

CHAPTER II

TESTS AND TEST CONSTRUCTION

For this literature review, the focus is on language tests and distinctions in testing, and discussion of the qualities of achievement tests, especially in the areas of test construction, validity and reliability. The literature review includes the following topics:

- Tests in Language Assessment
- Kinds and Purposes of Test
- Achievement Tests
- Key Distinctions in Testing
- Test Construction Process
- Validity
- Reliability

Tests in Language Assessment

Tests, one form of assessment, are widely used all over the world to assess learners' abilities. Tests are much more than the process of students answering the content in the exam paper or taking the test. Wharton (2004, p.1) says that "A test is not just a question paper or a set of specifications: far more importantly, it is what happens when real candidates interact with these". Testing is a process, which involves steps such as preparation of questions, students taking the test, teachers marking and reporting the results and giving feedback to students. Brown (2001, p.383) claimed that

“A test is first a method. It is a set of techniques, procedures and items that constitute an instrument of some sort that requires performance or activity on the part of the test taker and sometimes on the part of the tester as well.”

Kinds and Purpose of Test

Tests can be categorized according to the types of information they provide about the students' ability and the purposes of the tests.

Basically, there are four types of tests (Hughes, 2003; Heaton, 1990; Bachman, 1990; McNamara, 2000); proficiency tests, achievement tests, diagnostic tests and placement tests. These are discussed in more detail below.

Proficiency tests

Proficiency tests are designed to assess the general language ability of a student. Hughes (2003, p.11) cited that “proficiency tests are designed to measure people’s ability in a language, regardless of any training they may have had in that language.” Proficiency tests are not limited to any one course, curriculum or single skill in the language either (Brown, 2001).

According to Hughes (2003), the contents of proficiency tests are not based on the contents or objectives of language courses, but on a specification of what candidates have to be able to do in the language. A proficiency test usually consists of standardized

multiple-choice items on grammar, vocabulary, aural comprehension, reading comprehension and sometimes a sample of writing (Brown, 2001).

Generally, proficiency tests are administered to students from various language learning backgrounds. The International English Language Testing System (IELTS) and the Test of English as a Foreign Language (TOEFL) are standardized proficiency tests used globally.

Achievement tests

Achievement tests are used to measure how successful individual students, or groups of students, have been in achieving the objectives of particular courses. As the focus of this study is an achievement test, more discussion about achievement tests is provided later.

Diagnostic tests

Diagnostic tests aim to identify learner's strengths and weaknesses. These kinds of tests are usually conducted at the start or in the middle of a course. Diagnostic tests are very useful because they can provide information about students' language ability. Based on the results of a diagnostic test, teachers can decide what they should do in terms of teaching focus and objectives.

However, diagnostic tests require enough samples of language items to assess students' real ability and can be time-consuming and therefore sometimes impractical (Hughes, 2003).

Placement tests

The purpose of placement tests is to provide information that will help to place students at the stage (or in the part) of the teaching program most appropriate to their abilities.

They aim to help put the students in an appropriate course or class for their language level. A placement test is usually conducted when students enter a new school or university. These tests are useful for large institutions which receive many students.

Placement tests should be designed for particular situations, as a purchased proficiency test may not meet the particular needs and wants of the school. Hughes (2003, p.17) also mentioned that “the placement tests that are more successful are those constructed for particular situations”. In-house proficiency tests are more appropriate because learning contexts are different from one another and a test constructed without taking the target students into account may do more harm than good.

These types of test are usually used in language testing fields according to the purposes of the language programs and situational needs. As this study is concerned with evaluating an achievement test, it may be useful to provide more discussion on achievement tests.

Achievement Tests

This kind of test is developed to assess how much learners have achieved from their learning over a certain amount of time. McNamara (2000, p.6) wrote, “achievement tests accumulate during, or at the end of a course of study to see whether and where progress has been made in terms of goals of learning”. Moreover, this kind of test is directly related to the language courses as it helps those involved in programs to find out if the objectives have been successfully achieved or not. There are two kinds of achievement tests: final achievement tests and progress achievement tests (Hughes, 2003; Weir, 1993). Final achievement tests can be defined as tests to assess learners’ achievement and are administered at the end of the course of study. A progress achievement test is a test which is intended to measure the progress the students are making during the course (Hughes, 2003, p.13, 14).

Good achievement tests should have the following qualities:

- be valid (Hughes, 2003; Weir, 1993; Heaton, 1990).
- be reliable (Hughes, 2003; Weir, 1993; Bachman and Palmer, 1996).
- be practical (Hughes, 2003; Weir, 1993).
- use direct testing (Hughes, 2003; Weir, 1993; Heaton, 1990).
- measure what the program intends to teach (Bailey, 1996).
- be based on objectives and cover the syllabus (Hughes, 2003).
- use authentic tasks and texts (Bailey, 1996).
- have complete keys, scoring rubrics and marking schemes (Hughes, 2003; Weir, 1993; Bachman & Palmer, 1996).

There are some basic decisions that should be made when constructing a test. These basic decisions arise because of key distinctions in language testing. These distinctions are discussed below.

Key Distinctions in Testing

Direct and Indirect tests

There are two approaches to test construction: using direct and indirect tests. Direct tests require the test takers to perform exactly the skills that the test intends to measure (for example, asking the test takers to write an essay in a writing test). This kind of test is usually used for testing productive skills like speaking and writing. Hughes (2003, p.17) mentioned that “Direct testing is easier to carry out when it is intended to measure the productive skills of speaking and writing.” Indirect tests, on the other hand, assess the skills indirectly (for example, assessing students’ writing ability by their ability to edit sentences). Hughes (2003, p.18) explained that “Indirect testing attempts to measure the abilities that underlie the skills in which we are interested.” This kind of testing is usually used for testing reading ability, writing ability and sometimes for speaking ability.

Using indirect testing has some problems because we cannot always prove the relationship between the performance of the skills we are interested in measuring and the performance in the test. That is why direct testing should be encouraged whenever

possible. Direct testing can therefore enhance the validity of the test. Hughes (2003, p.17) points out the advantages of using direct tests as follows:

“Direct testing has a number of attractions. First, provided that we are clear about just what abilities we want to assess, it is relatively straightforward to create the conditions which will elicit the behavior on which to base our judgments. Secondly, at least in the case of productive skills, the assessment and interpretation of students’ performance is also quite straightforward. Thirdly, since practice for the test involves practice of the skills that we wish to foster, there is likely to be a helpful backwash effect.”

Discrete point and Integrative testing

Discrete point tests are designed to test one element at a time, item by item (for example, in a form of a series of items, each testing a particular grammar structure) (Hughes, 2003; McNamara, 2000). Discrete point tests will always tend to be indirect tests. McNamara (2000, p.14) stated that “ ‘a discrete point test’ could be achieved through constructing a test consisting of many small items all directed at the same general target-say, grammatical structure, or vocabulary knowledge.” Discrete points can be used for testing grammar, vocabulary and the four skills of language in isolation. McNamara (2000, p.14) also mentioned that multiple-choice items are the most suitable type for discrete tests. The discrete items can be useful for testing particular structures of language but cannot be appropriate for assessing learners’ ability of communication in real life situations, as learners have to use more than one language element in communicating. McNamara

(2000, p.14) also stated that “the discrete point tradition of testing was seen as focusing too exclusively on knowledge of the formal linguistic system for its own sake rather than on the way such knowledge is used to achieve communication.”

Integrative testing requires the test takers to use many language elements in the completion of a task (for example, dictation, writing compositions, making notes while listening to a lecture) (Hughes, 2003; McNamara, 2000). These kinds of tests are likely to be direct tests. They involve an integrated performance on the part of the language user. Integrative tests are useful for assessing how much learners can use language elements to achieve communication. However, integrative tests tend to be expensive, as they are time consuming and difficult to score, requiring trained raters, and in any case may be potentially unreliable (McNamara, 2000).

Norm-referenced and Criterion-referenced tests

Another distinction of language test is the two distinct ways of interpreting test scores: norm-referenced (NR) tests and criterion-referenced (CR) tests. The two primary distinctions between NR and CR tests are in their design, construction, and development, and in the scales they yield and the interpretation of these scales (Bachman, 1990). Bachman (1990, p.72) claimed that “Norm-referenced tests are designed to enable the test user to make ‘nominative’ interpretations of test results.” This kind of test relates one candidate’s performance to that of other candidates (Hughes, 2003). Test results are interpreted with reference to the performance of a given group, or norm group (Bachman,

1990). Significant examples of NR (Norm-referenced) tests are standardized tests such as TOEFL.

Standardized tests have three main characteristics (Gronlund cited in Bachman, 1990, p.74). First, standardized tests are based on fixed or standard content, which does not vary from one form of the test to another. Second, there are standard procedures for administering and scoring the test, which do not vary from one administration of the test to the next. Finally, standardized tests have been thoroughly tried out, and through a process of empirical research and development, their characteristics are well known (Bachman, 1990, p.74).

However, NR tests cannot describe directly what the student is capable of doing in the target language (Hughes, 2003). They just relate the individual's performance to other candidates. Criterion-referenced tests aims to classify test takers according to their ability to perform a language task or set of tasks satisfactorily or not (Hughes, 2003). Bachman (1990, p.75) describes CR tests as "designed to be representative of specified levels of ability or domains of content, and the items of parts will be selected according to how adequately they represent these ability levels or content domains".

CR tests offer two advantages for learners. First, these kinds of tests set meaningful standards for learners in terms of what the learners can achieve, which will not change even if tried on different groups of candidates, and secondly, they motivate the students to attain those standards. (Hughes, 2003).

Objective and Subjective scoring

This distinction is based on the extent to which scoring the response requires judgment on the part of the scorer. Bachman (1990, p.76) explained that “in an objective test the correctness of the test taker’s response is determined entirely by predetermined criteria so that no judgment is required on the part of the scorer”. Objective scoring does not need the scorer’s opinion to the students’ response. In this kind of test, an answer key is provided by the test developers and the scorers just need to give marks by reference to this. Examples of objectively scored items are multiple-choice and matching.

Subjective testing requires scoring by opinion and judgment on the part of the scorer (for example, essays, letter writing, short answers). Hughes (2003, p.22) mentioned that there are degrees of subjectivity in testing. The degree of subjectivity would vary between, for example, short answers for reading passage, and scoring of compositions. The latter would require a greater degree of judgment on the part of the scorer. Subjective tests also need specific marking schemes to score the written or oral response of the test takers.

Objective scoring can be helpful for enhancing the reliability of a test, as it does not need judgment from the scorer. On the other hand, subjective scoring should be carefully structured in order to obtain reliability of the test results (Hughes, 2003).

The distinctions discussed above represent decisions that have to be taken as part of the test construction process. It is this test construction process that will be described in the following section.

Test Construction Process

There are two general approaches to constructing achievement tests. The first approach is basing tests directly on the syllabus or on the materials used. This is known as the “syllabus-content approach” (Hughes, 2003, p.13). Hughes added that this approach could be considered fair because the test contains only what the students have been taught. However, he also points out that if the syllabus is badly designed or the materials used are badly chosen, the test results can be misleading. The performance on the test may not reflect the achievement of course objectives, for example, the course objective is to develop speaking skills but if the course content and the test itself do not provide students with enough chances to speak, test results will not reflect students’ achievement in terms of course objectives. Teachers should be careful not to base tests totally on course content because it can lead to “teaching to the test”. The primary purpose of teaching is to encourage students to learn, not just to pass examinations.

The other approach mentioned by Hughes to develop achievement tests is to base the test directly on the course objectives. Hughes (2003, p.13) claimed, “It compels course designers to be explicit about objectives and it makes it possible for performance on the test to show just how far students have achieved those objectives”. In order to achieve the objectives, the syllabus and materials must be consistent with the objectives of the course.

There are certain stages that are vital in constructing a test (Hughes, 2003; Alderson et al, 1995). When developing an achievement test, seven stages are required to be taken into consideration in order to develop a good achievement test.

1. Stating the objective: It is necessary to make clear the purpose of the test before constructing it. Then, the kind of test, the abilities to be assessed, the possible backwash, and the value of the results should be discussed (Hughes, 2003).
2. Writing specifications for the test: Alderson et al. (1995, p.9) claimed that “A test’s specifications provide the official statement about what the test tests and how it tests it.” A set of specifications for the test should be written at the outset after the purposes and objectives of the test have been decided. The test specification will include information on content, test structure, timing, medium/channel, techniques to be used, criteria level of performance, and scoring procedure (Hughes, 2003; Alderson et al, 1995). This specification is needed for a range of different groups such as the test developer, test editors, test takers, teachers, admission officer of the program and, for public tests, publishers who wish to produce textbooks related to the tests (Alderson et al, 1995).
3. Writing and moderating test items: After writing the specifications of a test, the test developers can start writing the test items. Choices have to be made. The test developers should choose widely from the whole area of content. Especially in achievement tests, the test writer should be careful to write the items to cover the syllabus and meet the objectives of the test. Items should always be written with the specifications in mind. Alderson et al (1995, p.40) wrote that “for achievement tests, it is clearly important that those who write the test know what is reasonable to expect students to have covered at the particular stage in learning and also how far students

have actually progressed through the curriculum.” There are many kinds of test item types like multiple-choice, gap filling, ordering tasks, cloze etc. Each item type has both strength and weakness. Therefore, the item writer should choose the test item type that is most suitable for the students and to reach the objectives of the tests. Then the items should be moderated. Moderation is the investigation of the proposed items by (ideally) at least two colleagues who are not the test developers (Hughes, 2003). This is to try to find weaknesses in the items and remedy them.

4. Informal trialling of items: After the test items have been moderated, the test should be trialled (Hughes, 2003; Alderson et al, 1995; McNamara, 2000). It is recommended that the items must be presented in the form of a test to native speakers. Items that prove difficult for the native speakers almost certainly need revision or replacement. The items that have survived moderation and informal trialling on native speakers should be put together into a test, which is then administered under test conditions to a group similar to that for which the test is intended (Hughes, 2003). The pretesting of the test is necessary as it can help the examiners to know how difficult the test items are, whether the items really work or not, and whether the test is appropriate for the students’ level or not (Alderson et al, 1995). Problems in administration and scoring should be noted.
5. Analysis of the results of the trial: After trialling, the results should be analyzed for reliability of the items and the test. If there are flaws in some items, necessary changes should be made (Hughes, 2003; Alderson et al, 1995).

6. **Validation:** The final version of the test can be validated. It is necessary for the test to be validated before administering it (Hughes, 2003; Alderson et al, 1995; Bachman, 1990). Discussion on the validity will be presented in the next section.

7. **Training of examiners and administrators:** It may be necessary to train the examiners and the administrators before administering the test. Alderson et al (1995, p.105) pointed out that “the training of examiners is a crucial component of any testing programme”. A handbook that contains the rationale of the test, description of the test, sample items, an explanation of how test scores are to be interpreted, training materials and details of test administration should be produced for the test takers, test users and staff. Using the test handbook and other materials, all staff who will be involved in the test process should be trained (Hughes, 2003; Alderson et al, 1995).

This can be a basic framework for test construction. When constructing a particular test, some points from this stage may be omitted but all the stages are important in constructing a test. However, there are two more important test qualities that every test should have; the validity and the reliability of the test. These are also essential components of test construction process. A discussion on validity and reliability as they relate to tests is given below.

Validity

Validity is a vital quality for every test. Validity is defined by Henning (1987, p.89) as “the appropriateness of a given test or any of its component parts as a measure of what it is purported to measure. A test is said to be valid to the extent that it measures what it is supposed to measure. It follows that the term *valid* when used to describe a test should usually be accompanied by the preposition *for*. Any test then may be valid for some purposes, but not for others.”

A test should test what the writer wants it to test (Hughes, 2003; Brown, 2001; Weir, 1993). There are various aspects of validity described in language testing literature including content validity, face validity, construct validity, concurrent validity and predictive validity. All these forms of validity are important keys in developing language tests depending on the purpose and kind of tests. However, according to Alderson et al (1995, p.171), these aspects of validity are in reality different methods of assessing validity. They point out two main ways of assessing validity in language testing; internal validity and external validity, and define these two terms as “...internal validity relates to studies of the perceived content of the test and its perceived effect, and external validity relates to study comparing students’ test scores with measures of their ability gleaned from outside the test.”

There are ways of assessing internal and external validity. Face validity, content validity and response validity are used for assessing internal validity whereas concurrent validity and predictive validity are use for assessing external validity.

Among these types of test validity, the content validity of assessing internal validity is particularly relevant for validation of achievement tests which is measuring learners' progress. However, all aspects of validity should be examined in order to find out the validity of the test.

Content validity

"A test is said to have content validity if its content constitutes a representative of language skills, structures, etc. with which it is meant to be concerned" (Hughes, 2003, p.26). The test samples should be selected carefully based on the course objectives and syllabus. Heaton (1988, p.160) defines content validity as depending on a careful analysis of the language being tested and of the particular course objectives. Content validation can be done by 'experts' making judgments in a systematic way. One common way is to analyse the content of a test and to compare it with the test specifications (Alderson et al, 1995). Therefore, in order to find out whether a test has content validity or not, the test should be checked to see if it includes enough samples based on the test specifications.

Face validity

If a test looks as if it measures what it is supposed to measure, then it can be said to have face validity (Hughes, 2003, p.33). If it is a speaking test, the test should include requirement for test taker's oral responses. According to Heaton (1988), face validity is one basic factor for tests as it can provide a quick and reasonable guide and it can maintain student's motivation. Learners need to be convinced that the test is really testing

what it claims to test. Only then, will they take it seriously. Alderson et al (1995, p.173) claimed, “Tests that do not appear to be valid to users may not be taken seriously for their given purpose”. Face validity concerns the appeal of the test to the lay judgment, judgment of the students, their parents, and members of the public (Davies, 1990). The concept of face validity is closely related with content validity because both are concerned with the test content (Brown, 2001).

Predictive Validity

Predictive validity concerns the degree to which a test can predict students’ performance in the future. It looks forward to student’s future performance in new learning programs or places of employment. Placement tests before a course would be judged to have predictive validity if they result in successful placement of students in suitable classes (Heaton, 1990; Hughes, 2003).

Concurrent validity

The value of a test can be determined by comparing its results with the results of another test that is thought to be a valid measure of the same or similar activities. If the outcome of the two tests is similar, it can be said that the test in question has concurrent validity. Thus, concurrent validity can be defined as a result of comparison of the test result with that of another well-established and valid test which is administered roughly at the same time (Heaton, 1988; Cohen, 1994).

Construct Validity

A test has construct validity if it can measure certain specific characteristics in accordance with some theory of language behavior and learning (Heaton, 1988). The term “construct” is defined by Hughes (2003, p.31) as “any underlying ability or trait which is hypothesized in a theory of language ability”. Every issue in language learning and teaching involves theoretical constructs (Brown, 2001). Cohen (1994, p.40) claimed that, “construct validity examines whether the instrument is a true reflection of the theory of the trait being measured”. Construct validity is related to research. Theories of language testing and assumptions are to be validated and supported by construct validity. When the theory is validated, the test result can be considered accurate and dependable.

Validity in scoring

It is vital to consider that not only test items should be valid but also the scoring of the responses must be valid too (e.g., when scoring reading comprehension, the scoring of the responses should not take into account spelling or grammatical errors). The rating of every item should be valid (Hughes, 2003, p.33).

Here are some recommendations for enhancing validity:

- A test should limit itself to measuring only what it is intended to test (Weir, 1993; Hughes, 2003).

- A test should measure what it is supposed to measure (Hughes, 2003; Bailey, 1996).
- A test should have full specifications and the test content should be a fair reflection of the specifications (Hughes, 2003).
- A test should involve realistic language activities performed under appropriate conditions (Weir, 1993).
- Test content should constitute a representative sample of the language skills, structures, etc. with which it is meant to be concerned (Hughes, 2003).
- A test should cover the syllabus or what has been taught (Hughes, 2003; Heaton, 1988; Weir, 1993)
- The test should be direct (Hughes, 2003).
- Scorers should score what the test is intended to measure (Hughes, 2003; Bachman, 1997).

As discussed above, validity is an essential component in language testing, so how to assess the validity is also important. In assessing validity, it is important to make clear the purpose of the assessment. Alderson et al (1995, p.170) mentions, "If it is to be used for any purpose, the validity of used for that purpose needs to be established and demonstrated." It is better for tests to be validated in more than one way. Alderson et al (1995, p.171) state that "... the more different 'types' of validity that can be established, the better, and the more evidence can be gathered for any one 'type' of validity, the better." They also provide procedures for evaluation of validity. These

can be used in assessing different types of validity in language tests and summarized in Table 1.

Table 1: Procedures for assessing different types of validity

<u>Types of Validity</u>	<u>Procedures for Evaluation</u>
Face Validity	Questionnaires to, interviews with candidates, administrators
Content Validity	<ul style="list-style-type: none"> a) Compare test content with specifications/syllabus. b) Questionnaires to, interviews with “experts” such as teachers, subject specialists, applied linguists. c) Expert judges rate test items and texts according to precise list of criteria
Concurrent Validity	<ul style="list-style-type: none"> a) Correlate students’ test scores with their scores on other tests. b) Correlate students’ test scores with teachers’ ranking. c) Correlate students’ test scores with other measures of ability such as students’ or teachers’ rating.
Predictive Validity	<ul style="list-style-type: none"> a) Correlate students’ score with their scores on tests taken some time later. b) Correlate students’ test scores with success in final exams. c) Correlate students’ test scores with other measures of their ability taken some time later, such as subject teachers’ assessments, language teachers’ assessments. d) Correlate students’ scores with success of later placement.
Construct Validity	<ul style="list-style-type: none"> a) Correlate each subtest with other subtests. b) Correlate each subtest with total test. c) Correlate each subtest with total minus system. d) Compare students’ score with students’ bio data and psychological characteristics. e) Multitrait-multimethod studies. f) Factor analysis.

Reliability

Reliability is as important as validity. It is also an essential quality of a test. "Reliability is a quality of test scores, and a perfectly reliable score, or measure, would be one which is free from errors of measurement" (American Psychological Association, cited in Bachman, 1990, p.24). A reliable test produces consistent results. Generally, two components contribute to the reliability of a test, test factors and scoring factors.

Test factors

The test factors play a vital role in test reliability. Test factors include the degree of ambiguity of the test items, restrictions on freedom of response, clarity of instructions, the quality of layout, length of the tests and students' familiarity with the tests (Hughes, 2003). These test-related factors are within the control of test developers though not all the factors can be restricted. However, it is important for the test developers to do the best they can to increase test reliability.

Scoring factors

The reliability of a test also depends on how the test responses are scored. Scorer reliability refers to the consistency of the results, the type of tests, and the experience and quality of the scorers. Hughes (2003, p.43) states, "If the scoring of a test is not reliable, then the test results cannot be reliable either." Objective tests can have high scorer

reliability, as they do not require any form of judgment from the scorer. However, in subjective tests, a high reliability cannot be expected. Therefore, when including subjective scoring items in test, the test developers should carefully structure marking schemes for the examiners and more than one examiner should score the exam papers.

Recommendations to reduce unreliability in tests include:

- The test should use direct testing (Hughes, 2003; Weir, 1993; Heaton, 1988).
- The test instructions should be clear and unambiguous for all learners (Davies & Pearse, 2000).
- The test should be free from any kind of error (Hughes, 2003).
- The test should be able to be scored objectively (Bachman, 1990; Hughes, 2003).
- The test should have clear and specific scoring directions and comprehensive marking schemes (Weir, 1993).
- The test should be an appropriate length (Hughes, 2003; Weir, 1993).
- A test should be moderated before administering it (Hughes, 2003; Madsen, 1983).
- A test should be well administered. Tests should be administered in the same way whoever may be in charge or wherever it takes place (Bachman, 1990; Hughes, 2003; Bachman and Palmer, 1996).

As reliability is an important factor in language testing, it is vital to know how the reliability of a test can be assessed. Many methods can be used in assessing reliability of the test. Hughes (2003) stated that it is possible to quantify the reliability of a test in the

form of a reliability coefficient. Reliability coefficients allow us to compare the reliability of different tests. The ideal reliability coefficient is 1. A test with a reliability coefficient of 1 would provide the same results for particular groups regardless of the time of testing. Reliability coefficients should be as near as 1 as possible to indicate high reliability. There are two methods that can be used in calculating the reliability coefficient, the test-retest method and the split-half method.

The test-retest method estimates indicate how consistent test scores are over time (Bachman, 1990). This method needs to have two sets of scores for comparison. To get the necessary data, the same test should be given twice to the same group of people. The reliability is the correlation between the scores of the two tests. If the results are consistent over time, the score should be similar (Hughes, 2003; Bachman, 1990).

The split-half method estimates internal consistency of the test. Bachman (1990, p.172) stated, "Internal consistency is concerned with how consistent test takers' performances on the different parts of the test are with each other." This method also requires two sets of scores for comparison. In this method, the test is divided into two halves and the degree to which the scores on these two halves are consistent with each other is determined (Bachman, 1990). In using this method, it is necessary to make sure that the two halves have equal means and variances and are independent of each other (Bachman, 1990; Hughes, 2003).

In the construction process and evaluation of a test, the validity and reliability of the test must be evaluated. The validation of language tests should be based on a detailed description of both the abilities to be measured and the facet of the test methods (Bachman, 1990). A framework of description should be used to measure the validity of

the test. One study using this kind of framework to measure language test validation has been reported by Bachman, Davidson, Ryan, and Choi (1989). In this study, a content and task analysis of the “Test of English as a Foreign Language”, “Test of Written English”, “Speaking Proficiency in English”, and the “First Certificate in English” was carried out to investigate the comparability of these tests. Weir and Wu (2006) conducted a case study on general English proficiency tests to investigate the extent to which three forms of the test are parallel in terms of two types of validity evidence: parallel-forms reliability and content validity. Both reliability and validity of the test should be evaluated when constructing a test in order to produce a good quality test to assess the learners’ ability accurately.

Summary of the Chapter

The literature review has given insight about tests, the important qualities of tests and the test construction process. Tests as one form of assessment have been widely used in language programs to assess learners’ abilities. There are four different types of tests, proficiency tests, achievement tests, placement tests and diagnostic tests. The function of each kind of test differs according to their purposes and the situational need. Following this, some distinctions in language testing field that are significant and important for language test development were introduced. Then, a brief discussion on the achievement test included the function, the purpose and the qualities of a good achievement test. This chapter also provided discussion about the test construction process and the stages that need to be considered in developing a test. Consideration of validity and reliability is

part of the test construction process. These are essential qualities that every test should have. The chapter therefore discussed those concepts in some detail.

PAYYAP UNIVERSITY