CHAPTER II
LITERATURE REVIEW

2.1 Introduction

This chapter presents the review of the main issues concerned with the reading assessment. Particularly, the main factors in this study, i.e. nature of reading comprehension, test authenticity, test delivery medium, computer-adaptive tests and the attitudes of test takers towards test authenticity and test delivery mediums are explored. Therefore, all the principles and concepts obtained are applied to the research design and research instruments.

2.2 Reading assessment

2.2.1 The movement of language assessment

Language assessment has made a significant move from previous assessment models particulary from the discrete-point approach whereby tests are designed to assess learners' proficiency by demonstrating knowledge of particular structural elements of the target language.

Read (2000) suggests that the movement to current language testing design is probably caused by the criticisms of the discrete-point approach to language testing. Importantly, the discrete-point approach is criticized for the difficulty it poses in general statement about a learner's ability on the basis of test scores. Being proficient in a second language is being able to exploit that knowledge effectively for various communicative purposes; thus learners are required to show that they can use language appropriately in their own speech and writing, rather than by just demonstrating that they understand the meanings. Generally in normal language use, context significantly influences the way of using language so that in communication situations, it is quite possible for adept learners to compensate for lack of knowledge of particular words by making the best use of limited lexical resources. Similarly, with regard to reading comprehension, readers do not have to understand every word in order to satisfactorily extract meaning from a text.

Some words can be ignored, while the meaning of others can be guessed from contextual clues, background knowledge of the subject matter and so on.

The acceptance of these criticisms has led to the adoption of the communicative approach to language testing. Today's language proficiency tests are based on tasks simulating communication activities that the learners are likely to be engaged in outside of the classroom (Read, 2000).

In test design, Bachman and Palmer (1996) and McNamara (2000) suggest that the test design will move away from formal standardized tests made up of multiple choice items to measure students' knowledge of a content area, towards what is variously known as alternative, performance-based or standards-based assessment, which includes judging students' ability to perform more open-ended, holistic and real-world tasks within their normal learning environment.

Empirically, the movement requires a test to relinquish the discrete-point approach for the communicative approach, where test takers are likely to perform more open-ended test tasks. Similarly, Bachman (1990) mentions that the purpose of language testing is seen as allowing language testers to make inferences about learners' language ability. Learners need to be able to draw on that knowledge effectively for communicative purpose under normal time constraints.

Generally speaking, the current design of language tests will apply a more communicative approach. The test takers will be assessed for communicative purposes through more open-ended tasks.

2.2.2 The nature of reading comprehension

Reading comprehension can be considered as a very necessary skill in order to help people survive and be successful in the society. Moreover, many reading teachers also see reading as a means of exploring and expanding human potential (Devine, 1986). In order to potentially assess reading comprehension skill, understanding of the nature of reading has to be explored.

Educators and scholars have considered the nature of reading for many decades. Some may see reading as a process while many may see it as a product. Alderson (2000), suggests that reading can be seen as process and product. As a process, reading is the interaction between a reader and the text. During the process, presumably, many things are happening such as the reader looking at print, deciphering in some sense the marks on the page, deciding what they mean and so on.

Evidently, many different things can be going on when a reader reads. The process is likely to be dynamic, variable, and different for the same reader on the same text at a different time or with a different purpose in reading. It is even more likely, then, that the process will be different for different readers on different texts at different times and with different purposes (Alderson, 2000). Therefore, understanding the process of reading is presumably important to an understanding of the nature of reading, but at the same time it is evidently a difficult thing to do. The process is normally silent, internal, and private.

Alderson (2000) points out that it is possible to see reading as a process, and to examine the product of that process. Any theory of reading is likely to be affected by the emphasis that is placed on a process or a product. However, a product is easier to investigate than a process.

Reading seen as a process can be described by three general approaches to comprehension theories. Devine (1986) summarizes them as the bottom-up, top-down, and interactive approaches as follows:

1. Bottom-up approach: This approach is a serial model, where the reader begins with the printed word, recognizes graphic stimuli, decodes them to sound, recognizes words and decodes meanings. This approach was typically associated with behaviorism and with phonics approaches to the teaching of reading that argue that children need to learn to recognize letters before they can read words, and so on. In this traditional view, readers are passive decoders of sequential graphic–phonemic–syntactic–semantic systems, in that order (Alderson, 2000).

2. Top-down approach: This approach emphasizes the importance of the schemata, and the reader's contribution, over the text. Goodman (1982), for example, calls reading a psycholinguistic guessing game, in which readers guess or predict the text's meaning on the basis of minimal textual information, and maximum use of existing, activated knowledge. Smith (1971) claims that non-visual information transcends the text, and includes the reader's experience with the reading process, knowledge of the context of the text, familiarity with the structures and patterns of the language and of specific text types, as well as generalized knowledge of the world and specific subject matter knowledge.

3. According to Alderson (2000), the reading comprehension process can be seen as a combination of interactive models, in which every component in the reading process can interact with any other component, be it higher up or lower down. Processing, in fact, is now thought to be parallel rather than serial (Grabe, 1991). Rumelhart's (1977) model, for example, incorporates feedback mechanisms that allow knowledge sources (linguistic as well as world knowledge) to interact with visual input. In his model, a final hypothesis about the text is synthesized from multiple knowledge sources interacting continuously and simultaneously. Stanovich (1980), on the other hand, has developed an interactive compensatory model in which the degree of interaction among components depends upon knowledge deficits in individual components, where interaction occurs to compensate for deficits. Thus, readers with poor word recognition skills may use top-down knowledge to compensate.

Moreover, for the process of reading, Alderson (2000) also points out that there is an increasing tendency to see reading as a sociocultural practice since reading is not an isolated activity that takes place in some vacuum. Reading is usually undertaken for some purpose, in a social context, and that social context itself contributes to a reader's notion of what it means to read, or, as recent thinkers tend to put it, to be literate.

Recent research (Barton, 1994) shows the richness of the social world within which literacy events take place. For example, shopping lists are written and used. Moreover, reading may be the result of writing; for example, the use of shopping lists may lead to some form of writing–taking notes while reading an academic textbook, writing an essay after re-reading the notes. Reading will often be accompanied by talking like reading aloud a snippet from a newspaper in order to discuss political bias or the performance of a football team.

Regarding the product perspective, Alderson (2000) states that the product is the result of the reading process. Although different readers may engage in different reading processes, the understandings they end up with will be similar.

In recent years research efforts tend to focus on understanding the reading process since teachers of reading are trying to improve the way their students comprehend reading texts. However, a product approach is still in use today since the product measurement method is not as complex as the process one. According to this approach, researchers typically design tests of understanding of particular texts, administer the tests to suitable informants, using particular research designs, and then inspect the relationship between the results of the tests and variables of interest (Alderson, 2000). Therefore, in the current trend of reading research, the inevitable pendulum swing in research and teaching fashions have led to revived interest in the product of reading.

It can be concluded that reading can be seen as both a process and a product. If it is seen as a process, the interaction between the reader and the text becomes important to consider. However, since the process is silent, internal, and private, it is different for different readers on different texts at different times and with different purposes.

In brief, the reading process can be described through three approaches: bottom-up approach; top-down approach and interactive approach. In terms of the bottom-up approach, the recognition of letters comes before the ability to read words. Regarding the top-down approach, the knowledge of the context of the text, familiarity with the structures and patterns of the language and of specific text types, as well as generalized knowledge of the world and specific subject matter knowledge are needed to allow a reader to comprehend the text. However, reading can also be seen in terms of an interactive model because every component in the reading process can interact with any other component, higher up or lower down. Another important trend is to see reading as a sociocultural practice. Therefore, in assessing reading comprehension more than the process or the product of reading needs to be considered. The readers' purpose and social context, which can contribute to the reader's notion of what it means to read, should be understood.

2.2.3 Levels of reading comprehension

In order to understand reading comprehension, levels of reading comprehension are explored. Thrasher (2000) identifies three levels of reading comprehension as follows:

First, the reader examines the words of the author and determines what is being said, and what information is being presented.

Second, the reader looks at the relationships between statements within the materials and from these intrinsic relationships derive various meanings. The intrinsic relationships the reader perceives are colored and influenced by his or her previous knowledge of and experience with the topic in question.

Third, the reader takes the product of the literal meaning of the texts, i.e. what the author has said and the interpretative meaning of the texts, i.e. what the author meant by what he said and applies it to the knowledge already possessed, thereby deepening understanding. At the applied level the reader selects intrinsic relationships produced at the interpretative level of comprehension and synthesizes them with concepts that are the product of previous knowledge and experience.

It can be concluded that reading comprehension can be usefully divided into three levels: the literal, the interpretative, and the applied.

Similarly, Alderson (2000) also classifies levels of understanding of a text into a literal understanding of text, an understanding of meanings that are not directly stated in the text, and an understanding of the main implications of the text.

Mohamad (1999) further specifies elements in the three levels of reading comprehension. These specified elements can be used as the construct to be measured for reading comprehension tests. The three levels are presented in the hierarchy from the least to the most sophisticated level of reading.

Level one (Literal): measures what is actually stated. The elements measured in this level are:

Facts and details

Rote learning and memorization

Surface understanding only

Tests in this category are objective tests dealing with true / false, multiple-choice and fill-in-the blank questions. Common questions used to illicit this type of thinking are who, what, when, and where questions.

Level two (Interpretive or inferential): measures what is implied or meant, rather than what is actually stated. The following elements are measured in this level:

Drawing inferences

Tapping into prior knowledge / experience

Attaching new learning to old information

Making logical leaps and educated guesses

Reading between the lines to determine what is meant by what is stated

Tests in this category are subjective, and the types of questions asked are open-ended, thought-provoking questions beginning with expressions like why, what, if, and how.

Level three (Critical): measures what was said (literal) and then what was meant by what was said (interpretive) and then extend (apply) the concepts or ideas beyond the situation.

In summary, a consideration of the nature of reading must include recognition of frequency made distinctions among levels of meaning and understanding in and from the text. Thus, test writers must also consider the level of meaning that they believe readers ought to get out of a particular text when assessing how well they have understood the text in question.

2.2.4 Construct of reading comprehension tests

In order to design a test, the relationship between theories of reading, reading in the real world and the assessment of reading are important aspects that should be considered.

Foremost are the theories of reading, since what matters at the end of the test is the extent to which test developers can generalize from the assessment procedure or test to reading performance in the real world. Test developers have to make use of the theories because they want to know something about readers' reading ability or reading behavior beyond the test situation which is often referred to as generalizability of test results.

One way of addressing the issue of generalizability is to take an abstract notion of reading ability, a theory of reading, and then to seek to operationalize this theory in the tests. Constructs come from a theory of reading, and they are realized through the texts and the tasks they require readers to perform, the understandings they exhibit and the inferences the test developers make from those understandings, typically reflected in scores. This approach seeks generalizability by appealing to the theory. The test is generalizable to the extent that it adequately reflects a theory and the extent to which that theory is correct, and gives an adequate account of what is involved in reading. This tradition has a long, respected and influential history.

Secondly, the real world is an important factor to be considered. Alderson (2000) mentions that in reading tests, test developers have to consider how generalizable the conclusions can be from one setting to another. Moreover, when assessing reading ability, test developers should be interested in how well that performance approximates or predicts how test takers will read in other settings, or how well that performance can explain reading in other settings.

In terms of the assessment of reading, Alderson (2000) mentions that reading assessors are interested in knowing how well the test takers read, and how well they read something for some purpose with what degree of effectiveness, if not pleasure. However, the test methods are a source of potential bias in the measurement. What test developers actually measure on a test is the result of an interaction between ability and test method. The use of appropriate test methods can enhance the validity of the inferences, provided that they are appropriately chosen and explicitly related to the constructs.

The medium of text presentation is also considered a crucial aspect (Alderson, 2000). According to the implications for computer-based testing, test developers must be aware that it is possible to transfer texts from paper to the computer screen. Ideally, computer-based tests of reading would concentrate on testing the ability to read texts designed for reading from monitor screens like texts put up on the Internet. If a wider range of texts is required, as in the computer-based TOEFL, then it is crucial that a close attention be paid to the interface and screen design features, and that is likely to include the relationship between the verbally presented piece of information and the graphically presented one.

It can be concluded that in order to increase the generalizability of the reading comprehension tests, or the ability to make conclusions from one setting to another, the real world must be considered as an important factor to apply to the test. Another important factor is the mediums of text presentation since the process of reading texts via different mediums may affect test takers' comprehension. Therefore, the real world and the mediums of text presentation can be important factors that test developers have to consider when developing reading comprehension tests.

Consequently, the constructs of reading are based upon a model of reading and the factors that affect reading in so far as they are relevant to the assessment of the constructs. Test developers will concentrate on matters relating to readers. While it may be theoretically possible to include text variables in the construct, no model of reading that they have discussed explicitly distinguishes the ability to process one sort of a text from the ability to process other sorts. Certainly, different linguistic features of a text have implications for the sorts of knowledge and abilities that readers need, but it would include all possible linguistic variables in the constructs.

Alderson (2000) mentions that any variables having an obvious impact on either the reading process or its product needs to be taken into account during the test design or validation. If the reading process or product varies according to such influences, and if such influences occur in the test or assessment procedures, then this is a risk to the validity of the tests. The variables such as readers' background and topical knowledge, their cultural knowledge and their knowledge of the language in which the target texts are written, the linguistic knowledge, and all linguistic and metalinguistic levels are clearly relevant to the constructs.

The linguistic threshold clearly varies by task: the more demanding the task, the higher the linguistic threshold. What makes a task demanding will relate to variables like text, topic, text language, background knowledge and task type. A strict definition of the second-language reading construct might exclude linguistic knowledge per se, as well as first-language reading ability.

Most models of reading make a reference address to numerous skills required in reading. At the very least, therefore, students should be tested on a range of relevant skills and strategies, with the results possibly being provided in the diagnostic, profile-based format.

The relationship between items testing reading and those testing what might be called linguistic skills is something to be borne in mind when designing reading tests. Depending on the view of what reading involves, test designers may well feel that tests should focus on reading constructs, not irrelevant linguistic skills.

It can be concluded that test designers need to define carefully what the construct is, and to what extent it includes, or should exclude. This will relate to the purpose for which they are designing the test.

Moreover, the likelihood that different readers will have different emotional responses, different understanding and different affective responses during reading and their effects on comprehension and interpretation have important implications for the testing and assessment of reading (Alderson, 2000). These effects may be included in the reading constructs. Inevitably, the anxiety created by many testing settings will result in a different performance under different conditions. Therefore, the scores need to be interpreted accordingly. It may be that informal assessment procedures in non-threatening environments might result in qualitatively better performances than test-based assessments.

Even if it is true that testers are limited in what they can actually test, it is clearly important that what tests measure be as little contaminated as possible by the test method, and that the results of a reading test be generalizable to non-testing situations as much as possible.

2.2.5 Trends in reading assessment

The influence of the language testing approach allows some considerations in reading test design. As Alderson (2000) stresses the importance of testing in relation to real world contexts, reading should be assessed within a number of situations. That is, reading assessment should be done in relation to uses such as inferences about reading ability, decisions about individuals that are made based on information from tests or assessment procedures. A number of real world tasks are needed for the assessment of reading.

Bachman and Palmer (1996) also propose the application of real world context in test development as the target language use (TLU) domain. Since they identify three main components in test development: design, operationalization and administration, they suggest ways to integrate TLU domain in each component as follows.

In the test design stage, test developers produce a design statement which covers the purpose of the test, a description of the TLU domain and task types, the characteristics of test takers, a definition of the test constructs, a plan for evaluating test usefulness, an inventory of available resources, and a plan for their allocation and management.

At the operationalization stage, testers select, specify and write. They produce a test blueprint with details of the test structure such as number, salience, sequence and relative importance of parts and number of tasks and test task specifications (specifications of purpose, definition of constructs, setting, time allotment, instructions, characteristics of input and expected response and scoring method). Finally, the administration stage includes collecting feedback, analyzing and archiving the tests and producing test scores. Although this is a useful guideline, it involves a rather extensive overlap within and across component stages and with the initial description of test task characteristics.

Alderson (2000) points out the importance of testing situations, noting that test developers should propose to illustrate the framework of test specifications by identifying TLU and test task characteristics for a number of assessment settings and purposes.

In terms of available approaches for reading assessment, Alderson (2000) identifies two broad categories. One is the analytic approach: to seek to test whether readers successfully engage in, or master, those aspects of the process which testers consider being important. Thus, testers might seek to devise test items which explore whether a reader can successfully deduce the meaning of unknown words from context.

The testers might devise tasks that require readers to scan rapidly through a number of headlines in order to identify the ones that are relevant to a particular need or topic. In other words, the testers seek to isolate and identify components of the reading process relevant to the purpose for which one is testing.

The other broad approach is to recognize that the act of assessing itself risks disturbing parts of the process one is wishing to assess, and to acknowledge that individual readers may not need to engage in a particular activity in order to read successfully. Such an approach would entail seeking to simulate as far as possible the conditions in which one is interested such as reading newspapers in order to get an overview of the day's events, scanning TV guides in order to plan the evening's viewing and then assess whether the reader had successfully completed the task. The assumption is that if the task is successfully completed, then the reader will engage in the processes of his or her interest.

In brief, an understanding of the reading nature can be used as a guideline to develop a reading assessment. Moreover, testing to real world contexts is considered an important issue for developing a test. The current reading assessments are likely to be conducted within situations. The two other important issues in test design are operationalization and administration. Therefore, the target language use (TLU) domain is taken into account in these three main components in the test development: design, operationalization and administration.

As authenticity and medium of text presentation are previously mentioned as important factors affecting reading comprehension, a research question is raised here concerning whether test authenticity and test delivery medium can have any effect on test takers' English reading proficiency. Therefore, the mediums used to deliver the tests are investigated in order to observe their effects on the test takers' reading ability. Moreover, this study will also focus on the reading comprehension tests which possess different degrees of authenticity.

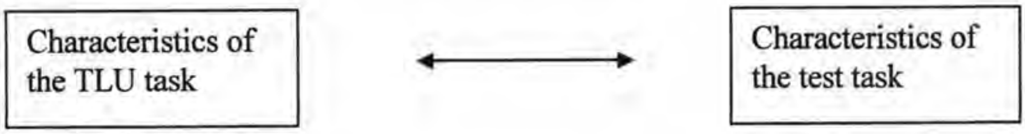
2.3 Test authenticity

Authenticity is considered a critical quality of language tests and most language test developers implicitly consider authenticity in designing language tests (Bachman and Palmer, 1996, Huhta et al, 1997). In order to apply authenticity to language tests, its concepts or definitions, characteristics, advantages and other aspects should be touched upon considered in more detail.

2.3.1 Definitions of authenticity

In order to justify the use of language tests, one needs to demonstrate that performance on language tests corresponds to language use in specific domains other than the language test itself. One aspect of demonstrating this pertains to the correspondence between the characteristics of the target language use (TLU) tasks and those of the test tasks. This correspondence is the heart of authenticity (Bachman and Palmer, 1996). Therefore, a test task whose characteristics correspond to those of TLU tasks can be described as relatively authentic.

Thus, Bachman and Palmer (1996) define authenticity as the degree of correspondence of the characteristics of a given language test task to the features of a TLU task. The relationship is shown in Figure 2.1.



Characteristics of the TLU task

Characteristics of the test task

Figure 2.1: Authenticity

(Bachman and Palmer, 1996:23)

In defining authenticity in terms of task characteristics, it is not a radical departure from the current testing practice. Rather, this approach may provide a more precise way of incorporating the notion of authenticity into the design and development of language tests.

In support of the earlier, Ingram (2003), stresses that authenticity should be considered when the test is designed because it relates the test task to the domain of language use to which test developers want their test score interpretations to generalize. Thus, authenticity provides a means for investigating the extent to which score interpretations generalize beyond the performance on the test to language use in the TLU domain, or to other similar non-test language use domains.

Differently, Mueller (2003) views the definition of authenticity regarding the real life approach. He defines authenticity as a form of assessment in which students are asked to perform real world tasks that demonstrate meaningful application of essential knowledge and skills. Wiggins (1993) supports Mueller's definition and says that authenticity helps to create worthy problems or questions of importance, in which students must be encouraged to apply knowledge to carry out performances effectively and creatively. These are either replicas of, or analogs to the kinds of problems faced by adult citizens and consumers or professionals in the field.

Moreover Stiggins (1987) cited in Mueller (2003) defines an authentic performance assessment as one which calls upon the examinee to demonstrate specific skills and competencies, that is, to apply the skills and knowledge they have mastered.

In conclusion, the notion of authenticity accords with the real life approach (Wiggins, 1993, Mueller, 2003). In this aspect, the performances in real world tasks are focused while other scholars (Bachman and Palmer, 1996, Ingram, 2003) see the more practical definition of authenticity. They define that it is a degree of correspondence of the characteristics of a given language test task to the characteristics of a TLU task. In other words, the test which is considered authentic allows a test taker to perform a real world task requiring the application of mastered knowledge.

2.3.2 Characteristics of authenticity

Bachman and Palmer's characterization of authenticity (1996) has implications for how to design, develop, and use language tests as follows:

1. Authenticity is relative, not absolute so that it is considered as relatively more or relatively less authentic, rather than authentic and inauthentic.

2. The relative authenticity of a test task cannot be determined just by looking at it. Three sets of characteristics must be considered: those of the test-takers, of the TLU task, and of the test task.

3. Certain test tasks may be relatively useful for their intended purposes even though they are low in authenticity.

4. In either designing new tests or analyzing existing tests, the estimates of authenticity are only guesses. Test developers can do their best to design test tasks that they believe will be authentic for a given group of test takers, but they need to realize that different test takers may process the same test task in different ways, often in ways they may not anticipate.

5. The minimum acceptable levels specified for authenticity will depend on the specific testing situation. These qualities must be considered essential to language tests if they reflect current views about the nature of language use, language learning, and language teaching. At the same time, however, the minimum acceptable levels for authenticity must be balanced with those for the other test qualities.

Authenticity depends upon how the test developers define the constructs of language ability for a given test situation. Authenticity has to do with the relevance of the test task to the TLU domain, and is thus related to the traditional notion of content validity. This also provides a basis for specifying the domain to which the test developers want to generalize the test scores. This can be obtained by investigating the construct validity (Bachman and Palmer, 1996).

Therefore, authenticity is relative and can be determined by looking at the test takers, the TLU task, and the test task. Authenticity of a test can change across groups of test takers because the estimation of authenticity is only guesses. Test developers will design authentic tests according to their judgements. Because of this, the acceptable level of authenticity depends on each situation.

2.3.3 Advantages of authenticity

Authenticity is considered an important test quality for many reasons. Authenticity links to construct validity since investigating the generalizability of score interpretations is an important part of construct validation. According to Wagner (2002) authenticity can contribute to construct validity. Authentic tasks that provide representative coverage of the content and processes of the construct domain have realistic settings or close simulations of real life language use and can minimize sources of invalidity. If authentic tasks that are sufficiently representative are used, then the score interpretation of the assessment can be generalized to non-test language situations.

Another reason for considering authenticity important is its potential effects on test takers' perceptions of the test. Test takers and test users tend to react to a language test because of the perceived relevance to a TLU domain, of the test's topical content, and the types of tasks required. It is this relevance that can help promote a positively affective response to the test task and can help test takers perform their best.

It can be said that authenticity possesses a lot of advantages for language tests. The scores obtained from authentic tests can be interpreted in order to generalize beyond the performance on the tests to the language use in the TLU domain. Authenticity can also contribute to construct validity. When taking an authentic test, test takers and test users tend to perceive that the test is relevant to a TLU domain.

Regarding real life approach, Mueller (2003) cites the advantages of authentic assessment which have allowed it to become more popular in recent years as follows:

Authentic assessments are direct measures.

Since students must be able to use the acquired knowledge and skills in the real world, assessments have to be able to tell if test takers can apply what they have learned to authentic situations. If a student does well on a test of knowledge, it might also be inferred that the student could apply that knowledge. But this is rather an indirect evidence. The direct check for the knowledge to apply can require test takers to use what they have learned in some meaningful way. Authentic assessments will provide the most direct evidence.

Authentic assessments capture the constructive nature of learning.

Because knowledge cannot simply be fed, students' own interpretation of the world needs to be constructed. Thus, in assessing students' ability they cannot be asked to repeat back information they have received. They must also be asked to demonstrate that they have accurately constructed meaning about what they have been taught. Furthermore, they must be given the opportunity to engage in the construction of meaning. Authentic tasks not only serve as assessments but also as vehicles for such learning.

Authentic assessments integrate teaching, learning and assessment.

Authentic assessment encourages the integration of teaching, learning and assessing. In the authentic assessment model, the same authentic task used to measure the students' ability to apply the knowledge or skills is used as a vehicle for student learning.

Authentic assessments provide multiple paths to demonstration.

Students have different strengths and weaknesses in how they learn. They are different in how they can best demonstrate what they have learned. In a traditional assessment model, multiple–choice questions do not allow for much variability in how students demonstrate the knowledge and skills they have acquired. On the one hand, these multiple–choice questions ensure consistency and comparability among the candidates. On the other hand, multiple-choice testing favors those who are better test takers and does not give students any chance to illustrate their opinions and creativity.

It can be concluded that authentic assessments have become popular in recent years because they can provide the most direct evidence. Applying this approach to assessment, students are given the opportunity to engage in the construction of meaning. Also authentic assessment can measure students' ability to apply knowledge. Lastly, students can best demonstrate what they have learned.

Though authenticity provides many advantages, a lot of tests still lack this quality because of limitations which will be discussed below.

2.3.4 Limitations of authenticity in language tests

Although authentic tests can be more immediately informative and interpretable for assessing the candidate is potential to perform in a real life situation to the full extent of his ability, authentic tests also possess some limitations. Ingram (2003) warns that authentic assessments may be impractical, particularly when assessing a large number of candidates using the language in real life. Thus, most testers focus on the types of tests that they are able to manage in the testing room. They statistically relate either the items or the results to real life performance through various forms of validity.

As for the test procedures, design, construction and performance of item types and items related to real language performance or real language behaviour, may create difficulties in observing a learner in real life situations. Moreover, testers may have difficulty controlling the language in such situations, where elicitation and observation of maximum language performance is needed.

Therefore, the main limitation of authentic assessments is the impracticality of observing a large number of candidates using the language in real life. Testers may also face problems in controlling authentic situations.

2.3.5 Differences between authentic assessment and traditional assessment

The lack of authenticity in language tests can bring about a kind of test which is called "traditional". This traditional assessment possesses a lot of different perspectives from those of authentic assessment. Mueller (2003) makes a comparison between the assumptions of the two approaches to assessment (authentic assessment and traditional assessment) as follows:

1. Traditional Assessment (TA) can be referred to as the forced-choice measures of multiple-choice tests, cloze, true-false, or matching tasks and the like that have been and remain common in education. Students typically select an answer or recall information to complete the test. These tests may be standardized or teacher-created. They may be administered locally, statewide, or internationally.

2. Authentic Assessment (AA) springs from the following reasoning and practice:

1. A school's mission is to develop productive citizens.

2. To be a productive citizen, an individual must be capable of performing meaningful tasks in the real world.

3. Therefore, schools must help students become proficient at performing the tasks they will encounter when they graduate.

4. To determine if it is successful, the school must then ask students to perform meaningful tasks that replicate real world challenges to see if they are capable of doing so.

Mueller (2003) distinguishes traditional assessment (TA) from Authentic Assessment (AA) according to their attributes. TA's as well as AA's vary considerably in the forms they take. But, typically, along the continuum of attributes, TA's falls more towards the left end of the continuum and AA's falls more towards the right. Figure 2.2 illustrates Mueller's continuum of Traditional Assessment and Authentic Assessment.

Traditional	**Authentic**
Selecting a response	Performing a task
Contrived	Real–life
Recall/ Recognition	Construction/ Application
Teacher–structured	Student–structured
Indirect Evidence	Direct Evidence

Figure 2.2: The continuum of traditional assessment and authentic assessment

(Mueller, 2003: 3)

Each attribute of TA and AA can be compared as follows:

Selecting a response to performing a task: On traditional assessments, students are typically given several choices (e.g., multiple-choice; true or false; matching) and asked to select the right answer. In contrast, in authentic assessments, students are asked to demonstrate understanding by performing a more complex task usually representative of more meaningful application.

Contrived to real-life: It does not occur often enough in life outside school to be asked to select from four alternatives to indicate proficiency at something. TA asks test takers to demonstrate proficiency in a short period of time. On the other hand, AA asks test takers to demonstrate proficiency by doing something.

Recall/ Recognition of knowledge to construction/ application of knowledge: Well–designed traditional assessments (i.e., tests and quizzes) can effectively determine whether or not students have acquired a body of knowledge. Furthermore, they are often asked to recall or recognize facts and ideas, and propositions in life, so tests are somewhat authentic in that sense. However, the demonstration of recall and recognition on tests is typically much less revealing about what test takers really know and can do than when they are asked to construct a product or performance from facts, ideas and propositions. Authentic assessments often ask students to analyze, synthesize and apply what they have learned in a substantial manner, and students create new meaning in the process as well.

Teacher–structured to student–structured: After completing traditional assessment, what a student can and will demonstrate has been carefully structured by the person who develops the test. A student's attention will be focused on and limited to what is on the test. In contrast, authentic assessments allow students more choice and construction in determining what is presented as evidence of proficiency. Even when students cannot choose their own topics or formats, there are usually multiple acceptable routes towards constructing a product or performance. Obviously, assessments are more carefully controlled by the teachers who offer advantages and disadvantages of the test tasks. Similarly, more student–structured tasks have strengths and weaknesses that must be considered when choosing and designing an assessment.

Indirect evidence to direct evidence: Even if a multiple–choice question asks a student to analyze or apply facts to a new situation rather than just recall the facts, the student selects the correct answer. At best, the results can allow inferences to be made about what that student might know and might be able to do with that knowledge. The evidence is very indirect, particularly for claims of meaningful application in complex, real world situations. Authentic assessments on the other hand, offer more direct evidence of application and construction of knowledge.

It can be concluded that for AA the test tasks are more complex and are usually representative of more meaningful application. In terms of real life assessment, test takers are asked to demonstrate proficiency by doing something. AA often asks students to analyze, synthesize and apply what they have learned in a substantial manner, and students create a new meaning in the process. Importantly, students are allowed to determine what is presented as evidence of proficiency in a test. Finally, AA offers more direct evidence of application and construction of knowledge.

As for TA, item types are provided in order to ask test takers to select the right answer. The tests contrive means of assessment to increase the number of times the test takers can be asked to demonstrate their proficiency in a short period of time. The tests often ask the test takers to recall or recognize facts and ideas. Moreover, what a student demonstrates in the tests is carefully structured by the teacher or the person who develops the test. Lastly, TA is indirect, particularly for the claims of meaningful application in complex real world situations.

However, Mueller (2003) suggests that the teacher does not have to choose between AA and TA. It is likely that a mix of both approaches will best meet test developers' needs. It might be best to assess in a traditional manner to prove that students master the knowledge and to investigate that students are able to apply that knowledge in a real context which could be demonstrated through an authentic assessment.

2.3.6 Authenticity in reading comprehension tests

According to the attributes of test authenticity suggested by Mueller, these attributes can be probably applied to reading comprehension tests as follows:

1. Performing a task: Test takers are required to perform more complex reading tasks which are meaningful in the real world. Hence, they allow to demonstrate their understanding in a more open-ended way similarly their real life. They have a chance to write down their own answers without forced choice. And the answers will be corrected with more acceptable answers.

2. Real-life: The reading tasks tend to occur in everyday life or when they graduate. During an authentic reading test, a test taker, therefore, has more meaningful purposes for reading such as reading for a job application, reading for recreation and reading for obtaining important data to prepare an academic report.

3. Construction/ Application: During taking an authentic reading test, it is assumed that test takers will apply understanding to replicated real world tasks meaningfully. For example, they use the information obtained from classified jobs to write an appropriate application letter in order to be accepted as an applicant.

4. Student-structured: In an authentic reading test, test takers themselves are involved in the test design. The reading topics and the characteristic of reading tasks are obtained from test takers' need survey. Therefore, all the reading materials and test tasks are relevant to their real life.

5. Direct Evidence: Since real world situations are integrated with reading test tasks, the score obtained from the test allows inferences to become or provide more direct evidence of real life application. The test results such as reading classified jobs can reflect test takers' ability in using classified jobs for job application.

According to the definition of test authenticity as detailed by Bachman and Palmer (1996), authenticity can be assured in reading tests by allowing test takers to get involved in test design. Hence, they can give ideas of what kinds of reading purposes and tasks they would encounter in their real life. Due to the fact that the TLU task and the test task can change across groups of test takers, authenticity can be estimated by judging. Alderson (2000) mentions that research results have not been gathered in ways that reflect the real world and the real readers' purposes when test developers ask test takers to take a test of reading. It has been found that test developers are rarely doing this for the reader's purposes. In fact, test developers impose on the test their own purposes. And even if test developers try to simulate real world purposes for reading a given text, the fact remains that the ultimate purpose of the event is to evaluate the readers' ability to read. This is rarely the purpose for which real readers read real texts.

It can be briefly concluded that in developing a reading test, test developers should allow test takers to take part in deciding the reading topics and characteristics of test tasks. Therefore, a passage whose topical content and task characteristics can be matched with the kinds of topics and material the test takers may read outside the testing situation.

2.3.7 Reasons of lacking authenticity in language tests

Though authenticity is considered a critical attribute of language tests, many language tests still lack this attribute. Ingram (2003) warns that lack of authenticity in language tests leaves a serious gap between the test and real life language use. The main reasons for lack of authenticity in language tests include:

Impoverished contexts within traditional approaches to test design, even though language is known to be heavily situation dependent;

The limited range of situations which is possible to include in tests that are largely constrained by paper-and-pencil presentation and response modes;

The disparity between the test situations and real life authentic language situations;

The pre–determined and limited content of tests that have been statistically standardized in order to ensure statistical validity and reliability. The content and language elicitation modes of such tests are commonly limited to those that can be controlled and adjusted in accordance with statistical requirements;

The inability of pre–determined tests to match test responses with individual candidates' needs, interests, experiences, proficiency levels, and other personal characteristics, i.e., the lack of adaptiveness of such tests, and hence their inability to accurately identify particular skills; and

The failure of most tests to present their results in ways that allow their ready interpretation in terms of candidates' real life language ability. For most tests, there is a double gap between the test and real life ability for many reasons. First, there is a gap between test items and real life use of the language. Second, there is a gap that the end user has to bridge between how the results are expressed (in a numerical score or an abbreviated behavioral description) and the language demands of real life language use situations.

The lack of authenticity in language tests is, therefore, caused by the impoverished contexts within the language, the limited range of situations in the test, the disparity between the test situations and real life situation, the use of statistical validity and reliability in a test which determines and limits the content of tests, the needs to match the test with individual candidates' personal characteristics, and the attempt to present the results in ways that allow the ready interpretation in terms of candidates' real life language ability.

2.3.8 Studies on authenticity in language tests

Since Bachman and Palmer (1996) state that the task characteristics (they use the term task in place of test method and characteristics in place of facets) are always going to affect test scores to some extent, it is impossible to eliminate the effects of task characteristics. Moreover, it is necessary to control them as much as possible so that the test will be appropriate for what they are intended for. Therefore, most test developers have the aim to understand and be aware of what characteristics can be varied, and how they can be varied to best tailor tests to make them appropriate for specific test takers.

Studies which focus on the effects of authenticity in language tests have been done by many researchers. The results obtained from these studies are also varied, but most of the findings suggest the influence of authenticity on language tests.

As the issue is the ways in which comprehension tests can be made more authentic, Wagner (2002) investigated two authentic variables: the mediums used to deliver the aural input in the listening test and test item types.

In this study, he explored the listening process with the aural input delivered the on video tape which is considered a more authentic test. Two types of test items used in the test were limited production item types considered a more authentic item type and multiple-choice item type, considered as a less authentic item type. Video tapes were used because video allows listeners to perceive and process nonverbal information. A model of L2 listening ability was hypothesized and operationalized, and an assessment instrument was created. This video listening test was then administered to 85 ESL students. The data from this test were then analyzed using reliability and exploratory factor analyses. The results seem to provide some evidence for the effects of the test method on the test takers' performance. The findings indicate that the limited production item type may be more suitable for testing a listener's ability to comprehend inferential information, while multiple- choice item type may be better suited to assess a listener's ability to comprehend explicitly stated information.

Similarly, Lynch's study also shows the importance of authenticity in reading tests. Lynch (2003) studied authentic, performance-based assessment in ESL/EFL reading instruction. A reading exercise was implemented as a representative instructional model and it was also used to inform a valid performance-based reading comprehension test. The performance-based reading assessment tasks in this study was considered authentic since they involved reading for authentic purpose and the selection of texts depended on the students being assessed as well as the specific domain characteristics of the context within which they were expected to perform such as academic or work. This study used an authentic performance-based assessment focusing on assessing the aspects of the reading process.

The study was conducted by introducing and practicing the exercise intensively in class in order to have students employ them in self-study. By working through the exercise several times in class, the teacher could demonstrate and model each of the tasks as well as feedback.

Lynch reported that the advantage of performance-based testing resides in its potential to create and maintain positive washback on the teaching and learning process. Such positive impact on the instructional process is not, however, a function of performance-based testing. This can only derive from a comprehensively valid interaction between the nature of the instruction preceding evaluation and the actual performances being assessed. Validity must be grounded in a range of interrelated factors which can all be subsumed under a general notion of construct validity. In order to maximize the validity of performance-based assessments, both test designers and teachers need to be aware of these factors and their interaction. Construct validity and its interrelated aspects are discussed and applied to the authentic performance-based testing of reading comprehension. In brief, the findings indicate that authenticity can allow more validity in performance-based reading comprehension tests.

Although the results from the research studies support the importance of authenticity in language tests, there are some studies that conclude otherwise, based on reported perceptions of test takers in the studies. Lewkowicz's study yields the same findings.

Lewkowicz (2000) examined some dimensions of authenticity in language tests and reported the results of three empirical investigations. In the first two, she asked a number of non-native speakers of English and native speakers of English to judge whether the listening passages played to them were pedagogic or authentic and why. In the third study, students were asked to reflect on a traditional multiple–choice test and a more integrated authentic test, and to judge, for example, which better indicated their ability to use the language in the real world. These studies reported that even experienced respondents could find it difficult to distinguish between pedagogic and authentic listening texts, and that test takers can hold widely differing views about what makes a test authentic.

The question of authenticity has been given much prominence and it is now considered one of the key attributes of any test, on a par with reliability and other aspects of validity. It does not mean that present testing philosophy requires all tests to be authentic, but rather it proposes that in designing any test a desired level of authenticity needs to be considered.

It is argued that in designing a language test, the quality of authenticity may not be as important as previously considered. If it is not possible to distinguish authentic tests from the inauthentic ones, then it may not be necessary for testers to place a great deal of weight on ascertaining a desired level of authenticity. Other attributes such as test practicality, in reality, may be more important in designing appropriate tests which are acceptable to the test taking community.

Although the results obtained from Lewkowicz's study implies that it is not necessary for all tests to be authentic, she recommends that in designing any test, a desired level of authenticity has to be ascertained. However, the results from Wagner's study and Lynch's study support the use of authenticity in language assessment. Regarding Wagner's study, the aural input in the listening test delivered through the use of the video which is considered a more authentic test may lead to more reliable and valid listening assessment. Similarly, Lynch's study suggests that authenticity can be used to inform more valid performance-based reading comprehension tests.

Authenticity should be considered in designing language tests as it can also give rise to a more reliable and valid test. The researcher, therefore, is interested in investigating whether authenticity can affect test takers' English reading proficiency. This study focuses on reading proficiency skills because there are still a limited number of studies investigating authenticity in reading tests. Moreover, the test takers' attitude towards test authenticity is also investigated.

2.4 Mediums of test delivery

The medium of test delivery can be an important variable affecting test takers' performance as Alderson (2000) mentioned. This is able to affect test takers' reading ability.

The three main types of the test delivery medium which are used nowadays are the traditional test (paper-and-pencil tests), tests delivered by computer (or standalone tests), and tests delivered by the Internet.

2.4.1 The characteristics of mediums of test delivery

Language tests have been conducted for many decades and not only are the testing principles focused on, but the technology for supporting test development is also considered in order to establish more efficient tests. Most traditional tests are delivered in the form of paper-and-pencil. However, the computer and the Internet are replacing paper-and-pencil tests. Due to the fact that more and more information is now available on computer screens, especially with the development of the Internet, the World Wide Web, and the use of computer–based self–instructional materials, test developers, therefore, should attempt to make use of these technologies in order to develop new mediums of testing (Alderson 2000).

In order to investigate whether these different test delivery mediums can affect the test takers' performance, the important characteristics of each medium are explored.

2.4.1.1 Tests delivered by paper-and-pencil or paper-based tests (PBT)

The important characteristics of PBT can be described mainly through the testing environment, the test rubric and the nature of the input.

In terms of testing environment, especially for standardized tests, the environment has to be prepared consistently with the specifications in the test blueprint (Bachman and Palmer, 1996). However, a comparison of the testing environment of PBT with that of other standardized tests delivered by the computer, reveal little difference since these standardized computer-based tests are also administered in formal situations. Both use proctors to supervise test administration and manage the physical condition of the exam room.

In terms of the test rubric, the scores on the paper-and-pencil test are obtained by summing the correct answers. Each item response is scored as either correct or incorrect since typically, varieties of paper-and-pencil tests are multiple-choice, true-false, cloze completion or question-answer types of items. (http://www.elckiev.org/toefl /toefl_en.php#6) Regarding the nature of the input, the input is presented visually in the receptive mode.

2.4.1.2 Tests delivered by computer or computer-based tests (CBT)

Much reading does take place on screens due to the increased use of the word–processor, the use of e-mail, access to the World–Wide Web, computer–based instruction and even computer–based testing. There are all increasingly important elements of literacy. And it is probably true that future generations will be much more comfortable reading from screen than current generations, who are still adapting to the new media (Alderson, 2000). Therefore, it is necessary to consider the role of information technology in the assessment of reading.

CBT is one of the test–delivery mediums which is widely used. In order to understand the characteristics of CBT, the testing environment, test rubric and the nature of input are considered.

Regarding testing environment, CBTs are conducted in the test center in which the computers and other facilitators are provided. Tests are administered under strict supervision and security measures. A test administrator or supervisor is authorized to dismiss test takers from a test session. If test takers take excessive or extended unscheduled breaks during the test session, test center supervisors are required to strictly monitor unscheduled breaks and report test takers who take excessive or extended breaks (http://www.gre.org/cbttest.html).

In terms of test rubric, CBTs offer the opportunity to be adaptive; that is, they are tailored to test takers' performance level and provide precise information about their abilities using fewer test questions than traditional PBT. At the start of each section, test takers are presented with test questions of middle level difficulty. As they answer each question, the computer scores that question and uses that information, as well as their responses to any preceding questions and information about the test design, to determine which question is presented next. As long as they respond correctly to each question, questions of increasing difficulty typically will be presented. When test takers respond incorrectly, the computer typically will present them with questions of lesser difficulty. Their next question will be the one that best reflects both their previous performance and the requirements of the test design. This means that different test takers will be given different questions. The statistical characteristics of the questions answered correctly and incorrectly, including the difficulty levels, are taken into account in the calculation of the score. Therefore, it is appropriate to compare scores of different test takers even though they received different questions (http://www.gre.org/cbttest.html).

In terms of the nature of the input, the input may be presented either aurally or visually, in the receptive mode. The question types used in CBT are varied, including, matching, ordering, click on a word, click on a sentence, click on a paragraph, and insertion.

In the reading section, in addition to the traditional multiple-choice questions, a number of new types of questions are presented, namely click on a word, click on a sentence, click on a paragraph, and insertion. The majority of these new types of questions measure the same language ability as PBT, particularly in the case of TOEFL, but the inserted questions measure something different.

The characteristics of new item types: click on a word, click on a sentence, click on a paragraph, and insertion are described as follows:

Click on a word is used to measure knowledge of vocabulary. In the PBT, this language skill is examined by the multiple-choice questions. In the CBT, test takers are to click on the word with a certain meaning or one that refers to a certain pronoun.

Click on a sentence measures the ability to find certain information in the given passage. The PBT uses the traditional type of the multiple choice question asking the candidates to choose the number of the line that contains the right answer. In the CBT this ability is measured by asking the test takers to click on the sentence that contains the answer, instead of choosing among the four lines.

Click on a paragraph measures the ability to find a paragraph where a certain topic or the main idea is discussed. In the CBT a topic or the main idea is offered, and the test-takers are to click on the paragraph where this topic or idea is discussed.

An insertion question is a new type of question that asks the test takers to determine the place in the text where a certain sentence is to be inserted. In this type of question they must realize the succession of ideas in the text, determine the type of information in the sentence that is to be inserted and decide where to insert it. The ability to realize the sequence of ideas in the text and to decide whether a certain sentence correlates with it is not measured in the PBT TOEFL (http://www.elckiev.org/toefl /toefl _en.php#6)

In brief, the CBT must be conducted where computers and other facilities are provided. The test administrators and other staff are required to strictly monitor the testing processes. A computer-adaptive test allows different test takers to be given different questions since the difficulty level of each item depends on the response of the previous item. Therefore, it is appropriate to compare scores of different test takers even though they received different questions. The important characteristics of CBT is that because of technology, there are various types of items used such as matching, ordering, click on a word, click on a sentence, click on a paragraph, and insertion.

2.4.1.3 Tests delivered by Internet or web–based tests (IBT)

Radical improvements in assessment will derive from advances in three areas: technology, measurement, and cognitive science (Bennett, 2001). New technology is likely to be the most influential of the three areas in the near future. This might be due to their generality and capacity to improve assessment. The technological advancements revolve primarily around the Internet. Alderson (2000) found that the contribution of the Internet is in assessment, particularly in reading tests. He foresees the future tests on the Internet will make available a range of media and information sources that can be integrated into the test, thereby allowing the testing of information accessing and processing skills, as well as opening up tests to a variety of different input texts. These days there are a lot of proficiency tests which are WBTs or being developed to be WBTs in order to be available on the Internet, such as, QuestionMark, Cambridge, IELTS, TOEFL, TOEIC, Computerized placement tests (LOPT, TOPE) (Prapphal, 1997).

Bennett (2001) suggests that the important characteristics of Internet-based tests or web-based tests (IBTs or WBT) are being interactive, broadband, switched, networked, and standards-based. Each characteristic can be described as follows:

"Interactive" means that a task can be presented to a student and can quickly respond to that student's actions.

"Switched" means that test developers can engage in different interactions with different students simultaneously. In combination, these two characteristics (interactive and switched) make for individualised assessments.

"Broadband" means that those interactions can contain large quantities of information. For assessment tasks, that information could include audio, video, and animation. Those features might make tasks more authentic and more engaging, as well as allow the test developers to assess skills that cannot be measured in paper-and-pencil. The test developers might also use audio and video to capture answers, for example, giving examinees choice in their response modalities such as typing, speaking. However, this approach is relatively new and still needs exploring.

"Networked" indicates that everything is linked. This linkage means that testing agencies, schools, parents, government officials, item writers, test reviewers, human scorers, and students are tied together electronically. Such electronic connection can allow for enormous efficiencies.

Finally, standards-based means that the network runs according to a set of conventional rules that all participants follow. The interchange of data can be assessed from a wide variety of computing platforms as long as the software is run on those platforms, like the Internet browsers.

According to these characteristics, the Internet can deliver a test efficiently on a mass scale. The test content can be sent to almost any desktop. It can retrieve the data immediately, process it, and make the information available anywhere in the world, anytime, day or night. On the other hand, paper delivery cannot compete with these capacities.

Fulcher (2000) points out a variety of reasons why the delivery of the traditional CBTs on the Internet is particularly interesting.

Firstly, the only software needed to take the tests is a standard browser. This can be loaded onto any type of computers, making the test delivery system truly platform independent. Another requirement is hardware, a reasonably fast processor, to download the information from the host server providing the test.

The second important advantage of the Internet as a means of delivery is that tests can be delivered to any machine linked to the Internet, at any time convenient to the provider and the client. The web also provides advantages in the flexibility of test design without the need to resort to third-party plug-ins. It is quite feasible, for example, to use the frames facility of the browser to divide the computer screen into windows, each of which contains a content page. Prompts may be set up on a series of frames that incorporate text, images, audio, and video, where computer links are reliable and quick. In fact, the flexibility of html in designing web pages makes it possible to design a range of novel task types through the imaginative combination of multimedia in a frames environment for low-stakes testing or research (Fulcher, 2000).

Another advantage of delivering tests over the web is that links can be established to more information. Facilities, databases, and libraries are recommended for web-based testing in academic programs. Tests are self-contained, have watertight units, and involve the use of information from the outside world, to any degree the test designer wishes to incorporate it.

This potential can be used to increase the authenticity of some testing activities. In computer-based testing via the Internet, innovation is possible where there is flexibility over the format and content of the prompt. However, it is not easy to be innovative in the area of item types. Most Internet browsers support multiple-choice, pull-down menu and constructed response item types, and combinations of these. For example, multiple pull-down menus can provide matching or sequencing items.

Constructed response items may be of two types: limited constructed response where a word or short phrase is required, and which is automatically scored against a template, and extended constructed response, which must be e-mailed to human raters for scoring. In this respect, little has changed since Alderson (1988 cited in Fulcher, 2000) found it difficult to design innovative item types for computer based language tests.

There are still other limitations in using the Internet to deliver tests. As Bennett (2001) points out, the Internet is not being built to serve the needs of large-scale assessment. Fulcher (2000) provides the reasons why on-line tests are available only as low-stakes quizzes. This is mainly because large scale high-stakes test delivery over the web faces serious security problems. If security issues associated with the transfer of information over the Internet are not solved, it is unlikely that testing organizations will use the web.

Moreover, Fulcher (2000) raises other limitations from a measurement perspective. Internet testing raises many questions that still need to be investigated. At present, there is not enough research evidence for the introduction of novelty, except for using it in low stakes testing and research. Measurement and ethical questions must be addressed in relation to the Internet, when developing CAT listening tests with the video instead of the audio.

In summary, we are currently in a situation where innovation and flexibility of the Internet are possible but its implementation will not be a pressing concern until its conceptual problems have been thoroughly studied.

2.4.2 Comparison between paper-based tests (PBT), computer-based tests (CBT) and Internet-based tests (IBT)

Because the differences of the test delivery mediums may affect test takers' performance, the differences between paper-based, computer-based tests and Internet-based tests are provided in the following table.

Table 2.1: The Comparison between Paper-based Tests and Computer-based Tests

Paper-based Test	Computer-based Test	Internet-based Test
Test Content Form		
1. Test Taker Response		
The item type mostly used is multiple-choice, particularly for error analysis questions.	The test has some types of questions that are not used in the paper-based test such as clicking on a word, clicking on a sentence, etc.	Audio, video and animation can be included in the test tasks. Audio and video can be used to capture answers, for example, giving examinees choice in their response modalities such as typing, speaking. These features might make the tasks more authentic and more engaging.
2. Test Input		
The test is in the form of paper-and- pencil.	The test has the capacity to provide simulations of real-life situations, three dimensional graphics, voice-activated responses, on-screen calculators, and split screens that show reading passages and questions at the same time.	The software needed can be loaded onto any type of computer. Prompts may be set up on a series of frames that incorporate test, images, audio, and video, where computer links are reliable and quick. A range of novel task types through the imaginative combination of multimedia in a frame environment can be designed.

Administration		
In terms of administration, walk-ins are not permitted.	Walk-ins are permitted if space is available.	In terms of administration, testing agencies, schools, parents, government officials, item writers, test reviewers, human scorers, and students are tied together electronically. Therefore, the network runs a set of conventional rules that all participants follow.
Test Algorithm		
All test takers do the same test within the same time length.	(Only if adaptive) The test saves time by matching questions, and the order of presenting matches the ability of each test taker. The test taker does not waste time on the questions that are too easy or too hard.	The tests can be delivered to any machine linked to the Internet, at any time convenient to the provider and the client. Therefore, it provides the flexibility of test design without the need to resort to third-party plug-ins.
Score Reporting		
The score reporting takes time.	The test allows faster score reporting. In some cases scores are on the screen at the end of the test.	The test results can be obtained immediately.

(http:www.fiu.edu/~biology1/grad/pagestoefl.pdf.)

It can be concluded that there are several differences between computer-based tests, paper-based tests and Internet-based tests. The first one is the item types used.

In computer-based tests and Internet-based tests, there are more types of items used such as click on word, sentence, paragraph and insertion. In terms of administration, computer-based tests and Internet-based tests are more flexible since walk-ins are permitted if space is available. Computer-based tests and Internet-based tests are also able to tailor the test items to each test taker. This kind of testing is called computer-adaptive testing, whereby each test taker faces different test items requiring different time length.

2.4.3 The advantages and limitations of paper-based tests (PBT), computer-based tests (CBT) and Internet-based tests (IBT)

2.4.3.1 The advantages and limitations of PBT

Though much information nowadays is simply processed on screen, many readers prefer to print out tests and process it at leisure (Alderson, 2000). This may also affect the use of PBT, particulalyr in reading tests. The advantages of PBT over CBT can be mentioned as follows:

1. People tend to prefer reading from print to reading on screen. This might be caused by the potential fatigue effect due to screen glare. Moreover, screen fonts and character types can have a marked effect on ease of reading (Oltman 1990).

2. Many test developers recommend that test takers should not be asked to process text of more than a single screen in length (Alderson, 2000). Therefore, longer passages are more effectively used on PBT.

The limitations of PBT are also various, as follows:

1. In PBT, there is no possibility of recording response latencies and time on texts or tasks.

2. It would be difficult for PBT to present the tests that are tailored to readers' ability levels. Therefore, the items in PBT may frustrate test takers by presenting them with items that are too difficult or too easy.

In brief, the advantages of PBT are having less fatigue effect due to screen glare or the screen fonts and character types. Moreover, longer passages are more effectively used on PBT. However, there are also many limitations. It lacks the ability to record response latencies and time on texts or tasks of the test-takers. It is also difficult for PBT to tailor the test to readers' ability levels.

2.4.3.2 The advantages and limitations of computer-based tests(CBT)

Due to the development of computer technology, CBT can provide advantages which enhance the development of language tests, particularly reading tests. Moreover, there are many opportunities for exploitation of the computer environment which do not easily exist with paper-and-pencil tests (Alderson, 2000). Below are the discussions concerning the advantages and limitations of CBT.

1. The possibility of recording response latencies and time on texts or tasks opens up a whole new world of exploration of rates of reading, of word recognition, and so on which are not available, or only very crudely, in the case of paper-based tests. The computer is able to capture every detail of a learner's progress through a test, such as which items are consulted first, which are answered first, and in what sequence, and which clue facilities are used. The possibilities are almost endless and the limitation is more likely to be in our ability to analyze and interpret the data than our ability to capture data (Alderson 2000).

2. The development of diagnostic tests of skills could be facilitated by the computer. Tests can be designed to present clues and hints to test takers as part of the test-taking procedure. This can be monitored not only to understand the test-taking process, but also to examine the response validity of the answers. Only the items that the student indeed engages in during the testing process are used. The unintended processing of items, if it is detected, can be used for diagnostic purposes.

3. Computer-based tests of reading allows the possibility of developing measures of rate and speed, which may prove very useful, especially in the light of recent research into the importance of automaticity.

4. Alderson (2000) mentions the use of computer-based adaptive tests (tests whose items are adjusted in difficulty during the test performance) for offering opportunities to a more efficient test of reading, but also to present tests tailored to readers' ability levels. The tests do not frustrate test takers by presenting them with items that are too difficult or too easy. They are also considered learner-adaptive tests because the candidates can decide whether to take an easier or a more difficult next item based on their own performance or indeed based upon the immediate feedback that such an adaptive-computer test can provide.

5. Another advantage of delivering tests by the computer is the ease with which data can be collected, analyzed and related to test performance. This may well enable test users to gain greater insights into what is involved in taking reading tests, and in turn, this

might lead to improvement in test design and the development of other assessment procedures.

6. Ingram (2003) mentions that for the virtual community the target language can be created on a computer screen and the global community can access it via technology. It is important to understand how real language learning takes place in real life and how natural acquisition processes can be drawn and enhanced through appropriately designed language teaching. Then, language assessors have to think creatively and imaginatively about how to use the capacities of modern technology to maximize the efficiency and effectiveness of learning that draws on our understanding of real language learning.

However, CBT has many limitations.

1. The most obvious problem for CBT reading is that the amount of text displayed on screen is limited and the video monitor is much less flexible when it comes to allowing readers to go back and forth through the text. Alderson (2000) recommends that test takers should not be asked to process a text of more than a single screen in length. Anyway, more research is needed to support this recommendation.

2. Screen reading is more tiring, slower, and influenced by a number of variables that do not affect the normal print. For example, colour combinations, more space between words, larger font size are important variables (Piamsai, 2006).

3. In terms of the implications of CBT, Alderson (2000) mentions that test developers must be aware that it is not simple to transfer texts from paper to computer screen. Ideally, CBT of reading should concentrate on testing the ability to read texts designed for reading from monitor screens. If a wider range of texts is required, as in the computer-based TOEFL, then it is crucial that close attention be paid to the interface and screen design features. The relationship between verbally and graphically presented information has to be considered.

2.4.3.3 The advantages and limitations of internet-based tests(IBT)

The next generation of most standardized tests will be in the form of internet-based tests since this medium can incorporate an integrated-skills approach to simulate real life environment (http://www.uk.toefl.eu/toefl-sites/toefl-default/about-the-toefl/internet-based-testing). This latest test delivery medium can provide crucial benefits as follows:

1. Online registration and online score reporting can make it easier for test takers to register for the test, as well as receive their results.

2. Proprietary content can be delivered directly to test takers any time of day, anywhere in the world.

3. Video and high-resolution graphics can be provided.

4. Text can be presented in mixed formatting.

5. Changes, scoring data updates, and content code and key modifications can be loaded.

6. It is more flexible exam scheduling.

7. More test locations can be used.

8. Score reporting is faster.

9. Costs are reduced.

10. A wide variety of question types can be used.

(http://www.castleworldwide.com/tds_v5/services/delivery/internet-based-testing.htm)

However, there are limitations of internet-based tests that should be considered.

1. The testing software and the hardware on which the software runs might be costly.

2. In terms of test security, there might be a risk of cheating.

3. While Internet-based tests provide remarkable capabilities, they may be prone to slow-down and errors that are beyond the control of even the most sophisticated users. Accordingly, test reliability can be affected.

(www.fadvassessments.com/docs/papers/internet_testing_whitepaper.pdf)

It can be concluded that because of the development of computer technology, CBT can bring many advantages to language tests. First of all, the response latencies and time on texts or tasks can be recorded so that every detail of a learner's progress can be investigated. CBT is also useful to second-language acquisition (SLA) research because computers can be used to assist in the assessment of strategies in SLA studies. Moreover, the development of diagnostic tests of skills can be facilitated by the computer. CBT can also help develop measures of reading rate and speed. In terms of the computer-adaptive testing, it can tailor the test to readers' ability levels. The data from CBT can be collected, analyzed and related to test performance more easily than PBT. Lastly, the virtual community can benefit from the tests administered on a computer screen.

As for the limitations of CBT, it can be concluded that readers can only process one screen at a time, and scrolling forward and backward is more time-consuming and less efficient than turning over pages. Moreover, it is recommended that variables affecting test scores such as space on screen, color combinations, and font size should be considered.

IBT may be able to provide more advantages in such a way that test takers can take a test anywhere and any time. However this advantage brings some limitation. Test security and test reliability can be affected.

2.4.4 Studies related to computer-based tests (CBT) and paper-based tests (PBT)

The effects of test delivery mediums on test takers' performance have been explored by many studies. In terms of test delivery medium effects, the findings from the studies can be categorized into three groups: (1) test takers perform better in CBT, (2) test takers perform better in PBT and (3) test takers perform indifferently between the two mediums. However, there are a number of studies conducted to investigate the factors affecting test takers' performance when they take a test delivered by different mediums. These studies' findings are also included here separately.

1. The studies obtaining the findings that test takers perform better in CBT

The findings of Choi et al's study (2003) conducted with Grade 11 and 12 students taking both CBT and PBT reflect that overall test results show test takers to have scored significantly higher on the CBT than on the PBT. An examination of individual classes' results reveals that the two Grade 11 classes performed significantly better (i.e., at the 1% level) on the CBT than on the PBT results. The two Grade 12 classes were also higher, although only marginally on the CBT.

There are studies comparing the effect of CBT and PBT on the more open-ended test or writing test.

Russell and Plati (2001) studied the middle school students who took an essay test on computer. They found that not only most wrote longer essays but also performed better than a randomly assigned group taking the same test on paper regardless of whether their keyboarding speed was high or low. Similary, Wolfe et al. (1996) found that the secondary school students tended to increase essay length when writing on computer.

In terms of score, MacCann et al. (2002) found that students randomly assigned to test on computer received higher scores than those taking the same test on paper for either one or two of three essays, depending upon whether the essays were graded in their original forms or transcribed to the other form before being graded.

However, the findings from these studies used small, non-representative samples. Moreover, there is no study investigating particularly on reading comprehension test. Even so, the results obtained suggest that mode may have an impact on test score.

2. The studies obtaining the findings that test takers perform better in PBT

In CBT for language proficiency tests, Bunderson et al. (1989) offered an overview of studies on test equivalence and commented that in general it was found more frequently that the mean scores of PBT and CBT were not frequently found to be equivalent. The scores on tests administered on paper were often higher than those on computer administered tests.

Peak (2005) mentions that with respect to specific comparability research, there accumulated evidence has in the print that suggests that the computer may be used to administer tests in many traditional multiple-choice test settings without any significant effect on student performance. One exception to this finding is with respect to tests that include extensive reading passages; for these, studies have tended to show lower performance on computer-based tests than on paper tests (Mazzeo and Harvey, 1988; Murphy, Long, Holleran, and Esterly, 2000; Choi and Tinkler, 2002; O' Malley, Kirkpatrick, Sherwood, Burdick, Hsieh, and Sanford, 2005 cited in Peak, 2005). These differences may be due to issues related to scrolling and the strategies that students use to organize information (e.g. underlining key phrases). As students continue to become more familiar with reading on the computer and as computer interfaces begin to include tools to enhance student's reading comprehension, these differences may disappear.

According to this, there are studies (comparing PBT and CBT) on the responding, especially for the open-ended response. The findings obtained can be applicable to constructed responses in reading tests. Most of these studies reported better performance of test takers on PBT.

Horkay et al (2006) conducted a research on CBT for writing test used for the National Assessment of Educational Progress (NAEP) in the United States. This study focuses on analyzing 1) the differences in performances between the delivery modes, 2) the interactions of delivery mode with group membership 3) the differences in performances between those taking the computer test on different types of equipment (i.e., school machines vs. NAEP- supplied laptops) and 4) whether computer familiarity was associated with online writing test performance. The findings indicate that there is no significant mean score differences between paper and computer delivery. In terms of computer familiarity, it significantly predicted online writing test performance after controlling for paper writing skill. This result can suggest that, for any given individual, a computer-based writing assessment may produce different results than a paper one, depending upon that individual's level of computer familiarity.

The studies with students taking postsecondary admissions tests reported that in the essay section the scores were marginally higher on the paper form than on the computer form, after controlling for English language proficiency (Wolfe and Manalo, 2004). Bridgeman and Cooper (1998) studied a large group of business-school applicants who wrote GMAT (Graduate Management Admission Test) essays in each mode. The results indicate that students performed better on the paper version than on the computer version of the test.

Maulan (2004) mentioned that the use of computers as a medium of delivery in language tests is very important in this cyber era since more students experience computer-based learning in the classrooms. This is also in conjunction with the Malaysian government's intention to ensure computer literacy among school leavers and university graduates. Computer-assisted language testing, however, is still questioned especially concerning its validity by many educators. This study examined the effect of the computer as a test delivery medium on the English language performance in a writing test. A total of 85 diploma students from University Teknologi MARA participated in the study. The subjects were randomly divided into two groups: experimental and control groups. The experimental group consisting of 44 subjects did the computer-assisted writing test whereas 41 subjects in the control group sat for the paper-and-pencil writing test. Apart from the writing test, the subjects were also tested on their keyboarding speed. They also responded to a questionnaire, which investigated their level of computer familiarity. In this study, the data were analyzed using two tests of significance, the t-test for independent samples to either confirm or reject the null hypotheses and the Pearson Product Moment Correlation to observe the relationships. The results obtained from the t-test showed that the subjects performed significantly better in the paper-and-pencil writing test than in the computer-assisted test. However, the results from Pearson Correlation showed that there was no significant relationship between the subjects' computer-familiarity and performance on the computer-assisted test. According to the findings, it is interesting that there is evidence for the test takers performing better in PBT.

3. The studies obtaining the findings that test takers do not perform differently on the two mediums

Though many studies reported the differences in performance when taking a test through different mediums, there are studies on the effects of CBT and PBT in language proficiency tests which have reported similar results.

Shaw et al. (2001) and Thighe et al. (2001) investigated the equivalence of PB and CB forms of the Listening and Reading Modules of IELTS. Shaw et al.'s study (ibid) involved 192 candidates taking a trial version of CB IELTS shortly before a different live PB version of the test which was used as the basis for their official scores. The CB tests were found to be reliable and item difficulty was highly correlated with PB and CB versions (r= 0.99 for listening, 0.90 for reading). In other words, text format had little effect on the order of item difficulty. Correlations (corrected for attenuation) of 0.83 and 0.90 were found between scores on the CB and PB versions of the Listening and Reading Modules respectively, satisfying the criterion of 0.8 and suggesting that format had a minimal effect on the scores awarded.

Green (1998) and Maycock (2004) have conducted a series of studies since 2001 into the comparability of IELTS tests delivered on paper and by computer. Initial trials of the computer-based linear version of IELTS were encouraging, finding that test format had little effect on the order of item difficulty and finding strong correlations between scores on the CB and PB versions of Listening and Reading forms, suggesting that format had a minimal effect on the scores awarded.

Jorgensen (2003) (cited in http://www.harcourt.com/about/news /articles/092903_assessment_stanford_diagnostictests.p) compared the performance of the test takers on paper-based and computer-based tests of the English diagnostic test. He found that there was no significant difference between the scores obtained from both tests. The similarity of the test takers' performance in both paper-based and computer-based tests can be described in terms of test format. Both tests consist of multiple-choice questions that measure the examinees' ability to understand spoken English, understand English reading passages, and recognize correct English grammar. Both tests also require the examinees to write an essay.

However, there is a study which did not indicate whether the differences of test takers' mean score were in favor of the PBT or the Online test. This study was conducted by Wang (2004). He aimed at comparing the performance on the Stanford Diagnostic Reading and Mathematics Tests for students in grades 2-12. Overall, there were no significant differences in total test score means based on administration mode, mode order, or mode-by-mode order interactions. However, for level 6 (grade 9-12), the differences in the means based on administration mode and mode order were statistically significant for reading, and the differences in the means based on mode-by-mode order interactions were statistically significant for math.

In conclusion, although there are different findings from the medium effect studies, many studies have reported that the mediums do have effects on test takers' performance. In terms of test fairness, these findings should not be neglected since test takers may perform better on CBT or PBT. Wolfe and Manalo (2004), therefore, recommend that providing examinees with a choice of medium in fairness, particularly when high-stakes decisions will be made based upon the test results. Test takers should be able to take the test in the form with which they feel comfortable.

Apart from the effect of medium on test takers performance, there are many studies which have investigated the factors affecting CBT performance. One factor investigated in many studies is computer familiarity. There is a group of studies investigating whether the familiarity with computers affect online writing test performance. For example, Marcoulides (1989) investigated the effects of inequities in computer access and familiarity on CBT performance. The findings show that inequities in computer access and familiarity led to lower levels of confidence and higher levels of anxiety toward computer-based tasks. Affective responses are related to computer anxiety, and computer experience but they are also correlated with computer-based test scores at non-trivial levels.

Similarly, Wolfe, Bolton, Feltovich and Bangert (1996) conducted a study with tenth grade students with little or no experience using computers outside of school. The results show that they scored higher on pen-and-paper essays than on computer-written ones, whereas students with a lot of computer experience showed no difference in performance across modes.

Wolfe et al. (1996) studied students with less experience writing on computer. They found that these students were disadvantaged by having to take the test that way. Less experienced students achieved lower scores, wrote fewer words, and wrote more simple sentences when tested on computer than when they were tested on paper. Students with more experience writing on computer achieved similar scores in both modes, but wrote fewer words and more simple sentences on paper than on computer.

Russell (1999) controlled for reading performance of middle-school students in his study. The results show that students with low keyboarding speed were disadvantaged by a computer-writing test relative to students with similar low levels of keyboarding skill taking a paper test. The opposite effect was detected for students with high keyboarding speed, who were better on the computer than on paper examinations.

A factor that can affect test takers' performance is presentation characteristics Fulcher (1999) and Choi et al (2003) reported the effects of the computer screen in that the large screen can allow test takers to read more easily than a potentially small diagram or chart on the printed page. Moreover, less construct-irrelevant variance may be being caused by the screen than with the paper-based version. Similarly, as test takers know how long they have to answer each question due to the countdown timer on the screen, they can give an item their full attention.

Similarly, Bridgeman, et al. (2003) looked at the effect of variations in screen size, resolution, and item-presentation latency on test performance for SATI: Reasoning Test Items. These investigators randomly assigned 357 high-school juniors to a variety of item presentation conditions. Two tests were administered, one consisting of quantitative comparison questions and one of multiple-choice comprehension questions with associated reading passages. Bridgeman and his colleagues found no effect on math scores. Reading comprehension scores, however, were higher by about .25 of a standard deviation for students using a larger, higher-resolution display (1024 by 768) (low resolution = 640 by 480) than for students using a smaller, lower resolution screen. (The effects of screen size and resolution could not be separated in the analysis.) Finally, the only test feature rated as interfering by the majority of students was scrolling. Bridgeman et al. (2003) suggest that a prudent approach in Web delivery of high-stakes tests would be to attempt to have comparable scrolling across computer configurations. Fortunately, variation in item presentation can be controlled, at least to a substantial degree. One approach is to establish hardware and software standards to limit presentation differences.

In conclusion, various factors affecting test takers' performance in CBT and PBT include computer familiarity, presentation characteristics and also test type. For test familiarity, although there is evidence indicating that computer technology affects test takers' performance, if test takers develop their computer skills, the differences in performance from paper-based to computer-based tests tend to be negligible.

In terms of test type, test takers have tended to respond favorably to CBT when demands placed upon them were limited: such as when simply clicking on the answer in multiple-choice tests. Where demands were greater, such as having to type in missing words and phrases, and subsequently calling up that input to amend it, test takers tend to react much less favorably to the test, with a marked preference for a paper-based version of the test.

2.4.5 Trends in using computer-based tests (CBT) and paper-based tests (PBT)

A trend in taking the Test of English as a Foreign Language (TOEFL) in many countries is to use computers. This change is part of an evolutionary effort to create a new and better generation of English proficiency tests (http://www.dca.ca.gov/cba/exam.html). Moreover, among the many advantages of using the computer are the new opportunities for better English proficiency assessment. Computer-based tests (CBT) are also more responsive to the needs of test takers and score users. These reasons make CBT increasingly popular and widely used nowadays. However, TOEFL is still being developed to be available on the Internet in order to allow take takers to access through the test more conveniently. Therefore, the TOEFL CBT now seems to be being replaced by the latest version of TOEFL, Internet-based test (iBT).

The CBT reading tests which aim to measure the ability to understand non-technical reading material (the new tasks in CBT) are developed to promote test takers' greater involvement with the test. (http://www.dca.ca.gov/cba/exam.html). Although CBT is widely used in many countries, the paper-based tests will continue to be administered in many areas in Asia. However, once computer-based tests like TOEFL are introduced in a country, there is a tendency for the paper-based tests to be replaced soon (http://www.dca.ca.gov/cba/exam.html). According to the trend, CBT is replacing PBT. Moreover, many researchers are interested in investigating the effectiveness of CBT in order to prove that CBT is actually more effective than PBT in assessing language proficiency.

There are many trends and issues related to CBT. For example, the effects of test methods (CBT, PBT) and the item types used such as the multiple-choice technique and other types which are more innovative are interesting to explore. Alderson (2000) suggests more research should be conducted on mediums of test delivery to investigate how people process information. Moreover, he also urges that descriptions of how people interact with computers should be examined. The validity of computer-based tests of reading is also indicated as an area for further research since an analysis of target language use domains has to be considered in reading test construction. Brown (2001) also suggests future research on computers in language testing should include Computer-Adaptive Language Testing (CALT) items as a research topic.

Based on the reviews mentioned above, the area of interest of this study is the effects of the test delivery medium (CBT and PBT) in reading tests. Moreover, authenticity in the test is also the focus. In addition, the attitudes of test takers towards these two mediums will be investigated in order to understand more about their performance on the tests.

2.5 Computer-adaptive testing

As previously discussed, one important capacity of computer-based tests is that they are adaptive. To effectively measure test takers in a mass-administered test, the test must contain items with difficulty levels which match the range of test takers' abilities. That is, the test should contain some easy items for the less proficient candidates and have some difficult items for the more proficient ones.

Discriminating among the candidates, the test must contain a broad range of item difficulties to suit the proficiency range of the population to be tested (Wainer et al.,1990). Since most examinees' abilities seem to lie in the middle of the continuum, mass tests tend to match this by having most of their items of moderate difficulty with fewer items at the extremes. Computer-adaptive testing (CAT) can prevent proficient test takers from taking too easy items before reaching the items that provide substantial information about their ability. This can reduce the chance of careless errors induced by boredom.

CAT is therefore developed in order to provide a test, which possesses a capacity to tailor test items according to the ability of a particular test taker. The following section describes CAT characteristics, Item Response Theory, the system design and operation of CAT, types of CAT, CAT and paper-and-pencil tests, CAT reporting scores, and CAT of reading proficiency.

2.5.1 Characteristics of Computer-Adaptive Testing (CAT)

Computer-adaptive language tests are a subtype of computer-assisted language tests because they are administered at computer terminals or on personal computers. Computer-adaptive testing has three characteristics: (a) the test items are selected and fitted to the individual student~~s~~ involved, (b) the test is ended when the student's ability level is located, and, as a consequence, (c) computer-adaptive tests are usually relatively short in terms of the number of items involved and the time needed (Madsen, 1991).

Because of the characteristics of CAT mentioned above, CAT development has to rely on Item Response Theory (IRT). Lord (1980) suggests that a combination of IRT and the concept of a flexilevel test be used to create a test specifically designed for an individual student.

Larson and Harold (1985) describe the flexilevel procedures that they are used to roughly determine the general ability level of the student within the first few test questions. Then, based on item response statistics, the computer selects items which are suitable for the student's particular level and administers those items in order to get a more finely tuned estimate of the student's ability level. This flexilevel strategy eliminates the need for students to answer numerous questions that are too difficult or too easy for them. In fact, in a CAT, all students take the tests that are suitable to their own particular ability levels so each student has different test items.

2.5.2 Item Response Theory (IRT)

In many educational and psychological measurement situations, there is an underlying variable of interest that psychometricians refer to as an unobservable, or latent trait. These traits cannot be measured directly as height or weight can be, since the variable is a concept rather than a physical dimension. The generic term is used within IRT to refer to such latent traits so that in academic areas, one can use descriptive terms such as reading ability and arithmetic ability. According to this, a primary goal of educational and psychological measurement is, therefore, the determination of how much of such a latent trait a person possesses. If one is going to measure how much of a latent trait a person has, it is necessary to have a defined scale of measurement with the numbers on the scale, and the amount of the trait that the numbers represent in a task (Baker, 2001).

Thus, IRT is established to measure unobservable traits. IRT is a statistical framework in which examinees can be described by a set of one or more ability scores that are predictive, through mathematical models, linking actual performance on test items, item statistics and examinee abilities (http://edres.org/scripts /cat/catdemo.htm).

Moreover, Hulin, Darsgow and Parsons (1983 cited in Bachman 2004) explain that IRT relates characteristics of an item (item parameters) and characteristics of individuals (latent traits) to the probability of a correct response. In other words, it relates a test taker's performance on a given test item, and hence of a set of items, or test, to his or her level of ability.

In terms of parameter types, Baker (2001) describes three uses in IRT as follows:

"a-parameter" refers to the discrimination parameter. This parameter is the proportion to the slope of the item characteristic curve (ICC).

"b-parameter" refers to the difficulty parameter. It is the point on the ability scale.

"c-parameter" refers to the guessing parameter. It is the probability of getting the item correct by guessing.

The number of parameters included can allow a variety of IRT models. One parameter or the Rash model refers to the inclusion of only the b-parameter, while the two-parameter model means the inclusion of the a-parameter and the b-parameter. The three-parameter model, therefore, includes all three parameters, a-, b- and c- parameters (Baker, 2001).

In order to apply IRT in CAT, the following important assumptions should be considered.

1. Dichotomous data: The type of data used in IRT tends to be dichotomous data in which the examinee receives a score of one for a correct answer; an incorrect answer receives a score of zero. Polytomous data or the data from free-response items are difficult to use in a test. In fact, they are difficult to score in a reliable manner. As a result, most tests used the IRT consisting of multiple-choice items. They are scored dichotomously: the correct answer receives a score of one, and each of the distracters yields a score of zero. Items scored dichotomously are often referred to as binary items (Baker, 2001).

2. Unidimentiality Assumption: IRT is able to make stronger predictions about individuals' performance on individual items, their levels of ability, and about the characteristics of individual items. In order to incorporate information about test takers' levels of ability, IRT must make an assumption about the number of abilities being measured. Most of the IRT models that are currently being applied make the specific

assumption that the items in a test measures a single, or unidimensional ability or trait, and that the items form a unidimensional scale of measurement (Baker, 2001).

3. Local Independence: IRT models also make the technical assumption of local independence since item response models assume that individual test items are locally independent. A technical assumption in IRT implies essentially that a test taker's responses to two different items of the same difficulty are statistically independent, or uncorrelated (Bachman, 1990).

In conclusion, IRT is a theory that describes test takers by relating their performance on a test item and a set of items to their level of ability. The important assumptions of IRT are that the data used should be dichotomous (though there are now Rasch models that can handle polytomous data, dichotomous data is the most typical type of data used to create CAT); the items in a test measures a single, or unidimensional ability or trait; and each item is independent or uncorrelated.

2.5.3 System design and Operation of CAT

Wainer et al. (1990) give a description of testing algorithms or a set of rules specifying the questions to be answered by the examinee, and their order of presentation. All tests are administered following some testing algorithms. A testing algorithm is most conveniently described in three obvious parts:

1. How to START: What is the first item presented to the examinee?
2. How to CONTINUE: After each response, what is the next item?
3. How to STOP: When is the test over?

An adaptive algorithm selects the item(s) to be administered, and specifies when the test is over, based on properties of the examinee and/ or the item responses.

According to the CAT algorithm, there are two distinct types of item parameter estimates identified in CAT systems (Wainer et al.,1990). The first type is an initial calibration. This type of item parameter is related to the How to START algorithm since a CAT cannot begin until it has an extensive and calibrated item pool. An initial calibration in which responses are solicited from examinees only to items not yet calibrated onto the scale is what usually occurs when a CAT program begins.

The second type is on-line calibration, in which examinees give responses to both new items and to items with previously estimated parameters. On-line calibration might be chosen not to be too hard or too easy for an examinee and then seeded in at an appropriately approximate place in the test. However, when it is accomplished, the outcome of each examinee's testing would be a response vector to items whose parameters are already essentially known, and some responses to new items whose parameters must be estimated.

The term On-Line Calibration, as it is used in adaptive testing, refers to estimating the parameters of new items that are presented to examinees during the course of their testing with previously-calibrated items. This kind of calibration is, therefore, in the How to CONTINUE part.

In terms of How to STOP or stopping rules, an adaptive test can be terminated when a target measurement precision has been attained, when a predetermined number of items has been given, or when a predetermined amount of time has elapsed. Any of these rules may be used in its pure form or a mixture of them can be used (Wainer et al., 1990).

All the three parts of CAT algorithm (starting, continuing, and stopping) can be illustrated by Figure 2.3 which shows the structure of an adaptive test.

Adaptive Test Logic

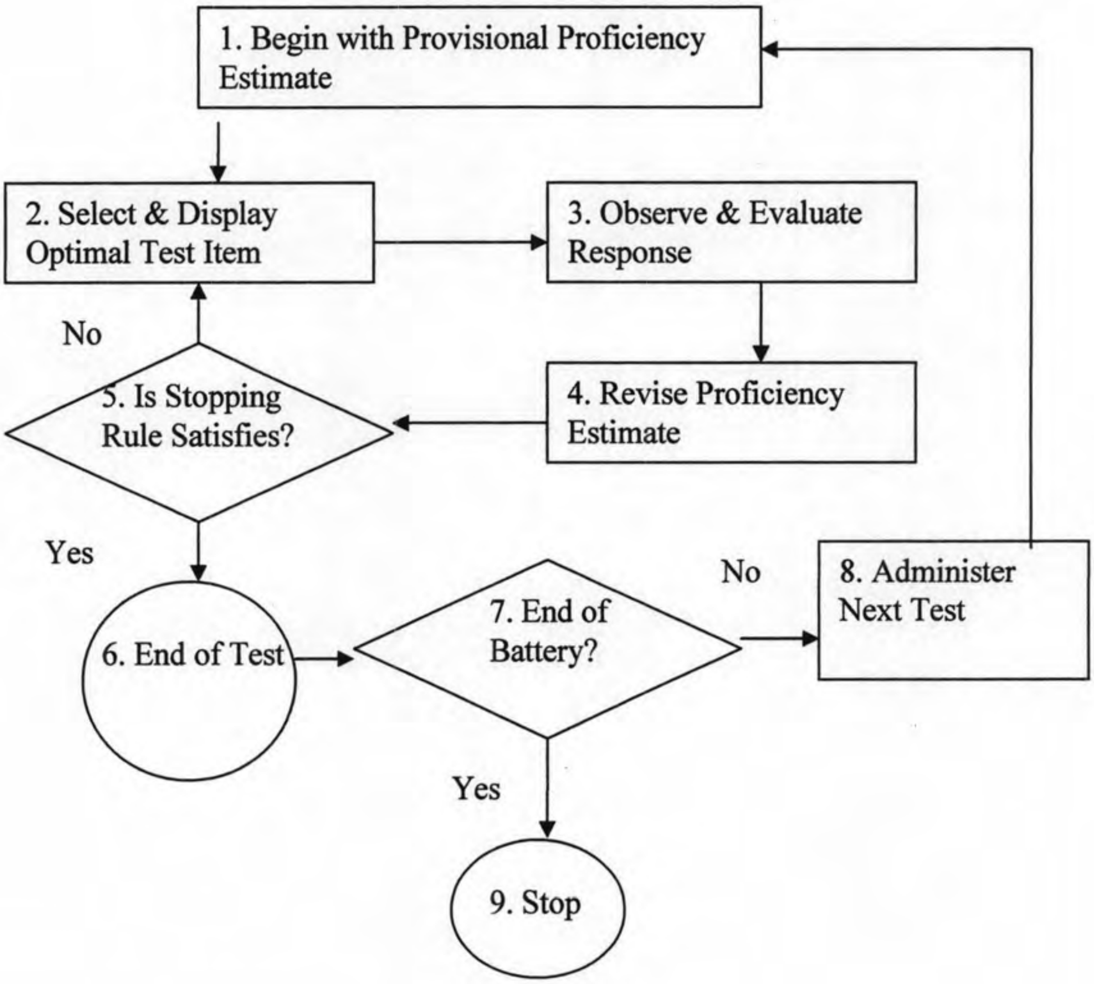

1. Begin with Provisional Proficiency Estimate
2. Select & Display Optimal Test Item
3. Observe & Evaluate Response
No
4. Revise Proficiency Estimate
5. Is Stopping Rule Satisfies?
Yes
No
8. Administer Next Test
7. End of Battery?
6. End of Test
Yes
9. Stop

Figure 2.3: A flowchart describing an adaptive test

(Wainer et.al., 1990: 108)

The flowchart begins with a Provisional Proficiency Estimate. After the response of the first item (initial calibration) is evaluated, the continuing step is activated process. This step includes the Select & Display Optimal Test Item, Observe & Evaluate Response and Revise Proficiency Estimate. This step can occur repeatedly until it can closely estimate the true proficiency of the test-taker. Finally, the test taker will come to the last step. If there is no subtest, the test will come to an end. If the test is composed of many subtests, the three steps will occur again.

2.5.4 Types of CAT

Hambleton et al. (1991) note that item selection strategies for adaptive testing can be broken down into two types: two-stage strategies and multistage strategies.

2.5.4.1 The two-stage strategies can be implemented without the use of a computer. As its name implies, ability estimates are obtained via a two-stage procedure: the examinee completes a routing test and is then directed to another set of several tests that have been "constructed to provide maximum information at certain points along the ability continuum. Ability estimates are then derived from a combination of scores from the routing test and the option test" (Hambleton et.al. 1991: 348)

Wainer et al. (1990) mention the three steps of testing algorithms and a two-stage test as follows:

1. How to START: Answer question #1.
2. How to CONTINUE: Answer the next sequentially numbered question until the end of the routing test is reached. (Somehow, the second-stage test is chosen here.) Answer the next sequentially numbered question.
3. How to STOP: Stop after answering the last question. (Wainer et al., 1990)

However, one problem remains. As Hulin, Drasgow, and Parsons (1983 cited in Bachman, 2004) point out, the two-stage test is minimally adaptive. If the second-stage test is incorrectly assigned due to unusual performance on the routing test, there is no way to recover. Given that the whole point of this enterprise is to make the test short, we are then left with scores on two very short and inappropriate tests; this can lead to very poor measurement. The answer seems obvious; multistage testing, or a CAT that adapts more often than once but less often than every item. In two-stage testing, the routing test could be considered a testlet, and each of the second-stage tests would be considered testlets.

- **Testlets:** A testlet is a group of items related to a single content area that is developed as a unit and contains a fixed number of predetermined paths that an examinee may follow (Wainer and Kiely, 1987). Two-stage testing would then be adaptive testing with a fixed number of testlets administered to each examinee and the usual kinds of adaptive decision rules applied between the administration of the first and second testlet. The only real difference, from a psychometric point of view, between a testlet and a conventional test item is that the testlet usually allows responses in several categories, and most test items are scored in only two categories: correct and incorrect. There may be many score categories on a testlet as there are response patterns to the items within the testlet (Wainer and Kiely, 1987); or there may be fewer, such as the number of summed scores of the items within the testlet (Thissen et. al., 1989). Scoring tests comprised of testlets does not present a problem for item response theory, which includes a number of models for items (in this case, testlets) with more than two possible responses.

Wainer and Kiely (1987) proposed the use of testlets in CAT to solve several practical problems endemic to CATs constructed at the level of individual items. One of these classes of problems involves context effects. The IRT model assumes that the items are changeable, that is, each item may be used at any time, with any combination of other items, without any effects of the previously administered items on the response probabilities for the current item. When all of the items in a large item pool may be presented in any combination, in any order, it is difficult to be sure that no item contains information that can be used to answer any other, and so on. If the items are prepackaged as testlets, each item carries its own context with it (Wainer and Kiely, 1987). Context effects are not completely avoided, but they can be markedly reduced by the application of the skill of test developers in the construction of the testlets. Test developers also have other skills and knowledge about the placement of items on tests that a CAT algorithm may lack. For instance, it may be best in certain kinds of tests to use some specific difficulty ordering, as is done when the test begins with easier items and proceeds to more difficult ones. Such fixed orders may be embedded within testlets.

In summary, it is easier for test developers to create good testlets for an adaptive system than it is for them to create good items. Moreover, it is not clear whether an adaptive test needs to adapt after each item. Because a two-stage test can perform almost as well as a fully adaptive test (Lord, 1980), it would seem that a test that adapts three or four or five times could provide both the precision and efficiency of adaptive testing and the control over item placement that test developers desire. A CAT based on testlets follows essentially the same algorithm as an item-based CAT, with the replacement of items with testlets:

1. How to START: Specify an initial estimate of proficiency; this specifies an initial testlet.
2. How to CONTINUE: Estimate proficiency ($\hat{\theta}$) after each testlet. Choose the remaining testlet that is most informative near $\hat{\theta}$ to be administered next.

3. How to STOP: Stop when the precision of $\hat{\theta}$ is adequate, or when some number of testlets has been administered. (Wainer et.al, 1990)

2.5.4.2 Multistage item selection strategies are much more complex than two-stage strategies since they involve a branching decision after each item is answered. Multistage strategies include either fixed-branching or variable branching decisions.

Fixed-branching

- The multistage fixed-branching models: The models allow all examinees to start at an item of medium difficulty and, based upon a correct or an incorrect response, pass through a set of items that have been arranged in order of item difficulty. After having completed a fixed set of items, either of two scores is used to obtain an estimate of ability: the difficulty of the (hypothetical) item that would have administered after the nth (last) item, or the average of the item difficulties, excluding the first item and including the hypothetical n + first item (Hambleton et al., 1991).

- A stratified-adaptive test: A second model of fixed-branching is referred to by Hambleton et al. (1991) as a stratified-adaptive test. This type of test has items stratified into levels according to their difficulty. Branching then occurs by difficulty level across strata and can follow any of a number of possible branching schemes. The fixed-branching model is similar to what Henning (1987) refers to as the step ladder test in which a specified number of items are held in the bank at each of a specified number of proficiency or achievement steps. During such a test, the computer selects an appropriate item at a predetermined number of steps above or below the previous item based on a correct or an incorrect answer to the previous item.

Fixed-branching or tree structure or pyramidal tests involve the placement of the test items in a branching tree in advance, depending on the response to each item. A different branch (or item) is chosen to be presented next, usually a more difficult item for a correct response and an easier item for an incorrect response (Wainer et al., 1990).

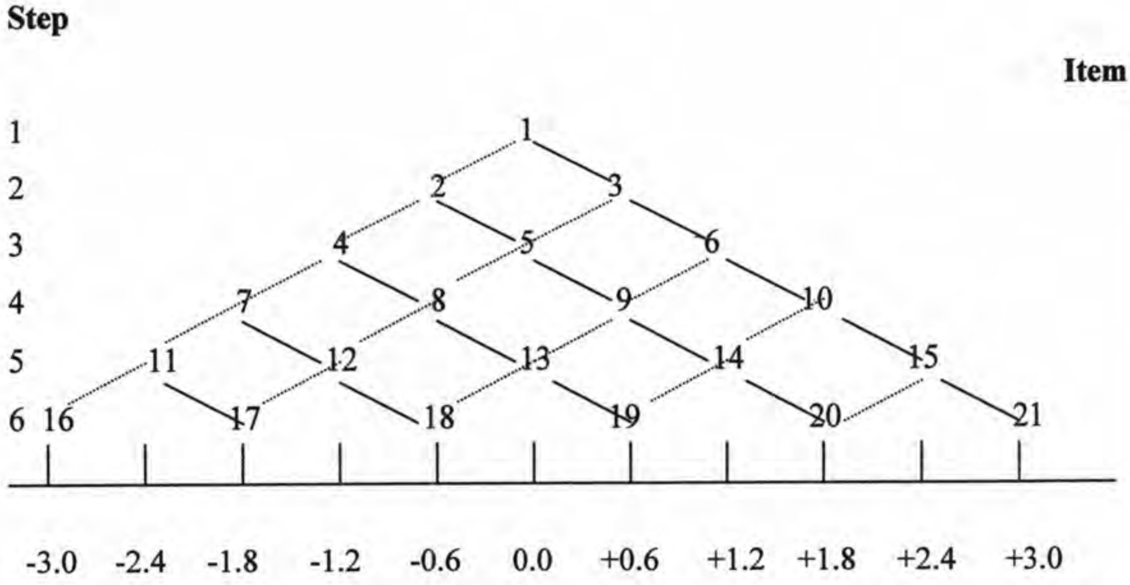


Figure 2.4: Constant step size pyramidal models

(Hambleton and Swaminathan, 1985: 86)

Figure 2.4 shows the constant 6-step pyramidal model. The horizontal axis represents the difficulty (b-parameter) of the items. The figure illustrates that the difficulty values fall between –3.0 - +3.0. This horizontal axis is equally divided into 10 sections (the difference of the difficulty level of each section is 0.6). The items in the same vertical line possess the same level of difficulty.

Wainer et al. (1990) mention that presenting the same number of items to each examinee in a fixed-length test has the advantages of being easy to implement. Also, item usage rates can be predicted more precisely. Examinees would be measured with varying degrees of precision; however, the least precision can be expected for extreme examinees because their first few items are probably not particularly informative in the neighborhood of their proficiencies. The performance of a fixed-length test may be evaluated through computerized simulation of the CAT.

In terms of the termination of a fixed-length test, Hambleton et al. (1991) mention that the decisions regarding when and how to terminate a test are vital to the success of computer-adaptive reading tests. They also outline several possible stopping rules as follows:

> *Another method involves setting a fixed number (not too large) of test items for the set of examinees. Testing time is (approximately) constant for all examinees, but the standard error of ability estimation will vary from one examinee to the next. In some applications, a minimum number of items which must be administered is*

> *specified, and then testing is continued until the measurement error associated with the ability estimate attains some prespecified acceptable level. This method often adds credibility to the testing in the minds of the examinees. Short tests are often viewed suspiciously by examinees.* (Hambletone et al., 1991: 249)

Variable branching

Variable branching item selection routines are based on a maximum likelihood estimation. An example of this type of branching is found in what Henning (1987) calls error-controlled tests. He explains their functioning as follows:

Following exposure to a specified set of introductory items, error-controlled tests employ a procedure such as unconditional maximum likelihood estimation in order to estimate the examinee's ability on the ability continuum. They then access and present the item in the bank that is nearest in difficulty to the estimated person ability, provided the item was not previously encountered. After each new item is encountered, a revised estimate of person ability is provided with an associated estimate of measurement standard error. The process continues in an iterative manner until the estimate of measurement error drops to a prespecified level of acceptability.

Hambleton et al. (1991) also outline stopping rules or the termination of variable branching as follows:

> *Several methods and combinations of methods are currently used. In one, testing is continued until some acceptable level of measurement error is achieved. In this way, ability estimates are all at the same level of measurement when testing is terminated (this parallels measurement within a classical test theory framework) though the number of items administered to each examinee will vary. It would also be possible to specify some acceptable but unequal levels of measurement precision for different ability levels. For example, a decision could be made that more precision is needed with middle abilities than for those at the extremes.* (Hambletone et al.,1991: 250)

2.5.5 Ways to compare CAT and PBT

Each computer-adaptive test that is generated from an item pool is really a form or an edition of the computer-adaptive test. Whenever scores based on different test forms are to be compared, it is necessary that they be equivalent in some sense. Statistical procedures, known as equating methods, have been devised to deal with the problems of achieving comparable scores (Wainer et al., 1990).

Lord (1980) specified four conditions that must be met in order to equate the two tests:

1. The two tests must measure the same construct. The same construct condition distinguishes true equating from scaling. In any population (P) scores on two tests, X and Y, can always be placed on the same scale through scaling, that is, they can be made to have comparable score distributions on P. In order for that scaling to qualify as an equating, X and Y have to be measures of the same construct. Tests X and Y need not be composed of unidimensional items, but they must be measuring the same dimension. This notion of content parallelism extends to CAT tests as well, in the form of content balancing.

2. The equating must achieve equity, i.e., for individuals of a given proficiency, the conditional distributions of scores on each test must be equal. The equity condition states that it must be a matter of indifference to the examinees whether they take X or Y. In order for equity to be achieved, the tests must be measures of the same construct or characteristic. Although the same construct is a prerequisite for equity, it does not ensure equity. Tests of the same construct may differ in terms of difficulty and other psychometric characteristics. For example, test X may be easier than test Y. If test X and test Y measure the same proficiency, examinees would choose to take the easier test X because they would get higher scores on it. Equating transformations are needed to produce equitable scores.

3. The equating, transformation should be invariant across populations on which it is derived. Population invariance is a requirement because equating transformations are one to one relationships between scores that should be unique and identical across populations. If population invariance is not achievable, it is probably because the tests are not measures of the same construct.

4. The equating transformation should be symmetric, i.e. the equating of Y to X should be the inverse of the equating of X to Y. Symmetry is essential because the same score on X should match up with a given score on Y regardless of whether X is equated to Y or Y is equated to X. The experienced statistician who is uninitiated to the world of equating sometimes makes the mistake of thinking that regression can be used to derive equating functions. Regression does not work as an equating method because it violates the symmetry condition, unless there is a deterministic functional relationship

between X and Y. In general, the regression of Y onto X and the regression of X onto Y provides two different one-to-one relationships between X and Y. In other words, because these two regressions are not inverse functions of one another, neither is an equating function.

Wainer et. al.(1990) also suggest ways to compare CAT with paper-and-pencil batteries.

It is frequently desirable to compare the precision of measurement obtained with a CAT with that obtained with a paper-and-pencil battery; this may be for purposes of comparing scores on a new CAT with an older paper-and-pencil form, or it may be that the same battery is administered to some examinees in adaptive form and to others in the form of a fixed test. If the paper-and-pencil test is IRT scored, there is no problem. In this case, both the CAT and paper-and-pencil form have information curves, and these may be directly compared to determine the measurement precision of both tests at any level of proficiency.

Because the CAT must be scored on a scale using IRT, if the paper-and-pencil test uses summed test scores and indices of precision developed from the traditional theory, only an incomplete comparison of the precision of the two tests is possible. In this case, the only indices of precision available for the paper-and-pencil test are internal consistency estimates of reliability. Only the corresponding indices of precision of the CAT (marginal reliability, and overall test-retest and alternate form reliabilites) may be compared to the equivalent values for the paper-and-pencil test.

According to the suggestion, the way to compare the two tests can be specified by the following steps:

1. In order to make the two tests able to measure the same construct, the test specification is established. Each item in the test specification will be specified for the dimension it measures. After the two tests are written, the experts will be asked to validate whether each item can measure the dimension specified in the test specification. According to this, the content contained in the two tests are balancing.

2. In terms of achieving test equity, after the test items in the two tests are piloted and analyzed, the level of item difficulty should be similar or equal so that the scores obtained from each test can be equal.

3. Since population invariance is a requirement to make the two tests equal, proof that the two groups of subjects are not significantly different in terms of their language ability must be obtained.

4. Since the equating transformation should be symmetric, the ability score (θ) obtained from Maximum Likelihood Estimation will be transformed to a raw score by using the Test Characteristic Curve (TCC) which is obtained after analyzing the test items by using the 3-parameter model.

Subsequently, the TCCs of the two tests will be compared in order to find out whether the same score on one test can match up with a given score on another test.

The steps for comparing the two tests mentioned earlier will be applied to the comparison of the tests in this study.

2.5.6 CAT Reporting Scores

For computer-adaptive tests, the scores that are appropriate for an adaptive test placed for the purpose of reporting scores are mentioned by Angoff (1984) as follows:

The proficiency (θ) score. Because the very nature of the adaptive testing process robs the simple number right score of its limited meaningfulness, another scale is necessary. The mathematical model that guides the adaptive testing process provides such a scale, the proficiency or θ scale. This scale is a property of the mathematical model known as item response theory (IRT), described earlier. Any test that is composed of items that have been fit by some IRT models can produce scores on the proficiency scale. This is true for conventional paper-and-pencil tests as well as computer-adaptive tests. The difference between the two types of tests is that adaptive tests require the proficiency scale or some derivative thereof, whereas the conventional test can suffice on the number-correct scale. Adaptive tests require a scale that is not tied into a particular set of items because adaptive test scores are based on so many different item sets.

By convention, θ scores are frequently placed on a metric that has a mean of 0 and a variance of 1 in some reference population, namely the population on which item parameters have been obtained. As seen later, other scalings of the proficiency scores are permissible, and may in fact be desirable.

The item pool score. Scores on the θ metric can be transformed via IRT formulae onto other metrics to produce scales that are more conventional in appearance. One such scale is the item pool score scale. The item pool score (IPS) is obtained by converting the θ score into an item true score for each item via,

$$P_j(\theta) = c_j + (1 - c_j)/[1+e^{-a_j(\theta - b_j)}], \qquad \text{(Equation 1)}$$

(Wainer et al.,1990: 139)

and then summing these item true scores across all items in the pool,

$$\varepsilon(\mathrm{IPS}) = \Sigma\, P_j(\theta)$$

(Wainer et al., 1990: 139)

In (Equation 1), a_j, b_j, and c_j are the item parameters of the three-parameter logistic function. The parameter b_j is the point on the θ metric corresponding to the inflection point of $P_j(\theta)$ and is interpreted as item difficulty; a_j is the item discrimination parameter and is proportional to the slope of $P_j(\theta)$ at the point of inflection; c_j is the lower asymptote of $P_j(\theta)$ and represents a pseudo-guessing parameter.

The minimum item pool score is the sum of the item $c_j(\Sigma\, c_j)$. The maximum item pool score is the number of items in the pool. The item pool score can be interpreted as the expected score that an examinee would obtain if given every item in the pool. It can be thought of as an expected score on a supertest.

Item subpool score. Often a particular subset of items from a pool is designed to represent a meaningful whole, and an expected score is sought for that subset. For example, certain items may form an already existing reference test for which scores have a well-known meaning. Such a subset of items might be particularly useful for equating CAT-based scores to paper-and-pencil test that are already placed on a well-established score scale. In situations like this, a subpool score composed of expected performance on that set of items can be obtained. The operations involved are identical to those used to obtain the item pool score except that only a subset of items is employed in the computation.

The minimum score on the item subpool is the sum of the c_j for the items in the subpool. The maximum item subpool scores are the number of items in the subpool. The item subpool score can be thought of as the expected score on the subpool.

Of the scores associated with the adaptive testing process, they are based on the proficiency score provided by the IRT model that underlies the testing process. Each of these scores defines a scale in itself. In addition, a variety of scales can be derived from these latter three scores.

2.5.7 CAT of reading proficiency

As mentioned above there are many types of CAT. In order to develop a CAT of reading proficiency Larson (cited in Chalhoub-Deville, 1999) suggests CAT procedures for testing reading proficiency. The four major factors identified by Lowe (1984) that impinge on receptive skills testing have to be considered. They are correction or scorability, production, content, and administration. Each of these factors can be managed well by employing computer-adaptive testing procedures.

Scorability: Testing specialists (Kaya-Carton et al. 1991) have tried to employ a variety of item types e.g. multiple-choice, multiple-choice cloze, free response cloze, scrambled-order sentences, cloze elide or cloze edit, open-ended, free-response, in a CAT format but found that they had to settle with objectively-scored items. In particular, the computer is very proficient and efficient at evaluating binary-choice (i.e. right or wrong) test items.

Production: A CAT reading test has the advantage over traditional paper-and-pencil versions in that each test is virtually unique, generating, as it was, multiple parallel test forms. This feature of CAT reduces dramatically the threat of test compromise. Additionally, it is possible to design the test to allow for piloting new test items during test administration. These new items, if they function properly, can be calibrated automatically and later inserted into the test item bank if desired. Also, other items when outdated or found problematic can be deleted easily from the item bank without affecting the performance of the test.

Item content: The content of test items is the single most important aspect of any test, including CAT proficiency tests. Items in reading proficiency measures must be valid, reliable and acceptable. To the extent that this is possible and feasible, they should contain samples of authentic language, and should be general enough to represent a variety of contexts. Computer-adaptive tests can be constructed to take into account item content considerations.

Administration: Traditional reading proficiency tests generally consist of objectively-scored items presented via paper-and-pencil testing formats. These types of items are just as easily incorporated into a computer-adaptive testing format. The computer excels in its ability to score an examinee's performance on such items and then branch to an appropriate subsequent item.

The designs and operation of CAT reading proficiency can be varied. Larson (1987) suggests a number of branching schemes specifically for selecting passages and items in an adaptive reading test. This type of test requires the examinees to answer a single item associated with each reading passage. Each reading passage is relatively short. Furthermore, Madsen (1991) suggests using a modified adaptive format in which the examinee would be required to answer three or four items after branching to a given passage. This he claims is preferable to requiring the examinee to read a lengthy passage, answer a single question, and then repeat the process again. Another approach, he suggests, is to have a relatively large number of items for each passage at varying levels of difficulty and branch within these items before going on to another passage. Kaya-Carton et al. (1991) posit the following alternatives for branching in a computer-adaptive reading test:

1. letting the examinee respond to items of a wide range of difficulty.
2. returning the examinee to the passage for another item of different difficulty level:
3. eliminating the passage from further testing; or
4. reducing the range of item difficulty associated with a given passage, letting the subject complete all items associated with the particular passage, and letting the computer calculate a passage mean and branch to another passage with a comparable mean passage score.

Zenisky et al (2003) raise concerns about local independence of test items which is also an important consideration relating to item context. Strict care must be taken to ensure that in the item sequencing, no item is dependent upon or related to another within a given test. Failure to do so will seriously affect validation of the test. In reading tests, Wainer and Kiely (1987) observed that the individual items on conventional reading comprehension tests may not satisfy the assumption of local independence because individual differences in familiarity with the various passage topics may induce higher dependence between questions within passages than across passages. However, Thissen et al. (1989) pointed out that a composite response to all of the items following each passage would be a single testlet response (with many alternatives), thus developing the test as a collection of those testlet responses; this would enable scoring to be handled smoothly with an IRT model for multiple-categorical responses. Further, Thissen et al. showed that the resulting testlet scores exhibited higher validity than scores derived at the item level.

Such results, and the test construction considerations described by Wainer and Kiely (1987), also suggest that at least some CATs are best constructed when adapted to the individual and scored as a set of testlets rather than as a set of a larger number of individual items. Both the test constructors and the psychological quality of the measurement of proficiency may benefit from this course of action. Further research in this area is clearly indicated; multistage testing has been neglected in favor of the consideration of individual-item-level CATs. However, both the computers and the IRT models have advanced to the point at which testlet-CATs are practical. Their performance should be considered in applied situations.

Based on the literature review in the previous section, this research aimed at investigating the effects of the mediums used to deliver a reading test. The type of computer-based test used as was a computer-adaptive test. Since another objective of this study was to compare the scores obtained from CAT with those from a paper-and-pencil test, a constant six-step pyramidal CAT was employed. Each examinee will return to the passage for another item of different difficulty level after every response. This type of CAT is used because the number and the content of the items can be controlled. In addition, the concern of local dependence can be solved by treating each test as a discrete item.

2.6 Attitudes towards test authenticity and test delivery medium

The goals of any foreign and second language studies are both linguistic and non-linguistic. The linguistic goals focus on developing the comprehension skills in the individual's ability to read, write, speak and understand the foreign and second language, and there are many tests available to assess these skills. On the other hand, non-linguistic goals emphasise such aspects as improved understanding of the other communities, desire to continue studying the language, an interest in learning other languages, etc. (Gardner, 1985). In order to understand test takers' performance, the non-linguistic goals that affect the performance should be investigated.

2.6.1 Factors affecting test takers' performance

In terms of factors affecting test takers' performance, Kunnan (1995) mentions that there are several kinds of factors which can cause different test performance. As Bachman (1990, cited in Kunnan, 1995) points out, communicative language ability, test method and test takers' characteristics are three important categories of influence apart from random factors. The third category is an interesting factor to be investigated because it can directly affect the test scores. The test takers' characteristics consist of cultural background, background knowledge, cognitive abilities, sex and age. Gardner and Clement (1990, cited in Kunnan, 1995) classify the following variables related to the test-takers' characteristics:

1. cognitive characteristics,
2. attitudes and motivation, and
3. personality attributes and the socio-structural perspective

Moreover, Saville (2000) identifies test takers' characteristics as sets of background factors that are hypothesized to affect second language acquisition and second language test performance. The factors can be grouped as strategic and socio-psychological and can be further divided into: (a) strategic factors: cognitive strategies; metacognitive strategies; and communication strategies, and (b) socio-psychological factors: attitude; anxiety; motivation; and effort.

Therefore, one of the important factors that can affect test takers' performance is test takers' characteristics which consist of many aspects. One important aspect of test-takers' characteristics is attitudes. Whittemore (2004) defines attitudes as a person's degree of favorableness or unfavorableness to a psychological object. Moreover, Baron and Byres (1994, cited in Whittemore, 2004) say that attitudes are about things from a solid object to ideas about oneself that are positive or negative evaluations of that object, stored in memory, and related to one another.

There is evidence from numerous studies showing the importance of attitudes. Gardner et al. (1976) cited in Kunnan (1995) and Lambert (1963, 1967) cited in Kunnan (1995) point out that attitudes and motivation can lead to successful second language achievement. Furthermore, Rand (1997) suggests ways to conduct tests that involve the demonstration of language skills. He mentions the importance of attitudes towards the test and suggests that positive attitudes of test takers towards testing should be created. This is because it can bring about better test results that both teachers and students desire.

Currently, students are afraid of tests because they view them as unfair, difficult, stressful and irrelevant to the course material studied. With positive tests, classroom motivation can be increased throughout the course, which in turn will lead to improved student performance.

Bachman and Palmer (1996) also suggest ways to create positive attitudes of the test-takers and these ways should be included in all phases of test development. Test developers should involve test takers in the design and development of the test by collecting information from them about their perceptions of the test and test tasks. Bailey (1999) also points out that if researchers feel that test-takers are involved in this way, they will perceive tests as more interactive and authentic, and will therefore be more motivated, which could lead to enhanced preparation and hence to better performance.

It can be concluded that one important factor that can cause different test performance is test takers' characteristics. Attitudes are one of the test takers' characteristics which can cause successful foreign and second language achievement. Positive attitudes towards testing can lead to better test results and can help improve student performance. Some ways to create students' positive attitudes towards testing are to involve them in the design and development of the test. Then, they will perceive tests as more interactive and authentic, resulting in more motivated, and better performance.

2.6.2 Attitudes towards test delivery mediums and test authenticity

After almost two decades of computers and the Internet being widely used, one might expect to see the increase of authenticity in language testing. However, authenticity is limited to some types of test formats such as multiple-choice and short answer. Ingram (2003) mentions that at a basic level, technology is used in testing reading comprehension in much the same way as it is used in testing listening comprehension with multiple-choice and drag-and-drop tests. However, there are many CD-ROMs and websites that offer interactive exercises with this function.

Regarding the attitudes of test takers, Pennington (1999) cited in Prapphal (2003) proposes that attitudes are one of the key elements in communicative computer-assisted language learning programs (CALL). Positive attitudes towards the computer may provide intrinsic motivation which can facilitate the students' writing process and self-expression as well as communication skills.

Although these technologies are widely used and useful for language assessment, there is evidence of limitations. Regarding reading on computer screens, there is some evidence showing that more extended reading tasks are harder to set on the computer. On-screen reading is difficult with longer texts. Research indicates that people read around 25% - 30% more slowly from a computer screen. Reading from computer screens is about 25% slower than reading from paper. Even users unaware of this fact usually report feeling fatigued when reading online texts. As a result, people don't want to read texts from computer screens. (http://www.useit.com/alerbox /9703b.html,1997)

Technology such as the computer and the Internet are widely used and they possess some advantages because they can increase the authenticity of the language learning experience and that of the language. However, there is some evidence from research showing a tendency of reluctance to read extensively from computer screens. Therefore, it is interesting to investigate the attitudes of test-takers towards reading tests delivered by the computer.

2.6.3 Studies related to the attitudes affecting language testing

As for attitudes towards testing, there are studies showing the effects of positive attitudes as perceived by test takers. Sturman (1996), who studied the washback effect which may affect learners' actions and/or their perceptions, used a combination of qualitative and quantitative data to investigate students' reactions to registration and placement procedures at two English-language schools in Japan. The placement procedures included a written test and an interview. He found that the students' perceptions of the accuracy of the placement process ie., the face validity of the results were statistically associated with their later satisfaction with the school, the teachers, and the lessons. It can be inferred from the results that when students have positive attitudes toward the testing procedures, face validity of the test tends to be high.

Brooks (2001) studied test takers' attitudes towards the computer-based TOEFL to investigate the attitudes of adult ESL learners towards the computer-based format. Both quantitative and qualitative results of the study suggest that the participants had a range of reactions to the computer-based format. They did recognise some of the benefits of this format of testing like the variety of possible question types, but at the same time they expressed some concerns about the new format, including reading long texts on the screen. Test takers indicated that the two computer-adaptive sections of the test seemed to match their level. In addition, the participants who felt comfortable using the computer had more positive attitudes towards the computer-format than the ones who were less confident in their computer abilities. In terms of implications for research, knowing how test takers feel about the mediums of test delivery can possibly help the development and implementation of CBT.

The results also showed that there were no differences between the scores obtained via computer-based and paper-and-pencil tests so long as students were motivated and testing conditions were equivalent. According to this study, it can be said that when a comparable testing environment is provided, different mediums of testing do not affect test takers' performance. The attitudes and motivation in doing both computer-based and paper-and-pencil tests are also equivalent.

Based on the literature review in the previous section, test takers' attitudes are one of the test takers' characteristics in test performance. Therefore, it is necessary to investigate how test takers feel about tests which are manipulated differently. The findings can help the development and implementation of test construction.

Conclusion

As there is a tendency to regard reading as a sociocultural practice, reading comprehension should not be assessed only in terms of process and product. The reader's purposes, and his or her social context should also be considered. Social contexts in the real world can be an important factor in a test since it can increase the generalizability of the reading comprehension tests, or the ability to make conclusions from one setting to another. Authenticity can be defined as the degree of correspondence of the characteristics of a given language test task to the characteristics of a target language use (TLU) task. The test which is considered authentic will allow a test taker to perform a real world task with knowledge gained prior to test taking.

One important characteristic of authenticity is that it is relative so we cannot judge whether the test is authentic or non-authentic, but it can be judged as relatively more or less authentic. Authenticity in a test can be determined by looking at the test takers, the TLU task, and the test task. More importantly, the authenticity of a test can vary across groups of test takers because the degree of authenticity is only estimated. Thus test developers can design authentic tests according to their judgment (Bachman, 1990). Because of this, the acceptable level of authenticity depends on the situation.

In recent years, authentic assessments have become popular because they can provide more direct evidence and students are given greater opportunity to engage in the construction of meaning. Also, authentic assessment can measure students' ability to apply knowledge. Lastly, students can best demonstrate what they have learned.

In developing a reading test, test developers are likely to choose a passage whose topical content matches the kinds of topics and material they think the test takers may read outside the testing situation. Numerous studies related to test authenticity indicate that in designing any test a desired level of authenticity needs to be ascertained. Allowing test takers to participate in the test design can obtain the TLU domain which is used to determine the degree of test authenticity. Moreover, some researchers have found that more authentic tests can lead to more reliable and valid assessment (Wagner, 2002 and Lynch, 2003).

Another important factor needing to be considered in test construction is the medium of test delivery. Alderson (2000) points out that in a reading test, the medium by which the text is presented is a crucial variable which can affect the measurement of the test takers' reading ability.

At present, innovation and flexibility in computer-based tests (CBT) is possible but test implementation is not feasible until the conceptual problems associated with such new mediums of test delivery are thoroughly studied. However, the development of computer technology and CBT can bring about many advantages to language tests. First of all, the response latencies and time on texts or tasks can be recorded so that every detail of a learner's progress can be investigated. CBT is also useful to second language acquisition (SLA) research because it can be used in the assessment of strategies in SLA research. Moreover, the development of diagnostic tests of skills could be facilitated by the computer. CBT can also help develop measures of reading rate and speed. In computer-adaptive testing, the tests can be tailored to suit the readers' ability levels. The data from CBT can be collected, analyzed and reported to test takers more easily than paper-based tests (PBT). Lastly, the virtual community can benefit from the tests administered via the Internet.

When comparing the characteristics of CBT with those of PBT, it can be concluded that they are different in terms of the item types used. In computer-based tests, test items involve tasks such as clicking on word, sentence, paragraph and insertion. In terms of administration, computer-based testing is more flexible since walk-ins are permitted if space is available. Computer-based tests can be tailored to suit each test taker. This type of CBT is called computer-adaptive testing. Such tests present each test taker with different test items and may vary in different testing time.

With regard to studies on CBT and PBT, it can be said that though the results obtained from the studies show that there is no significant difference between the mean scores obtained from CBT and PBT, there is some evidence indicating that computer technology affects test-takers' performance. Although there are many studies comparing test takers' performance on CBT and PBT, there is no study investigating the authenticity of those tests which particularly focus on reading skills.

In conclusion, the effects of test methods (CBT, PBT) and the item types using the multiple-choice technique and other types which are more innovative are interesting to explore. More research related to mediums of test delivery and its impact on test takers' performance and their attitudes to the tests are needed (Alderson, 2000). The validity of computer-based tests of reading is also important to explore since an analysis of target language use domains should be considered in reading test construction.

As for the variables affecting test scores, different readers have different emotional responses. Understanding affective response during reading and its effects on comprehension and interpretation has important implications for the testing and assessment of reading (Alderson, 2000). The effects from test takers' variables should be considered in the reading construct.

Another important factor in different test performance is test takers' characteristics. One of them is attitudes, a socio-psychological factor which can affect successful foreign and second language achievement. Positive attitudes towards testing can lead to better test results and can help improve students' performance.

The findings from many studies suggest that attitudes are one of the most influential variables that predict the rate and success of learning. Some studies (Sturman, 1996; Brooks, 2001) reported that the students perceive that they comprehend the passages more when they use an instructional program via computer. As testing is an activity which interacts with teaching and learning, the effect of attitudes towards language testing is also necessary to investigate.

In conclusion, this study will focus on reading proficiency skills since there are still a limited number of studies investigating particularly the authenticity of reading tests and the effects of test delivery mediums. This study will therefore construct reading proficiency tests with varying degrees of authenticity, relatively more authentic and relatively less authentic. Then these tests will be delivered via two mediums, computer and paper.

The aims of this study were thus to investigate the effects of test authenticity and test delivery mediums on test takers' reading proficiency in English. The test takers' attitudes are also examined in order to know how test takers feel about the tests which possess different degrees of authenticity delivered by different mediums. The findings obtained may help the development and implementation of test construction.