CHAPTER I
INTRODUCTION

1.1 Background of the study

The transitions and changes both in the theory and practice of language teaching potentially exerts a significant influence on language testing. Thus the end of the 70s saw the formal structural analysis of language provide the main focus for language teaching materials during this period, just as the structural syllabus generated by the structural approach in its various forms provided the main source for language test development. The main design principles for language tests of this kind were based on behaviorist psychology and structuralist theories in linguistics. Test items from this time, often referred to as the psychometric-structuralist era, are characterized by an emphasis on objectivity of marking. This was achieved by using carefully written discrete-point multiple-choice items (Spolsky, 1975).

Although such discrete-point multiple-choice tests are still used nowadays, their shortcomings have been criticized by many scholars. Carroll (1961) noted that a major limitation of discrete-point multiple-choice tests was that they tested only one element of language at a time. Importantly such tests did not reflect real language use in most cases. Similarly, Alderson and Hughes (1981) raised the question of test validity, pointing out that in the construction of the discrete-point multiple-choice tests, test writers could produce test items which satisfied measurement and linguistic criteria, but did not make any serious attempt to provide a valid external context. He mentioned that the construction of test items was based on a list of structures rather than on a reflection of language use in real life situations. However, since many language teaching methods were structural, the tests were considered valid. Their validity might now be questioned on the grounds that the language segments sampled for test items were neither adequate nor authentic and that the relationship between use and usage was left unexplored.

The communicative approach evolved out of a shift in language teaching, learning theory and methodology away from a predominantly structural focus to one that emphasized the importance of language in use. The communicative approach is based on the premise that the language to be taught should be related as closely as possible to the learner's immediate and future needs, that the learner should be prepared for authentic

communication, and that the language taught should have a high surrender value (Munby, 1978).

Accordingly there has been a shift in language testing, from the psychometric-structuralist framework, which emphasizes language structures, to communicative language testing, which concentrates more on language in use, the learners' future needs and authentic communication.

However, there are still ongoing changes in teaching. The constructivist approach is a current approach. Constructivists believe that learners are responsible for their own learning and must construct their own understanding by integrating new information to prior experiences; furthermore learners can learn, understand and remember best when they discover knowledge of their own exploration (Jonassen, 1995). More importantly, constructivists believe in purposeful knowledge which comprises three main features: authenticity, real practice and collaborative learning (Jonassen, 1991). The more opportunities learners have to interact using the language learned, the more they become fluent in their language use. As the teaching approach tends to move towards the constructivist approach, language in use, authenticity and real practice are thus the focus of language learning and teaching.

Since the focus of language teaching has shifted from language structures to language in use, trends in language assessment have also shifted. Due to the fact that discrete–point multiple-choice tests cannot reflect language use in real life, this kind of test format is considered invalid for the current teaching approach, which focuses on language in use, authenticity and real practice. Shohamy (1998) mentions that a paradigm shift in language testing caused by teaching approaches introduces new criteria for the validity of language tests. Consequential, systemic, interpretive and ethical concepts are a few of the new forms of validity, calling for the need to collect empirical data on the use of language tests. Such evidence may show that tests considered valid in the past may not be so if they are shown to have negative consequences.

In brief, the current trend in language assessment, which is influenced by the trend in language teaching is to reflect the current view of language and language use. Test writers should consider what is being tested and the kind of tasks or item types chosen as a means of testing. Particularly, in terms of test authenticity, test writers should construct a test which can be expected to show the influence of current ideas on what constitutes language ability and what exactly we are doing when we use language in our everyday life.

In terms of language skills, reading is probably recognized as the most important with its access to knowledge and information, and "worlds of ideas and feelings, as well as the knowledge of the ages and visions of the future" (Alderson, 2000: x). Importantly, for second language learners, the reading skill is very important in academic contexts. This gives rise to the challenge of exploring and understanding basic comprehension processes contributing significantly to implications for second language reading instruction (Lynch and Hudson, 1991). Similarly, the assessment of reading ability is of critical importance in a wide range of educational and professional settings, and the need for expertise in this area is widespread (Alderson, 2000).

Reading instructions have been developed for many decades and the constructivist approach is a current one which focuses on reading instruction. Constructivists see reading as social practices, which affect when you read, what you read, where you read, whom you read with, and of course why and how you read. Interacting with texts can involve practices as diverse as reading instructions, scanning a newspaper or reading an academic article. Learners need to understand that all readers construct meaning from texts differently, depending on their motivation, their background and even their state of mind (Wilson and Stacey, 2003).

It can be concluded that the current trend in reading assessment mirrors that in language teaching, via testing to assess language in use. Therefore, real world tasks or authentic tests become a crucial attribute of reading tests.

Accordingly, authenticity has become an increasingly crucial issue to be considered in developing language tests. The criticisms mentioned above have initiated and sustained a movement towards authentic assessments in a wide variety of learning domains.

In terms of test authenticity, McNamara (2000) points out that testing is about making an inference. Even when the test simulates the real world like reading a newspaper, it is not valued in itself, but only as an indicator of how a person would perform a similar or related task in the real world. Understanding testing involves recognizing a distinction between the criterion (relevant communicative behavior in the target situation) and the test. McNamara (2000) proposes Figure 1.1 to illustrate the distinction between a test and its criterion.

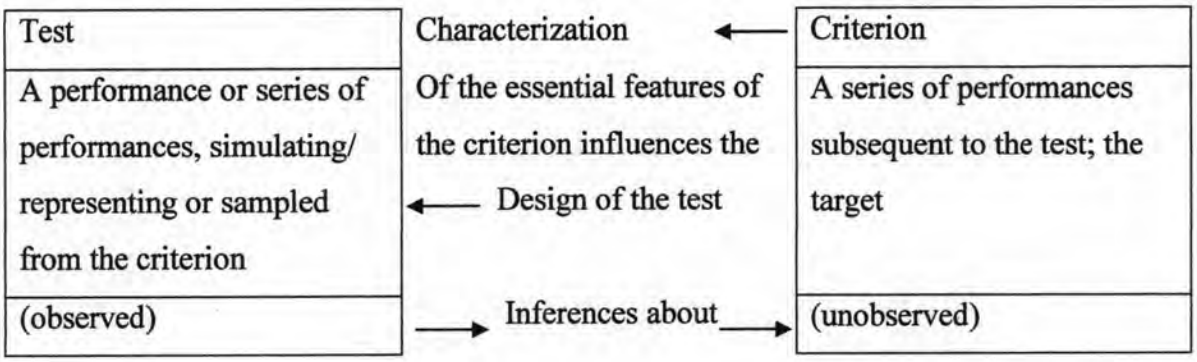

Figure 1.1: Test and criterion

(McNamara, 2000: 8)

It can be said that test authenticity is determined by the relationship between the criterion and test performance. Tests are based on theories of the nature of language use in the target setting and the way in which this is understood. This will be reflected in the test design.

The movement towards authenticity, therefore, has been an attempt to achieve a more appropriate and valid representation of student communicative reading competencies than that derived from discrete-point multiple-choice tests. In order to improve the authenticity of the test, McNamara (2000) suggests that from a practical point of view test design begins with decisions about the test content and what will go into the test. In fact, these decisions imply a view of the test construct, the way language and language use in test performance is seen, and the relationship of test performance to real-world contexts of use. In major test projects, articulating and defining the test construct may be the first stage of test development, resulting in an elaborated statement of the theoretical framework for the test.

Similarly, Bachman and Palmer (1996) in their discussion of construct validity also mention that the test construct allows more test authenticity by drawing on the validity theory of the great educational assessment thinker, Samuel Messick. Bachman and Palmer have examined the implications of Messick's work for language testing in a series of landmark papers and texts. Their approach focuses on an understanding of the nature of communicative language ability underlying test performance, and the relationship between test design and the contexts of future test use, which they define as the Target Language Use (TLU) domain or situations.

Moreover, Bachman and Palmer describe construct validity as pertaining to the meaningfulness and appropriateness of the interpretations that we make on the basis of test scores. When we interpret scores from language tests as indicators of test takers' language ability, a crucial question is to what extent can we justify these interpretations? The clear implication of this question is that as test developers and test users we must be able to provide adequate justification for any interpretation we make of a given test score. That is, we need to demonstrate or justify the validity of the interpretations we make of test scores, and not simply assert or argue that they are valid.

In order to justify a particular score interpretation, the construct that we want to measure, should be defined. The construct can be considered to be the specific definition of an ability that provides the basis for a given test or test task and for interpreting scores derived from this task. The term construct validity is, therefore, used to refer to the extent to which we can interpret a test score as an indicator of the ability (ies), or construct (s), we want to measure. Construct validity also has to do with the domain of generalization to which our score interpretations generalize. The domain of generalization is the set of tasks in the T [arget] L [anguage] U [se] domain to which the test tasks correspond. At the very least we want our interpretations about language ability to generalize beyond the testing situation itself to a particular TLU domain (Bachman and Palmer, 1996).

Accordingly, test authenticity is determined by the construct of the test. In order to increase the test authenticity, the target language use (TLU) domain has to be defined so that the scores obtained from the test would be able to generalize to the target situation of the language use.

Moreover, in defining the TLU domain, Bachman and Palmer (1996) suggest that the response format is another thing to consider in test design in order to increase test authenticity. The response format is the way in which candidates will be required to interact with the test materials, or the way in which the candidate will be required to respond to the materials. In terms of the response format, McNamara (2000) mentions that most of the traditional reading tests use limited types of test methods; these tests are usually used to check comprehension by asking test takers questions about what they have read. This is the most common way of testing reading comprehension but it is not the only way. Other testing techniques, which correspond to realistic tasks, should be used.

In conclusion, test authenticity is the attribute that should be considered in order to construct a test. The authenticity can be considered both in terms of the test construct

which includes the consideration of the test criterion or the TLU domain and the test performance. Moreover, the test construct defined and the authentic response format are also important features of authenticity. Because of the reasons mentioned above, test authenticity will be investigated in this study. Specifically, the researcher wants to examine whether authenticity can affect test takers' reading scores.

Rapid developments in computer technology have had a major impact on test delivery. Many important national and international language tests, including Test of English as a Foreign Language (TOEFL), are moving towards computer-based testing. Stimulus texts and prompts are presented not in examination booklets but on the screen, with candidates being required to key in their responses. However, the advent of Computer-based Tests (CBT) has not necessarily involved any change in the test content, which may remain quite conservative in its assumptions, and often simply represents a change in test method (McNamara, 2000).

Similarly, Ingram (2003) also mentions the importance of technology like computers in the development of language assessment since computers and the Internet have been widely used. One may see their substantial use to enhance language teaching and learning and, in particular, to increase the authenticity of the language learning experience and of the language that is tested. Madsen (1991) also agrees that computers involved in language test delivery can allow for maximum creativity and communicative expression on the part of the candidate, while making allowance for the still rather primitive state of the art as far as productive language skill correction via computers is concerned. Because of this, delivering tests via computers can increase test authenticity.

With the higher capacity of computers in evaluating test scores, computer-based tests are replacing paper-based tests. Cohen (1994) points out that modern technology has been employed in computer-based testing and in a subtype of computer-based testing, namely, computer-adaptive testing which involves the use of the computer as a vehicle for assessment instead of paper-and-pencil. Regarding the construction, it is more convenient to create computer-based tests using supported programs.

The computer-adaptive language test is uniquely tailored for each student. Madsen (1991) describes briefly its characteristics and advantages. The adaptive or tailored computer testing accesses a specially calibrated item bank and is driven by a statistical routine which analyzes student responses to questions and selects items for the candidates that are of appropriate difficulty. Then, when a specified standard error of measurement level has been reached, the exam is terminated. The psychometrically sound tailoring

process in computer-adaptive tests can provide for a more effective measure of language proficiency.

Item Response Theory (IRT) is applied to analyze the test items in computer–adaptive tests. According to Baker (2001), the central concept of IRT is the notion that persons can be placed on a scale on the basis of their ability in a given area, and that items measuring this ability can be placed on the same scale. Thus, there is a single scale which measures both difficulty and ability simultaneously. It is via this scale that the connection between items and respondents occurs.

The computer–adaptive testing (CAT) possesses many advantages. Madsen (1991) notes the convenience of providing exam results immediately: the benefit of accurate and consistent evaluation, diagnostic assistance to teachers and administrators, relief to test writers, and swift access to banks of test items. Test types other than multiple-choice questions can also be administered such as a cloze test which requires words to be typed into blanks in a prose passage. Furthermore, the tedious problem of deciphering student handwriting is eliminated.

The experimental findings reveal their superiority to paper-and-pencil tests in terms of reliability and validity, particularly when relatively few items are administered and there is a substantial reduction in time for the exam (Sukamolson, 1993). Similarly Madsen (1991) reported that over 80% of students required less reading items normally administered on the paper-and-pencil test.

Computer-adaptive tests of grammar and vocabulary have long been available, and recently similar tests of listening and reading skills have been developed (McNamara, 2000). However, in the Thai context, especially at King Mongkut's Institute of Technology North Bangkok, there is no research investigating the effects of the adaptive tests on reading performance. This study, therefore, aims to investigate the effects of different test delivery mediums (computer-based and paper-based) on test takers' reading scores.

In terms of the impact or washback of a test, Alderson and Wall (1996) mention that one needs to consider the impact of a test not only on teaching and learning but also on attitudes, material and effort. Attitude is considered an important factor which is able to affect language learning (Gardner,1985). Regarding testing, Rand (1997) mentions that positive attitudes towards testing can translate into better test results. Bachman and Palmer (1996 cited in Bailey, 1999) suggest ways to create positive attitudes of test takers. One is that test takers of particular tests should be included in all phases of test

development and design. Another is by collecting information from them about their perceptions of the test and test task. Bailey (1999) points out that researchers feel that if 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 said that test takers tend to have positive attitudes towards tests which they perceive as related to them. Accordingly, the test related to them will be considered as more authentic and able to create positive attitudes. Perceptions of authenticity not withstanding, perceptions of unfairness and stress, which are caused by the medium used to deliver the tests, also tend to affect the test takers' attitudes.

Because the computer, particularly in the case of computer-adaptive testing, can deliver tests tailored to the particular abilities of the candidates, it makes sense to use the very limited time available for testing to focus on those items that are just within, and just beyond a candidate's threshold of ability (McNamara, 2000). This is unlike traditional linear tests where all candidates take all the questions on a test, some of which are so easy for some candidates that they provide little information on their abilities, while others are too hard to be of use. Test takers who take the computer-adaptive test tend to have a more positive attitude towards the test because the items taken will not be too easy or too hard for them.

Of course computer-adaptive testing has limitations. Firstly, Madsen (1991) and Kiratibodee (2006) report on test taker anxiety as a drawback of tests delivered by computer. Another drawback is potential bias in computerized exams due to unfamiliarity with new technology. Interaction with the computer may thus be a stressful experience for some. Regarding reading tests delivered by computers, McNamara (2000) raises the questions about the importance of different kinds of presentation format delivered by computers. Thus, questions about the impact of computer delivery still remain.

In addition, attitudes of test takers towards both test authenticity and test delivery medium may be related to the impact or washback of language testing. Therefore, the effect of attitudes towards language testing is examined in this study.

In conclusion, test authenticity and test delivery medium may affect test takers' ability. This study will focus on reading tests since reading skill is very important particularly in academic contexts and there are still a limited number of studies investigating authenticity in reading tests. Furthermore, the use of computer-adaptive tests

of reading skills is a relatively recent development (McNamara, 2000). In the Thai context, there is a lack of research in this area.

First year students at King Mongkut's Institute of Technology North Bangkok who specialize in the fields of science and technology participate in this study. Reading skill is considered very important to them as previously mentioned and once students study science and technology, they spend a good deal of time reading technical material, analyzing it, and responding to it (Spretnak, 1983).

In terms of English language learning at King Mongkut's Institute of Technology North Bangkok, the content of English textbooks used in the past indicated a mixture of that found in typical materials of the Grammar-Translation Method and Audio-Lingual Method. The reading skills, in fact, were mostly limited to skimming and scanning. The texts used in the textbooks were taken directly from some academic textbooks in the students' field of study and not designed specifically to teach reading comprehension. The traditional reading tests used were composed of comprehension questions in the multiple choice format. Later the Department of Languages, Faculty of Applied Arts replaced the coursebook used for the foundation courses with commercial textbooks designed particularly for language in use (Scrivener, 2005). The current textbook, Straightforward, provides a series of communicative activities which expose students to extensive and varied language use. Although the textbook and teaching methods have become more communicative, the reading test tasks are still in multiple-choice format without consideration of language use in context. Because of this, there is a need to develop a test which is more relevant to real-life language use of students.

In addition, since the committee of the Testing and Translation Center at the Faculty of Applied Arts realize the importance of technology used in language assessment, there is a policy to develop an English standardized test administered by means of computer (Faculty of Applied Arts: King Mongkut's Institute of Technology North Bangkok, 2005).

Because of the reasons mentioned above, there is a need to develop a more authentic reading test delivered by means of computer. Thus, the aims of this research study are 1) to investigate the effects of test authenticity and test delivery mediums on test takers' reading proficiency and 2) to explore the attitudes of the test takers towards test authenticity and test delivery mediums. This study is conducted with first year students at King Mongkut's Institute of Technology North Bangkok. All of them study science and technology. Reading skill is the skill they tend to use most frequently both in their

academic context and daily life. Therefore, the Department of Languages, the Faculty of Applied Arts, provides particular reading courses to fulfill these needs. The findings obtained will help to improve the reading assessment of those courses and the standardized tests created by the Testing and Translation Center.

1.2 Research questions

1. Can test authenticity and test delivery mediums have any effects on test takers' English reading proficiency and what are their effect sizes?
2. What are the test takers' attitudes towards test authenticity and test delivery mediums?

1.3 Objectives of the study

1. To investigate the effects of test authenticity and test delivery mediums, and their effect sizes on test takers' reading English proficiency;
2. To examine the test takers' attitudes towards test authenticity and test delivery mediums.

1.4 Statement of hypotheses

1. The mean score obtained from the authentic reading comprehension test is significantly different from that obtained from the inauthentic reading comprehension test at the significant level of .05.

   ($\overline{X}$ Authentic $\neq$ $\overline{X}$ Inauthentic)

2. The mean score obtained from the computer-based reading comprehension test is significantly different from that obtained from the paper-based reading comprehension test at the significant level of .05.

   ($\overline{X}$ CBT $\neq$ $\overline{X}$ PBT)

3. There are significant interaction effects between test authenticity and test delivery mediums on students' reading proficiency at the significant level of .05.

   ($\overline{X}$ Authentic CBT $\neq$ $\overline{X}$ Inauthentic CBT)

   ($\overline{X}$ Authentic PBT $\neq$ $\overline{X}$ Inauthentic PBT)

   ($\overline{X}$ Authentic CBT $\neq$ $\overline{X}$ Authentic PBT)

($\overline{X}$ Inauthentic CBT ≠ $\overline{X}$ Inauthentic PBT)

($\overline{X}$ Authentic CBT ≠ $\overline{X}$ Inauthentic PBT)

($\overline{X}$ Inauthentic CBT ≠ $\overline{X}$ Authentic PBT)

1.5 Scope of the study

1. Population

The population of this study is first year undergraduate students at King Mongkut's Institute of Technology North Bangkok in the second semester of the academic year 2006. The students study in the Faculties of Engineering, Applied Science, Technical Education, Industrial Technology and Management and Agro-Industry. Typically, most are male, aged between 18 – 22.

2. Computer-Adaptive Test (CAT) Model

The model used in this study is a fixed-branch computer-adaptive test. All the items in the test are placed in a branching tree in advance. A different item is chosen to be presented next, a more difficult item for a correct response and an easier item for an incorrect response according to the fixed-branch established.

The CAT created in this study is a content-based CAT. Therefore, an incorrect response to a previous item will lead to an easier item measuring the same dimension. On the other hand, a correct response to a previous item will introduce a more difficult item measuring a different dimension.

3. Investigated Skills

This study only investigates general English reading comprehension skills.

4. Authenticity

The characteristics of authenticity in this study are determined by the format of the test tasks. In the authentic test, short answer, gap-filling and information-transfer tasks are used.

5. The Target Language Use (TLU) domain (Bachman and Palmer, 1990) used to specify the characteristics of authenticity in this study is obtained from a survey conducted with English language instructors, currently-enrolled students, instructors of the other faculties at King Mongkut's Institute of Technology North Bangkok and employers of graduates from King Mongkut's Institute of Technology North Bangkok.

6. Item Response Theory Model

The 3-parameter model is used in this study. The 3 parameters or three characteristics of the item consist of a-parameter (discriminating power), b-parameter (difficulty), and c-parameter (guessing). The test items are analyzed by the X-CALIBRE program (a program which is able to analyze the 3-parameter model). According to this, a test taker's ability will be estimated by means of Maximum Likelihood Estimation.

1.6 Assumptions

Students who participate in this study are familiar with using computers and they put their best effort in doing the tests.

1.7 Definition of terms

1. Test authenticity: Test authenticity in this study refers to the Authentic Reading Tests which consists of authentic tasks.

Authentic test tasks are tasks with the following characteristics:

1. meaningful and relevant to the test takers (Messick, 1996)
2. contextualized (Brown, 2004)
3. corresponding to the TLU domain (Bachman and Palmer, 1996)
4. using a more real life format (Alderson, 2000)

The reading test tasks that lack the above characteristics are considered inauthentic reading tests in this study.

2. Test delivery mediums consist of two modes:

1. Conventional paper-and-pencil test administration
2. Computer-based administration

The former uses the Authentic English Reading Comprehension Conventional Paper-and-Pencil Test (ACON) and Inauthentic English Reading Comprehension Conventional Paper-and-Pencil Test (ICON).

ACON is a paper-and-pencil test consisting of 3 reading texts and 21 items for each text. The item types used in this test are short answer, gap-filling and information-transfer. The test items are presented in sequence, arranged from the easiest item (lowest b-parameter) to the most difficult items (highest b-parameter).

ICON is different form ACON in terms of the item type used. All the items in this test are written in the form of multiple-choice format.

The latter employs the Authentic English Reading Comprehension Computer-Adaptive Test (ACOM) and Inauthentic English Reading Comprehension Computer-Adaptive Test (ICOM).

ACOM is a computer-based test consisting of 3 reading texts and 21 items for each text. The items are from ACON. The item types used in this test are short answer, gap-filling and information-transfer. The test items are tailored according to the test-takers' ability by using constant six-step pyramidal model of CAT.

ICOM is different form ACOM in terms of the item type used. All the items in this test are written in the form of multiple-choice format.

3. Computer-based reading comprehension ability is the scores derived from ACOM and ICOM.

4. Attitudes are the opinions of the students on the two independent variables (test authenticity and test delivery mediums) which are obtained from the retrospective method using the semi-structured interview.

1.8 Significance of the study

If the hypotheses are accepted, the following results will be obtained.

1. In terms of theoretical significance, the findings will contribute to more meaningful constructs of a reading comprehension test. Test developers have to consider the target language use in real life in the test items.
2. This study can contribute to theoretical knowledge in the area of language assessment regarding the predictive validity of the reading comprehension. This is because the test scores obtained from the test which is related to target language use in real life may predict some future behavior of the test takers.
3. The findings which are about the effects of test authenticity and mediums of test delivery will provide information for educational practitioners to develop more effective reading comprehension tests to test takers.
4. The test takers' attitudes towards test authenticity and means of test delivery will be able to show the possible causes that influence the students' performance on the test items. Therefore, test developers could pay attention to those causes when constructing reading comprehension tests.

However, if the hypotheses are not accepted the educators and researchers can gain insights in the following aspects:

1. Target language use or test authenticity may not be an important aspect which affects the test takers' scores.
2. Target language use domains have to be carefully defined.
3. Test authenticity and mediums of test delivery may not be important variables in assessing reading comprehension.
4. The test takers' attitudes towards test authenticity and mediums of test delivery may not be a crucial factor in reading performance.