Chapter 2: Literature Review

This chapter is divided into two main parts. The first part is a discussion of anxiety. It begins with a definition of what anxiety is in general. Then, facilitating and debilitating anxiety are explained. It continues with an explanation of the three types of anxiety and defines the concept of foreign language anxiety. The second part consists of an overview of testing. It begins with an explanation of the concepts of measurement and testing. Next, the five principles that a test should follow are presented. The chapter ends with a discussion of the relationship between FLA and testing.

2.1 Anxiety

Spielberger (1983) defines anxiety as "the subjective feeling of tension, apprehension, nervousness, and worry associated with an arousal of the automatic nervous system" (cited in Horwitz et al., 1986, p. 125). According to the National Institute of Mental Health, 18% of the adult population in the United States (40 million) suffers from some type of anxiety disorder. Anxiety can negatively affect a person's social or academic life. Academically, students with anxiety have more difficulty performing in the classroom. Anxiety can occur in subjects such as mathematics and foreign language courses. The latter has been shown to interfere with the language learning process (Licet-Kiami, 2012).

2.1.1 Facilitating and debilitating anxiety

Anxiety can be both positive and negative for students who are learning a foreign language (Alpert and Habert, 1960, cited in Scovel, 1978). Facilitating anxiety is the 'good' type of anxiety that motivates learners to learn the foreign language. This type of anxiety can incite students to work hard and study more for a quiz or an exam. On the other hand, debilitating anxiety encourages learners to 'flee' the learning task. Students with

debilitating anxiety usually are not able to concentrate on a task. It also makes the learner take an avoidance behavior. A potential effect of this type of anxiety is that students may not be able to produce anything during an oral exam. If a student does not produce anything, his/her performance cannot be adequately measured. Therefore, the score that he/she is given on an oral exam may be unreliable. Test reliability is discussed in more detail in the second part of this chapter.

2.1.2 Categories of anxiety

In psychology, anxiety is generally divided into three categories: trait, state, and situation-specific anxiety (Asgari, 2013). Trait anxiety is a part of an individual's personality and thus it is a predisposition to become anxious in any situation. This type of anxiety has been shown to negatively affect cognitive functioning and memory as well as causing someone to be evasive (Eysenck, 1979, cited in MacIntyre and Gardner, 1991). On the other hand, state anxiety is not a permanent characteristic of an individual's personality (MacIntyre and Gardner, 1991). Spielberger (1966) defines it as a "transitory state or condition of the organism that varies in intensity and fluctuates over time" (p. 12). For example, the anxiety experienced before taking an exam is a case of state anxiety. This type of anxiety is likely to decrease once the student has started taking the exam. The last type of anxiety, situation-specific anxiety, is defined by MacIntyre and Gardner (1991) as "trait anxiety limited in a given context" (p. 90). Situation specific anxiety persists over time within a given situation such as speaking in public, taking an exam, or participating in a language class. A particular type of situation specific anxiety is FLA (MacIntyre and Gardner, 1991, Horwtiz et al., 1986).

2.1.3 Foreign language anxiety (FLA)

MacIntyre (1999) defines FLA as "apprehension experienced when a situation requires the use of a second language in which the individual is not fully proficient" (p. 5). Horwitz et al. (1986) define FLA in a similar fashion as "distinct complex of self-perceptions, beliefs, feelings, and behaviors related to classroom language learning arising from the uniqueness of the language learning process" (p. 128).

For Horwitz et al. (1986), the following three components provide a useful description of FLA. The first one is fear of negative evaluation. It refers to "apprehension about others' evaluations, avoidance of evaluative situations, and the expectation that others would evaluate oneself negatively" (Watson and Friend, 1969, p. 449). Fear of negative evaluation is not limited to the student's teacher but also the student's classmates. In an oral exam, anxious students may not want to say anything because they might fear that their classmates or teacher may think of them as incompetent. Again, if a student does not produce anything, their oral performance cannot be adequately measured.

The second component is communication apprehension. According to Horwitz et al. (1986), communication apprehension is a "type of shyness characterized by fear of or anxiety about communicating with people" (p. 127). Communication apprehension raises an individual's level of anxiety when communicating with others because he or she fears that his or her performance is constantly being monitored not just by the instructor but also the classmates.

The third component of FLA is test anxiety, a type of anxiety that results from the fear to fail an exam (Horwitz et al., 1986). Horwitz et al. (1986) assert that oral tests have the capability to provoke both, test and communication apprehension anxiety in anxious students but I believe that oral exams also incite negative evaluation. Apprehensive

students might not want to say anything because they may be afraid to say something incorrectly and that their instructor or classmates think that they are incapable of performing as they should.

However, not all students suffer from the same amount of anxiety; for some, the feelings associated with FLA are more intense and thus debilitating while for others, the effect is less so. Horwitz et al. (1986) were the first to propose that FLA can be measured on a scale. They developed the *Foreign Language Classroom Anxiety Scale* (FLCAS). The FLCAS can be found in appendix A. This scale was first administered to university students who were in an introductory Spanish class at the University of Texas. The FLCAS showed an internal reliability with an alpha coefficient of 0.93 indicating that the FLCAS is a very reliable instrument in measuring a student's level of anxiety. The same scale was used in the present study.

Since its development and because of its high internal reliability, the FLCAS has been widely used in a large number of research projects studying the relationship between FLA and other variables. The variables that have been studied include age (e.g., Donovan and MacIntyre, 2005; Bailey, Onwuegbuzie, and Daley, 2000), gender (e.g., Park and French, 2013; Ezzi, 2012; Capan and Simsek, 2012), language level (e.g., Marcos-Llinás and Garau, 2009) and testing. However, only three studies have been carried out on performance on an oral test (Phillips, 1992; Wilson, 2006; Hewitt and Stephenson, 2012). The latter is a replication of Phillips' study. Wilson's paper was also partially based on Phillips' study using the same oral test and rubric. Since only these studies have been done, I believe it is important to add more to the discussion and provide results using different instruments because it is significant to have a better understanding of FLA in order to demonstrate that this condition exists and that it needs to be taken more seriously by

language teachers. It is also important to study FLA and its relationship with testing because none of the three studies previously mentioned have questioned the reliability of the results of the oral test they have used. Therefore, an important aspect of the present study is testing, which is explained in detail in the following section.

2.2 Testing

Testing helps administrators, evaluators, and teachers to confirm informal assessments of the learning process and to compare students to each other as well as to external criteria over time (Douglas, 2010). Specifically in the language learning context, testing helps instructors make inferences about the students' abilities in the target language. Tests are administered for a number of different reasons. Whether they are to admit students to a certain university program, to identify in which language course they should be placed, or to examine their abilities in a language class, tests should be used ethically and they must be trustworthy so that the inferences drawn from them are as accurate and fair as possible. When choosing or designing a test, three aspects should be taken into consideration: purpose, method, and justification (Douglas, 2010).

2.2.1 Aspects of language tests

Since the primary objective in language testing is to make inferences about the students' abilities in the target language, the purpose of the test should be clearly stated, especially in high-stakes contexts. High-stakes testing refers to tests upon which significant decisions about the test-takers are based (Brown and Abeywickrama, 2010). High-stakes tests can be especially important for students because they can affect their future. If a student is admitted to or rejected from an academic program, for example, this can affect the rest of his or her life or if a student does not pass a required language course because he

or she failed the test, it could determine whether he or she stays in the program and keeps his or her scholarship (Douglas, 2010).

The method used is another important consideration when adopting, adapting, or creating a test (Douglas, 2010). Method refers to the type of test or instrument evaluators choose when testing students such as true-false questions, multiple choice questions, fill-in-the blank questions, or questions that require open answers, an essay, or even a project.

The justification of a test questions whether the chosen instrument meets the needs of the evaluator and stake-holders. It also raises questions about the consistency and quality of the test. Therefore, justification is mainly concerned with the reliability and validity of the test. These two concepts are explained in more detail in the following sections about the five principles of a good a test.

But what exactly is a test? Brown and Abeywickrama (2010) state that a test is a "method of measuring a person's ability, knowledge, or performance in a given domain" (p. 3). Since measuring is a keyword in this definition, it is important to discuss this idea in more detail. Measurement, such as assigning test scores, refers to the process of quantifying students' observed performance (Brown and Abeywickrama, 2010). Bachman (1990) advocates differentiating between students quantitative and qualitative performance. Quantitative performance concerns the assignation of numbers. Quantification has the advantage of providing exact descriptions of students' performances as well as easier comparisons between students. Qualitative descriptions provide a more individualized feedback about a student's performance such as marginal comments on a written assignment or oral feedback on a student's speaking performance.

A test is supposed to measure unmistakably a specific construct. Bachman and Palmer (1996) state that the construct is the ability that evaluators want to test. Fluency as a

component of oral proficiency is an example of a construct. To assure that a test measures a specific construct, the test must follow the following quality principles: reliability, validity, practicality, authenticity, and washback. These concepts, based on the traditional testing paradigm, are further explained below.

2.2.2 Principles of a good test

2.2.2.1 Reliability

Reliability refers to the degree accurate measurement of students' abilities a test provides (Brown and Abeywickrama, 2010). For example, if a student takes a test and receives a certain score the first time and then takes it again a second time and receives a very different score, the test is unlikely to be reliable. A reliable test must be consistent. Moreover, a reliable test must give clear directions for scoring, has uniform rubrics for scoring, and lends itself to the consistent application of those rubrics by the scorer. Therefore, if two or more persons score the same test, the results should be very similar. This refers to what is known as inter-rater reliability. The higher it is, the more reliable the test is. In addition, a reliable test contains tasks that are clear to the test-taker.

There are a number of reasons related to the test itself that turn out to be unreliable: the task might be too difficult or unfamiliar, the instructions are unclear, or there may be multiple correct answers. There are several test-taker factors that can contribute to a test being unreliable. Test-taker variables include temporary illness, fatigue, and physical or psychological factors such as anxiety. These factors may "make an observed score deviate from the true score" (Brown and Abeywickrama, 2010, p. 28) which means that the test may be unreliable. The relationship between FLA and testing is discussed in more detail at the end of this chapter.

2.2.2.2 Validity

Validity concerns the "degree of appropriacy of the inferences we make on the basis of test performance" (Douglas, 2010, p. 10). Brown and Abeywickrama (2010) suggest that the validity of a test can be found in four types of evidence.

The first type is content-related validity. For a test to have content-related validity, the achievement that the test measures need to be clearly defined. For instance, if an instructor wants to test students' speaking abilities in asking and answering questions, then students can be given a test that requires them to orally ask and answer questions in a given context. But if students were asked to write a paragraph in this topic, then the test would lack content validity.

The second type is criterion-related evidence. Criterion validity has to do with the extent to which the test measures the criterion to be evaluated. The best way to demonstrate that a test has criterion validity is to compare the results of an exam with the results of some other assessment of the same criterion. For example, if students are tested on the use of simple past in English, the results of a test given by a teacher can be compared to another assessment such as commercially produced tests in another textbook.

The third type of evidence is construct validity. Brown and Abeywickrama (2010) assert that construct validity asks the question: Does this test actually tap into the theoretical construct as it has been defined? For instance, if a teacher wants to test fluency but only takes into account rhythm and hesitation but leaves out speed and other elements of fluency, then the test is lacking construct validity since it is not assessing all of the components of fluency.

Consequential validity, also known as impact, is the fourth evidence of validity. Consequential validity includes the test's accuracy in measuring what it is intended to

measure as well as the social consequences of a test's interpretation and use. Bachman and Palmer (1996) state that the consequential validity of tests needs to be considered from a macro and micro level. At the macro level, the effect can be on society and educational systems. For example, if an assessment reform changes formal assessment to ongoing assessment of projects and only some students can be coached in those projects because only a few families can afford to pay for coaching, then the test lacks validity. At the micro level, the effect can be on the individual students who take the test. For instance, if a test result is not interpreted correctly and adequately, consequences can range from the students' failing the test to failing the language course and, in some cases, being unable to complete graduation requirements.

Face validity is an additional aspect of consequential validity. Face validity has to do with the effect the format of the test has on students. If students feel that the test does not look like it is testing what students were told it would be testing, this can lead to confusion or distraction and can therefore lead to invalid results (Bachman, 1990). Teachers should try to avoid making tests that look very different from what students expect the test to be like.

2.2.2.3 Practicality

Brown and Abeywickrama (2010) stress that a practical test stays within the budgetary limits, i.e., it is not too expensive to administer. Also, an exam is practical when it can be finished by the test-taker within an appropriate amount of time. A test that takes too long to complete is impractical. Additionally, practical tests have clear directions for administration and properly make use of available human resources. A practical exam does not exceed available material resources and considers the time and effort involved not just

for the design and administration of the test but also for scoring, meaning that it does not take hours and hours for a test to be scored.

2.2.2.4 Authenticity

Bachman and Palmer (1996) state that authenticity is "the degree of correspondence of the characteristics of a given language test task to the feature of a target language task" (p. 36). An authentic test contains language that is as natural as possible. It has items that are contextualized and not isolated. It also includes meaningful, relevant, and interesting topics for the test-takers. Authentic tests that offer tasks that simulate real-life situations are important because the purpose of any test is essentially to assess the abilities in authentic contexts.

2.2.2.5 Washback

Washback refers to the effect the test has on test-takers (Brown and Abeywickrama, 2010). A test that offers beneficial washback to students, gives feedback that will be relevant to them, and, in the best case scenario, promotes continued learning is a good test.

2.3 Foreign language anxiety and testing

The scarce literature on this topic suggests that there is indeed a relationship between FLA and testing. The findings are discussed in this section.

Philips (1992) studied the effect of FLA of students that were studying French and their performance on an oral exam. The author also intended to find out what highly anxious students were experiencing when taking an oral exam. The findings showed a significant negative correlation between anxiety levels and the score of the oral test ($r = -.40$, $p < .01$) indicating that students who showed higher levels of FLA consistently received

lower grades on the oral exam compared to students with lower levels of anxiety. Phillips also found that even though the teacher took all the commonly accepted precautions (e.g., practice communicative roles plays) to make sure students felt as comfortable as possible, students found the oral assessment an unpleasant experience. Students reported feeling highly anxious.

A replication of Phillips' (1992) study was carried out two decades later by Hewitt and Stephenson (2012), this time with students that were studying English. In this research project, the authors corroborated that there is indeed a relationship between anxiety and oral test scores ($r=-.49$, $p<.01$). In his doctoral dissertation, Wilson (2006), obtained the same result ($r=-.49$, $p<.01$) between FLA and English learners in a Spanish university. The fact that these three studies found a significant negative correlation between FLA and oral test scores implies that oral tests are not a reliable measure to evaluate oral language performance in students suffering from FLA. Therefore, if there needs to be an oral exam, it must be carefully designed taking into consideration the special needs of students that suffer from FLA.

This chapter discussed the most important concepts that are fundamental in order to understand the bases of the study. The next chapter, Methodology, explains each step that was taken to answer the three questions of the present study.