UNIVERSITY OF EL SALVADOR
SCHOOL OF ARTS AND SCIENCES
FOREIGN LANGUAGES DEPARTMENT

UNDERGRADUATE RESEARCH:

VALIDITY OF PAPER AND PENCIL TEST GIVEN TO STUDENTS OF THE B.A IN TEFL IN THE SECOND ACADEMIC YEAR, SEMESTER I - 2011

IN ORDER TO OBTAIN THE DEGREE OF:

LICENCIATURA EN IDIOMA INGLES OPCION ENSEÑANZA

PRESENTED BY:

BLANCA DORIS ESCOBAR CHAVEZ EC05020

ANA DEYSI GOMEZ DIAZ GD99006

ADVISOR:

LIC. MIGUEL ANGEL CARRANZA MsE

SAN SALVADOR, EL SALVADOR, CENTRAL AMERICA, FEBRUARY 8 2012

AUTHORITIES OF THE UNIVERSITY OF EL SALVADOR

ING. MARIO ROBERTO NIETO LOVO
RECTOR

MTRA. ANA MARIA GLOWER DE ALVARADO
ACADEMIC VICE RECTOR

IN PROCESS TO BE ELECTED
ADMISTRATIVE VICE RECTOR

DRA. ANA LETICIA ZAVALETA DE AMAYA
GENERAL SECRETARY

AUTHORITIES OF SCHOOL OF ARTS AND SCIENCES

LIC. JOSE RAYMUNDO CALDERON MORAN
DEAN

MTRA. NORMA CECILIA BLANDON DE CASTRO
VICE-DEAN

MTRO. JULIO CESAR GRANDE RIVERA
SECRETARY

AUTHORITIES OF THE FOREIGN LANGUAGE DEPARTMENT

MTRO. JOSE RICARDO GAMERO ORTIZ
HEAD OF THE DEPARTMENT

RICARDO GARAY SALINAS, M Ed.
GENERAL COORDINATOR OF
THE DEGREE PROCESSES

LIC. MIGUEL ANGEL CARRANZA MsE

ADVISOR

ACKNOWLEDGES

God has been the light on my path. He has always been with me in everything I have done. So, that is why I want to thank that an important stage in my life has finally ended. All glory and thanks belongs to him.

After god I want to thank to my dear parents Juan Gómez and Francisca Díaz de Gómez who gave me strength and patience to reach this achievement that finally did it. You have been my moral and psychological support in good and bad moments too. So I appreciate everything you have done for me and this is the result of the struggle we did together since the beginning. For instance very grateful with you.

At the same time I want to thank my dear sisters Sara Gómez and Karla Gómez for helping me and support during these years. Thank you very much sisters.
My niece could not miss Andréa Gómez because you are responsible for this work also thanks Marcy.

I want to say thanks to my partner of this thesis because she was able to finish this work with me. She was very responsible, and punctual enough to work during this period of time together. I am really thankful with you Doris Escobar.

Thanksa lic. Miguel Angel Carranza for taking the time and share your knowledge in the development of this work.

ACKNOWLEDGES

I am grateful above all to God because every single blessing he has done in my life especially, during this long and sacrificed pathway that I gone round to crown my major.

I would like to thank also to my mother for her comprehension and love, she has been my angel every moment I have needed help. My mother is the one I owe this great dream become true. She supported me in my difficult moments, when I felt down, when I thought of giving up; her words were the engine that moved me to go ahead and pursue my dreams. God bless you my beautiful and blessed mom! Teresa Chavez.

To my younger sister Sonia Escobar, I want to thank her because she was there when I needed some pieces of advice. She gave me consolation when I felt sad. She also gave me her opinion in any task I was required to do during the process of this thesis.

I want to express my gratitude to my parents and brothers who supported me in an economical way. They trusted on me since the moment I decided to study my major and I hope you are proud of me now that I have reached my goal. Thank you my loves.

To my baby Sofia Segovia to whom with her birth showed me how wonderful the life is with her by my side. Having her with me I knew the sacrifice of being a mother and a student at the same time. I took and accepted the challenge God sent me by heart.

I owe my thanks to my partner of thesis Deysi Gomez for her comprehension and patience to work together. She demonstrated me that it is totally possible to find the right person to make a good team. A person who fights with you even in the worst moments for achieving a dream you have in common. She helped me to comprehend that with dedication and sacrifice everything can be done better.

My gratitude goes also to all my professors, for the knowledge they shared with me during my stay in the university, in my major. I learned not only the target language but also the necessary tools that a good teacher must have. Little by little I was getting from them the best qualities that have helped me to grow as a person and especially as a professional in the teaching area.

TABLE OF CONTENT

INTRODUCTION

To assess students' proficiency is to make sure that they have become able to perform or display a skill, an ability or knowledge on a required subject. Testing students proficiency in the acquisition of English as a second or foreign language must also fulfil the objectives by deciding important matters when actually testing them such as: what should be tested, the types of items to use, how long it should take and whether a textbook assessment or a teacher-made test should be used. All of these matters are important because a test should be designed in a way that includes all the contents taught in class and that it accomplishes all the requirements needed to be valid.

Tests involve students writing down their responses to questions or problems. They constitute a written evidence for teachers to be able to gather a large quantity of information from students' learning. Tests are formal ways to measure students' progress or achievement. Those can be short quizzes, mid-term exams, or final term exams, all with the purpose to know the extent students have understood the classes. As a consequence, tests are important for getting information about students learning, the first important decision when preparing to assess students' achievement is to identify the information and skills that will be tested.

A valid achievement test is one that provides students a fair opportunity to show what they have learned from instruction. The most important contents of a test should be based on the objectives and actual instruction that took place during classes, so that students have a fair chance to demonstrate what they have learned when they are taking

tests. For instance, at the moment of preparing a test, a teacher must take into consideration that it has to be well designed in order to get real outcomes from students' acquisition and learning of the new language. Teachers should also know that their evaluation system must be according to the objectives stated in the program, and that any test carried out must have a good degree of validity in order to evaluate and assess the students learning in a correct way.

Since there are no previous studies about validity in testing carried out in our context, researchers consulted some authors and they also interviewed teachers from the FLD in order to enrich the gotten information to describe the phenomena of validity, remarkable characteristics of a valid test and the aspects that can invalid a test as well.

CHAPTER 1. RESEARCH TOPIC

1. OBJECTIVES

I. GENERAL OBJECTIVE:

- To know the degree of validity that tests designed by teachers of the FLD have.

(1.1) SPECIFIC OBJECTIVES:

- To analyze the tests designed by teachers from the FLD of English Pronunciation, English Grammar I, Intermediate Intensive English II and General Didactic courses.
- To reflect upon the quality of evaluation system in the four courses in the Foreign Language Department based on the validity, reliability, practicability and beneficial effects on the process of teaching-learning.
- Identify to what extent or degree teachers from the FLD take into consideration the validity of the tests at the moment of designing them.
- To recognize the different degree of validity of tests designed by teachers.

(1.2) RESEARCH QUESTIONS

***Research Area*:** Validity in Testing English as a Foreign Language**.**

***Research Topic*:** Validity of paper and pencil tests given to students B. A. in TEFL in the second academic year, semester I/2011.

***General research question*:** What is the degree of validity in paper and pencil tests given to students B.A in TEFL, in their second academic year semester I/ 2011?

SPECIFIC QUESTIONS

1. Are face, content and construct validity included in each tests design?
2. Are the topics included on the list of contents to evaluate in accordance to the program?
3. Are the written tests appropriate to measure students' proficiency?
4. Are the tests designed, their form and implementation of the evaluation system in accordance to students' level?
5. Are the tests designed by teachers good enough to be carried out in the FLD?
6. Are the instructions on the tests clear enough?

(1.3) STATEMENT OF THE PROBLEM

For many years, tests have been used to grade the domain and the ability of pupils in any area. Testing is an important part of the teaching and learning process, since it is integrated into daily classroom teaching. It is done to be part of the evaluation process and not just to be the culminating event of the program. John Clark (2003) points out, that for carrying out an evaluation teachers have to take into account the activities taught during the class. For instance, assessment includes testing, evaluation and measurement. It uses both summative and formative evaluation to monitor students' progress and structure future assignments that may be beneficial for them. Moreover, it is generally used to make a particular educational decision. Likewise, the objective and learning targets should match, and the selected assessment techniques should accommodate the specific needs of the learner to whom it is applied.

Identifying specific learning targets can help a teacher encourage the use of specific domains such as cognitive, affective and psychomotor. Specific objectives are also necessary to meet the standards that most states set for their students. In addition, they ensure that students have gained the knowledge that was originally intentioned. That is the reason why it is necessary to gauge how much students have learned and through assessments them, the teacher is able to determine whether or not pupils have gotten what was taught during the course or program. In the University of El Salvador specifically in the Foreign Languages Department teachers have to design their own tests and know how to apply new techniques in order to improve their tests. For example, they include not only

their opinions, feelings and overviews of the evaluation process, but also take into account the students' comprehension of the contents and courses they are attending.

The evaluation instruments are an essential part that teachers of the FLD apply in the teaching methodology; they have to find the most suitable way to plan them every time they are in charge of a course. The implementation of test in each subject enables teachers to see whether or not students are prepared to deal with more complex tasks in other subjects during several stages of the major. It is a good opportunity for all students to demonstrate their skills and abilities required by the course objectives. According to John Clark (2002) at the moment of evaluating students' competence and performance in the acquisition of a second language, teachers are to include two important factors: "First and foremost, teachers are to design tests or different instruments of evaluation, according to students' needs". It means that teachers must recognize that a good test should be valid, and that students' comprehension of contents studied in the class, should be reflected on their scores.

For those reasons, one of the main purposes of researchers in charge of this study is to find out the level of validity that tests have. So that, there are three important areas in this study: Face Validity, Content Validity and Construct Validity. On tests that FLD teachers are carrying out, in the light of programs of the four subjects: Intermediate Intensive English II, Grammar I, English pronunciation and Didactic I. Furthermore, researchers want to know different opinions experts have about those tests. So, the main purpose of this research is to study the tests that teachers in the second year at the FLD implement to assess their pupils and also to analyze the entire validity data gathered through primary and

secondary sources to conclude this study. Thus, this study is focused on the validity of tests and quizzes given to students of B. A. in TEFL (Teaching English as a Foreign Language), during the first semester of year 2011.

If these tests are not well designed or they are given without a specific purpose students can be puzzled and as a result this could diminish their effort and confidence, because they will feel they are unsuccessful at the moment of taking those tests. Such situation is frustrating and it affects the students' academic performance causing drop outs, or in some cases having students take the same subject more than twice. Therefore, the poor design of tests can damage the students' scores so; teachers have to pay attention in the correct design of them. "It cannot be denied that a great deal of language testing is of very poor quality. Too often language test have a harmful effect on teaching and learning, and too often they fail to measure accurately whatever it is intended to measure" (Hughes 2000:01). The worst of all is when students fail even in a third opportunity and consequently they have to change major.

For all these reasons, the study is meaningful to carry out, since its aim is to discover if the tests and quizzes administered at the FLD meet the correct degrees of validity or if those tests follow the statements written on the program for each subject because, sometimes they are totally different with what was described on them and this can affect the validity on those tests. Finally, in a test face validity, content validity and construct validity must be included in the design a good test based on the moment of testing students' proficiency in the acquisition of English as a second language in the FLD especially in the B.A TEFL in the University of El Salvador.

CHAPTER 2

II. THEORETICAL FRAMEWORK

Tests are a valuable sources of assessment information through which teachers make generalizations or predictions about their students' performance on similar unobserved tasks, Marzano (2006) to understand better this point it is really important to define the term *assessment.* That is the process of identifying, gathering and interpreting information about students' learning. The central purpose of assessment is to provide information on students' achievement and progress and set the direction for ongoing teaching and learning. The process of collecting, synthesizing, and interpreting information to classroom decision making gathering about students, instruction and classroom climate. This is because, they are not expected to find the same practice as they have seen in the class; they are provided with a sample where they have to apply what they have studied in class. Teacher-made tests are very useful to measure large quantities of information and help the teacher to make decisions about what the students have learned or what it is needed to reinforce. They are also necessary to give the students a grade.

Testing, as it is defined by specialists, "It is an activity whose main purpose is to convey (usually to the tester) how well the tester knows or can do something" (Ur, Penny 2002; 33) in this situation the test-taker has the chance to demonstrate a skill, ability or knowledge considering the nature of the test. He might succeed or fail depending on the degree of mastering of the contents he has. At the end of the test, the facilitator will be able to determine the right answers or the wrong ones through a proper grading process, and he will have gained a formal and concrete evidence of the student's performance or

competence. On the other, the student is able to confirm whether or not he has the capacity to fulfill the requirement established in the objective of a specific subject.

What is a test? A test, in plain words, is a method of measuring a person's ability or knowledge in a given domain. However, sometimes, they are not very objective to reflect what students really know, or how well students have learned from the instruction, and may lead to a misinterpretation of the students' learning. There is a great relation between the assessment data and the resulting teacher decision as well as between the content of the program and the objectives of the written evaluations given to students.

(2.1) ***KIND OF TESTS***

There are many kind of tests which have a specific purpose or a particular criterion to be measured. Therefore, it is important to have test categorization to select a suitable one. The information obtained from test concerning the process of learning a language will probably vary from situation to situation. There are four types of tests which are appropriate according to the students' necessities: proficiency test, achievement test, diagnosis test and placement test.

(2.1.1) ***Proficiency test***

They are designed to measure students' abilities in the language according to the experience they may have had in that language. These types of test are not based on the objectives or contents of a specific course, but on the previous experience the student has obtained. In addition, these tests are useful for a particular purpose. A proficiency test might be useful to determine whether or not student has the capacity to follow a course of study. A proficiency test is not intended to be limited to any one course, curriculum or

single skill in the language. These have traditionally consisted of standardized multiple-choice items on grammar, vocabulary, reading comprehension and sometimes a sample of writing. One of the most common standardized proficiency tests is the TOELF (test of English as a foreign language) (Brown, 2001)

(2.1.2) ***Achievement test***

These types of tests are the most common in second language classroom, for they measure student's progress in their own learning. In contrast, to proficiency test, achievement test are directly related to language courses; the specific purpose is to determine how successful students have been in achieving the objectives of the course. These tests can be helpful divided in two types: final achievement test and progress achievement test. The former is administered at the end of the course and it usually includes all the topics or chapters taught during a period of time. In change, progress test measure the progress students have had at the end of a unit or topic taught. They usually determine how successful students have been in the mastering of specific abilities studied.

(2.1.3) ***Diagnostic test***

These are usually useful to identify students' strength and weakness. They are important at the beginning of courses to realize about students' needs in order to create profiles about their abilities and knowledge. They might be extremely helpful when teachers want to develop an excellent long-term plan of instruction and more clear content on the objectives. Nevertheless, diagnostic tests are almost never applied at the beginning of the courses, since they are difficult to design.

(2.1.4) ***Placement test***

As its name suggests it provides information to place students in a stage or level according to their current knowledge and abilities. They are useful to indicate students the current capacity they have towards the language. And important point in placement tests is that it can be presented as either as an oral test or as a written test, depending on what it is measured in the student. In general, there is a combination of the two tests, oral and written, to find out what is the correct level of the student to place. Certain proficiency tests and diagnostic tests can act in the role of placements tests, whose purpose is to place a student into an appropriate level or section of a language curriculum or school. A placement test typically includes a sampling of material to be covered in the curriculum. (Brown, 2001)

(2.2) ***Types of test items***

There are two basic types of paper-and-pencil test questions: selection items and supply items. As they suggest, selection items require the students to select the correct answer from among a number of choices, while supply items require the student to supply or construct his or her own answer. Within the general category of selection items are multiple-choice, true-false and matching questions.

(2.2.1) ***Multiple-choices items***

Multiple-choice items consist of a stem, which presents the problem or question to the student, and a set of options, or choices, from which the student selects an answer. The multiple-choice format is widely used in achievement tests of all types, primarily to assess learning outcomes at the factual knowledge and comprehension levels. However, with

suitable introductory material, this format can also be used to assess higher level thinking involving application, analysis, and synthesis.

Here are examples of multiple-choice items:

- Complete the statement or question. Choose the correct answer.

1. What _____ right now? Are you busy?

a. do you do
b. are you doing
c. you can do

2. _____ bread do we have in the cabinet?

a. Is there any
b. How many
c. How much

3. What time _____ lunch?

a. did Frank have
b. Frank had
c. was Frank

(2.2.2) ***True-false items***

The true-false format requires students to classify a statement into one of two categories: true or false, yes or not, correct or incorrect, fact or opinion. True-false items are used mainly to assess factual knowledge and comprehension behaviour, although they also can be used to assess higher level ones. The main limitation of true-false questions is their susceptibilities to guessing. *Sample of true-false test item*

Guess which of the following statements are true or false?

1. Apples, not caffeine, are more efficient at waking you up in the morning.

2. A pack-a-day smoker will lose approximately 2 teeth every 10 yrs.

3. People do not get sick from cold weather; it's from being indoors a lot more.

4. When you sneeze, all bodily functions stop, even your heart!

5. Only 7 per cent of the population is lefties.

(2.2.3) ***Matching items***

These consist of a column of premises, a column of responses and directions of matching the two. The matching exercise is similar to set of multiple-choice items, except that in a matching question, the sample set of options or responses is used for all the premises. Its main disadvantage is that it is limited to assessing mainly lower level behaviours. (Brown, 2001)

Sample matching test item

Directions: On the line to the left of each painting style, write the letter of the statement that best explains the style. There is one more definition than painting styles.

Painting Styles	**Explanation**
1. Abstract expressionism	**a.** style which combines naturally unrelated events, images, objects, or situations in a dreamlike scene.
2. Surrealism	**b.** style which makes geometric shapes of color that interact subtly with the backgrounds of similar intensities
3. Pop art	**c.** style in which the artists use a spontaneous method for creating art.
4. Gothic art	**d.** style that depicts objects or scenes from everyday life and employs techniques of commercial art.

(2.2.4) ***Supply items***

Supply items consist of short-answer and completion (also called fill-in the-blank) items and essay questions.

Short-answer and completion items

These items are very similar. Each presents the student with a question to answer. The short-answer format present s the problem with a direct question (e.g., what is the name of the first president of El Salvador?), while the completion formant presents the problem as an incomplete sentence (e.g., the name of the first president of El Salvador is_____). In each case, the student must suppose his or her own answer. Typically, the student is asked to reply with a word or phrase, number or sentence, rather than with a more extended response. Short-answer question are fairly easy to construct and diminish the like hood that students will guess answers. However, they tend to assess main factual knowledge or comprehension.

*(2.2.5)****Essay items***

Essay items give students the greatest opportunity to construct their own responses, making them most useful for assessing higher level thinking processes like analyzing, synthesizing and evaluating. The essay question is also the primary way teachers assess student's ability to organize, express, and defend ideas. The main limitations of essays are that they are time consuming to answer score, permit testing only of a limited among of student's learning, and place a premium or writing ability. (Airisian, 2000)

(2.3) the table below presents both pros and cons for various test item types.

ITEM TYPE	***PROS***	***CONS***
Multiple choice	More answers options (4-5) reduce the chance of guessing that an item is correct. Many items can aid in student comparison and reduce ambiguity. Greatest flexibility in type of outcome assessed: knowledge goals, application goals, analysis goals, etc.	Reading time increased with more answers. Reduces the number of questions that can be presented. Difficult to write four or five reasonable choices. takes more time to write questions
True/False	Can present many items at once. Easy to score. Used to assess popular misconceptions, cause-effect reactions.	Most difficult question to write objectively. Ambiguous terms can confuse many. Few answer options (2) increase the chance of guessing that an item is correct; need many items to overcome this effect.
Matching	Efficient Used to assess student understanding of associations, relationships, and definitions.	difficult to assess higher-order outcomes (i.e., analysis, synthesis, evaluation goals)
Supplied Response	Chances of guessing reduced. Measures knowledge and fact outcomes well, terminology, formulas.	Scoring is not objective. Can cause difficulty for computer scoring.
Essay	Less construction time, easier to write. Encourages more appropriate study habits. Measures higher-order outcomes (i.e., analysis, synthesis, or evaluation goals), creative thinking, and writing ability.	More grading time, hard to score. Can yield great variety of responses. Not efficient to test large bodies of content.

(2.4) How do teachers know if a test is good or not?

Before designing a test and then giving it to a group of students, there are a number of questions teachers need to ask. Is it administrable within a given constraints? Is it dependable? Does it accurately measure what teachers want it to measure? These questions can be answered through three classic criteria for testing a test: ***Practicality, Reliability and Validity.*** (Brown, 2001) that are described in the followings statements.

(2.4.1) PRACTICALITY

According to brown (2001) A good test should be practical. It is within the means of financial limitations, time constraints, ease of administration, and scoring and interpretation. A test that is prohibitory expensive is impractical. Besides, a test of language proficiency that takes a student ten hours to complete is impractical. A test that takes a few minutes for students to take and several hours for examiner to evaluate is impractical for most classroom situations. A test that can be scored only by computer is impractical if the test takes place a thousand miles away from the nearest computer. So the correct way to design a test has to include the practicality in all the senses to improve the criteria for testing a test.

(2.4.2) RELIABILITY

Reliability is the consistency of teacher's measurement or the degree to which an instrument measures the same way each time it is used under the same condition with the same subjects. In short, it is the repeatability of their measurement. A measure is considered reliable if a person's score on the same test given twice is similar. It is important to remember that reliability is not measured, it is estimated. A reliable test is consistent and

dependable. (Brown, 2001) Sources of unreliability may lie in the test itself or in the scoring of the test, known respectably as test reliability and scorer reliability. If teachers give the same test to the same subject or matched subjects on two different occasions, the test itself should yield similar results. There are two ways that reliability is usually estimated: test/retest and internal consistency.

(2.4.3)Test/retest

Test/retest is the more conservative method to estimate reliability. Simply put the idea behind test/retest is that you should get the same score on test 1 as you do on test 2. The three main components to this method are as follows:

1.) Implement your measurement instrument at two separate times for each subject;
2). Compute the correlation between the two separate measurements; and
3) Assume there is no change in the underlying condition (or trait you are trying to measure) between test 1 and test 2.

(2.4.4) Internal consistency

Internal consistency estimates reliability by grouping questions in a questionnaire that measure the same concept. For example, you could write two sets of three questions that measure the same concept (say class participation) and after collecting the responses, run a correlation between those two groups of three questions to determine if your instrument is reliably measuring that concept.

(2.4.5) VALIDITY

Validity is the extent to which a test measures what it claims to measure. It is vital for a test to be valid in order for the results to be accurately applied and interpreted.

(Brown, 2001) Validity refers to the results of a test and not to the tests itself. Validity is not determined by a single statistics, but by a body of a research that demonstrated the relationship between the test and the behavior it is intended to measure. And there are three types of validity that can help to clarify better the design of a good test.

(2.4.6) Differences between reliability and validity

Validity refers to the meaning of a test score or assessment result, whereas reliability is the consistency of a score or result. Consistency of assessment results is established over time, agreement among test items, and agreement between raters. A test that is not reliable cannot be valid; therefore, educators should evaluate reliability evidence before they even consider validity evidence.

To be valid a test has to meet three important characteristics: *face validity, construct validity and content validity.* The first one *face validity*, a test is said to have face validity if it looks as if it measures what it is supposed to measure (Hughes, 2000) for example, a test which pretended to measure pronunciation ability but which did not require the candidate to speak might be thought to lack face validity. For instance, face validity is hardly a scientific concept, yet it is very important. A test which does not have it may not be accepted by candidates, teachers, education authorities or employers. The second one *construct validity* has to do with the design or structure of a test. A test, part of a test or a testing technique is said to have construct validity if it can be demonstrated that it measures just the ability which it is supposed to measure. The word *construct* refers to any underlying ability or trait which is hypothesised in a theory of language ability. In addition, construct validity examines whether test performance reflects an underlying construct or set of related variables. And finally, *content validity*, has to do with the inclusion of proper samples of

the relevant topics seeing in class. In other words, a test is said to have content validity if its content constitutes a representative sample of the language skills, structures, etc. with which it is meant to be concerned. It is obvious that a grammar test, for instance, must be make of items testing knowledge or control grammar. The test would have content only if it included a proper sample of the relevant structures. Just what are the relevant structures will depend, of course, upon the purpose of the test. A comparison of test specifications and the test content is the basis for judgment as to content validity (Hughes, 2000).

These three characteristics of the test validity will guide the interest of this research. *Validity* is concerned with whether the information obtained from the test leads to make correct decisions about the students' learning. Such a point is discussed by authors like Gilbert Sax and Hughes (2000:01) who argue about the necessity to design a good test, a valid test let students show what they know, however, that does not reflect their entire knowledge of the studied matter. A test let students reinforce what has been taken place in class, since they prove themselves how much they acquired. A test cannot be an instrument to punish any misbehaviour in the class, or to make pupils fail and stumble in their academic performance, it is only made to measure students´ knowledge of the class up to that moment. What is really required for a reliable test is to have a purpose, to reflect how much students have achieved, and to let the teacher grade the pupils fairly.

CHAPTER 3 METHOD

III. METHODOLOGY

To verify the level of test validity, samples of quizzes and tests have been gathered from those used in the second academic year during semester I and II/ 2011 of the B. A. in TEFL. These tests were obtained once they have been graded by teachers. These tests and quizzes have been analyzed, in the light of their respective programs or syllabuses, in order to see if the design of contents and objectives in the syllabuses match with the written instruments the teachers carry on during the evaluation process. Also a checklist has been used by the researches to determine the construct validity of each single instrument. The checklist is very useful to discriminate possible gaps or problems which can affect the validity of a test. Interviews have been done to teachers in charge of these four subjects: Intermediate Intensive English II, Grammar I, English pronunciation and General Didactics. Some specialists in testing design have been also interviewed to know the degree of validity teachers manage and how they put into practice when designing a test.

(3.1) THE STUDY POPULATION

The population for this study is students of B. A. in TEFL who are in the first term of this academic year, at the University of El Salvador in the Foreign Languages Department. Of all the students' groups as possible candidates for this study, a 15% sample is being used to gather the information related with the research topic.

(3.2) THE STUDY SAMPLE

During the current semester I/ 2011 around 160 students are attending the B. A. in TEFL in the second academic year, but students from modern language are not included in this research. For this reason, the gathering of data for this project has been done in the respective courses being attended by students in English Pronunciation, English Grammar I, General Didactics and Intermediate English II. Students' tests have been asked at randomly to provide the first body of the data.

CHAPTER 4 RESULTS

IV. DATA ANALYSIS

This research is based on the degree of validity that tests provided by the teachers of the FLD have. To find out this, it was necessary to make a deep analysis of those tests. For this reason, the topic was divided into three main areas: *face validity, content validity and construct validity*. This was done with the purpose of recognizing the level of validity the tests had. And the tests that were checked by specialist and researchers had a low degree of validity and most of them do not have face validity, neither construct validity nor content validity. Also tests designed had problems, and the worst of all is that they do not measure the real level students are attending. Moreover, during the analysis of the sample tests, it was evident that the group of teachers does not design appropriate tests, because many of them have never received any kind of training in this field.

The consequences of those tests or bad design fall on students grades, owing to most of the time many of those pupils misunderstand the real aim of the exams. Also, to be a good teacher means to be aware that the evaluation system, made at the beginning of the course, must be followed to measure exactly what is intended to measure. In this case, students' proficiency of the target language as a main aspect. In addition, teachers need to recognize that in a course the evaluation system must be tied with the course objectives that have been stated in the program and what they want to reach at the end of the course is to help to improve students learning. Therefore, if teachers from the FLD want to become effective teachers, they must be trained to contribute with the learning process for the

students that are and will study this major. Especially, because in this major students need to develop the four macro skills and then teachers should know that every single test has to be made in a professional manner in order to assess them in all the areas, since they are the ones who play a crucial role during the process of the acquisition of a second language.

CHAPTER 5 FINDINGS

V. TEACHERS' INTERVIEWS ANALYSIS

To start with, this is going to be a deep description about the analysis of the opinions gathered through the interviews administered to some teachers from the FLD. This study includes the analysis of experts' opinions and researchers' own analysis of the results. It is important to mention that to collect all the data, a questionnaire was designed in order to interview the teachers, who were in charge of the subjects studied in this project. Another instrument used in the measurement of test validity was a checklist which was designed taking into account the three main areas that were included in the topic: *Face Validity, Construct Validity and Content Validity*. To have better point of view about the studied tests, researchers also analyzed the same exams that experts had analyzed and criticized before, they got the same observations, and the results were almost similar in both cases.

In addition, a group of teachers was chosen at random among those who had taught the subjects involved in this research: Intermediate Intensive English II, English Pronunciation, Grammar I and General Didactics. This group of teachers was informed about the purpose of the interview. Then, the interview was scheduled with each teacher. Next step was to know how they evaluated their students formally or informally according to the evaluation system they have. Most of them said they provide the traditional forms of evaluation which include written quizzes, mid-term exams, and final exams and incorporate the assessment of the four macro skills or a specific skill, depending on the subject they teach.

The way teachers evaluate students in an informal assessment is that they take into consideration students' participation in class, some oral tasks, the performance and proficiency they have when attending classes, and in general students' responsibility towards their own learning. It is important to say that only one teacher in particular said that he makes a diagnostic test to measure if students have a domain of every single topic studied in the class. If they show low understanding of it, the teacher looks for extra information to help them to improve their learning, by doing this, he not only helps students to become more accurate in the target language, but also contributes to increase their proficiency on it.

Another aspect being asked was about the characteristics that make a good test. In this point, teachers' opinions were that a good test is the one that reflects exactly what has been taught in class. Also, if it covers with all the requirements established in the program. Moreover, during the interview teachers were asked about previous knowledge about validity and reliability and it was found out that, only two of them knew exactly what those words meant. "Validity is a complex concept, yet it is indispensable to the teachers' understanding of what makes a "*good*" test" (Brown, 2001). Teachers also pointed out that those two elements are quite important to take into consideration at the time of designing a test because it must always have a good grade of validity.

Teachers were also questioned if they used teacher made tests or software generate tests. And all of them responded that they use those kinds of tests but trying to improve them because, in their opinions, textbooks and software generated test are too easy. So that, according to these answers, they were asked if they considered that these types of tests

indeed measured what they were supposed to measure. Most of them said that actually those ones "do not measure the proficiency of students and they do not allow students to think in a critical way, besides that, it is not a challenge for them due to the degree of difficulty that they present and sometimes those exams are not related to the objectives of the course". Surprisingly, one teacher said he does not ever do them, because it differs in the type of subject he is in charge of. He also added that **"**it is better to design our own tests because in that way we know exactly what we are evaluating.**"**

Some teachers adapt new forms of designing tests; they also stated that some other teachers implement textbook tests for quizzes as well as software generated tests. In this section they were asked about their opinions on these types of tests and they said that they are not totally related to the real instructions taught during classes. For instance, it is not a fair opportunity for students to demonstrate the domain or knowledge they have really learned during classes. So, those decisions are sometimes misunderstood by people who are experts in the area of tests design. It is considered as an inappropriate way to accomplish with an important role as a teacher, since they can damage the learning process of their students which means carrying out tests with a low grade of validity. A reason for mistrusting test is that very oftenly they fail to measure accurately whatever it is that they are intended to measure (Hughes, 2000). In this case it is very important to design each test because students can fail even thought they have studied, the real problem could be the bad design or instruction of the test.

Teachers were asked if they design their tests with a high level of validity and reliability aspects and how they knew that. Teachers said that their tests had good levels of

validity and reliability, because they design them themselves, they are careful to include what they had taught in the class. "Three types of validation are important in the role as a classroom teacher: content validity, face validity and construct validity". (Brown, 2001). In contrast, only one teacher mentioned that he does not know if his tests have validity or reliability. Teachers said that because validity and reliability are important traits in the tests, they have to take these into account every time they are going to administer a test to their students.

Another main point was FLD teachers' training or formal preparation on test design and evaluation system. Only two interviewed professors said that they had a previous training on how to design good tests. The rest of them answered that they have never taken any training or course on how to design valid and reliable tests, but they have read a lot of information about this issue in order to help themselves to manage it at the moment of carrying out a test. Moreover, researchers wanted to know if teachers are satisfied with the paper and pencil materials they administer to their students. Teachers responded “yes" most of them only use traditional assessment and they try to make all students use the language in a communicative way.

Teachers were asked how they consider tests affect students and almost all of them had the same opinion toward this issue. They said students get nervous at the time to take an exam, and this affects students a lot, since it prejudices them at the time to give a logical answer. When they get afraid, they forget everything though they have the enough knowledge of every single topic being asked on it and sometimes they dedicate a lot of time to get ready for it, but they are blocked and get bad grades. It might be that students create

barriers in their minds when doing the tests, especially when they know that it is a difficult subject and they can fail it. One teacher said that if she observes that a student is really nervous, she asks him/her not to do the exam until another day when she/he feels prepared and less nervous. "When students are prepared for a test it is the teacher's responsibility not only to help them get their English level required, but also familiarize them with the kinds of exam items they are likely to encounter, and give them training in how to succeed" (Salkind, 2000). In accordance to all the teachers from the FLD, a test can affect students in a negative or positive. Next, they said if the test is really poorly done everybody approves it or if it is so difficult most of them fail, so these are some factors that can affect pupils too.

(5.1) ANALYSIS OF EXPERTS' OPINIONS

It was essential to count with the support of experts in the Testing field. This is because they shared opinions and thoughts with the researchers about what they consider are the best elements to make a good judgment of designing a valid and reliable test. The characteristics it should posses to match with the course objectives and the goals the teachers planned at the beginning of the semester. Taking into account that teacher' decisions can contribute or diminish the learning of the students every time they design a test.

This group of highly experienced professor (experts) helped to achieve the goal of holding a better overview about the paper and pencil instruments that teachers from the FLD applied to their students during the first semester of the current year. Experts' opinions not only assisted to clarify many doubts that emerged at the beginning of this project, but also contributed to notice which are the weakest areas that teachers in charge of

those groups have when preparing their tests in any subject. Especially, because many of them are in charge of those groups year after year. Before looking for experts' opinions, a checklist was designed carefully to be used as researchers' instrument to gather almost all the vital data to support their research. A questionnaire was also made in order to make the teachers' interviews and to complete the other part of this work. These two instruments contained real statements.

A group of four experts was chosen to accomplish the analysis of the tests, since they are recognized because of their experiences at the design of paper and pencil tests; besides of the quality that the exams have when they make them. On the other hand, a determined number of exams with their respective checklists were given to the experts group. Each group of tests was divided into subjects. These subjects analyzed were: English grammar I, English pronunciation, Intermediate Intensive English II and General Didactics (which was taught in Spanish) related to validity in testing. The validity of every test was divided into three main areas: *Face Validity, Content Validity and Construct Validity.*

Moreover, in each area a group of eight to ten statements were written to criticize each respective exam and detect errors or misunderstanding that it could have. If there was something wrong, experts may add something to help to improve them. A determined space were left in each checklist in which all the experts were to suggest and write any important comment to facilitate the enrichment of this research. At the same time, one checklist was modified, because one subject included in this research was in Spanish, and the others were in English, so it contained some changes due to the things that the translation implied.

Therefore, with the purpose of reviewing the tests, experts were scheduled with a period of two weeks to conclude with their analysis, and those ones were the outcomes gotten by the researchers shared out in subject.

(5.2) GRAMMAR TEST ANALYSIS

This is a brief description about the results gathered in English grammar I tests. First of all, one point that the grammar expert claimed was that every single test must be designed from the easiest to the more complex section of items; however, in one exam there was a mixture of all the items being this an untidy test. He also expressed that a test should be organized according to the type of items included on it. This means that selections items have to be in a section (true or false, multiple choice and items matching) and supply items must be in another section (short-answer and complementation and essay items). Besides, the expert expressed that in one exam there was only one kind of items and for him it was not well designed, because they do not allow students to think in a critical way at the time to do an exam. And they help students to get only good grades due to the facility of them. According to the expert's opinion, it is not adequate in an academic level. There should be a blend of items to make the test more challenging for students.

He also added that the instruction in these exams was not clear enough. He discovered that some topics were not related to the contents stated in the program, and for instance, the objectives of the course were not connected with the real ones. These exams did not have established objectives that helped to understand the tests better. He also claimed that, there was a very large paragraph and students did not have the enough time to

complete it. He said this because each grammar course count with an hour per day and to complete the exam in that period of time it was a little hard for students according to his criterion. It means that almost all of them did not finish it and consequently, these teachers error were reflected in students' grades.

Finally, this teacher in charge of the analysis of this subject argued that in the FLD, English grammar tests are poorly designed; and in some cases these kinds of tests can confuse students who can misunderstand the purpose of the teacher who makes the exam. "Test items that do not reflect the important topics of instructions are not valid indicators of student's achievement" (Airisian, 2000). In accordance with researchers' own experiences, this type of tests can prejudice students' performance, due to the low degree of validity and reliability that they present.

(5.3) ENGLISH PRONUNCIATION

According to the expert in charge of the analysis of English Pronunciation, these tests that teachers from the FLD carried on are well designed. He pointed out that these kinds of exams achieve most of the requirements to take into consideration to be a good test. But, in his opinion there are only some wrong things on them that should be mentioned to help teachers in charge of this subject to improve the application of well designed tests such as: the teacher in charge of making the test did not write clear instructions, and it could make students to feel confused at the time to have the exam in their hands. A second thing was that, the test said that students should read a passage, but it was not included in

the exam, it means that the exams were incomplete. A third thing to argue is that they did not have any percentages.

Exams did not state a word to be read by students at the time of listening that part. The expert also mentioned that, in his own opinion, these tests had a good level of validity and reliability and those have some few observations but they are appropriate for being handed over in the FLD. Furthermore, he said that this particular subject does not require much critical thought, due to this subject is based on the development of the listening skill and the good pronunciation of the new language. For this reason, the items implemented in them are according to the listening part that teachers in charge of this subject can give to their students.

(5.4) INTERMEDIATE INTENSIVE ENGLISH II

The next expert helped by giving a brief explanation about the test designed by teachers of the TEFL. However, his words were good enough to clarify the way some teachers should design future tests. For example, he mentioned that some tests he checked were so perfect that could not be designed in the FLD. The expert claimed that some items were gotten from a textbook and also from software generated tests. In some tests he found out some extracts of TOELF, teachers just copied and pasted them. But, he added that this kind of issues can be used by beginner teachers because those one have no sufficient experience; but for those who have more than 10 years of experience in working in the FLD is not recommendable.

They should work as a team and take their time in order to design the evaluation system for enhancing every single test that is passed in the FLD. This is because, it is eminent that when teachers do not have time to prepare an exam or use the easiest way to do it, that looks bad design and done in a rush way. The expert suggested that professors should take a training or preparation on how to design a good test including in them a good level of validity and reliability. Last, some tests were not related with the ones described in the program. Finally, some statements were not checked by the expert because in his opinion it was like grammar section included in English tests. Tests were poorly designed. In addition, the expert and researchers' thoughts were that in those kind of tests is very difficult to find out any mistake because they were almost perfect, in spite of their contents were different from the program.

(5.5) GENERAL DIDACTICS

The next expert gave his opinion about this subject. He was more emphatic and he said these kinds of tests were poorly designed for an academic level. They were confusing; the items used were poorly constructed; so there were some problems. He said those tests did not have enough items. Another trouble was the use of ambiguous items in the same test, one test did not have face validity because the subject is General Didactics in Spanish and in the test was included some grammar structures, and other topics that were not stated in the syllabus. Also, the items lead to the correct answers and did not allow students to think in a critical way. Another point was that the instructions were not clear, they were ambiguous. "Test items should be brief, clearly written, and free of ambiguous words so that comprehension is not an issue" (Airisian, 2000). To make students' comprehension

easier a test has to be clear enough to give the students the best opportunity to put into practice their domain in a topic.

Besides that, there were not distracters; the test had a lot of responses related with the same item. According to the expert the person who designed these kinds of tests had a purpose that the entire group passed the subject with that kind of instruments used. Moreover, these tests were not related with the contents established in the program and they did not have any objective to achieve. For instance, the tests carried on in this didactics course were not valid enough to be included as a paper and pencil instrument tests, according to the expert in the subject. The researchers had the same opinion about the tests checked because they were poorly designed and tests had a very low level of validity in the three main areas of the research although these tests were designed in Spanish.

(5.6) RESEARCHERS' OWN ANALYSIS

To conclude this work, a brief analysis of the results obtained by researchers was done in order to make more credible the outcomes gotten from different resources that this study was based on. To gather all the essential data first of all, people in charge of this work had the goal of including their own point of views. Researchers analyzed the same quantity of tests that experts checked one by one and the instruments in order to analyze the four subjects in this study. After that, researchers compared their results with the experts' ones to be able to discover similarities or differences as well. In spite of that it helped to have a better and clear opinion concerning with the teachers' tests.

In this analysis it was found that most of the teachers from FLD used software generated test. And they are administering tests to their students with low and intermediate levels of validity and also poor design of tests. According to experts' and researchers' opinions when they criticized each test by using a check list experts wrote some observations for every single test they checked and researchers analyze their outcomes and these were the results researchers found. For instance, in some cases this kind of tests can prejudice student's scores in any manner. Among some characteristics to mention about this kind of test are: tests were too long, some of them did not have clear instructions, tests do not have any relation with the subjects taught by teacher and the course objectives were different from the program. Also, teachers of the FLD have never been trained in the issue of designing good tests and it could be observed at the time to make a quick review to their tests.

CHAPTER 6 DELIMITATIONS AND LIMITATIONS

VI. DELIMITATIONS

This work took place at the University of El Salvador in the Foreign Language Department, specifically in the English courses students of B. A. in TEFL are taught in the first and second term of this academic year.

(6.1) LIMITATIONS

This work was related to the lack of previous study regarding to validity in testing. This research is a compilation of primary data. It is important to remark the difficulties researchers have had to develop this research due to the contradictions found to design it properly. Some limitations were related with time assigned to teachers interviewed, they gave a specific date and time, however, when researchers dropped by their offices they were busy or were not there. Besides that, in those days, the University of El Salvador was carrying out an election process, there were some teachers in meetings almost always, and so they did not have time to give any interview. So that, researchers did not count with their support in the development of this work because they were very busy in others matters. In contrast, researchers had the collaboration of some experts to give their opinion, the limitation found here was that some of them took more than two weeks to check the instruments assigned although they did it for free.

CHAPTER 7 RECOMMENDATIONS AND CONCLUSIONS

VII. RECOMMENDATIONS

The development of this project was very important because paper and pencil instruments include some elements that must be treated in a careful way. Some are the students' outcomes in the grades, the skills they have gotten and the domain in a respective subject, that is, all these elements will be fulfilled depending on the teacher's wrong or good design of the test. But, according to this work researchers have found several anomalies described by some experts in tests designed by teachers at the FLD. There are some recommendations for teachers who want to improve their teaching and evaluating methods.

- Teachers should determine when to use a long or short exam according to the time students' have to complete it. It is important that teachers from the FLD take into account the time that a group of students has to solve tests. Sometimes this could affect them in a given way, because according to some students' outcomes, sometimes they can not fulfil a test because they do not have the enough time to conclude it.

- (Teacher made test are designed to evaluate students on what has been taught with respect to the curriculum whereas standardized test are designed to test on the subject matter without regard to standard reference or curriculum) Standardized tests have an advantage about teacher made test because, people who design tests could pilot them to discover the level of their validity. So, it means that the teachers' staff in the FLD

should do that before administering the tests and it that way they could be almost sure that there will not be any inconvenience for students to accomplish their academic performance at the time of responding an exam.

- If teachers work as a team at the beginning of the semester, it is an advantage because they are going to have an agreement about the matter of testing students. It is recommendable that all of them administer the same test, but if some of them make some changes on it, they should suggest a second or third option of test to avoid confusion at the time to compare teachers' tests.

- In some academic institutions there is a person in charge of checking the tests before these are administered to students. The FLD should have a person or committee who has enough experience in designing tests in order to suggest or make recommendations for improving the evaluation system.

- Teachers should take trainings on how to design a good test in order to improve the teaching and learning process in the FLD. It is really necessary not only in this department it should be included in all the majors. It could help them in the designing of good tests taking into account the students' level and the creation of appropriate tests. All of this done totally related with the course objectives stated in each program.

- Teachers in charge of the most important subjects as didactics should encourage students to pay attention in the validity of test, design of tests, type of items to use why to use them and also vocabulary to put into paper.

- In the University of El Salvador specifically in the Foreign Language Department there should be a subject related to evaluation and testing in order to support the lack of knowledge in testing in El Salvador to be the first university in having this necessary and important matter.

(7.1) CONCLUSION

Testing language skills is a crucial part of language teaching. Teachers' obligations are both challenging and demanding at the time of making an evaluation of students' proficiency in the acquisition of a second language. Testing gives significant data about pupils' true level as well as a chance for themselves to commit with their own learning process. Taking a test demand to study and get prepared for facing such duty. Thus, the designing of a good test is important because through it, the teacher can get information needed from the learner concerning to their learning. Furthermore, constructing a good test can be time consuming and a difficult task. It is hard to understand why something so essential to the learning process has not been virtually well developed.

Teachers do not take any course to be trained in this crucial subject. As a matter of fact, in the University of El Salvador there is not any available training in this field and not even in the country. For this reason, it is important and necessary that teachers take courses in assessments, because as it has been observed many of them do not have the certain criteria against which test can be measured in this case validity, reliability and practicability or how to create meaningful tests. "today tests designer are still challenged in their quest for more authentic, content valid instruments that simulate real-world interaction while still meeting reliability and practicability criteria" (Brown, 2001) So that, a test or another evaluation instrument has to be connected to the real instruction studied during the classes, and in that way they can provide learners a fair opportunity to give the best abilities they have gotten in their learning process.

During this analysis, it was noticed that almost all the teachers from the FLD use the software generated tests to make evaluations to their students. Also it was found that they only make some modifications and changes to them in order to these tests seem as if they have been designed by the teachers. The implementation of a good evaluation system must be according to the level students are attending, in this case the evaluation system carried out by teachers from the FLD is really poor and it does not reach the level of validity and reliability that it should have. Since teachers are the ones in charge of these important aspects, they must be prepared to provide their students good tests that accomplish with all the vital requirements to be an efficient test. They should be able to measure students' proficiency and accuracy in the second language; it means that, a test that could contain a good degree of validity and reliability.

As a result, teachers of the FLD do not have a pattern or any guidelines of how to design their own tests; almost all of them have not been trained in this area. The ones, who have received any training, claimed that they have the knowledge but they apply textbook tests, because it is an easy way to assess their students, though they know that those tests do not allow students to think in a critical way. But, those kinds of tests implemented by this group of teachers are poorly done this was said by experts and sometimes they affect students in a negative way. They get bad grades and sometimes fail the subject, owing to the fact that those instruments were bad designed and they had low and intermediate level of validity. They were bad constructed not covering all the contents studied in the class and in the worst case, it is included contents that are not stated in the course objectives; so it could be said that those tests have a low and intermediate validity level but they are careless designed.

APPENDIX

UNIVERSITY OF EL SALVADOR
SCHOOL OF ARTS AND SCIENCES
FOREIGN LANGUAGE DEPARTMENT
GRADUATION WORK 2011

TOPIC: VALIDITY IN TESTING

CHECKLIST

OBJECTIVE: To analyze the validity of paper and pencil instruments of INTERMEDIATE INTENSIVE ENGLISH I; given to students of the second year of the TEFL for the term 01/2011.
COURSE: INTERMEDIATE INTENSIVE ENGLISH II
EXPERT: _____DATE:_____
INSTRUCTIONS: Read each item and put a check where you deem appropriate.

Completely disagree	Disagree	Neutral	Agree	Completely Agree
1	2	3	4	5

FACE VALIDITY	**1**	**2**	**3**	**4**	**5**
1. The test looks like an English test.					
2. The test looks carefully constructed and well organized.					
3. All the items have comprehensible vocabulary according to student's level.					
4. It does not have ambiguous questions and ambiguous instructions.					
5. The test is legible.					
6. There is enough space between questions to facilitate easy reading and writing.					
7. The test is clearly doable within the allotted time limit.					
8. The test is practical and the instructions governing it are simple and direct.					
9. There is any evidence to think the test was prepared hastily.					
10. Stated items are easy and comprehensible in order to avoid confusion.					

Comments:_____
_____
_____

INSTRUCTIONS: Read each item and put a check where you deem appropriate.

Completely disagree	Disagree	Neutral	Agree	Completely Agree
1	2	3	4	5

CONTENT VALIDITY	**1**	**2**	**3**	**4**	**5**
1.The test constitutes a representative sample of any language skill					
2. It includes a proper sample of a relevant structures taught in the course.					
3. The structures are relevant according to the purpose of the test.					
4. The language macro skills are specifically covered in the test.					
5. The items on the test represent the entire range of possible items it should cover.					
6. The test is correlated to the course objectives or learning standards and benchmarks.					
7. The questions are fairly represented in the universe or domain detailed in the program.					
8. The items are arranged from simple to complex.					

Comments:_____
_____
_____

CONSTRUCT VALIDITY	**1**	**2**	**3**	**4**	**5**
1. The items reflect the contents of the structures.					
2. The items are consistent with the development of the test.					
3. The items really measure a specific skill.					
4. The test actually taps into the theoretical construct as it has been designed.					
5. Every skill in the test has been included in an appropriate way.					
6. The test format is developmentally appropriate for the level of the students.					
7. The items are clear and direct and require a definite task or response.					
8. The format for each question is suitable to the learning outcome.					

Comments:_____
_____
_____

ENGLISH GRAMMAR I TEST ANALYSIS

	FACE VALIDITY	1	2	3	4	5	6	7	8	9	10
1	*COMPLETELY DISAGREE*										
2	*AGREE*	3	4		2		3		3		
3	*NEUTRAL*	2	3	3	3	1	5	4			3
4	*DISAGREE*	2		4	2	6		3	4	2	4
5	*COMPLETELY AGREE*									5	

The level of validity in face validity is 40% low

	CONTENT VALIDITY	1	2	3	4	5	6	7	8
1	*COMPLETELY DISAGREE*				1				
2	*AGREE*				6	3		6	6
3	*NEUTRAL*	5	3	3		4	3	1	1
4	*DISAGREE*		4	4			4		
5	*COMPLETELY AGREE*	2							

The level of validity in content validity is 30% low

	CONSTRUCT VALIDITY	1	2	3	4	5	6	7	8
1	*COMPLETELY DISAGREE*								
2	*AGREE*					5			2
3	*NEUTRAL*	4	4	3	6	2	4	3	3
4	*DISAGREE*	3	3	4	1		3	4	
5	*COMPLETELY AGREE*								

The level of validity in construct validity is 35% low

ENGLISH PRONUNCIATION TESTS ANALYSIS

	FACE VALIDITY	1	2	3	4	5	6	7	8	9	10
1	*COMPLETELY DISAGREE*										
2	*AGREE*										
3	*NEUTRAL*					2					
4	*DISAGREE*	4	4		1	5		4	3	3	1
5	*COMPLETELY AGREE*	3	3	7	6		7	5	4	4	6

The level of validity in face validity is 70% high

	CONTENT VALIDITY	1	2	3	4	5	6	7	8
1	*COMPLETELY DISAGREE*								
2	*AGREE*								
3	*NEUTRAL*								
4	*DISAGREE*	4	2	1	1	4	2	2	1
5	*COMPLETELY AGREE*	3	5	6	6	3	5	5	6

The level of validity in content validity is 60% high

	CONSTRUCT VALIDITY	1	2	3	4	5	6	7	8
1	*COMPLETELY DISAGREE*								
2	*AGREE*								
3	*NEUTRAL*								
4	*DISAGREE*	3	3	2	2	1	1	1	2
5	*COMPLETELY AGREE*	4	4	5	5	6	6	6	5

The level of validity in construct validity is 65% high

INTERMEDIATE INTENSIVE ENGLISH II TESTS ANALYSIS

	FACE VALIDITY	1	2	3	4	5	6	7	8	9	10
1	*COMPLETELY DISAGREE*									1	
2	*AGREE*		5				1				
3	*NEUTRAL*		2	2			1	#		1	
4	*DISAGREE*			1			1		1		1
5	*COMPLETELY AGREE*	7		4	7	7	4		6	5	6

 # No answer

The level of validity in face validity is 70% high

	CONTENT VALIDITY	1	2	3	4	5	6	7	8
1	*COMPLETELY DISAGREE*				2		7	6	
2	*AGREE*					1		1	
3	*NEUTRAL*				#	#			#
4	*DISAGREE*		1		1				
5	*COMPLETELY AGREE*	5#	6	7	2	4			3

 # Expert said 2 tests were about grammar

The level of validity in content validity is 40% intermediate

	CONSTRUCT VALIDITY	1	2	3	4	5	6	7	8
1	*COMPLETELY DISAGREE*	6	6	7	1			7	7
2	*AGREE*					3			
3	*NEUTRAL*					#			
4	*DISAGREE*	1					1		
5	*COMPLETELY AGREE*				2		6		

 # Expert said 2 tests were about grammar

The level of validity in face validity is 30% low

GENERAL DIDACTICS TESTS ANALYSIS

	FACE VALIDITY	1	2	3	4	5	6	7	8	9	10
1	*COMPLETELY DISAGREE*		2		2					1	2
2	*AGREE*			2					2	1	
3	*NEUTRAL*							2			
4	*DISAGREE*	2				2	2				
5	*COMPLETELY AGREE*										

The level of validity in face validity is 30% low

	CONTENT VALIDITY	1	2	3	4	5	6	7	8
1	*COMPLETELY DISAGREE*					1			
2	*AGREE*					1			
3	*NEUTRAL*	1	2		2				
4	*DISAGREE*	1		2			2	2	2
5	*COMPLETELY AGREE*								

The level of validity in content validity is 20% low

	CONSTRUCT VALIDITY	1	2	3	4	5	6	7	8
1	*COMPLETELY DISAGREE*		1			2	2		2
2	*AGREE*	2	1		2				
3	*NEUTRAL*			2				2	
4	*DISAGREE*								
5	*COMPLETELY AGREE*								

The level of validity in face validity is 20% low

Researchers own results

ENGLISH PRONUNCIATION TESTS ANALYSIS

	FACE VALIDITY	1	2	3	4	5	6	7	8	9	10
1	*COMPLETELY DISAGREE*										
2	*AGREE*										
3	*NEUTRAL*	1	1	1	2	2	1	1	1	1	1
4	*DISAGREE*	3	3	3	1	5	1	3	3	3	1
5	*COMPLETELY AGREE*	3	3	4	4		5	4	3	3	5

The level of validity in face validity is 75% high

	CONTENT VALIDITY	1	2	3	4	5	6	7	8
1	*COMPLETELY DISAGREE*								
2	*AGREE*								
3	*NEUTRAL*	1	1	1	2	1	1	1	1
4	*DISAGREE*	3	2	2	1	3	2	2	1
5	*COMPLETELY AGREE*	3	4	4	4	3	4	4	5

The level of validity in content validity is 55% high

	CONSTRUCT VALIDITY	1	2	3	4	5	6	7	8
1	*COMPLETELY DISAGREE*								
2	*AGREE*								
3	*NEUTRAL*	2	1	1	1	1	2	2	2
4	*DISAGREE*	3	3	2	2	1	1	1	2
5	*COMPLETELY AGREE*	2	3	4	4	5	4	4	3

The level of validity in construct validity is 55% high

INTERMEDIATE INTENSIVE ENGLISH II TESTS ANALYSIS

	FACE VALIDITY	1	2	3	4	5	6	7	8	9	10
1	*COMPLETELY DISAGREE*										
2	*AGREE*		4	1			1				
3	*NEUTRAL*		2	2	2	2	1	5	2	1	1
4	*DISAGREE*	2	1	1				2	1		1
5	*COMPLETELY AGREE*	5		3	5	5	5		4	6	5

The level of validity in face validity is 75% high

	CONTENT VALIDITY	1	2	3	4	5	6	7	8
1	*COMPLETELY DISAGREE*				2		5	6	
2	*AGREE*					1	2	1	
3	*NEUTRAL*	2	1	1	4	2			4
4	*DISAGREE*		1	1	1				
5	*COMPLETELY AGREE*	5	5	5	3	4			3

The level of validity in content validity is 45% intermediate

	CONSTRUCT VALIDITY	1	2	3	4	5	6	7	8
1	*COMPLETELY DISAGREE*	4	5	5	2			5	5
2	*AGREE*	1	1	1	3	3		1	1
3	*NEUTRAL*		1	1		4	1	1	1
4	*DISAGREE*	1					1		
5	*COMPLETELY AGREE*				2		5		

The level of validity in face validity is 35% low

GENERAL DIDACTICS TESTS ANALYSIS

	FACE VALIDITY	1	2	3	4	5	6	7	8	9	10
1	*COMPLETELY DISAGREE*		1		1					1	1
2	*AGREE*		1	1	1				1	1	1
3	*NEUTRAL*			1		1	1	1	1		
4	*DISAGREE*	1				1	1	1			
5	*COMPLETELY AGREE*	1									

The level of validity in face validity is 35% low

	CONTENT VALIDITY	1	2	3	4	5	6	7	8
1	*COMPLETELY DISAGREE*								
2	*AGREE*					1			
3	*NEUTRAL*	1	1	1	1	1	1	1	1
4	*DISAGREE*	1	1	1	1		1	1	1
5	*COMPLETELY AGREE*								

The level of validity in content validity is 25% low

	CONSTRUCT VALIDITY	1	2	3	4	5	6	7	8
1	*COMPLETELY DISAGREE*								
2	*AGREE*	1	1	1	1	1	1	1	1
3	*NEUTRAL*	1	1	1	1	1	1	1	1
4	*DISAGREE*								
5	*COMPLETELY AGREE*								

The level of validity in face validity is 25% low

ENGLISH GRAMMAR I TEST ANALYSIS

	FACE VALIDITY	1	2	3	4	5	6	7	8	9	10
1	COMPLETELY DISAGREE										
2	AGREE	2	3	1	3	1	3	1	4		1
3	NEUTRAL	2	4	3	3	1	4	3		1	3
4	DISAGREE	2		3	1	5	1	3	3	2	3
5	COMPLETELY AGREE	1								4	

The level of validity in face validity is 45% low

	CONTENT VALIDITY	1	2	3	4	5	6	7	8
1	COMPLETELY DISAGREE				1				
2	AGREE		3		5	4		5	5
3	NEUTRAL	4	2	3	1	3	3	1	1
4	DISAGREE	2	1	3			4	1	1
5	COMPLETELY AGREE	1		1					

The level of validity in content validity is 35% low

	CONSTRUCT VALIDITY	1	2	3	4	5	6	7	8
1	COMPLETELY DISAGREE								
2	AGREE	1	1	2		4	1		2
3	NEUTRAL	3	3	3	6	2	3	4	3
4	DISAGREE	3	3	2	1	1	4	3	2
5	COMPLETELY AGREE								

The level of validity in construct validity is 40% low

IX. REFERENCES

- Airasian, Peter W. (2000) *Assessment in the Classroom. McGraw-Hill Higher education.*
- Bailey Kathleen M. (2000) *Learning about Language Assessment Dilemmas Decisions and Directs*, Heinle & Heinle publisher.
- Benavidez, Cesia Jemimma (2004) *A Comparison between the theory and the practice on the evaluation process of the FLD at the University of El Salvador*. May campus may 2004.
- Brown, A. Douglas (2001) *Teaching by Principles*, 2nd edition San Francisco State University.
- Hughes, Arthur (2000) *Testing for Language Teachers*. Cambridge University press.
- Lazaraton, Anne (2002) *A quality Approach to the Validation of oral Language Test*. Cambridge University.
- Marzano Robert J. (2006) *Classroom assessment and grading that work. Association for supervision and curriculum development.* Alexandria, Virginia USA.
- Najarro Barahona, Mayra. (2009) *Evaluation and Testing in the Foreign Language Department of the UES*. Main campus, March 12th 2009.
- Salkind J. Neil (2000) *Exploring Research* fourth edition, Prentice Hall, Upper Saddle River, New Jersey.
- Sax, Gilbert (2001) *Principles Educational Psychological.* Wadswoth publishing company.

(9.0) WEBLIOGRAPHY

- http://www.rubrics4teachers.com/
- http://www.gecdsb.onca/d&g/onlinepd/assessment%20&%20evaluation/paper%20and%pencil.htm.
- http://imoed-forum.blogspot.com/2009/11/test-validity
- www.es.scribd.com/doc/30315471/classroom-assessment-and-grading
- http://psychology.about.com/od/researchmethods/f/validity.htm
- http://www.experiment-resources.com/validity-and-reliability.html
- http://fcit.usf.edu/assessment/basic/basicc.html
- http://www.learningandteaching.info/teaching/assessment.htm

X. DEFINITIONS

- Ability: What one has learned over a period of time from both school and nonschool sources; one's general capability for performing task.
- Achievement: What one has learned from formal instruction, usually in school.
- Assessment: The process of collecting, synthesizing, and interpreting information to aid classroom decision making; includes information gathered about pupils, instruction, and classroom climate.
- Checklist: A written list of performance criteria associated with a particular activity or product in which an observer marks the pupil's performance or each criterion using a scale that has only two choices.
- Construct validity: refers to the extent to which operationalizations of a construct (e.g. practical tests developed from a theory) do actually measure what the theory says they do.
- Content validity: evidence involves the degree to which the content of the test matches a content domain associated with the construct.
- Criterion-referenced grading: Determining the quality of a pupil's performance by comparing it to preestablished standards of mastery.
- Curriculum: The skill, performances, attitudes and values pupils are expected to learn from schooling: includes statements of desired pupil outcomes, descriptions of materials and the planned sequence that will be used to help pupils attain the outcomes.
- Diagnose: Identify specific strengths and weaknesses in pupil's past and present learning.
- Evaluation: Judging the quality of goodness of a performance or a course of action.

- FLD: Foreign Language Department.
- Face validity: is an estimate of whether a test appears to measure a certain criterion; it does not guarantee that the test actually measures phenomena in that domain.
- Instruction: The methods and process by which pupil's behaviour are changed.
- Items: Questions or problems on an assessment instruction.
- Level: The grade level at which a particular commercial test should be administrated to pupils.
- Objectives: Statements that describe what pupils are expected to learn from instructions.
- Performance assessment: Observing and judging a pupil's skill in actually carrying out physical activity or producing a product.
- Proficiency : The quality of having great facility and competence
- Standardized tests: are tests on which all students answer the same questions, usually in multiple-choice format, and each question has only one correct answer. They reward the ability to quickly answer superficial questions that do not require real thought.
- Teacher made tests: are test and other measures that are planned, assembled, written or otherwise prepared by teachers for use with particular group of students.
- TEFL: Teaching English as a Foreign Language.
- Validity: refers to the degree to which a study accurately reflects or assesses the specific concept that the researcher is attempting to measure.