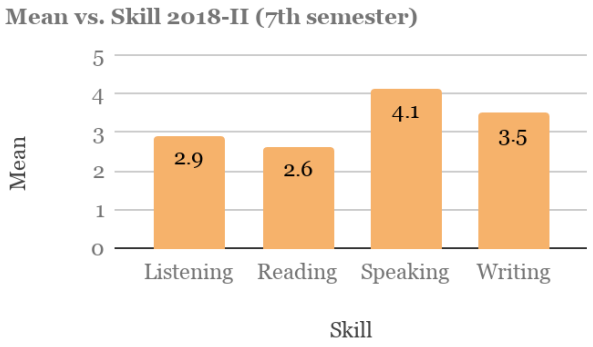

Chart 9. Comparison between language skills' means from 7th semester test (2018-II)

The data tabulation indicates that, in seventh semester, listening and reading were the skills with the lowest scores, whereas speaking was the skill with the best performance followed by writing. These results align to the ones obtained from fourth semester. However, the mean scores from seventh semester students regarding speaking showed marks above 4.

As it can be observed in the charts from both fourth and seventh semester, there is a tendency towards high speaking scores. Additionally, it was found that students performed better in terms of productive skills, whereas they showed lower ability regarding the receptive skills. It has been argued that the exposure to receptive skills may lead to the development of the productive ones. Sreena and Ilankumaran (2018) highlight that "without listening, no speaking is possible, without reading no writing is possible. So, the four skills go in pairs." (p. 670). Likewise, Segura (2013) refers to the correlation between productive and receptive skills that commonly surrounds ESL learners claiming that "in the same way that a good writer is a good reader, a good speaker is also a good listener" (p. 62). The fact that speaking in the Semaforización test has showed higher results than its opposite skill (listening) is a phenomenon that should be explored in detail in order to guarantee that the scores students obtain in the speaking part are accurate and reliable. In this sense, considering the importance of establishing whether such grades are reliable or not, this research will investigate the scores' reliability from the analysis of self-agreement of the raters.

4. Methodology

4.1 Type of Research

The nature of this research project is mixed-method since it involved the collection and analysis of both quantitative and qualitative data. In recent years, the use of this type of research has increased in the educational field, and Fraenkel, Wallen and Hyun (2012) argue that it helps researchers to understand the reasons behind the relationship between variables. For the collection and analysis of quantitative data, document analysis and statistical procedures were used. On the other hand, a retrospective verbal protocol was conducted in order to collect qualitative data, which was then analyzed through the content analysis technique. This data allowed the researchers to explore the possible factors influencing raters' reliability.

4.2 Type of study

This study followed an exploratory design, which purpose is to investigate a problem or phenomenon that has not been extensively studied, and to establish variables that are significant for subsequent analysis of the relationships among them (Fraenkel et al., 2012). The present research sought to explore the phenomenon identified after the analysis of scores from the five implementations of the Semaforización test, which is represented by the high scores obtained on the speaking part of the test. In order to identify the possible reasons behind this event, intra-rater reliability was estimated and the factors that could be affecting this quality of assessment were identified.

4.3 Context

This research project took place at the Universidad Tecnológica de Pereira (UTP), which is a state University located in the urban area of Pereira, the capital city of Risaralda. Based on the statistics of the University, the UTP began operations in 1961 with only one faculty; currently, the university offers 106 academic programs, including the Licenciatura en Bilingüismo con Énfasis en Inglés (LBI) program. The LBI was created in 2004 under the name of "Licenciatura en la Enseñanza de la Lengua Inglesa" with the purpose of contributing to the enhancement of English language in the region and Colombia. However, in 2017, the program was reformed in order to meet the requirements of the accreditation process, and as a result, it received the name of "Licenciatura en Bilingüismo con énfasis en inglés". At the present time, the LBI is an undergraduate program that lasts 10 semesters and has 605 students and 39 professors.

The curriculum of the program offers different courses that are classified within one of these areas: English-Spanish, Pedagogical-Linguistic, Research, Technological, and Intercultural. During the first three years, students are immersed in the development of the second language through ten English courses that belong to the English-Spanish area. One of them is the Upper Intermediate English course which is offered in the fourth semester of the program. In this

course, learners develop their linguistic, sociolinguistic and pragmatic competences through different tasks where the four English language skills are considered. At the end of the semester, students enrolled in this course are required to take a proficiency test in order to see if they have reached a minimum level of B1.2 according to the CEFR.

4.4 Setting

The Semaforización test is a proficiency test designed and administered by the LBI program, which seeks to assess whether students have reached the expected proficiency level according to the semester they are enrolled in. Students from the program have to take this test two times. The first moment is when students are taking the Upper Intermediate English course in fourth semester; here, learners have to demonstrate a B1.2 level of English competencies. The second test, designed to prove B2 language proficiency, takes place at the end of the seventh semester with students from the Initiation to the Teaching Practicum course. The Semaforización test is divided into four sections, which are related to the four language skills: Listening, Reading, Speaking, and Writing; the "Language Use" component is also assessed together with the Reading section. For the development of this research project, the speaking test scores from fourth semester were analyzed.

4.5 Participants

The participants involved in this research were, in the first place, two professors in charge of scoring the speaking part of the Semaforización test that was taken by students from fourth semester. And secondly, 10 students whose spoken performance was assessed by the two professors. The method used to select the sample for both students and raters was purposive sampling, which according to Fraenkel et al. (2012) consists of selecting the kind of sample that the researchers believe will satisfy the requirements of the study.

4.5.1 Raters.

To carry out the speaking part of the Semaforización test, two professors from the LBI program are required to play the roles of interlocutor and rater. The interlocutor is in charge of interacting with the test-takers and guiding them through the completion of the test. Such person has to introduce and explain the tasks, and ask the questions established for each of these tasks. On the other hand, the person whose responsibility is to assign a score to the candidates of the test is the rater. Although the interlocutor does not assess the students directly, there are some cases in which s/he supports the rater in the decision-making.

Since this research sought to estimate the intra-rater reliability of the speaking section of the Semaforización test, the professors who were assigned as assessors of that section were the main participants. In total, there were four raters in charge of assessing fourth semester students' speaking skill. Those raters were professors working at the LBI program who had previously performed the role of evaluators and interlocutors on the speaking test. Two out of four

professors were purposively selected to participate in the study based on their experience as raters of the test.

4.5.2 Students.

The present study was conducted with students from fourth semester of the LBI who were taking the Upper Intermediate English course. Their mother tongue is Spanish, and they were studying English as a second language (L2) with the aim of becoming English language teachers in the future. As they were working on achieving a B1.2 English level, the lessons from the aforementioned course focused on the development and improvement of their linguistic competences. The LBI program offered weekly training sessions where students interacted with the different types of items and tasks that constitute each of the skills sections from the Semaforización test. The sampling group consisted of 10 students (5 students per rater) who represented the 11% of the total population, which was 93.

4.6 Researchers' Roles

In order to establish the roles performed by the researchers, it is fundamental to understand them from the two different methodologies. Fraenkel et al. (2012) highlight that "the ideal researcher role in quantitative research is that of a detached observer, whereas qualitative researchers tend to become immersed in the situations in which they do their research" (p. 11). In this sense, the quantitative role was present at the moment of collecting and analyzing the numerical data (scores from the Semaforización test), which implied a non-participant observation. As for the qualitative role, the researchers were involved in the situation while they conducted the verbal protocol. Therefore, the researchers played the role of interviewers which required them to actively listen to the information that was being provided by the raters, and to maintain the conversation focused on the objectives of the study (Dickinger, 2007).

4.7 Data Collection Methods

Bearing in mind that this study follows a mixed method approach, two different data collection instruments were employed. The major source of information came from quantitative data which was gathered through a document analysis of fourth semester students' test scores. On the other side, qualitative data was collected from the application of a verbal protocol.

4.7.1 Document Analysis.

According to Fraenkel et al. (2012), document analysis is considered as a method of data collection in which documents are reviewed in order to extract information from them. Documents include any kind of "written, visual and physical material relevant to the study at hand" (Merriam, 1998, p. 112); examples of these include: newspapers, lesson plans, portfolios, charts, and attendance lists.

4.7.1.1 Numerical records.

The sub-category of documents in which this study was focused is “numerical records”. Fraenkel et al. (2012) point out that this category involves numerical data such as test scores in printed form. However, although the scores analyzed were not printed, they were stored in a digital format.

The scores considered for the development of the study were the ones taken from the speaking section of the Semaforización test administered at the end of the fourth semester. For this reason, the speaking results of 10 students (5 students per rater) enrolled in the Upper-Intermediate English course were collected right after the raters observed the performance and assigned a total of points based on the analytic rubric they were given. Additionally, speaking scores from a second rating process were gathered with the specific purpose of correlating them with the results from the test implementation. Such second rating was carried out by the raters who were then asked to listen to the recordings of the five students they assessed during the implementation of the test.

4.7.2. Retrospective verbal report.

The retrospective verbal report is a data collection technique in which participants have to orally report their thoughts and make assumptions about some actions taken during the performance of the task (Afflerbach & Johnston, 1984). This type of protocol usually occurs right after the subjects have finished doing the task. In the present study, this technique was used as a means of systematizing the rater’s perceptions about the consistency of scores, and the factors that could be affecting the level of intra-rater reliability on the speaking test.

In conducting the verbal protocol, both raters were asked to attend a meeting with the researchers in which they had to do the second rating, answer a series of questions and compare the rubrics from the first and second scoring procedure. The first rater attended the meeting two weeks after the Semaforización test implementation, whereas the second rater had the meeting five weeks after the test. The meeting started with a review of the research project and the events that took place the day of the test administration, followed by the reading of the protocol procedures (see Appendix B). After that, the raters were asked to listen to the recordings of five of the students they assessed the day of the Semaforización test and then, to give them a score once again using the same rubrics in a blank printed format. Once the audios and earphones were tested, the raters proceeded to assess the students. As soon as they finished the scoring process, the researchers asked 5 questions regarding the raters’ background and experience with the test. In this part, both raters had the opportunity to expand on their answers. The following step was the comparison between the rubrics from the test implementation and the ones obtained after the second rating performed at the beginning of the meeting. In order to do so, the researchers gave the raters two highlighters and showed them the two rubrics from each of the five students; they had to highlight the differences and similarities with different colors and provide their insights

about it. Finally, the raters were asked some questions about different factors that could have influenced their decisions when giving a grade.

4.8 Data Analysis

Due to the mixed method nature of the study, the data obtained was analyzed following the specifications of instruments found both in qualitative and quantitative approaches. For the analysis of test scores and rubrics, statistical methods and correlations were used, whereas content analysis was applied to the verbal protocol.

4.8.1 Statistics.

As explained before, statistics is a key feature in the analysis of quantitative data (Dörnyei, 2007). This study is no exception since the data collected from test scores was statistically analyzed. However, before analyzing the data, it was necessary to prepare it. Dörnyei (2007) suggests to do a cleaning of the data, which consists on spotting and correcting mistakes that are the result of human error on the data collection stage. In the present study, test scores were revised by the researchers so that the correlation yielded accurate values. For the purpose of selecting the most suitable correlation method to analyze the data, some professional help was requested. However, the results obtained were not statistically validated by an expert on the field.

4.8.1.1 Test scores correlation.

In the first instance, test scores were analyzed through a correlation method. Dörnyei (2007) points out that this statistical procedure is used in order to establish “the relationship between variables” (p. 223); such variables, for this specific case, were the students’ scores given by the professors in the implementation of the test and the scores obtained on the second grading session that was conducted two weeks after the test implementation. The correlation coefficient ranges from -1 to +1, being +1 the estimate of a high relationship or, in this case, high intra-rater reliability. 0 represents no relationship between the variables, whereas -1 means an inverse relationship (Dörnyei, 2007). Values lower than -.40 or +.40 suggest that the relationship, either positive or negative, is weak. On the contrary, values higher than -.80 or +.80 suggest that there is a strong relationship (Brown, 1988). In the case of this study, the professors who obtained a correlation above +.80 were considered reliable raters.

The type of correlation used in this research to estimate the level of intra-rater reliability was the Pearson Product-moment Correlation Coefficient. Brown and Hudson (2002) suggest the following formula to calculate it:

$$r_{xy} = \frac{\sum(X - M_x)(Y - M_y)}{NS_xS_y}$$

r_{xy}= Pearson product-moment correlation coefficient
X = each score on variable X
M_x = mean score for variable X
Y = each score on variable Y
M_y= mean score for variable Y
S_x = standard deviation for variable X
S_y = standard deviation for variable Y
N = number of examinees who took both tests

For the application of this formula in the present study, variable *X* represented the scores from the first rating and variable *Y* represented the scores from the second rating. Likewise, *N* was the total number of students whose performance was rated two times. All of the information corresponding to the dimensions presented before was organized in Table 1.

The first column of the table (N) shows the number of students whose speaking performance was assessed twice. The second column (X) presents the scores obtained from the first rating, whereas the fifth column (Y) presents the scores from the second rating. The third (M_x) and sixth (M_y) columns display the mean scores of X and Y respectively. The distances between the mean and the students' scores for both the first and second ratings are calculated in the fourth $(X - M_x)$ and seventh $(Y - M_y)$columns. Finally, the last column $[(X - M_x)(Y - M_y)]$shows the results of the fourth column times the seventh column.

Table 1
Table template for the calculation of the Pearson Product Moment Correlation Coefficient.

1	2	3	4	5	6	7	8
N	X	M_x	$X - M_x$	Y	M_y	$Y - M_y$	$(X - M_x)(Y - M_y)$
1							
2							
3							
4							
5							

Note. Adapted from *Criterion-Referenced Language Testing* (pp. 158-159), by J. Brown and T. Hudson, 2002, Cambridge: Cambridge University Press.

4.8.2 Content Analysis.

Qualitative data cannot be directly measured; therefore, it is necessary to process or analyze such information through a technique that opens the possibility of turning it into quantifiable data. Content analysis is defined by Andréu (2000) as a technique that consists of interpreting any type of message encountered in different formats such as interviews' transcripts, videos, texts and paintings. This technique requires a systematic and objective reading process that enables the researcher to identify and develop appropriate categories.

The procedures used for analyzing the retrospective verbal protocol were adapted by the researchers from the content analysis steps proposed by Fraenkel et al. (2012). The first step was to establish the objective for which the content analysis procedure was conducted. As stated in the research questions of the project, one of the objectives was to identify which were the factors that could affect their reliability in the test. Secondly, the terms related to the factors were defined. As the researchers did not know exactly which were the factors influencing the raters, only a couple of possible terms where defined beforehand. To carry out the following step, which is locating the data, a verbal protocol that allowed to identify the factors influencing reliability was conducted and later transcribed. Next, the researchers had to decide on the specific units to be analyzed; in this case, the sentences from the transcripts of the interviews were considered. The fifth step consisted of developing a sampling plan. As the purpose was to identify the factors affecting rater's reliability, the verbal protocol was only applied to the two professors in charge of assessing the speaking section of the test. After that, the coding categories were formulated; in order to do so, the researchers pre-established two categories (experience, tiredness) as factors that could affect reliability. The other categories arose from the analysis of the transcripts. Subsequently, with the aim of ensuring the validity and reliability of the data analysis, each researcher analyzed both transcripts from the rater's interviews and highlighted the data to support the categories. The researchers' outcomes were then compared and discussed. Finally, different procedures were used to analyze the data. For the categories previously established, the ones that were mentioned the most by both raters were considered for the findings.

4.8.2.1 Coding.

In order to organize and analyze the qualitative data, which derives from the retrospective verbal protocol responses, a coding process was necessary. Mackey and Gass (2005) state that coding requires researchers to organize and analyze all the collected data with the aim of transforming it into quantifiable information. In the case of this research study, the whole verbal protocol was transcribed before the coding. Once the protocol was in the form of a text, the researchers highlighted in different colors the pieces of information that corresponded to the raters' perceptions about their scoring procedures, and the factors that could affect intra-rater reliability. Those factors were organized under the following categories: experience, lack of adherence to rubric criteria, rater-student relationship, physical conditions and rater's pressure

and responsibility. In addition, the information from the transcript was systematized by placing a code at the beginning of each line. The code that was used to organize the information was “L6-10R3”. The “L” plus the numbers indicate the number of the lines, and the “R” letter followed by a number indicate the rater.

Code	Description	Example
L#-#	Transcript’s lines	**L39-42**R5
R#	Rater	L2**R2**

4.9 Ethical Considerations

For the purpose of ensuring the protection of human subjects in this study, some ethical considerations were taken into account. Mackey and Gass (2005) emphasize the importance of providing the subjects with sufficient information about the research to be conducted. Bearing this in mind, the participants of this research, which included professors and students from the LBI program, were informed about the purpose, procedures, data collection instruments and possible benefits of the study. Participation in the project was completely voluntary, and individuals had the right to withdraw at any time. Likewise, no names of professors or students were disclosed under any circumstance so as to guarantee confidentiality.

The students participating in this research signed a consent letter which specified that they were going to be recorded during the speaking test, and that their grades would be used for analyzing the level of intra-rater reliability on the Semaforización test (see Appendix C). Professors, similar to students, signed a consent letter in which they were informed of the collection of the rubrics they used to score, the second rating process they would be requested to perform and the verbal protocol that would be conducted after it (see Appendix D).

5. Findings

5.1 Quantitative results of the correlation between the scores from the test administration and the scores from the second rating.

In order to conduct the analytical part of the quantitative data, the Pearson Product-moment correlation coefficient method was used. This correlation, as proposed by Luoma (2004), is commonly used to calculate intra-rater reliability in speaking assessments. In the present study, the variables considered for the calculation are five students' speaking scores obtained during the implementation of the Semaforización test, and the same students' scores obtained after a second rating of the same speaking performance.

The Pearson correlation formula and its descriptors are the following:

$$r_{xy} = \frac{\sum(X - M_x)(Y - M_y)}{N S_x S_y}$$

r_{xy}= Pearson product-moment correlation coefficient
X = each score on variable X
M_x = mean score for variable X
Y = each score on variable Y
M_y= mean score for variable Y
S_x = standard deviation for variable X
S_y = standard deviation for variable Y
N = number of examinees who took both tests

Table 2 and Table 3 below show eight columns that contain the required data for the completion of the formula. Column 1 displays the number of students whose speaking performance was assessed twice. Columns 2 (X) and 5 (Y) show students' scores obtained in the first and second rating respectively. Columns 3 (M_x) and 6 (M_y) present the mean of each set of scores. The distances between the mean and the students' scores for both the first and second ratings are calculated in Columns 4 ($X - M_x$) and 7 ($Y - M_y$). Finally, the last column [$(X - M_x)(Y - M_y)$] shows the results of the fourth column times the seventh column.

Table 2
Correlation between the first and second rating carried out by rater 1.

1	2	3	4	5	6	7	8
N	X	M_x	$X - M_x$	Y	M_y	$Y - M_y$	$(X - M_x)(Y - M_y)$

1	2.5	2.5	0	2.25	2.15	0.10	0
2	2.5	2.5	0	2	2.15	-0.15	0
3	2	2.5	-0.5	2	2.15	-0.15	0.07
4	2.75	2.5	0.25	2.25	2.15	0.10	0.02
5	2.75	2.5	0.25	2.25	2.15	0.10	0.02

$S_x = 0.3$
$S_y = 0.1$
$\sum(X - M_x)(Y - M_y) = 0.125$
$NS_xS_y = 5(0.3)(0.1) = 0.15$
$r_{xy} = \frac{\sum(X - M_x)(Y - M_y)}{NS_xS_y} = \frac{0.125}{0.15} = 0.83$

The result of the Pearson Product-moment correlation coefficient *(Table 2)* shows that rater 1 is 83% reliable, which is interpreted as a high level of agreement (Dörnyei, 2007). As it can be observed, students got low grades that ranged between 2.0 and 2.75 on the speaking part of the Semaforización test. Likewise, the scores assigned after listening to the recordings ranged between 2.0 and 2.25. This shows that rater 1 had a good level of consistency given that the scores did not change significantly from one rating to the other. The data also manifests that the rater was more severe on the second rating as the mean score of Y (M_y) is 2.15, opposite to the mean score from the first rating (M_x) which is 2.5.

Table 3
Correlation between the first and second rating carried out by rater 2.

1	2	3	4	5	6	7	8
N	X	M_x	$X - M_x$	Y	M_y	$Y - M_y$	$(X - M_x)(Y - M_y)$
1	5	3.85	1.15	5	3.45	1.55	1.78
2	3.25	3.85	-0.6	2.5	3.45	-0.95	0.57
3	4	3.85	0.15	3.5	3.45	0.05	0.00
4	3	3.85	-0.85	2.75	3.45	-0.70	0.59

5	4	3.85	0.15	3.5	3.45	0.05	0.00

$S_x = 0.7$
$S_y = 0.9$
$\sum(X - M_x)(Y - M_y) = 2.96$
$NS_xS_y = 5(0.7)(0.9) = 3.15$
$r_{xy} = \frac{\sum(X - M_x)(Y - M_y)}{NS_xS_y} = \frac{2.96}{3.15} = 0.94$

The results of the correlation displayed in table 3 indicate that the level of reliability for rater 2 is 0.94, which according to Dörnyei (2007) is high. Although both of the evaluators were highly reliable, the value obtained after the correlation by rater 1 was 9 hundredths (0.09) lower than the 0.94 obtained by rater 2.

Another point of contrast are the scores provided by the raters in the first administration. While the second rater assigned varied grades that go from 3.0 to 5.0, rater 1 assigned more closely related scores (2.0 to 2.75). In the second rating process conducted with rater 2, the given scores ranged from 2.75 to 5.0. In spite of this, the scores provided did not vary that much as to negatively affect the reliability.

As it can be seen from the consistency analysis, the raters obtained a Pearson-product Moment correlation value of 0.83 and 0.94 respectively (see Tables 2 and 3) being rater 2 the most reliable. These values are presented in Brown (1988) as an ideal performance due to the fact that the correlations were higher than 0.80.

5.2 Factors that May Influence Intra-rater Reliability

After the analysis of the qualitative data, it was found that there are five factors that have a positive and negative impact on the reliability of the raters from the speaking part of the Semaforización test. The factors found were compared and contrasted with other studies related to language assessment. The factors described below are experience, lack of adherence to the rubric's criteria, rater-student relationship, physical conditions, and rater’s pressure and responsibility.

5.2.1 Experience in language teaching and in the Semaforización test leads to higher levels of intra-rater reliability.

5.2.1.1 English language teaching and language assessment experience certifies the level of rater’s reliability.

It is believed that experience assessing English language learners plays an important role in developing raters' reliability. In this study, experience in language assessment is determined

by the years the professor has dedicated to teach English as a second language. This decision is made considering that assessment is an essential part of this profession. In fact, when asked about their experience in language assessment, the raters agreed on the notion that being a language teacher contributes to gaining practice on such a field.

The reason why experience in English language teaching (ELT) is important for rater reliability on speaking tests relies on the work language teachers have to do when assessing students' productions. In the LBI program, the professors have to assess learners following a set of criteria that allows them to make more objective judgments and be more ethical. In addition, having several years of experience assessing spoken language helps professors realize that it entails a certain level of difficulty due to the fast-fading nature of speech. Both, the familiarity with assessment instruments and the awareness on the complexity of assessing speaking acquired through many years of experience, shape educators as raters. From the analysis of the interview responses, it was discovered that both raters had experience in ELT. By comparing the information gathered from the verbal protocol with the one obtained from the correlation coefficient, it can be argued that the raters' experience influenced their ability to provide consistent scores since both of them obtained a reliability value higher than 0.80. The following excerpts show the responses to the question about their language assessment experience:

> L7-10R1: In language assessment… I think the experience that I have had, well... being a teacher because being a (language) teacher... part of what we do in the classroom is also… well, assessment is part of it, of what we have to do. So, I have like almost 20 something years of experience, almost 28.
>
> L8-12R2: My experience… well, I think it goes back to my whole experience as a teacher, being a language teacher for about 15 years; and it's one role of every language teacher. But apart from that, during the last five-six years I have been working as an academic advisor for (...), and one of the main lines of the academic… let's say, support that we provide is on assessment"

In the previous extracts, the raters agree that the years they have dedicated to teach English have contributed to gaining experience in assessing language learners. The samples suggest that both raters view assessment as an integral part of the teaching process. As an example, in L7-10R1, rater 1 declares that part of what teachers have to do in the classroom is to assess students; likewise, in L8-12R2, rater 2 states that assessing language learners is one of the roles of the teachers. This process of teaching and assessing entails a period of time that varies from teacher to teacher. As shown in the samples above, the raters have more than a decade of experience in ELT and language assessment. Rater 1 reported to have around 28 years and rater 2 claimed to have around 15 years of experience.

According to Liskin-Gasparro (as cited in Fulcher, 2003), having experience as teachers helps evaluators to be more consistent on the application of scoring criteria. Taking into account that ELT experience enhances language assessment experience, teachers acquire familiarity with the scoring criteria and the tasks used to assess language productions. Such familiarity certifies rater reliability. In this sense, it can be argued that rater 1 and rater 2 were consistent at the moment of scoring learners' performance due to the fact that they are experienced professors who have been constantly required to assess students.

5.2.1.2 Experience in the Semaforización test increases score consistency of the raters.

Having previous experience either as an interlocutor or as a rater on the speaking part of the Semaforización test gives raters the possibility to provide more consistent scores. For the speaking section of this test, two professors are required to carry out the assessment of the students. One of the professors plays the role of interlocutor asking the questions of the test and guiding students throughout the completion of the tasks. On the other hand, the other professor is in charge of marking the performance of the candidate by following the descriptors of a rubric. This last professor is given the name of rater or evaluator.

Regardless of the roles played by the professors during the past examinations, they have acquired an understanding of the type of tasks that are included on the speaking section and what the expected response from the test-taker is. Similarly, they have gotten acquainted with the type of assessment criteria and descriptors implemented. The raters have both participated previously on the Speaking exam; however, the number of times they have played the role of evaluator varies. The following excerpts show evidence of the aforementioned aspect:

> L2-3R1: I have had the opportunity to… well, with this Semaforización I think it's like my third time (as interlocutor).
> L5-R1: (...) grading just this time.

> L5-6R2: Before these last examinations I think I have had that role (evaluator) a couple of times. Two times as the evaluator and once more as the interlocutor.

As it can be noted from the comments above, rater 1 has not played the role of evaluator before this test; on the contrary, she has been an interlocutor three times. As for rater 2, he has more experience as rater than as interlocutor taking into account that he has played such role twice before this administration of the test. After comparing this qualitative data with the results from the statistical analysis, it can be claimed that participating in the speaking part of the Semaforización test either as an interlocutor or as a rater has a positive impact on the reliability of the raters. However, it was observed that rater 2 obtained a higher value on the correlation (0.94) than rater 1 (0.83). Thus, it can be suggested that rater 2 was more reliable given the fact that he played the role of rater more times which gave him more familiarity with the specific

rubric of the speaking part. This is in accordance with Furneax and Rignall, and Shaw (as cited in Davis, 2015) who claim that knowing and understanding the "scoring system" during a particular period of time boosts raters' consistency. Having a proper understanding of the rubric allows the rater to develop a particular way of rating which helps him to focus on specific criteria at specific moments of the test. For instance, in the following comment it can be seen how rater 2 approaches the rating scale during the speaking test:

> L114-119R2: One thing that often happens is that… let's say, one student is talking about something and there is an item that you are paying more attention than the others. Like, ok, during this passage of the examination I want to pay attention to the pronunciation, so I focus on that for a while. Then, in another part of the evaluation, I try to focus on other things. Of course, the last part which is about interacting with the peers, I try to focus more on the interactive communication.

In L114-119R2, the rater states that he has established a pattern that allows him to assess specific criteria at different moments of the speaking test. This sample proposes that the rater has built a schema of the criteria and descriptors to be assessed. Some criteria may be reflected throughout the whole performance; however, some other may be only present during a specific task of the test and that is why it is relevant to know when to focus on that. To illustrate, in L114-119R2, rater 2 explains that he provides a score for "interactive communication" during the last task of the speaking test which is about having students to discuss an issue. The rater's internalization of the rubric can be attributed to the experience that he has in assessing the speaking skill on that particular test. Based on this, it can be stated that a high level of reliability can be achieved when raters get experience in the specific tasks of the test and internalize the rubric's criteria.

5.2.2 Difficulty in adhering to rubrics' criteria affects rater reliability.

Rubrics have been found to lead to a more reliable and less subjective assessment (Silvestri & Oescher, 2006). They list the criteria that must be referred to when carrying out assessment procedures. This nature of rubrics makes them a tool used for criterion-referenced assessment, which according to Aviles (2001), is an approach that seeks to measure students' performance based on pre-set criteria. In the case of the rubric implemented to assess students taking the speaking part of the Semaforización test, four criteria are included: discourse management, pronunciation, interactive communication, and grammar and vocabulary. Each criterion has a rating scale with three different values (0.5, 0.75 and 1.25) and descriptors providing indicators of the expected performance so that the raters know which score should be given to a test-taker (see appendix A).

In spite of the fact that rubrics have been addressed as assessment instruments that improve rater reliability, some studies conclude that using them does not guarantee that the rater

will achieve reliable scores (González, Trejo & Roux, 2015; Jeong, 2015). The analysis of the data revealed that the raters identified an issue with the design of the rubric that led to inconsistent scoring. The raters explained that they had to deal with situations in which the candidate's speaking performance did not fit into any of the values established on the rubric. This issue is addressed in the following lines:

> L18-19R1: Sometimes, it is like there is a lack there (on the rubric), like there should be something in the middle.
>
> L68-70R1: (...) but as I told you, maybe it was like in the middle. So, which grade should I give? This one or this one? Maybe it's not even 0.5 or 0.75; it could be like 0.60.
>
> L139-140R2: I think that in many occasions I found myself undecided. Like, I know the performance of the student is not on one of the items of the rubric neither in the other. And then, I think some in-between descriptors are necessary because I've found myself in that situation a couple of times.

In the previous samples, both raters state that there is a lack of values and descriptors on the rubric design and that in-between score points on the rating scale are necessary. In L139-140R2, rater 2 declares that in several moments he felt undecided about providing a score given the lack of values and descriptors to place students' performances. These lines evidence that the raters are not fully adhering to the parameters established in the rubrics; on the contrary, they are following their own mental criteria. Because of these internal criteria, raters are facing difficulty at the moment of translating qualitative descriptions or performance into quantitative intervals. When difficulty applying the rubric set for criterion-referenced grading is encountered, raters rely on norm-referenced grading. The following excerpt illustrates this assessment shift:

> L39-48R2: I tried to use the rubric, first of all, and the key words that differentiate the performance of students. I think it's inevitable to compare their performances. And I sometimes see myself comparing their performances and then I go back to what I should compare their performance to, which is the picture that I have in my mind of a speaker in this proficiency level that we are looking for, in this case the B1. So, I have to monitor myself sometimes cause I see myself like: "Oh well, if I compare these two students this might be 1.25". I don't know, but then I consciously go back to my reference which is the independent user B1. So, it makes it a little bit complicated cause I tend to go biased by their own performance and the comparisons I make.

Rater 2, in L39-44R2, explains that for assessing students' speaking skill, he tries to follow the rubric and focus on the descriptors that distinguish the performance of each test-taker.

However, although the rater is using a rubric throughout the whole process, he identifies a tendency to compare the performance of the students that are being assessed. Instead of assessing students' performance against the criteria established on the rubric, rater 2 followed a norm-referenced grading in which students are compared to one another in order to decide their scores. The rater highlights that he is constantly monitoring himself so that he can identify when he is comparing students among themselves and stop doing so. Nevertheless, based on his comments, it can be stated that he does not go back to the rubric criteria but to his own mental model or reference of a B1 language user.

In spite of the distinct purposes of criterion-referenced and norm-referenced assessment, "the overlap between these two approaches to assessment is often overlooked" (Burton, 2006, p. 74). Johnstone and Rubenstein (as cited in Burton, 2006) explain that the essence of criterion-referenced assessment is diminished if raters get influenced by the performance of other students in the same group as the norm-referenced assessment indicates. Although the purpose for including a rubric on a proficiency test is that of comparing the test-taker performance with the criterion, Orr (2002) points out that raters usually struggle to base their decisions on the rubric if they are no clear for them; as a result, raters end up making comparisons. Likewise, Lumley (as cited in Graham, Milanowski and Miller, 2012) found that "when evaluators are unable to decide between two score points, other extraneous factors often creep into their decision-making, such as (...) comparing the current subject with previously rated subjects." According to this information, it is very feasible that when rater 2 was undecided between two descriptors, he had opted for comparing two or more students' performances as a strategy to make a decision. Joe, Harmes, and Hickerson, (2011) add that when a rubric is perceived as complex by the raters, they do not adhere to it, but they follow their own criteria which may be different from the one described on the rubric. When rater 2 indicates that he "goes back" to the picture he has on his mind of a B1 speaker or to his reference of a B1 user, he is moving away from the test rubric and implementing his own criteria. The shift on the assessment approach used by the raters and their internal criteria results in lack of objectivity and reliability.

5.2.3 Rater-student relationship can hinder reliable scores.

Taking into account that the raters from the Semaforización test are at the same time professors from the LBI program, there is a possibility that the evaluator and the test-taker have had a teacher-student relationship in which the rater knew about the students' background and English language proficiency. From the analysis of the data collected, it was found that being aware of this information makes raters more subjective and, therefore, less reliable. This is best evidenced in the following samples:

> L225-236R1: Something important is that the students we were both evaluating were students of my partner's group, so she (the interlocutor) knew the students; I didn't know her students. I think we should evaluate people that probably we don't know. So, I think

> it will give… probably we don't involve emotions or feelings. For example, when we finished, the only thing she told me was like "oh, you know what? she is a very very good student" or "she is a very good student; however, she didn't answer correctly" or "however, she was nervous" or things like that. I think if we have less chances to involve feelings, probably it would be more reliable… the way of grading would be more reliable. Something that I noticed, and something that I think it should be done is that we should not know students. Like, when you go and you take a standardized test, you don't know your evaluator, so maybe it would be something to think about and to improve.
>
> L193-96R2: I think it could influence a bit (knowing the test-taker) cause you always have this reference of the student as a user in other scenarios. And then, I guess, in some cases this could be unconscious; when there is something you miss, when there is a gap, I would fill it with the reference I have of the student.

From both extracts, it is evident that the raters find the familiarity with students as a disadvantage at the moment of scoring the candidates' speaking skill. However, the reasons provided by the raters vary according to their own processes. In the first place, in L225-236R1, rater 1 views rater-student relationship as a factor that may trigger emotions and keep raters from being reliable in their scoring procedures. For instance, the rater recalls how her partner, who played the role of interlocutor and knew all students, showed leniency towards the test-takers. Accordingly, rater 1 emphasizes on the idea of having evaluators to assess unknown candidates and suggests replicating the conditions of standardized tests in order to guarantee consistency on the scores. As for rater 2, instead of involving emotions, he asserts in L193-96R2 that he uses the students' background information and proficiency level as a reference of what they are able to do or not. The rater explains that when he misses a piece of information in the test-taker's performance, he unconsciously recalls the reference he has of the student so that he can imagine what would have been the word, sentence, reaction or response used by the person.

Mead (1980) considers that the familiarity between the evaluator and the student threatens the objectivity of the assessment process. In agreement with what rater 1 says, the author notes that when evaluators are required to assess someone they already know, they are prone to feel empathy towards the test-taker. On the other hand, Mead also alludes to rater 2 when explaining that it is likely that the raters rely on "preconceived notions" (p. 5) or relate previous production of the student in order to complement her performance. Finally, the author proposes that ratings should be carried out with students that are unfamiliar to the rater with the aim of minimizing the impact on rater's reliability.

5.2.4 The influence of physical conditions of the rating processes on rater's reliability.

In order to conduct this study, the raters selected to participate were asked to assess students in two moments, both with different physical conditions. In the first rating, which was the test implementation, evaluators assessed test-takers on the spot, meaning that they could observe students' body language and reactions when speaking. However, the second rating was carried out in a room where the researchers and the raters met. Here, raters were given an electronic device with the voice recordings of the students' oral performances made during the test implementation. This time, the raters did not have the possibility to observe the candidate's reactions. The analysis of the raters' responses given on the interview reveal that the difference in the physical conditions of the two ratings had an impact on the scores' variation.

Rater's reliability could have been affected during the first rating since evaluators had the possibility to observe students while speaking. External factors such as non-verbal communication (body language, gestures, and facial expressions) present in the students' performances could have interfered in their decision-making. On the contrary, during the second rating process, the raters listened to the candidates' responses without watching them. Since body language, gestures and facial expressions could not be perceived, raters were more likely to be objective. These insights can be found in the following data:

> L33-35R1: So, probably notice that if we compare, I was more strict today (...). So, probably having students in front of me, looking at them... like the way they were shaking. Probably it made me change my own… the score, yes.
>
> L93-99R2: Maybe it could've been the body language. I don't know, being there looking at the person. (...). We also communicate things when we gesticulate with our faces or we move our hands. That could've been a difference. And I think it's the only one (...) I mean… about the conditions in which these two results were obtained.

The raters explained that one of the reasons behind the difference in grades is attributed to the different physical conditions under which these grades were assigned. In L33-35R1 and L93-99R2, both raters highlight that having students in front of them and observing their non-verbal communication was a possible source of variation on the points given. For example, in L33-35R1, rater 1 argues that she was more strict during the second rating because she did not have the interference of those external factors. As for rater 2, he recognizes that changing the conditions in which the scores were obtained generated some kind of inconsistency. He is aware that through gestures and facial expressions people also communicate things, and such communication can complement students' oral performances.

As claimed by Mead (1980), assessment procedures carried out on the spot differ in some aspects from the ones made from voice recordings. The author explains that "there may be a

tendency to be more attracted to the enthusiasm and presence of students when rating on the spot than when rating tape recordings" (p. 17). This tendency may influence raters to be subjective when applying the scoring criteria leading to low reliability levels. The author's claims are in accordance with what the raters said about perceiving the students' performance differently when the conditions changed. It can be hypothesized, then, that using video recordings of the test-takers instead of tape recordings for the second rating would have evened the conditions and allowed raters to provide scores similar to the ones from the first rating.

5.2.5 The pressure and responsibility to give an accurate score may impact intra-rater reliability.

Most types of assessments require teachers to assign a grade to the learner in question. In the case of standardized tests, this grade has either positive or negative consequences on the academic life of the test-takers. For instance, the Semaforización test administered at the LBI program can hinder students from advancing on their studies if they do not achieve the expected proficiency level in the specific semesters. These conditions can make the raters in charge of assessing the speaking part of the test feel pressure to give an accurate grade that represents the real test-takers' language ability. However, this pressure can play an important role on the level of reliability as expressed by the two raters involved in this study:

> L131-132R1: I think it is like the responsibility you have. Like, what about if I give a wrong assessment?
>
> L162-164R2: I know that it means a great deal for the courses. And, well, that was basically the pressure. Like, it could have been very important for them to have an accurate grade.
>
> L222-223R1: That's why I was tired at the end… not because… because it was just like one hour, one hour and a half (the test implementation). No, it was the pressure.
>
> L172-173R2: I felt even more… let's say, compromised [sic][1] to give an accurate evaluation. So yes, there is a little pressure.

The samples L131-132R1 and L162-164R2 suggest that the raters are aware of the impact that the score they assign on the test has on the academic progress of the students, and therefore, they recognize the important role that they are playing as raters of the test. In L222-223R1 and L172-173R2, both raters manifest the pressure they felt during the rating process

[1] The expression refers to the word "committed" which is being pledged or obligated to do something.

carried out in the test implementation. They explain that the origin of such pressure derives from the responsibility they have and the repercussion of the scores they provide.

There are consequences associated with the use of standardized tests. Regarding test-takers, Shohamy (as cited in Brown, 2004) emphasizes that such types of tests “determine people's future education” (p. 21). In addition, Gronlund and Waugh (2008) state in their work that teachers should remember that learners’ motivation, course performance, autonomous learning and attitude towards schoolwork are affected by the assessment procedures they carry out. From the analysis of the data, it can be revealed that the raters considered the impact of their rating and that it made them feel pressure. Although theory to support the finding associated with pressure as a factor that can affect the raters reliability was not found, it is worth mentioning it in the present study as both raters agreed on the notion that the consistency of their scores could have been influenced due to the responsibility and pressure they felt during the assessment procedures.

6. Conclusions

The aim of this study was to estimate the level of intra-rater reliability on the speaking part of the Semaforización test and to explore the factors that could influence such reliability. In order to do so, the scores gathered in the speaking part of the test and a verbal protocol conducted with the raters were analyzed. A correlation coefficient method was used to analyze the numerical data or the scores, whereas a content analysis technique was implemented to extract information from the verbal protocol responses. As a result, it was found that raters had a high level of agreement and that there are five factors that can influence their level of reliability either in a positive or negative way.

To answer the first research question of this study: What is the level of intra-rater reliability at the proficiency speaking test? the Pearson-product Moment Correlation Coefficient was calculated. This correlation was estimated between the scores from the implementation of the speaking part of the test and the scores from the second rating. The results show that both raters were highly reliable since their levels of reliability were 0.83 and 0.94 (see tables 2 and 3). As pointed out by Brown (1988), all values higher than 0.80 can be interpreted as an ideal performance. The obtained values are consistent with Shohamy's (1983) earlier findings regarding rater reliability on an oral proficiency test. However, they are slightly higher than the coefficients reported in previous studies conducted on the same topic (Kang & Rubin, 2012; Sak, 2008).

As an attempt to answer the second research question: What factors influence the level of intra-rater reliability?, a verbal protocol was carried out with the raters from which five factors were identified. During the interview, the raters were asked questions about their teaching background and then, they were required to compare the scores they assigned to each student and to mention the reasons why they could have differed one from another. The findings suggest that experience is one of the factors that has a positive influence on the level of intra-rater reliability. The more experienced raters are, the more reliable they will be. The years of teaching experience and the experience with the speaking test help raters be familiarised with the use of rubrics, the tasks for assessing speaking, and the difficulties found at the moment of assessing oral performance. Furthermore, both raters claimed to have more than a decade of experience in language teaching and assessment. When comparing this information with the results obtained from the correlation coefficients, which were high, it can be argued that experience can increase rater reliability. In fact, these findings confirmed the ones presented by Kang and Rubin (2012); in their study, only novice raters participated, and the results from the correlation yielded moderate to low reliability levels. Another important observation that derives from the comparison of the quantitative and qualitative data is that rater 2 was 94% reliable, being more reliable than rater 1. It can be attributed to the fact that he had more experience as a rater of the speaking test, and, therefore, he was more familiar with the rubric used to assess oral performances. Bearing this in mind, if high levels of reliability are desired on oral proficiency

tests, it is crucial to choose raters with enough experience in language assessment. Given the case raters are not experienced, the administrators of the test should provide the necessary tools, such as training sessions, to achieve satisfactory levels of consistency.

The second finding related to the second research question is lack of adherence to the rubrics' criteria. The analysis of the interview responses revealed that both raters encountered difficulties with the rubric when trying to translate the qualitative information into quantitative intervals. This difficulty was addressed by the raters as an issue in the rubrics design in which there was a lack of in-between values and descriptors. As a consequence, the raters opted for comparing a student's performance against a previously rated student, or students from the same group. By doing this, raters were focusing more on norm-referenced assessment and leaving aside the criterion-referenced approach. Apart from comparing students, there was a tendency towards following an internal criteria. To illustrate, one of the raters claimed that he based his judgements on a mental reference he had about a B1 language user as it was the expected level test-takers had to achieve. By comparing students among them and following an internal criteria, the raters are negatively affecting their reliability. With the aim of diminishing this effect, raters at the LBI program should receive more training on the use of the rubric.

Another factor that could be extracted from the verbal protocol was the rater-student relationship. Knowing students and their language proficiency can prevent raters from being reliable. In the first place, knowing students can trigger raters' emotions and lead to subjective scoring as addressed by rater 1. On the other hand, rater 2 states that being familiar with the student affects reliability since there is a reference of the student's language proficiency and what the student is able to utter with the English spoken language. The rater can unconsciously base his scoring on that reference, which is not objective and leads to low reliability levels. Thus, in order to ensure that raters are consistent with their markings, it is necessary to avoid assigning raters to a group of test-takers to whom they are familiar.

The physical conditions under which the ratings of oral performances are conducted can be considered as another factor affecting intra-rater reliability. From the analysis of the rater's responses, it was evidenced that the difference in the conditions of the two rating sessions had an effect on the consistency of the scores provided. In oral proficiency tests, it is common to carry out assessments on the spot, meaning that the raters are able to observe student's body language, facial expressions and gestures while performing. However, in order to estimate rater-reliability, a second marking is required, and usually, raters have to assess test-takers once more based on tape recordings. This difference may influence evaluators into changing their own way of grading. In fact, one rater claimed to be more strict during the second rating given that test-takers were not present, and she could not observe their behavior. In this sense, it can be argued that assessing students on the spot leads raters to more subjective markings. This can be solved by rearranging the location of the raters during the test implementation. It is possible that if evaluators are not able to observe test-taker's body movements and gestures, they apply the assessment criteria in a more objective way. Additionally, the accuracy of the information

gathered through intra-rater correlations can be improved by ensuring the same conditions for both ratings. For instance, using video-recordings instead of tape-recordings on the second rating.

Finally, the last factor found is rater's pressure and responsibility. The scores given on the Semaforización test determine the academic progress of the students from the LBI program. As speaking is one of the skills assessed in this test, the score students obtain is of great importance. Raters are aware of the implications and the responsibility they have. For this reason, they feel pressure to give an accurate score. According to the raters who participated in this study, this pressure can have an effect on their level of reliability. Both raters obtained values higher than 0.80 on their correlation coefficient, which means that pressure did not affect to a large extent their scoring process. However, it could have been one of the reasons for not obtaining a level of reliability similar to +1.

Little literature on the factors that affect rater reliability was found apart from experience, fatigue, training and rubric implementation. Therefore, this study aimed at exploring other factors that impact the self-consistency of raters. From this exploratory study, lack of adherence to the rubrics' criteria, rater-student relationship, physical conditions and rater's pressure and responsibility were found. This outcome drew researchers' attention since the estimation of intra-rater reliability was higher than 0.80 in spite of having some factors that could have negatively affected the reliability.

7. Research and Pedagogical Implications

Some pedagogical and research implications derive from the development of this research project. The pedagogical implications aim at strengthening language assessment practices in the program where the study was conducted, whereas the research implications suggest future studies on the qualities of the Semaforización test in order to ensure that the process carried out is accurate and fair.

In the first place, this study shed light onto the need for proper education and training in language assessment. It was found that reliable assessment procedures cannot be guaranteed if there is no solid education on the knowledge, skills and principles of language assessment. Even if teachers have a large amount of experience in language teaching, some unreliable practices can be observed due to lack of awareness of the true purposes of creating, implementing and interpreting assessments. It is imperative, then, that teachers at all levels of education undergo extensive training where their language assessment practices can be developed and improved. This can allow teachers to monitor the procedures they are conducting and to analyze the impact they may have on the reliability of the scores they are providing.

In the same line, raters in charge of rating not only the speaking section but also the writing section of the Semaforización test should be constantly trained in order to achieve desirable rater consistency. Subjectivity and judgement errors can be reduced by adhering to the scoring criteria or rubrics. Thus, training raters to understand the rubric and to be more self-consistent is imperative to increase their reliability. Unfortunately, the results and effects of a training session do not last for long after it has been conducted, so it is necessary to hold training sessions before each administration in order for raters to calibrate their skills, and internalize and interpret the set of criteria established on the rubric (Lumley & McNamara, 1995). Rating training should not be a once in a life matter, but a permanent process. Lumley and McNamara (1995) suggest that during these sessions, raters should be introduced to the criteria and then a series of performances should be presented so that raters can score them. The performances selected for the training must illustrate different levels of proficiency and different issues that may arise during the assessment. Follow-up ratings can be included and estimations of the raters' reliability can be made in order to decide if a rater can successfully participate in the rating process.

Although the Semaforización test has been administered for around three years, this study is the first attempt to analyze a small amount of data drew from it. Bearing in mind that this test has such an important impact on the academic progress of the students in the LBI program as well as on their professional development, further research should be done in relation to the reliability of scores obtained in the other sections of the test. It is crucial to further explore this quality of language assessment so that the validity of the test can be guaranteed. In addition, future research in the LBI program should consider the potential effects of rater's cognitive processes when assessing students, specifically, when implementing rubrics to measure

productive skills. It allows for the possibility of identifying the processes that can hinder raters from giving accurate and reliable scores and the processes that facilitate and enhance the evaluators' decision-making.

The findings of this study also prompt future qualitative research on the different factors and raters' characteristics that have an effect either positive or negative on the rating process. For instance, during the development of this project, the raters' pressure and responsibility to assign an accurate score was identified as a factor affecting raters' reliability; however, studies which aim at exploring this matter were not found and, therefore, comparisons could not be done.

8. Limitations

It is worth mentioning that throughout the development of this research project, different constraints arose which had an impact on the interpretation of the findings. The first limitation encountered was related to the lack of prior research studies on the topic. To exemplify, several studies aimed at estimating the level of inter-rater reliability; however, there was scarce literature concerning intra-rater reliability in speaking assessments. In addition, studies related to the factors that affected the level of rater reliability focused mainly on rater's experience, tiredness and training, leaving aside other factors that also have an important impact. On the other hand, the students' sample size constituted a limitation since the sampling group chosen for the data collection and analysis was too small. It was decided to select a small sample considering that the raters were professors from the LBI program and their workload could have prevented them from participating if they were asked to score again a large number of students; nevertheless, when analyzing the data obtained from the two rating sessions it was found that a larger sample size was necessary in order to ensure a more reliable coefficient value. The last constraint faced by the researchers has to do with the instrument used to analyze the numerical data. The researchers' lack of experience on statistical calculations was a limitation at the moment of implementing the Pearson-Product Correlation coefficient. As a consequence, the researchers took more time than expected to analyze the quantitative data.

9. References

Afflerbach, P., & Johnston, P. (1984). On the use of verbal reports in reading research. *Journal of Reading Behavior*. 16(1). 307-322. doi:10.1080/10862968409547524.

Andréu, A. J. (2000). Las técnicas de análisis de contenido: una revisión actualizada. *Fundación Centro Estudios Andaluces, Universidad de Granada, 10(2),* 1-34.

Areiza, H. (2013). Role of systematic formative assessment on students' views of their learning. *PROFILE Journal, 15(2),* 165-183.

Aviles, C. B. (2001). Grading with norm-referenced or criterion-referenced measurements: To curve or not to curve, that is the question. *Social Work Education, 20(5),* 603–608.

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

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests.* Oxford: Oxford University Press.

Bolaños, I., Cerdas, G., & Ramirez, J. (2014). Intra-rater reliability and the role of experience: A comparative case. *Revista de Lenguas Modernas, 20(1),* 295-307.

Brown, H. (2004). *Language assessment: Principle and classroom practice.* New York: Pearson.

Brown, J. D. (1988). *Understanding research in second language learning: A teacher's guide to statistics and research design.* Cambridge: Cambridge University Press.

Brown, J. D., & Hudson, T. (2002). Reliability, dependability, and unidimensionality. In *Criterion-Referenced Language Testing* (pp. 149-211). Cambridge: Cambridge University Press. doi:10.1017/CBO9781139524803.007

Brown, J. D. (2011). *Testing in language programs: A comprehensive guide to English language assessment.* New York: McGraw Hill.Education.

Burton, K. J. (2006) Designing criterion-referenced assessment. *Journal of Learning Design, 1(2),* 73-82.

Chapelle, C. A, & Brindley, G. (2010): Assessment. In N. Schmitt (Ed.), *An introduction to applied linguistics* (2nd ed., pp. 247-267): Abingdon, Oxon: Hodder Education.

Chiedu, R., & Omenogor, H. (2014). The concept of reliability in language testing: Issues and solutions. *Journal of Resourcefulness and Distinction, 8(1),* 1-9.

Davis, L. (2015). The influence of training and experience on rater performance in scoring spoken language. *Language Testing, 33(1),* 117–135.

Dickinger, A. (2007). Explorative Research. In: A. Dickinger, ed., *Perceived Quality of Mobile Services,* 1st ed. Germany: Peter Lang.

Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies.* Oxford: Oxford University Press.

Fulcher, G. (2003). *Testing Second Language Speaking.* London: Longman/Pearson Education.

Giraldo, F. (2018). A diagnostic study on teachers' beliefs and practices in foreign language assessment. *Íkala: Revista de Lenguaje y Cultura, 23(1),* 25-44.

González, E. F., Trejo, N. P. & Roux, R. (2017). Assessing EFL university students' writing: a study of score reliability. *Revista Electrónica de Investigación Educativa, 19(2)*, 91-103.

Graham, M., Milanowski, A., & Miller, J. (2012). Measuring and promoting inter-rater agreement of teacher and principal performance ratings. *Center for Educator Compensation Reform,* p. 18-19.

Gronlund, N. E., & Waugh, C. K. (2008). *Assessment of Student Achievement (9th ed.).* Upper Saddle River, NJ Pearson.

Hughes, A. (2010). *Testing for language teachers* (2nd ed.). Cambridge: Cambridge University Press.

Jeong, H. (2015). Rubrics in the classroom: do teachers really follow them? *Language Testing in Asia, 5(6).*

Joe, J. N., Harmes, J. C., & Hickerson, C. A. (2011). Using verbal reports to explore rater perceptual processes in scoring: a mixed methods application to oral communication assessment. *Assessment in Education: Principles, Policy & Practice, 18(3),* 239–258.

Kang, O., & Rubin, Don. (2012). Intra-rater reliability of oral proficiency ratings. *The International Journal of Educational and Psychological Assessment, 12(1),* 43-61.

Lim, G. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing, 28(4),* 543–560.

Ling, G., Mollaun, P., & Xi, X. (2014). A study on the impact of fatigue on human raters when scoring speaking responses. *Language Testing, 31(4),* 479–499.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: implications for training. Language Testing, 12(1), 54–71.

Luoma, S. (2004). Assessing speaking. Cambridge: Cambridge University Press.

Mansoor, I., & Grant, S. (2002). A writing rubric to assess ESL student performance. *Adventures in Assessment, 14,* 33–38.

McNamara, T. F. (2000). *Language testing*. Oxford, England: Oxford University Press.

Mead, N. (1980). *Assessing speaking skills: Issues of feasibility, reliability, validity and bias.* Education Commission of the States (Report No. ECS-00-SL-56). Denver: Colorado.

MEN. (2014). *Colombia Very Well: Programa Nacional de Inglés*. Bogotá, CO: MEN.

Mertler, Craig A. (2001). Designing scoring rubrics for your classroom. *Practical Assessment, Research & Evaluation, 7(25).*

Moskal, B., & Leydens, A. (2000). Scoring rubric development: validity and reliability. *Practical Assessment, Research & Evaluation, 7(10).*

Orr, M. (2002). The FCE Speaking test: using rater reports to help interpret test scores. *System, 30(2),* 143–154.

Picón, E. (2013). La rúbrica y la justicia en la evaluación. *Íkala, revista de lenguaje y cultura, 18(3)*, 79–94.

Pineda, D. (2013). The feasibility of assessing teenagers' oral english language performance with a rubric. *PROFILE Journal, 16 (1)*, 181-198.

Qian, D. (2009). Comparing direct and semi-direct modes for speaking assessment: Affective effects on test takers. *Language Assessment Quarterly, 6(2)*, 113-125, DOI: 10.1080/15434300902800059

Rahmawati, Y. (2014). Developing assessment for speaking. *IJEE, 1(2).*

Richards, J., & Rodgers, T. (2001). *Approaches and Methods in Language Teaching.* Cambridge: Cambridge University Press.

Sak, G. (2008). *An investigation of the validity and reliability of the speaking exam at a turkish university* (Master's thesis, Middle East Technical University, Ankara, Turkey).

Segura, R. (2013). *The importance of teaching listening and speaking skills* (Master's thesis, Universidad Complutense de Madrid, Madrid, Spain).

Shohamy, E. (1983) Rater reliability of the oral interview speaking test. *Foreign Language Annals, 16(3)*, 219-222.

Silvestri, L, & Oescher, J. (2006). Using rubrics to increase the reliability of assessment in health classes. *International Electronic Journal of Health Education, 9,* 25–30.

Sreena, S., & Ilankumaran, M. (2018). Developing productive skills through receptive skills – A cognitive approach. *International Journal of Engineering & Technology, 7(4.36)*, 669-673.

Torres, S. (1997). Testing accuracy and fluency in speaking through communicative activities. *HOW, 5(1)*, 95-104.

Tsushima, R. (2015). Methodological diversity in language assessment research: The role of mixed methods in classroom-based language assessment studies. *International Journal of Qualitative Methods, 14(2)*, 104–121.

Universidad Tecnológica de Pereira. (2015). *Acuerdo 15 de 2015*. Retrieved from http://media.utp.edu.co/ilex/archivos/Acuerdo%2015%20de%202015.pdf

Universidad Tecnológica de Pereira. (2017). *Acuerdo 18 de 2017*. Retrieved from https://www.utp.edu.co/cms-utp/data/bin/UTP/web/uploads/media/secretaria/documentos/1489182312-Acuerdo%20No.%2018.pdf

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke: Palgrave Macmillan.

Wiggins, G., (1994). Toward Better Report Cards. *Educational Leadership, 52 (2)*, 28-37.

8. Appendix

Appendix A
Semaforización Test Rubric Sample

Criteria	**Rating**			**Pts**	**Comments**
Discourse Management	**1.25** Produces extended stretches of language with very little hesitation. Contributions are relevant and there is clear organization of ideas. Uses a range of cohesive devices and discourse markers.	**0.75** Produces extended stretches of language despite hesitation. Contributions are relevant and there is very little repetition Uses a range of cohesive devices.	**0.5** Produces responses which are extended beyond short phrases, despite hesitation. Contributions are mostly relevant, despite some repetition. Uses basic cohesive devices.		
Pronunciation	**1.25** Is intelligible. Intonation is appropriate. Sentence and word stress is accurately placed. Individual sounds are articulated clearly.	**0.75** Is intelligible. Intonation is generally appropriate. Sentence and word stress is generally accurately placed. Individual sounds are generally articulated clearly.	**0.5** Is mostly intelligible, and has some control of phonological features at both utterance and word levels.		
Interactive Communication	**1.25** Initiates and responds appropriately, linking contributions to those of other speakers. Maintains and develops the interaction and negotiates towards an outcome.	**0.75** Initiates and responds appropriately. Maintains and develops the interaction and negotiates towards an outcome with very little support.	**0.5** Initiates and responds appropriately. Keeps the interaction going with very little prompting and support.		
Grammar and	**1.25**	**0.75**	**0.5**		

Vocabulary	Shows a good degree of control of a range of simple and complex grammatical forms. Uses a range of appropriate vocabulary to give and exchange views on a wide range of familiar topics.	Shows a good degree of control of a range of simple grammatical forms, and attempts some complex grammatical forms. Uses a range of appropriate vocabulary to give and exchange views on a wide range of familiar topics.	Shows a good degree of control of a range of simple grammatical forms. Uses a range of appropriate vocabulary when talking about everyday situations.		
				Total Points	_/5

Appendix B
Retrospective Verbal Report

Verbal Protocol - Raters
Licenciatura en Bilingüismo con énfasis en inglés
Universidad Tecnológica de Pereira

Title of study
Estimating Intra-rater Reliability on an Oral English Proficiency Test from a Bilingual Education Program

Researchers
Name: Vivian Alejandra Aguirre Isaziga **Telephone:** 317 562 6203
E-mail: alejandraisaziga@utp.edu.co

Name: Daniela Montoya Ruiz **Telephone:** 317 525 7585
E-mail:dany.montoya26@utp.edu.co

Rater #:
Date:

To give continuity to the procedure of data collection which started in the Semaforización test, you will be asked to respond to a series of questions based on your own rating process. This verbal protocol is carried out with the aim of gathering information that will be necessary for the development of this research project which seeks to explore the reliability of raters in a proficiency speaking test administered at the "Licenciatura en Bilingüismo con énfasis en inglés" program. Your answers will be recorded and the information collected here will be confidentially handled. Individual names and other personally identifiable information will not be used.

Before conducting the interview, you will be asked to listen to the audios recorded during the implementation of the speaking part of the Semaforización test. For this part, you will have to give a score once again to the five students that participated in the research using the same scoring criteria.

1. Have you been a rater in the speaking part of the Semaforización test before?
    - How many times?

2. What is your experience in language assessment?

3. What is your opinion about the rubrics used for this test?

4. How did you manage to rate two or more students at the same time?
    - What is challenging about it?

5. How was the rating environment during the Semaforización test?
    - Did you identify any distractions?

6. What similarities do you find between these two rubrics? Can you highlight them? *(Researchers will point to the rubrics from the first and the second rating. This question will be done with each pair of rubrics.)*
    - Are there similarities in the comments you provided?

7. What differences can you identify between the rubrics? Can you highlight them? *(Researchers will point to the rubrics from the first and the second rating. This question will be done with each pair of rubrics.)*
    - Are there differences in the comments you provided?

8. If differences occurred, why do you think these happened?

9. Do you think tiredness could have influenced the differences between the scores in the first rating and in the second rating?

    If so, on a scale from 1 to 5, how tired did you feel during the assessment of the last group of students the day of the Semaforización test?

10. Do you consider that experience played an important role in the similarities or differences between the scores you provided? How?

Thank you very much for your time and help.

Comments and/or Observations:

_____

_____

_____

_____

_____

_____

Appendix C
Consent Form Students

Consent to Participate in Research Study - Students
Licenciatura en Bilingüismo con énfasis en inglés
Universidad Tecnológica de Pereira

Title of study
Estimating Intra-rater Reliability on an Oral English Proficiency Test from a Bilingual Education Program

Researchers
Name: Vivian Alejandra Aguirre Isaziga **Telephone:** 317 562 6203
E-mail: alejandraisaziga@utp.edu.co

Name: Daniela Montoya Ruiz **Telephone:** 317 525 7585
E-mail:dany.montoya26@utp.edu.co

Purpose of study
You are being asked to take part in a research study that seeks to estimate the degree of intra-rater reliability of the scores from the speaking part of the "Pruebas de Semaforización". Before you decide to participate in this study, it is important that you understand what the research will involve. Please read the following information carefully and ask the researchers if there is anything that is not clear or if you need more information.

Study procedures
Your participation will only involve being audio recorded during your performance on the speaking portion from the Semaforización test. No further interviews, observations or questionnaires will be conducted before, during or after the implementation of the test.

Potential benefits
There will be no direct benefit to you for your participation in this study. However, we hope that the information obtained from this research may help the Licenciatura en Bilingüismo con énfasis en inglés program to make decisions regarding the assessing instruments, the constructs of the test, and the rater's role on the reliability of the test.

Confidentiality
The records of this study will be kept private. In any sort of report we make public, we will not include any information that will make it possible to identify you. Research records will be kept

in a locked file; only the researchers and the professors will have access to the records. We will delete the recordings after they have fulfilled their purpose, which we anticipate will be after one month of its taping.

Voluntary participation

Your participation in this research is voluntary. It is up to you to decide whether or not to take part in this study. If you decide to take part, you will be asked to sign this consent form. After you sign the consent form, you are still free to withdraw at any time and without giving a reason.

Contact information

If you have any questions or concerns about this study or if any problems arise, please contact the researchers of the study by e-mail at iropt.rp@gmail.com, or by phone at 317-525-7585.

Consent

I have read this consent form and have been given the opportunity to ask questions.
I give my consent to participate in this study.

Participant's signature Date:_____

_____

Researcher's signature

_____ Date:_____

_____ Date:_____

Appendix D
Consent Form Professors

Consent to Participate in Research Study - Professors
Licenciatura en Bilingüismo con énfasis en inglés
Universidad Tecnológica de Pereira

Title of study
Estimating Intra-rater Reliability on an Oral English Proficiency Test from a Bilingual Education Program

Researchers
Name: Vivian Alejandra Aguirre Isaziga **Telephone:** 317 562 6203
E-mail: alejandraisaziga@utp.edu.co

Name: Daniela Montoya Ruiz **Telephone:** 317 525 7585
E-mail:dany.montoya26@utp.edu.co

Purpose of study
You are being asked to take part in a research study that seeks to estimate the degree of intra-rater reliability of the scores from the speaking part of the "Pruebas de Semaforización". Before you decide to participate in this study, it is important that you understand what the research will involve. Please read the following information carefully and ask the researchers if there is anything that is not clear or if you need more information.

Study procedures
Your participation will involve scoring the speaking part of the "Pruebas de Semaforización" for fourth semester students that will be taking part at the end of the current semester. During the implementation of the test, students' participation will be recorded considering that in order to complete this study, you will be asked to score the speaking part one more time after two weeks. This means that you will give a grade to 5 of the students that you scored during the administration of the test by listening to their recordings. Finally, you will be interviewed by the researchers to collect information about your perceptions of the assessment procedures.

Risks and discomforts
There is a particular discomforts associated with this research, which is the amount of time that you will have to spend listening to the students' recordings and giving a grade one more time.

Potential benefits

There will be no direct benefit to you for your participation in this study. However, we hope that the information obtained from this research may help the Licenciatura en Bilingüismo con énfasis en inglés program to make decisions regarding the assessing instruments, the constructs of the test, and the rater's role on the reliability of the test.

Confidentiality

The records of this study will be kept private. In any sort of report we make public, we will not include any information that will make it possible to identify you. Research records will be kept in a locked file; only the researchers will have access to them. If we record the interview, we will delete the recordings after they have been transcribed, which we anticipate will be within two months of its taping.

Voluntary participation

Your participation in this research is voluntary. It is up to you to decide whether or not to take part in this study. If you decide to take part, you will be asked to sign a consent form. After you sign the consent form, you are still free to withdraw at any time and without giving a reason.

Contact information

If you have any questions or concerns about this study or if any problems arise, please contact the researchers of the study by e-mail at iropt.rp@gmail.com, or by phone at 317-525-7585.

Consent

I have read this consent form and have been given the opportunity to ask questions.
I give my consent to participate in this study.

Participant's signature

_____ Date:_____

A copy of this consent form should be given to you.

Researcher's signature

_____ Date:_____

Appendix E
Permission Request Letter

June 22, 2019

Dear ---,

We are writing to request permission to use the test scores and rubrics from the Semaforización test for the development of our research project *"Estimating Intra-rater Reliability on an Oral English Proficiency Test"*. This project seeks to estimate the degree of intra-rater reliability on the speaking part from the aforementioned test, and it is carried out as a requirement for the degree of B.A. in English Teaching.

The scores will be used with two purposes: (1) to make a statistical analysis of the grades that correspond to the language skills assessed on the test and (2) to correlate 10 students' scores in order to estimate the level of intra-rater reliability. Similarly, the rubrics will be used with the purpose of comparing the scores assigned by the raters in the first and second rating.

To ensure confidentiality and protection of the information used, we assure you that this data will not be shared with third parties. In any sort of report we make public, we will not include information that will make it possible to identify the students or professors to whom the scores and rubrics belong.

Thank you for your time and consideration of this request.

Sincerely,

_____ _____

Daniela Montoya Ruiz Vivian Alejandra Aguirre Isaziga

_____

Director of the Licenciatura en bilingüismo con énfasis en inglés