UNIVERSIDAD AUTONOMA DEL ESTADO DE HIDALGO
INSTITUTO DE CIENCIAS SOCIALES Y HUMANIDADES

TESTING WRITING SKILLS

MONOGRAFIA
QUE PARA OBTENER EL GRADO DE
LICENCIADO EN
ENSEÑANZA DE LA LENGUA INGLESA

PRESENTA:
Alma Rosa Hernández Barona

Director:
MPhil. Hilda Hidalgo Avilés

Pachuca, Hgo. 2010

Acknowledgements

This *monografia* is a great effort that makes me feel a successful person because my profession is a fundamental aspect in my life. For that reason I am very grateful with the guidance provided by my supervisor, Hilda Hidalgo Avilés, who encouraged me with the development of this project giving me the whole possible sources, her patience, and mainly her most significant words to conclude it.

Moreover, it is important to mention my professors in Research Seminar, Mtra. Rosa María Funtherbuk Razo and Mtra. Bertha Paredes Zepeda who gave me their support during the course and they provided me with certain aspects to improve this work with their experiences.

I would like to thank the Universidad Autónoma del Estado de Hidalgo and the Instituto de Ciencias Sociales y Humanidades for giving me this great opportunity in order to get my degree to carry out another goal in my life that it will provide me with a better condition as a worker and the continuity of my professional training.

As well, I dedicate this document to the reason of my life, Selene, my little daughter, because she gives me the strength to be a better person in order to conclude with this challenge because I want to be her biggest reason and example in her life.

Finally, I give my regards and blessings to my parents who give me their support at every moment I need it. Their help was fundamental because I could work and develop this project at the same time. All the people mentioned in this monograph are special for me because I know that without their assistance and enthusiasm I would not have finished this work.

TABLE OF CONTENTS

CHAPTER III TESTING WRITING TECHNIQUES

LIST OF TABLES

Introduction

Testing is a universal practice in the world and it has been modified due to the different changes in Education. Testing helps teachers to verify people's abilities, strengths, weaknesses in a certain area and it can be carried out considering the level of the tested person by means of an instrument where results will be interpreted according to an established score (MacNamara, 2000).

Bachman and Palmer (1996) state that testing is a process with a variety of considerations which have to be taken into account in order to develop an appropriate test. Also they argue that this process has to be considered not only in the development of international and national tests, but also for classroom tests because it is where commonly language learning and teaching happens.

For that reason, testing influences teaching since it involves to follow a procedure when designing different types of tests which have to be directly related to the objective of the course syllabus, qualities, structure, techniques to be used, scoring processes, and interpretation of the results.

Test purpose is considered the most important characteristic in test design because the purpose will make explicit the use of the test (Bachman, 1990; Bachman and Palmer, 1996). Bachman (1990), Hughes (2002), Weir (1993), MacNamara (2000), and Bachman and Palmer (1996) consider reliability, validity, practicality, authenticity, interactiveness, impact and washback as general test qualities which contribute to test usefulness because they define aspects like stability of measurement; justification over the scores; correspondence between tasks and candidates' language abilities; the use and the implementation of the test in a particular situation;

and the different ways in which tests use affects. All these aspects are discussed throughout this document.

Kitao and Kitao (1996) consider writing as one of the most difficult skills in testing process because of two major problems that affect directly the matter of control related with the content and the scales that have to reflect precisely students' learning with numeral data.

Thus, as writing is one of the most difficult skills to master because it implies not only the correct use of language, but the development and presentation of thoughts in a structured way, this paper is developed in order to provide information about how testing writing is carried out based on different authors like Hughes (2002), Weir (1993), MacNamara (2000), Bachman (1990), Bachman and Palmer (1996), Kitao, and Kitao (1996), and Richards (2003).

This document will be helpful for further studies because it includes the most important facts that the testing writing process involves. In addition, it can be useful for English teachers in order to be informed about testing considerations and as a result of this, they can improve their tests design considering syllabus, methodology, techniques, tasks, and instruments used in the classroom as a whole to consider changes on their teaching methodology and plans.

This research is organized into five chapters. In chapter one, the concept of writing skill and some specifications which are contemplated when this ability is tested are included. Then, the concept of testing is extensively compared with other terms in order to clarify the function of this process and types of tests in order to emphasize test purpose. Chapter two presents an overview of general qualities and some important moderations in order to carry out effectively testing design. In chapter three some important testing writing techniques are presented, while chapter four shows the most important details of the instruments and marking procedure for testing in order to complement the process of design. Finally the conclusions of this document are provided.

CHAPTER I. WRITING AND TESTING

1.1 Overview

In this chapter, the concept of writing skill is provided in order to understand better the purpose of testing this skill. Then, some language test terms are clarified as well.

1.2 Conceptualizing Writing

Hedge (2000) claims that writing is the result of employing strategies to manage the composing process of a text. Such process involves a number of activities; for instance: setting goals, generating ideas, organizing information, selecting appropriate language, making a draft, reading and reviewing, then revising and editing. For that reason, writing often seems to be the hardest of the skills, even for native speakers of a language, since it implies not only a graphic representation of speech, but the development and presentation of thoughts in a structured way.

Shaughnessy (1977, mentioned in Hedge, 2000) supports the last consideration because he expresses “one of the most important facts about the composing process that seems to get hidden from students is that the process that creates precision is itself messy” (p. 2).

As well, Kitao and Kitao (1996) consider writing as a difficult ability because of the extended number of component skills involved which are presented below:

1. “Grammatical ability. This is the ability to write English in grammatically correct sentences.
2. Lexical ability. The ability to choose words that are correct and used appropriately.

3. Mechanical ability. The ability to correctly use punctuation, spelling, capitalization, etc.

4. Stylistic skills. The ability to use sentences and paragraphs appropriately.

5. Organizational skills. The ability to organize written work according to the conventions of English, including the order and selection of material.

6. Judgments of appropriacy. The ability to make judgments about what is appropriate depending on the task, the purpose of the writing and the audience" (p.2).

Writers need to develop aspects about orthography, spelling, the use of the correct function and form of words, their correct word order, use vocabulary and style correctly, clearness, support ideas or information, make the text coherent, background knowledge in order for our ideas to make sense and communicate what we want in a text. We also must create a context of what we are writing about, we cannot make reference to something or someone if we cannot point it personally. Additionally, it may try to communicate and include the possible answers to the reader's reactions or doubts.

As well, Richards (2003) sustains that learning how to write, mainly in second language acquisition, is one of the most challenging aspects because it requires an extensive and specialized instruction by the part of the instructor and largely when learners' objectives demand it.

Writing may be one of the most complex skills in language learning because second language learners face a big range of challenges. Consequently, it is fundamental to test written students' products in order to know if they are getting a real progress in the ability.

1.3 Testing Writing

Hughes (1989) mentions three issues that have to be taken into account when testing writing:

1 We have to set writing tasks that are properly representative of the population of tasks that we should expect students to be able to perform.

2 The tasks should elicit valid samples of writing (i.e which truly represent the students' ability).

3 It is essential that the samples of writing can and will be scored validly and reliably (p. 85).

When teachers test writing, they should select appropriate activities for each part of the test according to what they want to measure. For example, if they want to test describing activities, teachers should be focused only on that topic, and not to move away into some other aspect. Some writing techniques are developed in detail in chapter three.

As people have different abilities in different levels, the teacher should take a sample of each type of task, some learners may be better at filling a certain layout and some others in developing topics freely. By having a representative sample, we mean using various types of tasks because as Kitao and Kitao (1996) claim it is a new trend to test students on different types of tasks they have produced during a long period of time, rather than over one piece of writing on a particular occasion.

Regarding the second issue, the teacher should select specific aspects or topics from the content that will test only the ability of writing and nothing else. Also the teacher should select a new task for each time he or she is going to test a specific topic. This will provide a variety to students and will achieve greater reliability and therefore greater validity. Beside this, there should be a balance between what is desirable and what is practical; this means the teacher should apply tasks that provide reliable results, but at the same time they should be interesting and creative. It is important to determine the most relevant elements or topics of a course when a test is developed so that it will help students to use the language in their lives, because as Hughes (1989) states “one of the most important aspects is the content of the test because this should reflect what the students have been doing in a class or during a course and also the items in the test should be relevant in terms of real world language use” (p.21).

The content of the test will depend on the kind of syllabus teachers are using. Also, they will decide the level of importance of each selected element and the way these elements are going to be tested (Harris, 2001). Because as Hughes (1989) mentions “in language testing we are not normally interested in knowing whether students are creative, imaginative, or even intelligent, have wide general knowledge, or have a good reason for the opinion they happen to hold” (p. 90). Tests become isolated practices created without a context. Students are tested on what they know rather what they can do with the language. A further explanation of these qualities is given in chapter two.

Teachers have the obligation to write, administer and mark tests in order to see how our students are learning. For that reason, qualities and moderations, and criteria have to be considered also when designing a test, which will be explained in the subsequently chapters of this document.

Finally, another important aspect for testers is to be clear and specific, instead of giving vague comments which do not concrete to the students how to do improve the task. For example, Zamel (1985) suggests that teachers should not test writing as a finished product but as work in progress, that is, to test during the whole writing course rather than as a final task. For this reason, it is important for teachers to set the marking criteria, which is one of the most time-consuming aspects of testing but which can contribute to mark the test in a more objective way.

All these aspects become the framework that test design needs in order to be a fair practice for both students and the teacher. Weir (1993) agrees with Hughes' (1989) issues pointing out a framework in the design of writing tests as he sustains that the main objective of process writing is to know what the students do when they write. He provides an approach, which he calls "content-oriented approach", where students' writing "is an exploratory, generative, collaborative, recursive process rather than as a linear route to a predetermined product" (p. 130). This framework is focused on the effect of the text on a reader understanding from the "target discourse community" and writing favored in the process approach with writer-focused type instead of personal discovery.

Weir (1993) firstly considers what features of real-life performance and then how they have to be tested. Thus, he provides the next usual headings, which will be further described in chapter four, in order to be considered in the design of writing tests:

1. Conditions: Specifications which teachers might take into account in the design of a writing task of students in their own situation.
2. Operations: Procedures which might be included in writing tasks at various level of ability.
3. Quality of output: Criteria which teachers might use to make decisions on level of performance on the above operations under specified conditions.

Therefore, it is concluded that teachers have to follow a framework in order to design, administer and mark tests in order to know students' progress in the writing skill. Furthermore, they must not overlook what the program requires or what students need to know when marking the tests; they should also care about students' progress and observe if students are really competent in the ability because they show a real progress in writing process in real contexts.

In the next section some terms are given in order to understand better the function of testing and consequently some kinds of tests are explained emphasizing the testing purpose.

1.4 Understanding Testing

It is essential to understand broadly the function of testing and emphasize the difference between various concepts which are used as synonyms in the process of evaluation where testing is only one part of the procedure and for that reason; those concepts are clarified below in order to give the reader an extensively comprehension about the objective of this document.

1.4.1 Evaluation

When we hear this word immediately we relate it as a synonym of test, measurement, and assessment, but the truth is that all these language test terms have their own procedure which will determine, in the educational area, students' achievement or proficiency in a certain skill.

Weiss, (1972, mentioned in Bachman, 1990), cites that "evaluation can be defined as the systematic gathering of information for the purpose of making decisions" (p.22). Thus, it is understandable that evaluation is a process where other procedures are involved in this case: testing, measurement, and assessment. And it is automatically known when the author expresses *a systematic collection of information*, as the term system determines a group of associated actions which are orderly set to reach a goal (Kirster, 1999). As well, the words "making decisions" is related because it is obvious that the expected objective is to define if the evaluated person is proficient or not in the specific ability.

Likewise William (2006) says that "evaluation is the systematic acquisition and assessment of information to provide useful feedback about some object" (p. 1). This definition is similar to Weiss's as he sustains that "evaluation is a systematic acquisition of information", but William adds the term "assessment" that it is done after every evaluation because collecting data, not only helps to make judgments, but also making inferences, whether or not an assessment has to take place in order to improve future outcomes. Finally, the use of the ambiguous term "object" may possibly refer (according to institutional conditions) to a program, person, need, or activity.

Moreover, Saskatchewan (1997) states that “evaluation is the culminating act of interpreting the information gathered for the purpose of making decisions or judgments about students’ learning and needs, often at reporting time” (p. 1).

Thus, evaluation is a process that implies to take into account all the aspects that interfere in the teaching practice itself. Now we turn to define testing which is the focus of this document.

1.4.2 Testing

According to McNamara (2000), testing is a universal practice in the world that has been modified during the time in order to verify people’s abilities in a certain area. Commonly it is used for determining in numbers (scales from 0-10), letters (A, B, C), or judgments (Competent or Incompetent) the level of the tested person by means of an instrument (rubrics and exams) that takes part in the evaluation process in order to infer results according to a quantification system, a researcher or by the same teacher in the classroom (McNamara, 2000).

Caroll (1968, mentioned in Bachman, 1990) claims that “…an educational test is a procedure designed to elicit certain behavior from which one can make inferences about certain characteristics of an individual” (p. 20). In other words, Bachman supports that a test is a sample where there are different kinds of contents based on some specific acquired abilities that can be measurable in the current course. Therefore, the tester is capable of quantifying *individual’s behavior* in order to proof test takers' progress in the language skill based on serious rules given out by a system previously researched and justified.

Furthermore, Bachman (1990) states that language teachers regularly use tests

> "to help diagnose student strengths and weaknesses, to assess students' progress, and to assist in evaluating student achievement. Language tests are also frequently used as sources of information in evaluating the effectiveness of different approaches to language teaching. As sources of feedback on learning and teaching, language tests can thus provide useful input into the process of language teaching" (p. 3).

Tests have different uses which help not only the teaching and learning process but also the educational area. It is considered as only a part of the process of evaluation in the classroom, but its importance is relevant in order to identify and then modify various aspects of the same design of it, the course syllabus, and its impact on students' success.

Thus, testing is an important process in the learning of English as a Foreign Language and also in the Language Teaching, but some factors have to be considered in order to get valid results and also to motivate the learners to improve the tested skill in subsequently evaluations because the elaboration of tests has to be connected with the main purposes of the course, teaching and learning.

1.4.3 Measurement

According to Bachman (1990) "in the social sciences measurement is the process of quantifying the characteristics of persons according to explicit procedures and rules" (p.18). Likewise this author considers three important aspects: quantification, characteristics, and explicit rules and procedures.

Thus, Bachman considers measurement as a process where dimensions to some capacity are determined. In the *quantification* of it, some standard instruments such as nominal scales, ordinal scales, and interval scales are used to measure the correspondent student's level with the corresponding number, letter, or even a verdict.

While MacNamara (2000) argues that "measurement investigates the quality of the process of assessment by looking at scores" (p. 56). He considers only two steps in this procedure: Quantification and Checking for various kinds of mathematical and statistical patterns. This author centers his attention to the process of evaluation as a whole instead of *characteristics of persons*, in view of the *matrix* results, as he calls them, which are taken from various instruments filled by a certain group of tested people so that it could make available not only numbers and decisions, but also useful information that will help to improve all the components around the process.

1.4.4 Assessment

In Saskatchewan's article (1997) she cites that "Assessment is the act of gathering information on a daily basis in order to understand individual students' learning and needs" (p.54).

As well Black and William (2001) states that assessment "…is generally used to refer to all activities teachers use to help students learn and to gauge their progress" (p.2).

In other words, assessment is the information that supports teachers in planning and adapting for further instruction. This evidence is presented during the whole course and they could be products like portfolios, activities, exercises, reports, projects, presentations, and also tests which can be useful in providing all what teachers need to know about students' improvement.

During assessment phase, teachers select appropriate tools and techniques, then collect and gather information on students' progress. So, teachers must determine where, when, and how assessments will be conducted, and students must be consulted and informed.

As well, in this period teachers can enhance students' understanding of their own advance by involving them in gathering their own evidence and share their findings and facts with them according to the assessment purposes. And as a result of this, such participation makes possible for students to identify and set personal learning goals (Saskatchewan, 1997).

Thus, self-assessment promotes students' abilities to assume more responsibility for their own learning by encouraging self-reflection and encouraging them to identify where they believe they have been successful and where they believe they require assistance. Discussing students' self-assessments with them allows the teacher to see how they value their own work and to ask questions that encourage students to reflect upon their experiences and set goals for new learning (Villardón, 2006).

Peer assessment allows students to collaborate and learn from others. Through discussions with peers, middle Level students can verbalize their concerns and ideas in a way that helps them clarify their thoughts and decide in which direction to proceed.

The instruments for peer and self-assessment should be collaboratively constructed by teachers and students. It is important for teachers to discuss learning objectives with them. Together, they can develop assessment and evaluation criteria relevant to the objectives, as well as to students' individual and group needs.

Assessment data can be collected and recorded by both the teacher and the students in a variety of ways. Through observation of students, interviews or even talks with students, teachers can discover much about their students' knowledge, abilities, interests, and needs. As well, teachers can collect samples of students' work in portfolios and conduct performance assessments within the context of classroom activities. Now we turn to describe the purposes of tests.

1.5 Test purpose

Bachman (1990) states in his book *Fundamental Considerations for Language testing* that one of the most important steps to develop in "the development of language tests" and "the interpretation of their results" is to state the purpose of the test in order to give it a direction according to its use considering "two major" functions: the first one in the evaluation of the syllabus and the second one in making decisions on participants' abilities and in the selection of the teaching techniques" (p.54).

Later Bachman and Palmer (1996) add that the purpose of the test indicates its content which specifies a task or tasks of the language skill and based on "test results; it provides enough and significant information for deciding on tested people, teachers or administrators' implications, and programs" (pp. 88, 96). These authors point out the importance of emphasizing a *language task* instead of *language ability* because they believe that language ability is a synonym of language skill which involves listening, speaking, reading, and writing. Thus a task is an activity which is focused on a particular skill in a specific test technique. For example, in *Test of*

English as a Foreign Language (TOEFL) there is a grammar section with 40 statements divided in two parts. The first 15 sentences consider structure and the last 25, written expression. According to the authors the sub-ability is specified and the test technique used is multiple-choice where the tasks are complementation of sentences and identification of mistakes.

Thus test purpose is vital when designing tests because its effects will present their real use starting from their development until the obtained results. In other words, the purpose of the test has to be appropriate and connected with the applied teaching procedures, to carry out the stated objectives and matters of the programs, and to improve certain language ability, and to employ the right measurement for getting results. If something is wrong, the outcomes will determine changes in any of the last stated issues.

1.6 Types of tests

MacNamara (2000) and Weir (1993) highlight that the most common types of tests used in the classroom are achievement and proficiency. However, Bachman and Palmer (1996) call them *low- and high -skates* respectively.

Achievement tests or low-skates are those that are completely related to the process of instruction and scores. And proficiency tests or high-skates estimate future situations of language use with or without teaching instruction and where scores are not the matter, but a final judgment takes place. These terms will be developed in detail in the following section.

However, other authors identify other kinds of tests. For instance, Hughes (2002) proposes a category of four types of tests: proficiency, achievement, diagnostic, and placement tests which will prove useful decisions on particular objectives and for forthcoming tests new designs or adaptations.

Bachman (1990) and Bachman and Palmer (1996) suggest the next classification: selection, placement, diagnosis, progress and grading tests that are based only on "decisions about test takers" because they think that the results have to be organized according to the "inferences" related to the different areas where they have an impact such as "test-takers, teachers, administrators, and programs".

In the following lines a detailed explanation of proficiency tests, achievement tests, diagnosis or diagnostic tests, selections tests, and placement test is provided.

1.6.1 Proficiency tests

Proficiency tests involve meaningful learning in students. The progression that they have in the language is what they are going to use in real life according to their own perspectives (McNamara, 2000).

For Weir (1993) when designing this type of tests, it is important to guarantee that the examinations have a total relationship with real situations that can effectively grow up in the students' ability in one or the whole English language skills in order to face consequently events.

Hughes (2002) provides a similar definition but he adds that "proficiency tests are designed to measure people's ability in a language, regardless of any training they may have had in that language and their content is not based on the content or objective of language courses that people taking the test may have followed" (p.11). Rather, it is based on a specification of what candidates have to be able to do in the language in order to be considered proficient. That means that the test-taker is expecting to have “sufficient command of the language” for a particular personal purpose. For instance examples of these kinds of exams would be the Cambridge First Certificate in English examination (FCE) and the Cambridge Certificate of Proficiency in English examination (CPE). The function of such tests is to show whether candidates have reached a certain standard with respect to a set of specified abilities.

However Bachman (1990) and Bachman and Palmer (1996) call this kind of tests as *progress* and they consider a purpose of *formative* evaluation like MacNamara (2000) and Weir (1993) do.

1.6.2 Achievement tests

In this kind of tests previously teaching has to be considered because it has a close connection with it and its good or bad administration will be reflected on the tests effects. For that reason, the tester should think precisely about past developed activities, students’ level, language use, the proper criteria for the kind of learners, and specifically what the teacher wants to evaluate about the whole content.

These tests gather evidence during the course or at the end of it and they reflect only the acquired knowledge of one feature of grammar or vocabulary which is directly stated in the syllabus. In the case of achievement tests, there is not any future responses about the language use in real life because it is mainly focused on the objectives of the course and the reports are commonly given in grades according to the established periods of evaluation in the institution.

That is why Hughes (1993) classifies two kinds of achievement test: final achievement tests and progress achievement tests. The first ones are focused on knowing the usefulness of the syllabus and as it was said before, they are administrated at the end of each course. Progress achievement tests are in charge of quantifying students' progress and the results are based on their marks, and so while it is higher the score it is better the improvement.

Therefore these kinds of tests simply evaluate the obtained knowledge of every given topic of the program of the course and all the results are observable on grades. Bachman (1990) and Palmer (1996) name these instruments as grading tests and relate them to *summative* evaluation that is useful for knowing about students' achievement at the end of the academic module.

Portfolios, exams, exercises, and practices are a clear example of achievement tests. It is important to say that the approach needs an alternative assessment which will be given after the examination in order to clarify the students' lowest areas, so that the learner and the teacher can get benefits for subsequently measurements (McNamara, 1993).

1.6.3 Diagnosis or Diagnostic tests

Bachman (1990) and Bachman and Palmer (1996) state that diagnostic tests are used to identify learners' strengths and weaknesses focused on diagnosing a particular aspect of the language. Thus, they serve as indicators to know where a teacher has to start and work on with a new group of students. For example, if the objective program includes identifying grammatical errors, the diagnostic exam has to be centered on all types of exercises which provide the students' knowledge about that.

1.6.4 Selection tests

Bachman (1990) relates selection tests only with the educational area, situating the instruments which provide enough information about test-takers in order to make decisions about their entrance to a determined institution. He states that it is a sort of "readiness test". In other words, it is an exam that evaluates previous knowledge, academic capacity, and accomplishment that will presume if a student is ready to be part of the new institution. One exemplification of is the Test of English as a Foreign Language (TOEFL). Some students who want to continue their high school or colleges studies abroad, mainly in the United States, have to take this test in order to gain the expected score (over 550 points) and present their results to be accepted in the school they chose (Bachman, 1990).

1.6.5 Placement tests

Hughes (1993) says that "placement tests are intended to provide information that will help to place students at the stage of the teaching program most appropriate to their abilities. Typically they are used to assign students to classes at different levels. And Bachman (1990) supports this information adding that "in many language programs students are grouped homogenously according to the factors such as level of language ability, language aptitude, language use needs, and professional or academic specialization" (p. 16). These kinds of tests are administered for international evaluators when a student assumes to have a certain level in the English Language.

The last words can be supported with Bachman and Palmer's (1996) definition. They think that a student who has studied a foreign language and wish to be placed in other advanced level course in a foreign college, she might take a placement language test to decide if she has the appropriate level for coursing it.

Tests are categorized according to the information they provide. And this categorization, as Hughes (2002) calls it, will establish their particular objective and the design of new tests if it is necessary.

For that reason, principal qualities, which are important to take into account in the plan of tests, are given in the next chapter. Such qualities are: reliability, validity, practicality, authenticity, interactiveness, impact, and backwash.

CHAPTER II. GENERAL QUALITIES IN TESTING

2.1 Overview

Weir (1993) sustains that “all tests have to be moderated before being administrated” (p.19). That means they have to be firstly developed and monitored under some conditions. Thus, when a test is “balanced among the qualities of usefulness” the tester has to consider the purpose of the test, its effects, selection of the measurement, evaluated language ability, its connection with the program of the course, available sources, and impact over test takers and teaching. All of this will determine the exactly usefulness of the test and subsequently contribute appropriately on the expected results (Hughes, 2002).

There are a number of authors who provide a framework so that tests contain all these qualities. Bachman (1990) cites reliability and validity. Hughes (2002) suggests: reliability, validity and backwash; while Weir (1993) proposes the same qualities, but he adds practicality. MacNamara (2000) only mentions validity and washback. And Bachman and Palmer (1996) provide a larger number as they cite six of them: reliability, validity, authenticity, interactiveness, practicality, and washback. All of these terms will be described below.

2.2 Reliability

Commonly, it is thought that test usefulness provides a kind of “metric” result through it can be evaluated, but at the same time some other aspects can be taken into account according to the test development and use. That is why measurement provides several processes that specify the relationships between scores, and factors that affect them.

According to Bachman and Palmer (1996) “reliability is considered as consistency of measurement. A reliable test score will be consistent across different characteristics of the testing situation”(p.19).

Furthermore Bachman and Palmer (1996) state that if the construct definition focuses on a relatively narrow range of components of language ability, the test developer can reasonably expect to achieve higher levels of reliability than if the construct is complex, including a wide range of components of language ability, as well as topical knowledge.

As a consequence, reliability has to do with the consistency of measures across different time, test forms, raters, and other characteristics of the measurement context. The identification of potential sources of error involves making judgments based on an adequate theory of sources of error.

Test performance is affected by factors rather the quantified abilities. Some examples are poor health, fatigue, lack of interest, and test-wiseness and which totally affect students’ test development, and obviously they are not normally associated with language ability, and hence they are not characteristics that are measured with language tests.

When the effects are reduced, various factors maximize reliability. In other contexts “less these factors affect test scores, the greater the relative effect of the language abilities which are measured, and therefore, the reliability of language test scores” (Bachman, 1990, p. 160). In other words, 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 a measurement. A measure is considered reliable if a person's score on the same test given twice is similar. It is important to notice that reliability is not measured, it is only estimated.

There are several general classes of reliability estimates: inter-rater reliability, test-retest reliability, inter-method reliability, and internal consistency reliability.

Inter-rater reliability is the variation in measurements when taken by different learners but with the same method or instruments. Test-retest reliability is the variation in measurements taken by a single person or instrument on the same item and under the same conditions. This includes intra-rater reliability. Inter-method reliability is the variation in measurements of the same target when taken by a different methods or instruments, but with the same person, or when inter-rater reliability can be ruled out. When dealing with forms, it may be termed parallel-forms reliability. Internal consistency reliability assesses the consistency of results across items within a test.

Reliability may be estimated through a variety of methods that fall into two types: single-administration and multiple-administration. Multiple-administration methods require that two assessments are administered. In the test-retest method, reliability is estimated the correlation coefficient between two administrations of the same measure. In the alternate forms method, reliability is estimated the correlation coefficient of two different forms of a measure, usually administered together.

Single-administration methods include split-half and internal consistency. The split-half method treats the two halves of a measure as alternate forms. This "halves reliability" estimate is then stepped up to the full test length using a formula.

One important and contrary aspect to mention is that Weir (1993) considers “reliability is often connected with taking enough samples of a student’s work. The more evidence we have of a student's ability the more confident we can be in the judgments we make concerning this ability (p. 20).

2.3 Validity

Reliability does not imply validity. That is, a reliable measure is measuring something consistently, but it is not measured what it is wanted to be measuring. For example, while there are many reliable tests of specific abilities, not all of them would be valid for predicting job performance or the real improvement of an ability related with the content of a program. The investigation of reliability is concerned with answering the question, "How much variance in test scores is due to measurement error?" In order to estimate the relative proportion of error and reliable variance in test scores, it is useful the measurement theory as a basis for designing data collection and for analyzing and interpreting the results. Validity, on the other hand, is concerned with identifying the factors that produce the reliable variance in test scores. That is, validation addresses the question, "What specific abilities account for the reliable variance in test scores?" Thus, it might be said that reliability is concerned with determining how much of the variance in test score is reliable variance, while validity is concerned with determining what abilities contribute to this reliable variance (Bachman, 1990).

Authors such as Weir and Bachman consider validity as the most important quality before reliability. Weir (1993) states that "validity is the starting point in test task design" (p.19) because with this quality the idea of developing a test from the idea of what the writer wants to test is stated. And that is where the writer can explicit what it is going to be tested reflecting a realistic use of the particular ability to be measured.

That means, to involve credible language activities performed under appropriate conditions and formats selected should incorporate as many important real-life features as possible.

Therefore, examining the meaningful of test scores is demonstrating that they are not excessively affected by factors other than the ability being tested. If a reading is giving to the students and then it is asked them to perform an essay related with content of the reading their writing ability it is not going to be valid. If the activity has to be turned valid, it has to be directed as a development of a writing essay related with the topic of the reading where students have to give their own opinion. Another strategy could be to develop a letter, if the content permits it.

In examining validity, it is also concerned with appropriateness and usefulness of the test score for a given purpose. For example, scores from a test designed to provide information about an individual's vocabulary knowledge might not be particularly useful for placing students in a writing program. Thus while reliability is a quality of tests scores themselves, validity is a quality of test interpretation and use (Bachman, 1990).

Hughes (2002) reinforces the last concept, as he points out that validity is concerned with scoring. If a test is valid, not only the items but also the way in which the responses are scored must be valid. If we are interested in measuring speaking or writing ability, it is not enough to elicit speech or writing in a valid fashion. The rating of that speech or writing has to be valid too. For instance, overemphasis on such mechanical features as spelling and punctuation can invalidate the scoring of written work and so the test of writing.

For constructing validation is the on-going process as Bachman and Palmer (1996) state. It demonstrates that a particular interpretation of test scores is justified, and involves mainly building a logical case in support of a particular interpretation and providing evidence that justify that interpretation. Several types of evidence can be provided in support a particular score interpretation. For example by means of content and criterion related with utility.

The evidence of construct validity of interpretations will involve gathering several types of information, but this is particularly crucial when the construct is complex, including language knowledge.

Additionally MacNamara (2000) says that the test content forms a satisfactory basis for the inferences to make from test performance. The procedures are used to establish the relevance of what candidates are asked to do. In other words, it is expected to know whether performance on a general proficiency test can be used to predict performance in particular occupational roles, and vice versa. The problem is that the drawn inferences about candidates based on a test designed for one purpose are not necessarily valid for another unrelated purpose.

The second form of evidence of test's construction validity is the criterion which is related with the results on the test which agree with those provided by some independent and highly dependable assessment of the candidate's ability such as time, the complementation of the set of functions included in the objectives, and concerns about the degree to which a test can predict candidates' future performance, for example, the outcome or assessment in the course.

2.4 Practicality

For Weir (1993) practicality is “to be sure that the task we are using are the most efficient way of obtaining the information we need about the test takers” (p.22).

There is often a great deal of pressure on teachers to make tests as short and as practical as possible but this should never be allowed to put at risk test validity. It inevitably happens that, in the operationalisation of tests, certain authentic features of real life are sacrificed. The problem remains that the less direct the test, the more difficult it will be to translate test scores into behavioral specifications.

Bachman and Palmer (1996) say that “practicality pertains primarily to the ways in which the test will be implemented and, to a large degree, whether it will be developed and used at all. But they also state a contrary form. In the test development process the determination of usefulness is cyclical, so that considerations of practicality are likely to affect our decisions at every state along” (p. 35).

Thus, practicality is defined as the relationship between the resources that will be vital in the design, development, and use of the test and the resources that will be presented for these activities.

2.5 Authenticity

It is a critical quality of language test that has not generally been discussed in language testing. Authenticity thus provides a means for investigating the extent to which score interpretations generalize beyond performance on the test to language use in the command of it.

Another reason for considering authenticity to be important is because of its potential effect on test takers' perception of the test and, hence, on their performance (Bachman and Palmer, 1996).

Bachman (1990) states that language testers have used different terms to identify this test quality. They refers it as "pragmatic" (Oller, 1979), "functional" (Carol, 1980; Farhady, 1980), "communicative (Morrow, 1979; Wesche, 1981; Canale, 1983), "performance" (Jones, 1979; Courchene and Bagheera, 1985; Wesche 1985), and "authentic" (Spolsky, 1985; Shohamy and Reves, 1985) and all of them consider it to characterize the extent to which the task required on a given test has to be similar or real to language use. However, different specifications have to be taken into account in language use and some of them are adaptable to real life, but others not, and they are called "nonreal-life" language use which does not have any meaningful way such as topics, participants, contexts as given examples.

On the other hand, Richards (2003) denotes "authenticity" an important consideration when selecting or designing materials. Authentic texts are not always good model and teachers should be careful select and adapt them for being appropriately used. The minimum acceptable level of authenticity might be stated in two ways: in terms of task characteristics and in terms of expected perception on the part of test takers and test users.

2.6 Interactiveness

According to Bachman and Palmer (1996) interactiveness is the extent and type of involvement of a test task. The individual characteristics that are most relevant for language testing are the test takers' language ability: language knowledge and strategic competence. For example, a test task that requires a test taker to relate the topical content of the test input to his/her own topical knowledge is likely to be relatively more interactive than one that does not.

Therefore, in order to be able to make inferences about language ability, responding to the test task must involve the test taker's areas of language knowledge and her strategic competence.

2.7 Impact

MacNamara (2000) cites that test impact is "the wider effect of test on the community as a whole, including the school" (pp. 74-75). Bachman and Palmer (1996) classify this quality in two levels: micro level, in terms of the individuals who are affected by the particular test use, and a macro level, in terms of the educational system or society.

Consequently, whenever these tests are used, the context of specific values and goals, and our choice will have specific consequences for, or impact on, both the individuals and the system involved. Test takers can be affected by three aspects of the testing procedure: the experience of taking and, in some cases, of preparing for the tests; the feedback they receive about their performance on the test; and the decisions that may be made about them on the basis of their test scores.

In fact, the abuses of the assessment process by some institutions affect directly the veracity of these tests and also affect mainly to the community that really are interest on doing it in the correct form. For that reason, test impact turns complex and unpredictable because of the lack of real results.

2.8 Washback

MacNamara (2000) and Hughes (2002) define backwash or washback as the effect of tests on teaching and learning. On the other hand, Wall and Alderson (1993) use the term a “impact study”.

The results for establishing backwash are taken after the application of a test and that is when resulting is analyzed. Also this quality can be harmful or beneficial because the preparation for it can some to dominate all teaching and learning activities in course. And if the test content and testing techniques are at variance with the objectives of the course, there is likely to be harmful. Hughes (2002) and Davies (1968), agree that a proper relationship between teaching and testing is surely that of partnership and as Hughes states “the good test is an obedient servant since it follows and apes the teaching” (p. 2).

On the other hand, Bachman and Palmer (1996) think that this process takes place in and is implemented by individuals, as well as educational and societal systems, and society at large. So they conclude that washback, as they call it, can be best considered within the scope of impact. Thus in investigation of washback, one must be prepared to find a simply effect of testing on teaching.

As it was stated in these last lines, the usefulness of a test is the most important consideration in designing a test and all the mentioned qualities: reliability, validity, authenticity, practicality, interactiveness, impact or washback contribute to the fact. Thus, they cannot be tested independently so that an appropriate balance in the effectiveness of the designed test could be achieved.

Some other important characteristics have to be followed in the procedure of test construction because the content of a test may be specified along a number of dimensions which will be explained below.

2.9 Moderation

Weir (1993) considers tests have to be managed with general principals, which were firstly explained in this chapter, in order to be successfully directed and interpreted. As well, a group of specific procedures should be taken into consideration when designing a test and based on Murphy's (1979, mentioned in Weir, 1993) information, he provides the next features: level, candidates, appropriate sample, overlap, clear indication, questions and texts, timing, and layout in order to control the design of a test.

MacNamara (2000) gives a more detailed explanation. He calls this process as test specifications and he defines them as a set of instructions for creating the test. Their function is to force explicitness about the design decisions and the specifications will include information on such matters as the length and structure, type of materials with which candidates will have to engage, the source, the extent, instruments, and scoring.

Bachman and Palmer (1996) consider three stages in the design of a test. Such stages involve activities and products which are directly connected with the qualities of usefulness. These stages are:

1. Design. It is related with the design statement.
2. Operationalisation. It is focused on test structure
3. Administration. It is linked with scores and feedback.

They sustain that design is a linear process where qualities of usefulness as well, but adding that resource allocation and management has to take place too. Finally, Hughes (2002) presents content, test structure, timing, medium/channel, techniques to be used, criteria level of performance, and scoring procedures as principal writing specifications for the test.

Hughes (2002) adds that this process is a group team because it has to be developed at least by two colleagues in order to analyze the possible weaknesses in the proposed items in the exam and when an item is not probable to be changed, it needs to be rejected. But mistakes are not hoping to be found.

Subsequently, these specifications are defined below in order to understand their relevance and function in testing design process.

2.9.1 Level of difficulty

Weir (1993) states that task set has to be taken into account in an appropriate level of difficulty when a particular ability is tested. Teachers have to try and put easier tasks/items in order to encourage all students to try their hardest and show their best. If tests starts with the most difficult task the weakest will soon give up.

Hughes (2002) supports the last definition, but he identifies this specification as *operation* where tasks must be in accordance with students' knowledge and level. Bachman (1990) considers into the test organization a *sequence of parts* which may measure level of ability and degrees of difficulty ordered from easy to hard. But also he proposes at randomly sequence in order to introduce an element of control on the test takers' responses. Nevertheless, different candidates may answer items in the sequence presented, while others may not.

Thus, test developers have to identify a variety of sources and kind of information in order to set appropriate tasks for a particular area and kinds of students.

2.9.2 Discrimination

Test does not have to discriminate between candidates' performance at different levels of achievement (Weir, 1993). All kind of test might include some of the more difficult elements from the syllabus which will be achievable perhaps by only the best students in the class. The number of these items would have to be limited so that the other students do not feel unmotivated and these tasks have to be placed at the end of every test.

Hughes (2002) calls this specification as "addresses of text". It refers to the kind of people that the candidate is expected to be able to write or speak to (considering age and status); or the kind of students whom reading and listening material are primarily intended.

MacNamara (2000) considers *Item discrimination.* This specification addresses different aim for him: consistency of performance by candidates across items. When items get harder, teachers would expect those who do it most excellent on the rest.

He also adds that *poor item discrimination indices* are a signal that an item deserves revision. If there are a lot of items with problems of discrimination, the information coming out from the test is confusing because some items will suggest that all test takers will be relatively better in the tasks placed in the test. Thus this information does not make clear real candidates' abilities.

Additionally Bachman (1990) proposes an *item response theor*y which he defines as "a powerful measurement theory that provides higher resources for estimating both the ability levels of test takers (level difficulty) and the characteristics of test items (discrimination)" (p. 7). This assumption helps to moderate tests and measure those characteristics at the same time.

Therefore, this kind of moderation makes equilibrium between test content and students' performance. But also, it is vital to think about that test content has to be selected according to some other important characteristics such as the course syllabus, a point which is developed below.

2.9.3 Appropriate sample

Each test is a representative of a whole from any specific area. In achievement testing this is defined by the content of the course and the methodology that has been employed in the classroom. In spite of this, decisions have to be made in order to get the most relevant sample in which lexical or structural items should be selected. Teachers have to be critical about the choice, because they need to summarize important aspects within the syllabus structure to know if the objective of the program and students learning suit (Weir, 1993).

The important thing is to choose generally from the whole area of content in order to match validity and to get beneficial backwash. Teachers should not concentrate on easy elements they have to be concentrated on testing lengthily and randomly, although it is a possible option to include elements that are particularly important.

Weir (1993) affirms that "what is included in the test according to the syllabus it will be used for marking wider statements about a student's ability in relation to all that has been taught or learnt" (p. 23).

The number of test items or procedures used to measure any given objective depends on mastery and content. There should be a large enough sample of items to allow measurement and secondly the test must be emphasized in the course content that has to be the most valuable for the students (Cohen, 1996).

An excessive covering is not adequate. Many structures, skills or communicative tasks could not be charged in different parts of the test.

2.9.4 Overlap and Questions and texts

Mills and Stoking (1996) use the term *item overlap* to refer to “the extent to which one item may cue the correct response to another item or the extent to which two items depend on the same specific knowledge” (p. 294).

In this moderation, teachers should try to avoid making task overlong or repetitive. Also tests are supposed to avoid visual and mental overload (Weir, 1993).

Input, as Bachman (1990) defines it, may be presented aurally or visually for receptive or oral and written responses, for example. Regarding questions and texts, in general, teachers much avoid *interdependence of items*. One question should not be dependent on ability to answer another. Items must be independent in order to get a clear measure about the considered abilities to the test (Weir, 1993).

2.9.5 Clarity of task

Test task should be explicit; it has to give a clear sign of what the tester is asking. All tests must be carefully proofread in order to eliminate mistakes, and so candidates should be able to misinterpret the task.

According to Bachman and Palmer (1996), writing instructions describe entirely and plainly the structure of test, and consequently test takers will know how they must respond. Bachman and Palmer consider that some instructions are very general and apply to the test as a whole and other instructions are closely linked with specific test tasks.

Test instructions play a decisive role in test takers' performance because they understand the conditions under which the test will be taken, the procedures to be followed and the nature of the tasks they are to complete (The way of the task has to be developed). Madsen (1982, mentioned in Bachman, 1990), cites that vague or erroneous instructions and inadequate time allocation (which will be explained subsequently) create test anxiety, and hence, influences on test performance.

As a consequence, Bachman specifies four facets of instructions: (1) language; (2) channel; (3) the specification of procedures and tasks; and (4) the explicitness of the criteria for correctness.

"Language and channel: Instructions presented might be the test taker's native language or the language being tested, or both.

Specifications of procedures and tasks: The instructions generally specify both the procedures to be followed in taking the test and the nature of the test taker's tasks.

Explicitness of criteria for correctness: The criteria for correctness may be quite clear, as in a multiple-choice test of grammar, in which there is only one grammatically correct choice. In other tests, the criteria may be rather vague, as in a writing test in which the test taker is simply told to make her composition clear and well organized" (pp. 123-124).

Cohen (2001) adds that instructions should be brief, clear and unambiguous. And he suggests giving examples which may help students understand the task, but they may well hinder so they do not give the whole picture. If a new technique is to be employed, test takers should have been given sufficient practice before taking the test. In achievement tests, they should have practiced it in class.Timing has also to be taken into account as it was mentioned in this section.

2.9.6 Timing and layout

A reasonable amount of time must be provided so that test takers can complete the task presented. As it was exposed before, if too little time is made available, stress will not be eliciting students' performance. For that reason, it must be clear to candidates how much time should be spent on each part of a test (Weir, 1993).

Bachman (1990) adds to the last specification that the amount of time devoted to the test is likely to affect the test performance. In some tests, time limit allows test takers answer all the items or parts of them.

Another concern is the presentation of the test for candidates. Weir (1993) considers important to take into account the printed version of the test, so it has to be clear in order to avoid bad effects and also it is essential laid-out question paper must be well organized.

CHAPTER III. TESTING WRITING TECHNIQUES

3.1Overview

There are different techniques teachers can use to test writing skill. But for each task used it has to be considered two main things according to Weir (1993). The first one is referred as the *operation. That is, a* test might include in the writing task various levels of the ability; for example, describing a process. And the second one is related to the conditions under which the task is performed, or the criteria.

Regarding writing, Kitao and Kitao (1996) emphasize that it is important to classify tasks according to test takers' levels in the ability. They state indirect writing activities are suitable for beginners because of their limited requirements; for example, filling in the blanks. While, intermediate and advanced stages, their writing capacity should be tested with direct activities such as free writings connected with real-life events. Thus this chapter introduces some available techniques for testing writing.

3.2 Multiple choice

Teachers use multiple choice items when they want to test writing ability in order recognize sentences that are grammatically correct and it is useful for finding out the difficulties that the students have with certain areas of grammar.

However, according to Weir (1993), multiple-choice items present a problem because only one option has to be selected as the best answer between various options that function as distracters.

There are some features to construct multiple-choice items and they have to be based on:

1. "They should have only one correct answer.

2. Only one feature at a time should be tested. For example, two grammatical functions could not be tested at the same time, if tenses will be taken into account only that point has to be considered and no more.
3. Each option should be grammatically correct. For example,

   He ___________________________ to the movies every day.

   a) goes b) went c) go d) going
4. All multiple-choice items should be at a level appropriate to the proficiency level of the testers.
5. Items should be as brief and clear as possible.
6. Items are arranged in rough order of increasing difficulty. In other words, it has to be started with easy statements and finish with complicated ones" (LELI, 2002. p. 24).

This technique is one of the most common used because it is easy to mark as Harmer (2005) mentions. Multiple choice exercises are easy to score and require little time from teachers to do it. Unfortunately, this type of activity does not represent a writing task as a whole and as Weir (1993) sustains, this activity, such as true-false task, is distrust because the answers could be guessed or get the right answer eliminating the wrong ones. Therefore, multiple-choice improve only the training in test taking techniques rather than any enhance in language ability.

3.3 Error-recognition items

In error recognition items, teachers can use their students' errors that they normally make in their writings. Teachers can select some sentences, then four words or phrases and underlines them and from those sentences, the students will choose the correct one and this one can be considered as a kind of multiple choice item. For example:

1. Select the underlined word or phrase which is incorrect or unacceptable.

I **do** hope you **wouldn't** mind **waiting** for **such** a long time.

A B C D

As well Heaton (1990) describes another option where students are told there is a grammar mistake in each sentence and correct it over the word it occurs; so this type of item is more useful for testing errors because of the omission of articles. See the example below:

1. There is a mistake in grammar in each of the following sentences. Write the letter of that part of the sentence in which it occurs.

   Sun / is shining / brightly today, / isn't it?

   A B C D

3.4 Re-arrangement

For the re-arrangement technique, teachers provide students unscrambled sentences and ask the students to write each sentence in the correct order, with this type of item the teacher can test the order of adjectives, the position of adverbs, inversion, connectors, pronouns and other areas of grammar.

1. We live in a___________________________________.

old / big / wood made / house / black / scary

2. Not only _____________________________________.

   /the examination/ very difficult/ unfair/ was/ but/ it/ was/ also/

This last arrangement can be used for sentences and as well as for words and phrases, usually this kind of task is used for testing connectives and reference devices (Ramírez and García, 2010).

3.5 Changing Words

Changing words is different from the others mentioned before, because the teacher only asks the students to write verbs in the correct tense or in the correct voice. Two examples are given next in order to understand better the technique.

1. Researchers (1) to convince that a drug they (2) to test can improve the memory and that it (3) to be the forerunner of other drugs which eventually (4) to improve mental ability.

   1______________ 2______________ 3_______________ 4_______________

2. Students who were given the drop for a fortnight did considerably (1. well) in test than others. The test included the (2. memorize) of list of words as well as of (3. inform) from two messages transmitted at the same time.

1______________ 2______________ 3_______________ 4_______________

With this exercise students are asked to provide the verb tense which is far from testing writing because this exercise tests grammar (Ramírez and García, 2010).

3.6 Blank-filling or Gap filling

Kitao and Kitao (1996) assume that gap filling is one of the most controlled techniques in writing. Test takers are presented with a passage with blanks, and they have to fill them with the correct missed information.

This activity is a combination between reading and writing skills, and "sometimes it is a problem because it makes it difficult to decide what the scores really mean" (p. 3). Weir (1993) adds that as this task only involves understanding vocabulary and structure of the cohesion devices, writing is reduced because answers are not provided. Also he stresses that gap filling is sometimes considered a suitable format for testing, a "productive writing ability" in a very guided sense" (p. 139).

In the following item the students have to solve an exercise like the following:

A: What you like to order? 1__________

B: I have the fried chicken. 2__________

C: You like rice or potatoes? 3__________

D: Potatoes please. 4__________

E: What kind of potatoes would you? 5__________

F: Mashed, baked or French fries? 6__________

G: I like French fries. 7__________

3.7 Copying

Weir (1993) cites that copying technique is repeated according to the level of the students. It is the first task that produces writing because test taker has to be sure to write every letter contained in the sentence. For that reason, it is an appropriate technique for children. But when it is used, formal writing has to be employed because several consequences are going to be face according to the illegibility terms.

Copy the following:

Fawzia likes figs and grapes.

____

She doesn’t like orange juice.

____

She’s going to ride to the market.

____

3.8 Controlled writing

Ramírez and García (2010) state that “controlled writing refers to have control over what students write and subsequently, teacher can ensure that certain grammatical patterns and language functions are tested” (p. 18). Some of the forms that controlled writing can take a *transformation*; this type of item requires that the students re-write sentences according to a certain pattern. A similar way of using *transformation* is by giving the students a word in brackets instead of the beginning of the new sentence. A negative aspect of using this type of item is the lack of context it provides as shown in the example below.

In Australia, 87% of married couples have children (most)

____.

This exercise does not really test writing because of the using the word provided that limits again the task (Ramírez and García, 2010).

3.9 Broken sentences

The teacher can test the ability of writing sentences from a series of words and phrases by using broken sentences and the students have to make as many changes as possible in order to form good sentences by adding articles, prepositions, punctuation marks, and verbs in the correct tense. An important aspect to take into account when using this type of items is that the teacher should put the broken sentences in the form of a paragraph, a dialogue or a letter. For example:

1. *Take*/ drugs and stimulants / *keep* awake / while revise examination/ often *be* very harmful. / It *be* far better / *lead* / balance life a/ and *get* enough sleep/ every night/ There *be*/ limit/ degree and span/ concentration/ which you be capable / *exert*/ Brain / *need* rest/ as much body. / Indeed, / it *be* quality/ than quantity work / that *be* important. (LELI, 2002. p. 29).

Candidates only have to order the words. This exercise does not really test students' ability to write complete paragraphs.

3.10 Sentence and paragraph completion

A realistic task of controlled writing is sentence and paragraph completion because writing is integrated with reading comprehension, so by reading, students can complete a sentence or a dialogue.

Example:

Most of my students in my class where rather lazy and did not enjoy the curse. Some even stay away from school quiet often. Pauline, however,

____

____

____________________________ etc.

This technique is more challenging and it is a mixture of both reading and writing skills (Kitao and Kitao, 1996).

3.11 Editing

Weir (1993) considers *editing* as a testing writing technique or as Kitao and Kitao (1996) call it *making corrections*. This activity contains a number of errors in grammar, spelling and punctuation and so mistakes have to be corrected in the spaces provided.

Example:

“I am <u>an</u> student of <u>Inglish</u>

a English

Afterwards the test takers have to make corrections on the wrong statements, but as Kitao and Kitao (1996) assume, this kind of task does not represent a writing task as a whole because it is only edited. Writing is out of the perspective.

3.12 Form filling

Form filling is an important task for students because in real life they are asked to fill forms. If they want to get a job where English is the first language, they have to give information about themselves. When using notes and diaries the teacher can ensure that students are working on a similar written task and what they have written can be compared fairly with one another. Acordding to Kitao and Kitao (1996) this activity is a limited way of testing writing. They state that “the advantage of this task is that it is a least somewhat communicative, but the disadvantage is that it does not require any connected discourse or any use of language greater than lexical knowledge and a small amount of grammar” (p. 3).

Finally Weir (1993) mentions that these types of exercises are more communicative and are based on real life situations such as joining to a club or giving information for a survey.

3.13 Letter Writing

Letter writing is a common activity for testing the ability. The development of the letter has to be related to a condition which is stated in the instructions in order to reflect real result when they will be tested.

Also drawings and pictures are options to be integrated in the activity in order to give information about a situation the candidates are expected to write. This activity test reading and writing at the same time (Kitao and Kitao, 1996).

3.14 Open-ended essay test

Essay writing is one of the most common activities in writing ability for advance students and this is considered as the best writing test (Kitao and Kitao, 1996).

The development of essays varies from length and number of words to several sentences. Setting the task is a reasonably affair. Topics have to be general and heavily on candidate’s knowledge and imagination. Little guidance could be given in order to get assessment. The extent of the use of the language is an advantage. Sometimes candidates prefer to develop essays over an open-ended question in different ways. The only disadvantage is that “essays are very difficult in comparing performances, especially if the production of different text types is involved” (Weir 1993, p. 144).

These are some of the techniques that teachers can use when testing writing. They can choose them taking all the aspects discussed in chapter two.

CHAPTER IV. MARKING CRITERIA

4.1 Overview

In this chapter the importance of marking criteria in testing the writing process is provided in order to allow students to know criterion used before and after the administration of the exam. When language skills are tested, objective and subjective testing have to be taken into account.

4.2 Objective and subjective testing

Hughes (2003) argues that objective and subjective testing is only different in the process of scoring. The scoring is objective if it is required a judgment on the part of the scorer*.* On the other hand, if judgment is called for, the scoring is said to be *subjective*.

Objective in scoring is required after by many testers, not for itself, but for the greater reliability it brings. Objective writing may be very factual, as in lab reports, technical explanations, and legal records. Or it may be about ideas and problems, as in news articles, professional communications, or research analyses (p. 22)*.*

An objective test is a test that has right or wrong answers and so can be marked objectively. It can be compared with a subjective test, which is evaluated by giving an opinion, usually based on agreed criteria.

Objective items require students to select the correct response from several alternatives or supply a word or short phrase to answer a question. Some examples of objective items are multiple-choice, true/false, matching, and completion items.

Subjective items require students to write and present an original answer. This kind of procedure determines the content to be covered and testers make subjective decisions about the selection of the best items where best answers would be those on essays or multiple-choice tasks (Bachman, 1990).

In general, the less subjective the scoring, the greater agreement there will be between two different scorers (and the scores will be variable even when one person is tested with the same test paper on different occasions). However, there are ways of obtaining reliable subjective scoring, even of compositions.

The main difference between these two writing types is the fact that one is factual while the other is strictly based on opinions and perspective.

4.3 Rubrics

Rubric is a scoring tool that contains the aspects taken into account for a piece of work. These usually inform certain level of knowledge or progress expected from several levels of quality.

These instruments grade the different students' performances. For example, like how they perform in certain activities in which reading, listening, collaborative work and behavior are marked, so teachers can get a final grade. Another use is when grading a specific task, for example writing an essay; here the teacher can use rubrics to take into account punctuation, content, organization and spelling. Table 1 is an example of a rubric used to test an essay:

Table1. Rubrics used to test an essay

Fluency	5 Following style- very easy to understand –both complex and simple sentences – very effective 4 Quiet flowing style –mostly easy to understand –a few complex sentences –effective 3 Style reasonably smooth –not too hard to understand –mostly (but not all) simple sentences – fairly effective 2 Jerky style –and effort needed to understand and enjoy –complex sentences confusing – mostly simple sentences on compound sentences 1 Very jerky –hard to understand –cannot enjoy reading –almost all simple sentences –complex sentences confusing –excessive use of ‘and’
Grammar	5 mastery of grammar taught on course –only 1 or 2 minor mistakes 4 A few minor mistakes only (prepositions, articles, etc.) 3 Only 1 or 2 major mistakes but a few minor ones 2 Major mistakes which lead to difficult in understanding –lack of mastery of sentences construction 1 Numerous serious mistakes –no mastery of sentences construction –almost unintelligent
Vocabulary	5 Use of wide range of vocabulary taught previously 4 Good use the new words acquired –use of appropriate synonyms, circumlocution, etc. 3 Attempts to use words acquired –fairly appropriate vocabulary on the whole but sometimes restricted –has to resort to use of synonyms, circumlocution, etc. on a few occasions 2 Restricted vocabulary .use of synonyms (but not always appropriate) –effects meaning 1 Very restricted vocabulary –inappropriate use of synonyms seriously hinders communication.
Spelling	5 No errors 4 1 or 2 minor errors only (e.g. ie or ei) 3 Several errors –do not interfere significantly with communication .not too hard to understand 2 Several errors –some interfere with communication –some words very hard to recognize 1 numerous errors –hard to recognize several words –communication made very difficult

Source: Heaton, 1990, p98.

The first aspect is to know the features that teachers want to mark like fluency, grammar, vocabulary and spelling. Each feature has a grade from 1 to 5, for each one of them. The teacher may state notes that can help to remember what is taken into account when grading and tick the appropriate box as shown in the table below:

Table 2: Example of using a rubric for marking students' writing

	5	4	3	2	1
Fluency		√			
Grammar			√		
Vocabulary				√	
Content	√				
Spelling			√		

Rubrics are not only for teacher's use, but also for students, as a self-evaluation for example in developing a research, in which the student has to follow certain steps so they can check their own progress. And also these instruments allow peer evaluation where students collaborate and learn from others. Through discussions with peers, middle level students can verbalize their concerns and ideas in a way that helps them clarify their thoughts and decide in which direction to proceed.

An interesting variable when creating rubrics is to develop them along with the students, in this way students may feel that the aspects to be graded are chosen fairly by them and by the teacher. They also realize about the components and how an appropriate writing task has to be made before starting it.

A reason why teachers use rubrics is that students can realize what the teacher wants them to do, also help the students to be aware of their work and grade; finally rubrics can provide feedback because students can observe their mistakes or weaknesses in the ability. Also the use of rubrics may reduce students' anxiety and nervousness because they will know in advance what aspects of their writing are being tested.

Hence, Weir's (1993) assumption, stated in chapter two, corroborates that a test rubric should be friendly, understandable, explicit, brief, simple and accessible for test takers. Then, test rubric should not be more difficult than the text or task.

4.4 Scales

When teachers are going to mark students' writings, they should score only what the students have written, for this reason it is important that the teacher can find the best way to mark students' performance. There are two approaches to scoring: holistic and analytic.

4.4.1 Holistic scoring

According to McNamara (2000) rating or scoring requires raters "to provide separate assessment for each of a number of aspects of performance rather than to record a single impression of the impact of the performance as a whole" (p. 43). This means that the teacher only counts the number of errors that a student has made and gives him or her mark. An advantage of this scoring system is that it is very fast to score and that the teachers can check a piece of writing of a single student more than once. In order to make the mark reliable the teacher can give the same piece of writing to different teachers; this way, the reliability degree may be higher. On the other hand, a disadvantage for using this scoring system is that sometimes the errors that the students make have different levels of importance. For this reason, it is important to select the appropriate scoring system for each level and for the purpose of the test.

4.4.2 Analytic scoring

According to McNamara (2000) analytic scoring “requires the development of a number of separate ratings scales for each aspect assessed” (p. 44). This means that the teacher should identify the areas that she or he wants to mark, for example content, vocabulary, fluency, etc., and give to each area an appropriate mark. Some advantages of using this type of scoring system is that the teachers can focus only on the skill they want to mark, they only concentrate in some aspects of students’ performance and the fact that the teacher has to give different scores makes the total mark more reliable. On the other hand, the main disadvantage of the analytic method is the time that it takes to mark. A second disadvantage is that sometimes the attention can be diverted from the whole effect of the piece of writing because of the concentration on the different aspects.

Thus, as Hughes (1989) states “any scale which is used, holistic or analytic, should reflect the particular purpose of the test and the form that the reported scores on it will take” (p. 105), this means that the scoring scales should cover the purpose of the test and should be close to the teacher needs in order to make the students’ mark valid, also the scale should be adapted for the situation in which they are going to be used.

Conclusions

Testing writing may represent a difficult thing to do because it involves many aspects such as the test purpose, principles for testing, specifications, selection of testing techniques, and criteria, among others. In addition, teachers should focus their interest in order to know students' real progress in the language ability not in what students are not able to do.

Testing is a process that involves only the design of tests and the process must be based on a established framework, so that we as teachers can determine the different specifications which will help us to guide and design our tests. However, frequently, we as teachers overlook the fact that teaching and testing are closely related and most of the time we include activities that we have never been practiced in the classroom or we base our design on exercises provided in the textbook. These two "practices" do not really contribute to know our students' ability to write because they do not follow the aspects required for test design.

Teachers have to be aware of the different techniques they can use to test writing depending on the purpose of the test. This means that teachers can choose from a variety of methods to test this skill in order to be aware that there are ways of testing not only by the "free writing" where students write about unrealistic topics students have to make up and that far away from being authentic.

Another aspect teachers have to be aware of is the fact they have to set a marking criteria when testing writing in order to reduce subjectivity. This is very important because sometimes teachers assign a mark based on their "feelings" or on the mark they think students deserve based on their performance in the classroom. However, this practice must be not promoted because it is neither valid nor reliable. Thus, teachers must establish very clearly the aspects they want their students to mark and students must know them.

This research can be used by teachers and students because it describes the process involved in test design for the writing skill. It also provides a variety of techniques teachers can use to test this skill. However this research may be used as a reference and further research on this topic is suggested.

References

Bachman, L. F. (1990). Fundamental Considerations in Language Testing. Oxford: Oxford University Press.

Bachman, L. F. and Palmer, A. S. (1996). Language Testing in Practice: Designing and Developing Useful Language Tests. New York: Oxford University Press.

Black, P. & William, D. (2001). Inside the Black Box: Raising Standards Through Classroom Assessment. [Online]. Phi Beta Kappan. Available at < http://www.collegenet.co.uk/admin/download/inside%20the%20black%20box_23_doc.pdf > [June 13th 2010].

Cohen, A. D. 2001. Second language assessment. In M. Celce-Murcia (Ed.), *Teaching English as a second or foreign language*. 3rd Ed. Boston: Heinle & Heinle/Thomson Learning.

Harmer J. (2005). The Practice of English Language Teaching. England: Pearson.

Hedge, T. (2000). Teaching and learning in the language classroom. Oxford: Oxford University Press.

Hughes, A. (2002). Testing for Language Teachers. Cambridge: Cambridge University Press.

Heaton, J. B. (1990). Classroom Testing. Longman Keys to Language Teaching. Hong Kong: Longman.

Kirster, K. (1999). Webster's New World Dictionary. Third College Edition. U.S.A. Pocket Books.

Kitao, S. and Kitao, K. (1996) Testing Writing. ERIC Document.

LELI. (2002). Diseño y Evaluación de Exámenes en Inglés. UAEH.

McNamara, T. (2000). Language Testing (Oxford Introduction to Language Study ELT). Oxford: Oxford University Press.

Mills, C. N., & Stocking, M. L. (1996). Practical issues in large-scale computerized adaptive testing. *Applied Measurement in Education*. 9, 294.

Ramírez, B. and García, A. (2010). Testing Students' Writing: an exploratory study of current teachers' practice. UAEH.

Richards, J. C., 2005. Second Language Writing. 2n Ed. Cambridge: Cambridge University Press.

Saskatchewan, R. (1997). Assessment and Evaluation. In English Language Arts: A Curriculum Guide for the Middle Level (Grades 6-9). [Online]. Saskatchewan Education. Available at: <http://www.sasked.gov.sk.ca/docs/mla/assess.html> [April 15th 2010].

Villardón, L. (2006). Evaluación del aprendizaje para promover el desarrollo de competencias. Educatio siglo XXI. Universidad de Deusto.

William, M. K. (2006). Definition of Evaluation. In Introduction to Evaluation. [Online]. Trochim. Available in: <http://www.socialresearchmethods.net/kb/intreval.htm> [April 13th 2009].

Weir, C. J. (1993). Understanding and Developing Language Tests. Great Britain. Prentice Hall International.

Zamel V. (1985). *Responding to Students Writing*. Tesol Quarterly,19 (1), 79-101