CHAPTER TWO
REVIEW OF LITERATURE

This chapter reviews pertinent research on language testing theory and the development of language tests. It consists of two parts: Approaches to Language Testing, and Methods of Statistical Analysis. The first part presents a chronological review of four different approaches to language test development, namely, the Pre-scientific Stage, the Psychometric-Structuralist Stage, the Psycholinguistic-Sociolinguistic Stage, and the Communicative Paradigm. The fourth approach is further divided into six sub-sections dealing with the journey of the theory of communicative competence into the fields of language teaching and testing, the different models of communicative competence and the distinguishing features of communicative language tests. This part of the chapter ends with a brief discussion of eclecticism as the dominant scene in most language tests today.

The second part of the chapter focuses on two popular methods of statistical analysis, namely item response theory (IRT) and classical item analysis (CIA). Because the latter of these theories is the statistical method adopted by the researcher in the analysis of the 2005 test data, it will dominate the discussion and form the core of this part of the chapter. Special focus will be given to two statistical indices within the theory, namely, item difficulty index and item discrimination index, which will be used in calculating the item statistics of the test.

Approaches to Language Testing

In the past decades, language testers spent much time and effort in attempts to find valid and reliable tests that met the demands of emerging approaches to language teaching. The main focus was on construct or language ability since the purpose of language testing is to measure language ability in a valid and reliable way. The different approaches to language testing is enough evidence of these attempts to develop valid and reliable language tests. Below is a review of these approaches.

Approaches to language testing are classified into four main categories: the essay-translation approach, the structuralist approach, the integrative approach and the communicative approach (Heaton, 1988, p. 15). Heaton’s terminology, at least in part, reflects the general tendencies and practices in the allied field of language teaching. Spolsky (1976, cited in Miyata-Buddy & Langham, 2000) had earlier made a
chronological classification of the approaches into three main stages. Spolsky’s term for the essay-translation approach is ‘pre-scientific’; his term for the structuralist approach is ‘psychometric-structuralist’; and his term for the integrative approach is ‘psycholinguistic-sociolinguistic’. While Heaton’s classification is based on concomitant developments in language teaching, Spolsky’s classification reflects the general characteristics of the era. Both classifications, however, are chronological. These approaches, remarks Weir, are not “‘distinct’ or ‘pure’ types” (1990, p. 2). In practice, most tests contain elements of all these approaches either in the test format or the techniques of assessment.

The following section adopts Spolsky’s three-way classification because it provides the most common terminology in the language testing literature. However, the terminology for the fourth category, the communicative paradigm, is borrowed from Weir (1990). The discussion of these different approaches reveals how trends in language testing have tended to follow trends in language teaching, which in turn have tended to follow trends in linguistics.

The Pre-Scientific Stage

Spolsky (1995, cited in Miyata-Buddy & Langham, 2000) designates this era as ‘pre-scientific’ because he traces its early beginnings to the Chinese civil service exams two thousand years ago and its present form to the less distant 18th-century Cambridge BA final examination (Tripos). This approach stemmed from the old grammar-translation method to language teaching. Grammar Translation, as the name suggests, is an approach to language teaching that foregrounds grammar and translation. For the grammar-translation teacher, the sentence is the basic unit of language practice and the students’ native language is the medium of instruction. The teacher lists the target language vocabulary, along with their native language equivalents, and uses them in grammar structures which the students are supposed to ‘learn’. The grammar rules are, therefore, taught deductively. The teachers place a high premium on the learning of grammar rules through translation “into and out of the target language” (Richards & Rodgers, 2001, p. 6). Naturally, writing and reading are emphasized at the expense of listening and speaking.

Language tests in the pre-scientific era usually consisted of essay writing and translation exercises. The students were asked to translate grammatical structures from the target language into their native language or vice versa. The tests did not rely on linguistic theory, and reliability was considered less important than the production
of a test that “felt fair” (Spolsky, 1995, p. 356, cited in Miyata-Boddy & Langham, 2000, p. 76). Clearly, the tests were highly subjective and unreliable. In his research, Asaad (1984) characterizes testing within the grammar-translation framework by “(a) a total lack of concern for statistical information; (b) reliance on the teacher’s subjective judgment; (c) lack of theoretical foundations” (p. 35).

Grammar Translation dominated the field of language teaching, and testing by implication, for about a century. By the 1940s, a decade which saw an ‘explosion’ of theories, Grammar Translation gave way to other ‘theory-informed’ methodologies. Besides, the growing interest in speaking proficiency among linguists and educators also contributed to the decline of a method “for which there is no theory” (Richards & Rodgers, p. 7) and which underlined the rote-learning of “endless lists of unusable grammar rules and vocabulary” (p. 6).

**The Psychometric-Structuralist Stage**

Spolsky (1976, cited in Weir, 1990) described this stage as the ‘psychometric-structuralist’ era. The term is derived from the reliance of testing theorists at this stage on insights developed by structural linguists and psychometricians. Language testers working within this conceptual framework endeavor to synthesize the findings of structural linguistics and psychometrics in an attempt to produce ‘reliable’ tests of language proficiency. The tests tended to measure the candidates’ knowledge of the elements of language (grammatical system and vocabulary) using objective and reliable testing techniques.

This stage is most apparent within the Audiolingual approach to language teaching. Audiolingualism, as a theory of language learning, was itself “derived from a view proposed by American linguists in the 1950s – a view that came to be known as structural linguistics” (Richards & Rodgers, p. 54). According to structural linguistics the primary medium of language is the spoken medium. All languages are spoken but not all languages have a graphological system for representing the spoken form. Besides, children learn to speak first and not all adults can write. Speech is language, so goes the principal dictum of the structural linguists. Like structural linguistics, the practitioners of the Audiolingual approach argue that language skills are learned more effectively if the items to be learned are presented in the spoken form first. Secondly, structural linguists assumed that language learning entails mastering the elements of the language and learning the rules used to combine these elements, from phonemes, through morphemes, words and phrases, to sentences.
Audiolingual teachers also underline the mastery of the elements of language and argue that these are learned through habit-formation and analogy, which "involves the processes of generalization and discrimination" (Richards & Rodgers, p. 57). In short, this 'scientific' approach to language analysis, structural linguistics, offered the foundation for the approach to language teaching known as Audiolingualism.

In the field of language testing, Robert Lado’s book *Language Testing* (1961) is considered the classical exposition of the structural linguists’ approach to testing. Lado and his followers (e.g. Harries 1969) assumed that language is a system of habits that involves matters of form, meaning and distributions of structure at several levels, i.e. the sentence, clause, phrase, word, morpheme and phoneme. Each of these elements of language constitutes a variable that language testers should aim to test. Tests subscribing to this view tended to "atomize and decontextualize the knowledge to be tested, and to test aspects of knowledge in isolation" (McNamara, 2000, p. 14). This practice of testing separate aspects of language is known as ‘discrete point testing’. Lado, the spokesperson for discrete point testing, believes that language should be broken down, using structural contrastive analysis, into small testable segments. Each test item is intended to give information about the test taker’s ability to handle that particular item of language. As a result, there was a need for “statistical measures of reliability and validity” to be used by language testers, who resorted to the use of the multiple-choice technique, which “lends itself admirably to statistical analysis” (Heaton, 1988, p.16).

According to Weir (1990), the advantage of discrete point tests is that they cover a large number of items in a comparatively short time and yield data which are easily quantifiable. This approach brought about the development of large-scale tests which were economical, easy to score and which permitted group testing. Among the tests developed during this period were the TOEFL (1964) and The Michigan Test of English Language proficiency (1961).

But although discrete point tests were well-received and widely applied, the approach came under criticism because of “the defects of the construct they seek to measure” (Weir, 1990, p. 2). Consequently, the validity of discrete point testing was drawn into question. Oller (1979) outlined the deficiencies of the approach in terms of construct validity as follows:

What makes it ineffective as a basis for teaching or testing languages is that crucial properties of language are lost when its elements are
separated. The fact is that in any system when the parts interact to produce properties and qualities that do not exist in the part separately, the whole is greater than the sum of its parts.” (p. 212 [original emphasis])

In testing the candidates’ language ability, it is not sufficient to test their linguistic ability alone. Testing communicative ability besides linguistic accuracy is very important in developing a valid language test. Knowledge of the separate elements of the language systems “in fact counts for nothing unless the user is able to combine them in new and appropriate ways to meet the linguistic demands of the situation in which he wishes to use the language” (Morrow, p. 145).

The Psycholinguistic-Sociolinguistic Stage

In response to a feeling that discrete point testing was no longer felt to provide a sufficient measure of language ability, testing moved into global integrative testing or what Spolsky (1976, cited in McNamara, 2000) termed the ‘psycholinguistic-sociolinguistic’ era. According to Asaad (1984) this trend is influenced mainly by developments in psycholinguistics and sociolinguistics that had their effect on language teaching and eventually on language testing. At that time language teaching suffered from the inadequacies of the foregoing approaches (Grammar Translation and Audiolingualism) in developing the communicative ability of the learners. There was a quest for alternatives to the grammar-based approaches and this quest led to the growing interest in communicative language teaching. According to McNamara (2000), there was a need within the communicative approach to language teaching for tests to measure the candidates’ productive language skills, which “led to a demand for language tests which involved an integrated performance on the part of the language user” (p. 14).

Davies (2003) points out that the psychometric-structuralist stage builds easily on the pre-scientific stage but the psychometric-structuralist and the psycholinguistic-sociolinguistic stages do not flow smoothly one into another. Davies remarks that what distinguishes the dynamic of stage 2 (psychometric-structuralist) from that of stage 3 (psycholinguistic-sociolinguistic) was that the imperative behind stage 2 was primarily a method of organizing and analysing language tests, while what motivated stage 3 was more a belief about language. (p. 357)
This is why the move from stage two to stage three is, therefore, marked by a ‘paradigm shift’.

Integrative tests integrate knowledge of relevant systematic features of language (e.g. grammar, vocabulary, pronunciation) with an understanding of context. These tests are often designed to assess the learner’s ability to use two or more skills simultaneously. McNamara (2000) points out that integrative tests (e.g. oral interviews and the composition of written texts) and tests involving comprehension of extended discourse “tended to be expensive, as they were time consuming and difficult to score, requiring trained raters; and in any case were potentially unreliable” (p. 15).

Oiler, in the 1970s, worked out a solution to this problem. He proposed what came to be known as the ‘Unitary Competence Hypothesis’ (UCH). According to the UCH, performance on a given test (which Oiler termed ‘a pragmatic test’) depended on the same underlying abilities “which it is argued every learner possesses regardless of the purpose for which the language is being learnt” (Heaton, 1988, p. 16). Oiler (1979) defines a pragmatic test as

Any procedure or task that causes the learner to process sequences of elements in a language that conform to the normal contextual constraints of that language and which requires the learner to relate sequences of linguistic elements via pragmatic mappings to extralinguistic context. (p. 38)

Oiler’s pragmatic tests underlie his very active advocacy of the use of cloze testing and dictation as types of integrative tests. It was argued by Oiler that a cloze test was an appropriate substitute for a test of productive skills, like those mentioned above, because it required the candidate to integrate grammatical, lexical, contextual, and pragmatic knowledge in order to be able to supply the missing words. Heaton (1988) agrees with Oiler. He points out that cloze testing is a useful tool for testing general linguistic ability, including the ability to use appropriate linguistic expressions in the appropriate linguistic and situational contexts. According to Heaton, it is argued that “three types of knowledge are required in order to perform successfully on a cloze test: linguistic knowledge, textual knowledge, and knowledge of the world” (p. 17). The Cloze procedure is now popular not only in achievement tests, but also in proficiency, placement and diagnostic tests. Hughes (2003) recommends the use of cloze testing to measure the overall language ability. For him the cloze procedure
seemed “very attractive as a measure of overall ability” because cloze tests were “easy to construct, administer and score” (p. 189). Furthermore, these tests are economical and highly reliable.

However, cloze tests, which were claimed to integrate language skills in ways which are approximate to actual language use, are indirect and so their communicative validity remains suspect. McNamara (2000) points out that “further work soon showed that cloze tests on the whole seemed mostly to be measuring the same kinds of things as discrete-point tests of grammar and vocabulary” (p. 16). Hughes, who recommends the use of the cloze procedure to test overall language ability, identifies another problem with this technique. Hughes remarks that educated native speakers “varied quite considerably in their ability to predict the missing words” and that some of them “did less well than many non-native speakers” (p. 189). The result is that the validity of the procedure was brought into question.

The other major type of integrative tests advocated by Oiler is ‘dictation’. Johnson (2001) reports that dictation was “decried in the 1960s because it was unclear exactly what it was testing” but it was later shown to be “of value” and enjoyed a revival in the 1980s (p. 298). Oiler (1979) had earlier claimed that global integrative tests such as dictation went beyond the measurement of a limited part of language competence, as is the case with discrete point tests. For Heaton, the integrated skills involved in tests of dictation include “auditory discrimination, the auditory memory span, spelling, the recognition of sound segments, a familiarity with the grammatical and lexical patterning of the language, and overall textual comprehension” (p. 17). Hughes (2003) argues that the results obtained from the dictation tests are similar to those obtained from the cloze tests. However, in predicting the overall ability of the candidate, dictation tests have the advantage of measuring the candidate’s listening ability. Contrary to the arguments of Hughes, some applied linguists, according to Heaton, argue that dictation tests tend to encourage the student to focus on the individual sounds rather than on the text as a whole. This shift of focus from the text to individual sounds is “sufficient to impair the auditory memory span”, which makes it difficult for the students “to retain everything they hear” (p. 18). Dictation tests, though easy to construct and administer, are not easy to score.

Indirect integrative tests such as cloze and dictation did not escape criticism. Both tests have been found to measure only the candidate’s linguistic competence. They do not tell us anything directly about the performance ability of the students.
Morrow states that neither cloze nor dictation allows for spontaneous production by the candidate and that the norms followed are those of the examiner and not of the student himself. Weir (1990) is equally skeptical. He remarks that “neither testing procedure offers the possibility for oral or non-controlled written production” and points out that because the oral and written skills are very important, “some means of assessing them reliably in communicative situations should be found” (p. 6 [my emphasis]). Change was a foregone conclusion. In the words of Weir

The deficiencies in the type of information the ‘discrete point’ approaches of the psychometric-structuralist era and the more integrative approaches of the psycholinguistic-sociolinguistic era can provide bring about a need to investigate the ‘communicative paradigm’ to see whether this approach might prove more satisfactory”. (p. 6)

**The Communicative Paradigm**

The fact that discrete point and integrative testing only provided a measure of the candidate’s competence rather than performance brought about the need for an alternative approach. This was Communicative Language Testing. Heaton remarks that the communicative approach to language testing is sometimes linked to the integrative approach in that “both approaches emphasise the importance of the meaning of utterances rather than their form and structure” (p. 19). However, there are important differences between the two approaches. According to Heaton, communicative tests are concerned “primarily (if not totally) with how language is used in communication” and as such their aim is to “incorporate tasks which approximate as closely as possible to those facing the students in real life” (p. 19). Unlike integrative tests, communicative tests are more concerned with language ‘use’ than with language ‘usage’. Heaton defines language use as the way “people actually use language for a multitude of different purposes” whereas usage is defined in terms of the “formal patterns of language (described in prescriptive grammars and lexicons)” (p. 19 [original emphasis]).

During the last three decades the field of second/foreign language learning and teaching has witnessed a major transformation. The focus of language learning and teaching theories has shifted from concentrating on grammatical forms and the various ways these elements combine to form grammatical or well-formed sentences to the communicative aspects of language. This change was triggered by Chomsky’s
(1965) differentiation between competence and performance. The assumptions of, and reactions to, Chomsky’s linguistic theory are dealt with in the next section.

**Impact of the Theory of Communicative Competence on Teaching and Testing**

According to McNamara (2000), a new theory of language began to exert a significant influence on language teaching and language testing from the early 1970s onwards. This new theory was Dell Hymes’ theory of communicative competence, which “greatly expanded the scope of what was covered by an understanding of language and the ability to use language in context, particularly in terms of the social demands of performance” (p.16). The theory of communicative competence initiated a profound shift from a psychological perspective on language which views language as an internal phenomenon, to a sociological perspective which focuses on the external and social function of language.

Before discussing Hymes’ theory of communicative competence, Hymes’ own debt to Chomsky must be acknowledged. Hymes takes Chomsky’s dichotomy between ‘competence’ and ‘performance’ (Chomsky, 1965) as the point of departure for his argument that Chomsky’s grammatical competence is not a sufficient basis for a total view of communication. Before Hymes (1972), linguistic theory was dominated by Chomsky’s dichotomy between competence and performance. Chomsky (1965) used the word ‘competence’ to refer to the underlying competence of “an ideal speaker-listener, in a completely homogeneous speech community” (p. 3). This underlying competence refers to the knowledge of the language system which includes rules of morphology and rules of how linguistic elements can be combined to generate well-formed sentences. Performance, on the other hand, refers to the actual use of language (production and comprehension) in real time. Real time production/comprehension is constrained by a number of factors such as “memory limitations, distractions, shifts of attention and interest, and errors (random or characteristic)” (p. 3). Therefore, according to Chomsky, performance cannot reflect the underlying competence which his theory of transformational-generative grammar attempted to account for.

The linguistic community accepted Chomsky’s dichotomy but some linguists such as Hymes (1972) thought that the notions of ‘competence’ and ‘performance’ as described by Chomsky were somehow limited. In Hymes’ opinion, Chomsky’s linguistic theory makes linguists see performance as “an imperfect manifestation of an underlying system” (Hymes, p. 272) and its main objective is to reconstruct the
grammatical structure of a language through its use in performance. Valle (2002) makes an assessment of Chomsky and Hymes’ perspectives on the issue. According to Valle, Chomsky’s linguistic theory focuses on “the ideal speaker-listener” whereas Hymes “was more concerned with the difference in character, abilities, performance, and attitudes that make up the different members of the language community” (p.7).

For Hymes, Chomsky’s theory accounted for the grammaticality of sentences but not for the appropriateness of utterances in a specific social context and so it failed to account for the sociocultural dimensions of language use. According to Hymes, the fact that a normal child acquires knowledge of sentences not only as grammatical but also as appropriate is not accounted for in transformational grammar, which divides linguistic theory into two parts: linguistic competence and linguistic performance. To incorporate the socio-cultural aspect of language, Hymes introduces the new concept of ‘communicative competence’, which, according to Valle, can be defined as “the ability to use language for the purpose of communication in a way that is appropriate for a specific sociocultural context” (p.7). Research in the areas of language learning and acquisition has reinforced this dichotomy between knowledge about language rules and forms and knowledge that enables a person to interact communicatively in a given speech community.

These developments in linguistic theory manifested themselves in the field of second language teaching with a radical change of the methodological models. The introduction of the concept of communicative competence in linguistic theory set the basis for the new communicative approach to second language teaching, which advocated a radical change when compared with Audiolingualism. Valle sums up the differences between the two models thus:

The differences between the two models can be summarized in the following shifts in focus: from structure to meaning; from word and sentence-level context to discourse-level context; from accuracy to fluency; from rote structure drills to meaningful, real language use; from made-for-classroom materials to authentic materials; and from discrete point tests to integrative tests of communicative competence.

(p. 8)

The most comprehensive attempt to elaborate a theoretical framework for the communicative approach came from Canale and Swain (1980). In their paper, Canale and Swain examine “currently accepted principles” of communicative approaches to
second language learning, which leads them in turn to develop “a somewhat modified set of principles which is consistent with a more comprehensive theoretical framework for the consideration of communicative competence” (p.1). This led to the emergence of communicative language teaching (CLT) which marks the beginning of a major paradigm shift within language teaching in the 20th century, one whose ramifications continue to be felt today. According to Richards and Rodgers, communicative language teaching refers to a diverse set of principles that reflect a communicative view of language and language learning and that can be used to support a wide variety of classroom procedures. Richards and Rodgers (p. 172) provide the following outline of these principles,

- Learners learn a language through using it to communicate.
- Authentic and meaningful communication should be the goal of classroom activities.
- Fluency is an important dimension of communication.
- Communication involves the integration of different language skills.
- Learning is a process of creative construction and involves trial and error.

Johnson and Johnson (1998, cited in Richards & Rodgers, p. 173) identify five core characteristics that underlie current applications of communicative methodology. The first characteristic is that of “appropriateness”. Language use is constrained not only by grammaticality but also by appropriateness. Such factors as the setting, the roles of the participants and the purpose of communication all matter in the choice of the register appropriate to the particular communication situation. Thus a language learner “may need to be able to use formal as well as casual styles of speaking”. The second of these characteristics relates to the “message focus”. The focus of teaching in this approach is to replicate real-life communication situations by way of designing activities that require the students to share information, i.e. create and understand messages. “Psycholinguistic processing” is the third characteristic of the communicative methodology. Communicative activities seek to “engage learners in the use of the cognitive and other processes that are important factors in second language acquisition”. The last two characteristics are “risk taking” and “free practice”. The first of these encourages language learners to go beyond what they
have been taught by making guesses, errors and other communication strategies. Students learn best by taking risks. The other, and last, characteristic of the communicative methodology involves a shift from the practice of teaching individual skills to a more “holistic practice” that encourages “the simultaneous use of a variety of subskills”.

CLT finds its parallel in communicative language testing (Johnson, 2001). The developments in linguistic theory and language teaching which resulted from the emergence of and continuing interest in communicative competence have also influenced the language testing community. The first obvious change that took place was the development of tests focusing on aspects of communicative competence rather than linguistic competence which was traditionally regarded as the nucleus of language tests. New tests were developed in order to test the other competencies claimed to be part of a language user’s communicative competence.

It is universally recognized by language test researchers that the development of valid second/foreign language tests can only be achieved if there is a clear definition of what it means to know a language. Language tests attempt to measure how proficient speakers are in using a language and what type of knowledge they fall back on in the use of language. This justifies the need for a theory which captures and adequately explains the nature of a speaker’s knowledge of language. The definition of language ability has always preoccupied pioneers in the fields of language teaching and testing. According to Spolsky (1989), language tests involve “measuring a subject’s knowledge of, and proficiency in, the use of a language”. The concern of any theory of communicative competence is with the nature of this knowledge and proficiency. Therefore, language testers have to be very clear about the nature of knowledge of language in order for them to be able to develop “sound language tests” because “until you have decided what you are measuring, you cannot claim to have measured it” (p. 140).

Davies (2003) points out that Spolsky in his 1976 paper had recognized that this new construct of communicative competence was “starting to achieve orthodoxy” (p. 358). The new construct has been implemented in the development of language tests since then. The more recent, and expanded, views on second language teaching approaches have contributed to the current expansion in language test development. For example, the developmental work of Morrow (1979), Bachman (1990) and
Bachman and Palmer (1996) are based on the communicative teaching approaches proposed by theorists such as Widdowson (1979) and Canale and Swain (1980).

Before listing the features which distinguish communicative language testing from the previous approaches to language testing, it will be better to review the models of communicative competence on which it is based. According to Brindley (2002), the most influential models within communicative language testing are Canale and Swain (1980) and Bachman (1990), which is updated in Bachman and Palmer (1996).

**Models of Communicative Competence**

Hymes (1972) viewed communicative competence as the interaction of grammatical (what is formally possible), psycholinguistic (what is feasible in terms of human information processing), socio-cultural (what is appropriate in terms of social meaning or values of a given utterance), and probabilistic (what actually occurs) systems of communication. Hymes’ work was important as a rudimentary sketch of a communicative approach that could be concerned with the interaction of socio-cultural contexts and meaning. However, his theory did not consider other different components of communicative competence such as the knowledge of how language is used to perform communicative functions, and how these communicative functions can be combined according to rules of discourse. Such integrative theories of communicative competence are best represented in the work of Canale and Swain (1980).

**The Canale and Swain model.** The response to the challenge of communicative language testing was marked with “a continuing theoretical engagement” with the concept of communicative competence and with its implications for “the performance requirement of communicative language testing” (McNamara, 2000, p. 17). Part of this response took the form of attempts to “specify the components of communicative competence” and the role of these components in actual performance in language. The purpose is to “provide a comprehensive framework for test development and testing research and a basis for the interpretation of test performance” (p. 17). Canale and Swain provided a useful clarification of the terminology necessary for forming a definite picture of the ability to use language communicatively. Their work began in an attempt to “determine the feasibility and practicality of measuring what we shall call the ‘communicative competence’ of
students enrolled in ‘core’ (similar to general) French as a second language programs in elementary and secondary schools in Ontario” (Canale & Swain, p. 1).

The most influential feature of this model was “its treatment of the domains of language knowledge as including, in addition to grammatical competence, sociolinguistic competence (following Hymes), strategic competence and (subsequently) discourse competence” (McNamara, 1996, p. 61 [original emphasis]). The first three competencies, i.e. grammatical competence, sociolinguistic competence and strategic competence, according to Canale and Swain, combine to produce communicative competence. The first, the grammatical competence, includes “knowledge of lexical items and rules of morphology, syntax, sentence-grammar semantics and phonology” (p. 29). The second is sociolinguistic competence, which is made up of “sociocultural rules of use and rules of discourse” (p. 30). The third competence they proposed, strategic competence, is related to “verbal and non-verbal communication strategies that may be called into action to compensate for breakdowns in communication due to performance variables or to insufficient competence” (p. 30). McNamara (1996) points out that the fourth competence, namely, discourse competence, was added to the model by Canale (1983a). Canale updated the 1980 version of the model by the introduction of discourse competence as a fourth component of communicative competence. Discourse competence is defined as “basically concerned with above-sentence-level cohesion and coherence” (Davies, et al, p. 47).

The main implication of this model for communicative language testing is that since there is a theoretical distinction between competence and performance, the learner has to be tested not only on knowledge of language, but also on the ability to put it to use in a communicative situation. However, Canale and Swain use the term communicative competence in a fundamentally different way from Hymes. For them, this term is restricted to language knowledge. According to McNamara (1996), Canale and Swain (p. 7) deliberately exclude Hymes’ “ability for use”, the model of underlying capacities in performance, from their model of communicative competence. Consequently, Canale and Swain’s 1980 model

… lacks a notion of potential for use or underlying skill. They are quite clear about this: they argue that while performance may demonstrate such factors as volition, motivation, etc., they ‘doubt that there is any theory of human action that can adequately explicate ability for use’
Canale and Swain view ‘ability for use’ as part of communicative performance, which they define as “the realization of these competencies and their interaction in the actual production and comprehension of utterances” (p. 6). This means that they consider communicative performance as the demonstration or the realization of the language knowledge they have proposed.

Although Canale and Swain developed one of the most influential theoretical frameworks for teaching and evaluation in second language programs, there are some problems with their position. According to McNamara (1996), the first problem is that they include some aspects of ‘ability for use’ in their discussion despite the fact that they do not acknowledge it. McNamara points out that,

The difficulty is revealed in their discussion of strategic competence, which as we have seen refers to the possession of “coping” strategies (Canale and Swain, 1980: 31) in actual performance in the face of inadequacies in any of the other areas of competence. (p. 62)

It is clear that what is involved here is not only knowledge but also ability or skill, because ‘coping’ is an aspect of performance that involves certain capacities such as problem-solving and also certain personality factors such as preparedness to take risks, versatility, etc.

Another problem with this model is acknowledged by Canale himself. Canale (1983a, p. 12, cited in McNamara, 1996, p. 63) acknowledges that the model does not explain how the different components of communicative competence interact with each other and with other factors in real communication. In Canale’s own words,

… this theoretical framework is not a model of communicative competence, where model implies some specification of the manner and order in which the components interact and in which the various competencies are normally acquired.

Comparing the original 1980 jointly authored paper with Canale’s (1983a) paper, Canale and Swain could not agree on the role of the theory of performance (McNamara, 1996). In Canale (1983a) paper, a model of underlying abilities in performance is called for and thus reversing the position reached in the 1980 paper. “Canale prefers the term actual communication to the previous term performance, because of the possibility that the term will be understood only in the sense introduced
by Chomsky (1965) (McNamara, pp. 63-64). Canale (1983a, cited in McNamara, p. 64) distinguishes *actual communication* and the knowledge and skills underlying it in the following way:

Both knowledge and skill underlie actual communication in a systematic and necessary way, and are thus included in communicative competence .... This view is not only consistent with the distinction between communicative competence and actual communication but depends crucially on it; in particular, this notion of skill - how well one can perform knowledge in actual situations – requires a distinction between underlying capacities (competence) and their manifestation in concrete situations (actual communication).

Canale (1983a, 1983b. cited in McNamara, 1996, p. 64), introduces a fourth aspect of language knowledge, i.e. discourse competence, which concerns the ability to combine meaning and grammar in the production of a “unified” discourse. Canale argues that unity of a text is achieved through cohesion in form and coherence in meaning. McNamara remarks that Canale (1983b, p. 339) “extends the scope of strategic competence to include the ability ‘to enhance the rhetorical effect of utterances’” and that he (1983a, p. 11) “sees strategic competence as operating in relation to each of grammatical competence, sociolinguistic competence and discourse competence” (1996, 65 [original emphasis]). Despite Canale’s realization of the need for a theory of performance, other aspects of his argument, as outlined in his two papers written around the same time, remain inconsistent.

Further years of discussion and reflection led to a detailed reformulation of Canale and Swain’s framework. There has been a further specification of different components of knowledge that would appear to be included in communicative competence. Thus Lyle Bachman (1990), for example, has identified subcategories of knowledge within the broader categories of grammatical, sociolinguistic and discourse competencies.

**The Bachman (1990) model.** The second model to be reviewed here is Bachman (1990). This model attempts to consolidate the communicative competence theory into a more comprehensive model. It has taken into consideration the objections and suggestions that resulted from the discussion of communicative competence and presented a new model of communicative competence which
Bachman called the “Communicative Language Ability”. Bachman specifies different components of this communicative language ability (Figure 2.1).

**Figure 2.1.** Components of communicative language ability in communicative language use (Bachman, 1990, p. 85).

Bachman’s proposed framework of Communicative Language Ability is comprised of a wide range of elements and is more comprehensive than the Canale and Swain model. McNamara (1996) observes that Bachman’s model, which contains the fullest treatment of communicative competence, recognized the problems with Canale and Swain’s (1980) model and corrected them. It is “easier to understand if we see it as a refinement and elaboration of Canale and Swain’s work, rather than an altogether new departure” (p. 66).
Regarding knowledge of language, Bachman (1990) proposes a reorganization and recategorization of the components of this knowledge. The model consists of three components, namely, language competence, strategic competence, and psychophysiological mechanisms. Being an important model in communicative language testing, Bachman’s model warrants a closer look at the components of Communicative Language Ability. Below is a brief discussion of Bachman’s treatment of the components of language ability.

Bachman defines **Language Competence** as “control of the rules of usage and use” (1990, p. 105). It consists of organizational competence and pragmatic competence (figure 2.2). McNamara (1996) defines organizational competence as the subcomponent which

… includes the knowledge involved in creating or recognizing grammatically correct utterances and comprehending their propositional content (*grammatical competence*), and in organizing them to form texts (*textual competence*). (p. 68 [original emphasis])

And characterizes pragmatic competence by the inclusion of

… knowledge of the pragmatic conventions for performing felicitous language functions (*illocutionary competence*) and knowledge of sociolinguistic rules of appropriateness (*sociolinguistic competence)*. (p. 68 [original emphasis])

One important point of development in this model is that pragmatic competence is seen as a ‘multi-faceted construct’ and an indispensable part of Communicative Language Ability. Another important point of development is in the way discourse competence is handled. According to McNamara (1996), Canale’s unsatisfactory **Discourse competence** is broken up, and the elements (**Cohesion** and **Coherence**) redistributed. Cohesion goes to **textual competence**, part of organizational competence (the other part being **grammatical competence**, the two together representing ‘those abilities involved in controlling the formal structure of language’). Coherence is itself divided between **illocutionary competence** and **strategic competence**. (pp. 68-69 [original italics])
Bachman’s second component of language ability is **Strategic Competence**. As shown in Figure 2.1, strategic competence is not seen as part of language competence in this model of Communicative Language Ability, as was the case with Canale and Swain’s model. Canale and Swain considered strategic competence as having a compensatory function which is employed when the other competencies of the user are not adequate. Although Bachman acknowledges the awareness of Canale and Swain’s description of strategic competence, he criticizes the inability of their model to account for the mechanisms by which strategic competence operates. Bachman (1990) reconceptualizes strategic competence and views it “as an important part of all communicative language use, not just that in which language abilities are deficient and must be compensated for” (p. 100). He argues that strategic competence operates in three general areas, assessment, planning and execution. These areas are involved in assessing, planning and executing the most effective means of achieving a communicative aim. Strategic competence is considered as “a general ability, which enables an individual to make the most effective use of available abilities in carrying
out a given task” (p. 106), rather than an area of knowledge. The three areas in which strategic competence operates will be dealt with later under the Bachman and Palmer (1996) model.

Bachman’s third component of communicative language ability is *Psychophysiological Mechanisms*. These are the neurological and physiological processes involved in language use. These mechanisms “characterize the channel (auditory, visual) and mode (receptive, productive) in which competence is implemented” (Bachman, 1990, p. 108). The implementation of the other competencies through these channels and modes produces the four language skills: listening, speaking, reading and writing.

Bachman's (1990) framework, which is an extension of the earlier models, recognizes the issue of the interaction of the components of communicative competence. This framework attempts to “characterize the processes by which the various components interact with each other and with the context in which language use occurs” (p. 81). Strategic competence is the means by which this is handled in the model. This point is clearly demonstrated in Figure 2.1.

Bachman points out that test performance is affected by the characteristics of the methods used to elicit test performance in addition to the abilities to be measured. He provides some examples of how the characteristics of test methods affect test performance. For example, individuals may not be able to speak well in an oral interview if the interviewer is someone they do not know. They will tend to perform better if the interviewer is known to them. Some test takers, Bachman gives another example, may find a cloze test “intimidating” and “do poorly” but perform well on “a test consisting of the completion of individual sentences, or a multiple-choice test based on a reading passage” (p. 111). These characteristics are important for designing, developing and using language tests, because they are the only factors upon which test writers have some control.

Bachman’s (1990, p. 119) framework of test method ‘facets’ includes five major categories. These five categories are the testing environment, the test rubric, the nature of the input the test taker receives, the nature of the expected response to that input, and the relationship between input and response. These categories are listed in figure 2.3. This framework of test method facets will be explained in some detail in the next section along with the refinement of this framework made by Bachman and Palmer (1996).
1. FACETS OF THE TESTING ENVIRONMENT

Familiarity of the place and equipment
Personnel
Time of testing
Physical Conditions

2. FACETS OF THE TEST RUBRIC

Test organization
Salience of parts
Sequence of parts
Relative importance of parts
Time allocation
Instructions
Language (native, target)
Channel (aural, visual)
Specification of procedures and tasks
Explicitness of criteria for correctness

3. FACETS OF THE INPUT

Format
Channel of presentation (aural, visual)
Mode of presentation (receptive)
Form of presentation (language, nonlanguage, both)
Vehicle of presentation (‘live’, ‘canned’, both)
Language of presentation (native, target, both)
Identification of problem (specific, general)
Degree of speededness.

Nature of language
Length
Propositional content
Vocabulary (frequency, specialization)
Degree of contextualization (embedded/reduced)
Distribution of new information (compact/diffuse)
Type of information (concrete/abstract, positive/negative, factual/counter-factual)
Topic
Genre

Organizational characteristics
Grammar
Cohesion
Rhetorical organization
Pragmatic characteristic
Illocutionary force
Sociolinguistic characteristics

Restrictions on response
Channel
Format
Organizational characteristics
Propositional and illocutionary characteristics
Time or length of response

4. FACETS OF THE EXPECTED RESPONSES

Format
Channel (aural, visual)
Mode (productive)
Type of response (selected, constructed)
Form of response (language, nonlanguage, both)
Language of response (native, target, both)

Nature of language
Length
Propositional content
Vocabulary (frequency, specialization)
Degree of contextualization (embedded/reduced)
Distribution of new information (compact/diffuse)
Type of information (concrete/abstract, positive/negative, factual/counter-factual)
Topic
Genre

Organizational characteristics
Grammar
Cohesion
Rhetorical organization
Pragmatic characteristic
Illocutionary force
Sociolinguistic characteristics

Restrictions on response
Channel
Format
Organizational characteristics
Propositional and illocutionary characteristics
Time or length of response

5. RELATIONSHIP BETWEEN INPUT AND RESPONSE

Reciprocal
Nonreciprocal
Adaptive

Figure 2.3. Categories of test method facets (Bachman, 1990, p. 119).
A comment by McNamara (1996) on this model beautifully sums up the advantages of Bachman’s model over previous models of communicative competence. Bachman’s model, so remarks McNamara, “helps us develop the necessary clarity, specificity, and explicitness in stating the grounds for inferences about candidates’ abilities made on the basis of test performance, thereby also facilitating the empirical investigation of such claims” (p. 71).

Bachman’s two frameworks, the framework of Communicative Language Ability and the framework of test method ‘facets’, were followed by some attempts by Bachman and Palmer (1996) to reformulate the model. The Bachman 1990 and the Bachman and Palmer 1996 models of language ability have “extended and refined earlier models” and “has generated much empirical research into the nature of language ability and performance on language tests” (Bachman, 2003, p. 421).

The Bachman and Palmer (1996) model. As mentioned in the previous section, the Bachman (1990) model has been developed by Bachman and Palmer (1996). One of the most important developments in Bachman and Palmer (1996) model is the inclusion of “what were previously regarded as ‘non-language’ factors, such as personality and background knowledge” (Brindley, 2001, pp.141-142). The 1996 framework includes the test takers’ knowledge of the world (topical knowledge) and emotional memories that may influence their test performance (affective schemata). These developments on the earlier model are diagrammatically shown in figure 2.4.

The changes that have taken place in Bachman and Palmer’s model can be divided into minor changes (in the language knowledge) and more substantial changes (in the ‘ability for use’). According to McNamara (1996), the changes in the language knowledge are “nominal”. The earlier illocutionary competence, a subcategory within pragmatic knowledge, has been relabeled “Functional knowledge, but is otherwise unchanged” (p. 72 [original emphasis]). The changes in the ‘ability for use’, on the other hand, are more significant. McNamara observes that the overall model has been given “a new component, called affective schemata” or “affect” which is described as “the affective or emotional correlates of topical knowledge”, which in turn are characterized as “knowledge structures in long-term memory” (pp. 72-73 [original emphasis]). Bachman and Palmer (1996) justify the inclusion of affective schemata in their model of language ability as follows:
The affective schemata, in combination with the characteristics of the particular task, determine, to a large extent, the language user’s affective response to the task, and can either facilitate or limit the flexibility with which he responds in a given context. (p. 65)

An important advantage of the inclusion of affective schemata is that the affective response of language users may influence “not only whether they even attempt to use language in a given situation, but also how flexible they are in adapting their language use to variations in the setting” (p. 65).

Figure 2.4. Some components of language use and language test performance (Bachman and Palmer, 1996, p. 63).
The other significant change is that the other broad aspect of ‘ability for use’, strategic competence, has been reconceptualized as a set of metacognitive components or strategies which can be thought of as “higher order executive processes that provide a cognitive management function in language use, as well as in other cognitive activities” (p. 70). Bachman and Palmer reformulate the three general areas in which metacognitive components operate. These three areas of metacognitive strategies are relabeled “goal-setting”, “assessment”, and “planning” (Bachman, 1991, p. 684). The uses of these three areas are illustrated in Table 2.1 below.

Table 2.1

<table>
<thead>
<tr>
<th>Goal Setting</th>
<th>Assessment</th>
<th>Planning</th>
</tr>
</thead>
<tbody>
<tr>
<td>(deciding what one is going to do)</td>
<td>(taking stock of what is needed, what one has to work with and how well one has done).</td>
<td>(deciding how to use what one has)</td>
</tr>
<tr>
<td>Identifying the test tasks.</td>
<td>Assessing the characteristics of the test task to determine the desirability and feasibility of successfully completing it and what is needed to complete it.</td>
<td>Selecting elements from the areas of topical knowledge and language knowledge for successfully completing the test task.</td>
</tr>
<tr>
<td>Choosing one or more tasks from a set of possible tasks (sometimes by default, if only one task is understandable).</td>
<td>Assessing our own knowledge (topical, language) components to see if relevant areas of knowledge are available for successfully completing the test task.</td>
<td>Formulating one or more plans for implementing these elements in a response to the test task.</td>
</tr>
<tr>
<td>Deciding whether or not to attempt to complete the task(s) selected.</td>
<td>Assessing the correctness or appropriateness of the response to the test task.</td>
<td>Selecting one plan for initial implementation as a response to the test task.</td>
</tr>
</tbody>
</table>

According to Bachman and Palmer, strategic competence consists of metacognitive strategies, which are executive processes that enable language users to engage in goal-setting, assessment, and planning. This model has been recommended by a number of researchers in the area of language test development and use. Brindley (2002) points out that researchers have identified a number of useful features of the Bachman (1990) model and the subsequent Bachman and Palmer (1996) framework. McNamara (1996) indicates that an important significance of the 1996 model is that it models aspects of Hymes’ ‘ability for use’ which were not previously modeled. McNamara considers this “an important advance” because it is for the first time that an attempt has been made to “deal explicitly in a model of second language
communicative ability with the aspect of ability for use which relates to affective or volitional factors” (p. 74 [original emphasis]). Smith (1999) is no less appreciative. He remarks that in Bachman and Palmer (1996), language ability is constructed “in a way which is practical enough for the practitioners to make use of when designing language tests”. According to Smith, an important addition of the Bachman and Palmer model is the consideration of “how far test results can be affected by what testees bring of themselves into the test situation”. The model extends the factors that influence performance on a test to non-linguistic elements such as “affective schema, metacognitive strategies in approaching the test task, and topical knowledge beyond the level of language” (p. 220). Brindley (2001) comments on Bachman and Palmer’s (1996) inclusion of personality and background knowledge as parts of communicative ability as “an important development since it recognises the key role that personal characteristics may play in language performance and opens the way for the development of assessment procedures which attempts to build such factors into the assessment situation” (p. 142).

Another important feature of Bachman and Palmer (1996) model is their framework of language task characteristics. Bachman (1990) test method facets are renamed task characteristics and presented “in the form of a checklist that can be used as a guide for describing and comparing target language use tasks and test tasks” (Brindley, 2002, p. 461). According to Bachman and Palmer, the characteristics of tasks are of interest for several reasons. First, they provide a link between tasks in different domains – the domain of test tasks and the domain of non-test tasks – and permit us to select or design test tasks that correspond to language use tasks. Second, the characteristics of the test task will help determine the extent and ways in which the test taker’s language ability is engaged. Third, the degree of correspondence between the characteristics of a given test task and of a particular language use task will determine the authenticity of the test task, the validity of inferences made, and the domain to which those inferences will generalize. Finally, they can be controlled by the way language tests are designed and developed.

Bachman and Palmer’s framework of task characteristics can be used for both language use tasks and test tasks. This framework consists of a set of features for describing five aspects of tasks: setting, test rubric, input, expected response, and relationship between input and response as shown in Table 2.2 below.
Table 2.2

<table>
<thead>
<tr>
<th>Task Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Characteristics of the setting</td>
</tr>
<tr>
<td>Physical characteristics</td>
</tr>
<tr>
<td>Participants</td>
</tr>
<tr>
<td>Time of task</td>
</tr>
<tr>
<td>Characteristics of the test rubrics</td>
</tr>
<tr>
<td>Instructions</td>
</tr>
<tr>
<td>Language (native, target)</td>
</tr>
<tr>
<td>Channel (aural, visual)</td>
</tr>
<tr>
<td>Specification of procedures and tasks</td>
</tr>
<tr>
<td>Structure</td>
</tr>
<tr>
<td>Number of parts/tasks</td>
</tr>
<tr>
<td>Salience of parts/tasks</td>
</tr>
<tr>
<td>Sequence of parts/tasks</td>
</tr>
<tr>
<td>Relative importance of parts/tasks</td>
</tr>
<tr>
<td>Number of tasks/items per part</td>
</tr>
<tr>
<td>Time allotment</td>
</tr>
<tr>
<td>Scoring method</td>
</tr>
<tr>
<td>Criteria for correctness</td>
</tr>
<tr>
<td>Procedures for scoring the response</td>
</tr>
<tr>
<td>Explicitness of criteria and procedures</td>
</tr>
<tr>
<td>Characteristics of the input</td>
</tr>
<tr>
<td>Form</td>
</tr>
<tr>
<td>Channel (aural, visual)</td>
</tr>
<tr>
<td>Language (native, target, both)</td>
</tr>
<tr>
<td>Length</td>
</tr>
<tr>
<td>Type (item, prompt)</td>
</tr>
<tr>
<td>Degree of speededness</td>
</tr>
<tr>
<td>Vehicle (‘live’, ‘reproduced’, both)</td>
</tr>
<tr>
<td>Language of input</td>
</tr>
<tr>
<td>Language characteristics</td>
</tr>
<tr>
<td>Organizational characteristics</td>
</tr>
<tr>
<td>Grammatical (vocabulary, syntax, phonology, graphology)</td>
</tr>
<tr>
<td>Textual (cohesion, rhetorical/conversational organization)</td>
</tr>
<tr>
<td>Pragmatic characteristics</td>
</tr>
<tr>
<td>Functional (ideational, manipulative, heuristic, imaginative)</td>
</tr>
<tr>
<td>Sociolinguistic (dialect/variety, register, naturalness, cultural references and figurative language)</td>
</tr>
<tr>
<td>Topical characteristics</td>
</tr>
<tr>
<td>Characteristics of the expected response</td>
</tr>
<tr>
<td>Form</td>
</tr>
<tr>
<td>Channel (aural, visual)</td>
</tr>
<tr>
<td>Language (native, target, both)</td>
</tr>
<tr>
<td>Length</td>
</tr>
<tr>
<td>Type (selected, limited production, extended production)</td>
</tr>
<tr>
<td>Degree of speededness</td>
</tr>
<tr>
<td>Language of expected response</td>
</tr>
<tr>
<td>Language characteristics</td>
</tr>
<tr>
<td>Organizational characteristics</td>
</tr>
<tr>
<td>Grammatical (vocabulary, syntax, phonology, graphology)</td>
</tr>
<tr>
<td>Textual (cohesion, rhetorical/conversational organization)</td>
</tr>
<tr>
<td>Pragmatic characteristics</td>
</tr>
<tr>
<td>Functional (ideational, manipulative, heuristic, imaginative)</td>
</tr>
<tr>
<td>Sociolinguistic (dialect/variety, register, naturalness, cultural references and figurative language)</td>
</tr>
<tr>
<td>Topical characteristics</td>
</tr>
<tr>
<td>Relationship between input and response</td>
</tr>
<tr>
<td>Reactivity (reciprocal, non-reciprocal, adaptive)</td>
</tr>
<tr>
<td>Scope of relationship (broad, narrow)</td>
</tr>
<tr>
<td>Directness of relationship (direct, indirect)</td>
</tr>
</tbody>
</table>
Bachman and Palmer (1996, p. 47) emphasize that the purpose of this framework is to provide a basis for the development and use of language tests. This involves three activities:

1. describing target language use (TLU) tasks as a basis for designing language test tasks,
2. describing different test tasks in order to insure their comparability and as a means for assessing reliability, and
3. comparing the characteristics of TLU and test tasks to assess authenticity.

By stressing the importance of using test tasks which correspond to those used in real life, Bachman and Palmer narrow the gap between the artificial testing situation and how language will be implemented by the test takers in the future. Carr (2006) points out that the framework of test task characteristics proposed by Bachman and Palmer (1996) “provides a means for systematically describing various characteristics of tests and testing procedures. It is primarily intended as a tool for designing and constructing language tests” (p. 269).

In addition to proposing two frameworks of language ability and of test task characteristics, Bachman and Palmer (1996) introduce the concept of ‘test usefulness’. In their own words, the “most important consideration in designing and developing a language test is the use for which it is intended, so that the most important quality of a test is its usefulness” (p. 17). To achieve test usefulness, Bachman and Palmer propose six qualities, namely, construct validity, reliability, authenticity, interactiveness, impact and practicality. Each of these qualities is carefully explained in simple terms and with illustrative examples. Bachman and Palmer also argue that test developers need to find an appropriate balance among these qualities. This balance will vary from one testing situation to another because what constitutes an appropriate balance can be determined only by considering the different qualities in combination as they affect the overall usefulness of a particular test. These six qualities are listed in the form of a checklist (Appendix B). This checklist includes places to indicate the degree to which and how the quality has been satisfied in a particular test. These qualities have been defined under the Definition of Terms in the first chapter.

In the second part of their book, Bachman and Palmer turn their discussion of frameworks that can be used in test development to a specific set of procedures for developing useful language tests. According to Smith (1999), this part presents a
“very detailed and helpful step-by-step approach used to take the reader through the many steps of designing a test” (p. 220). Bachman and Palmer organize test development into three stages: design, operationalization and administration. For them this type of organization will give a better chance of monitoring the usefulness of the test throughout the development process and hence of producing a useful test.

Bachman and Palmer’s frameworks have been drawn on in the construction and validation of a range of assessments for a variety of purposes and populations (Chalhoub-Deville, 1997). According to Brindley (2002), a number of these projects have used Bachman and Palmer’s components of language ability as “a basis for identifying and describing aspects of language ability to be included in tests and assessment procedures” (p. 462). Brindley also observes that there are signs of recent adoptions of Bachman and Palmer’s framework of task characteristics as “a practical model for test construction (see, e.g., Alderson 2000; Douglas 2000)” (p. 462). Bachman and Palmer’s (1996) analytical rating scales have also been adopted by some researchers such as Amodeo (2000) in scoring oral communicative competence in second-language classrooms.

Being unanimously acknowledged as one of the most important models of communicative language ability, Bachman and Palmer’s (1996) model will be adopted by the researcher as the practical model in the construction of the admission test in the three stages of test development, namely, design, operationalization and administration.

**Distinguishing Features of Communicative Language Tests**

There have been several attempts to define necessary conditions in order for tests to qualify for membership in the paradigm of communicative language testing (e.g. Carrol, 1980; Weir, 1990; Bachman, 1990). According to Weir, it is important to be “as precise as possible about the skills and performance conditions for any tests which claim to assess communicative language ability” (pp. 10-11). Test constructors must identify the skills and performance conditions which are the most important features of language use in particular contexts. These features, if incorporated wherever appropriate, would result in a test task which reflects “the attributes of the activity in real life that it was meant to replicate” (p. 11). Morrow maintains that language use in communicative situation has a number of features which are not measured in conventional language tests. He mentions six such example features of communicative tests, namely, that communication is interaction–based, that it is...
unpredictable, that it has context, that it is purposive, and that the language is authentic and behavior-based (pp.149-150). Wesche (1981, cited in Valle 2002, p. 22) offers suggestions for making tests “more” communicative. These suggestions, according to Valle, involve the use of

... authentic, longer, contextualized texts in stimulus items or multiple-choice responses; asking listening and reading comprehension questions based on the understanding of the general meaning of the text; asking about the appropriateness of language in an exchange; basing scoring of oral and written production on getting the message across and appropriateness, as well as on grammatical accuracy.

Based on these and other views (e.g. Heaton, 1988; Johnson, 2001; Davies, 2003; Hughes, 2003) of what communicative tests should be like, the principal distinguishing features of communicative language tests are summarized below.

Communicative tests should have high content validity. If communicative tests are to be used to make inferences about how a test taker can function in an ordinary situation outside the test, these tests have to be an accurate reflection of that situation. This means that “the sample of language collected and the tasks the candidate is called upon to perform should be as representative as possible of the language and skills needed to function in the real life context” (Miyata-Boddy & Langham, p. 78). Therefore, the most important criterion for communicative tests is that they should be based on “precise and detailed specifications of the needs of the learner for whom they are constructed” (Heaton, 1988, p.19). Weir, whose book Communicative Language Testing (1990) is considered to be one of the most important books on communicative language testing in the literature on language testing, points out that if we want to “make statements about communicative language ability in real life situations”, we need to pay careful attention to the nature of the texts and tasks used in language tests. These texts and tasks should “accord as far as possible with the general descriptive parameters of the intended target situation”. The test tasks should pay particular attention to the language skills necessary for successful interaction in similar communicative situations. In addition, tests should “meet the performance condition of that context as fully as possible” (p. 11).

Communicative language tests need to be context-specific. The role of context as a determinant of communicative language ability is stressed and integrative approaches to assessment, as against decontextualized approaches, are advocated. To
measure language proficiency adequately in different situations, test developers must consider where, when, how, with whom and why the language is to be used and on what topics, and with what effect (Weir, 1990). Language cannot be meaningful if it is devoid of context: linguistic, discoursal and sociocultural. For Oller (1979), the higher the level at which language is contextualized, the more effective language perception, processing and acquisition are likely to be. Heaton asserts the importance of reflecting the culture of a particular country while constructing communicative tests. “Not only should test content be totally relevant for a particular group of testees but the tasks set should relate to real life situations, usually specific to a particular country or culture” (p. 20). In second/foreign language testing, especially at higher educational levels, the tasks included in the test should be a fair reflection of the type of tasks the candidate will be required to perform as part of the course itself.

Communicative language tests give paramount importance to the authenticity of test texts and tasks. Weir (1990) points out that if a test of communicative ability includes inauthentic tasks, there is a danger that “the method employed could interfere with the measurement of the construct we are interested in”. Instead of measuring communicative language ability, Weir warns, we could “end up measuring ability to cope with a combination of these skills in specified contexts” (pp. 11-12). Face validity is also related to the authenticity of tasks. Tests which appear to test “real life skills” gain acceptance more than those “which use formats such as cloze, which are not seen outside the test itself” (Miyata-Boddy & Langham, p. 78). Using tasks which the test takers might recognize also makes it easier to explain the test to them. Morrow points out that in communicative language tests, reliability, which is very important, will be “subordinate to face validity” (p. 151).

Tests of communicative language ability should be as direct as possible. In other words, “the tasks candidates have to perform should involve realistic discourse processing” (Weir, 1990, p. 12). In direct testing, the best way to check if test takers know how to write compositions is to ask them to write compositions. If we want to know how well they pronounce a language, they should be asked to speak the language. This means that the test takers have to perform both receptively and productively in relevant contexts. Weir also advocates the use of genuine texts and recommends that care be taken with regard to task length and processing in real time. According to Hughes (2003, p. 17), direct testing has the following attractions,
First, provided that we are clear about just what abilities we want to assess, it is relatively straightforward to create the conditions which will elicit the behaviour on which to base our judgements. Secondly, at least in the case of the productive skills, the assessment and interpretation of students’ performance is also quite straightforward.

The focus of communicative tests is more on “the function of utterances rather than their grammatical structure” (Johnson, 2001, p. 298 [my emphasis]). Communicative language tests are designed to test “the use of English in what are typically seen as non-linguistic or real-world tasks” (Davies et al., 1999, p. 26).

Communicative tests tend to treat the four skills separately, providing information about a learner’s performance in each skill. This practice is based on a view of language referred to as the ‘divisible competence hypothesis’, which is quite contrary to Oller’s Unitary Competence Hypothesis (UCH) mentioned earlier. The divisible competence hypothesis suggests that language competence has separate components that need to be tested separately. Communicative testing results in an attempt to obtain different profiles of a learner’s performance in the language. The result is what is known as profile reporting, where learners are given a separate score for each of the four skills, together with descriptions of skill levels achieved. Good examples of such tests are two tests widely used to assess English for the purpose of academic study. These are the British IELTS Test (International English Language Testing System) and the American TOEFL (Test of English as a Foreign Language). In addition to an overall mark, scores for each skill are given, and these individual scores may play a part in determining whether or not a student is accepted to study at an institution.

Communicative tests should be assessed qualitatively rather than quantitatively (Morrow, 1979). Morrow argues that candidates’ responses to communicative test tasks are more than simply right or wrong, and that they should be assessed on the basis of how far they have moved toward an approximation of the native speakers system. In communicative tests, the candidate’s responses can be converted into a numerical score using rating scales which are used to show the learner’s level of performance in the different skills tested. Tests should reveal the quality of the candidates’ language performance according to carefully drawn-up and well-established criteria. Therefore, communicative tests should be criterion-referenced rather than norm-referenced. Each candidate’s performance is evaluated
according to their degree of success in performing the language tasks set rather than comparing their performance with the performance of other test takers. Heaton points out that qualitative judgments are also “superior to quantitative assessment from another point of view”. Qualitative judgments, when presented in the form of brief written descriptions, are of “considerable use in familiarising testees and their teachers (or sponsors) with much-needed guidance concerning performance and problem areas” (p.19).

Finally, tests of communicative spoken ability should have certain characteristics. They should reflect normal spoken discourse and the candidates should be given a chance to initiate. Weir (1990) argues that “the tasks should be conducted under normal time constraints and the elements of unpredictability in oral interaction must be recognized, for authentic communication may lead the participants in unforeseen directions” (p.12). Morrow (p. 149) also points out that the “processing of unpredictable data in real time is a vital aspect of using language”.

Communicative language testing, like any other form of testing, has its own problems. There are problems with the content of the tests and with their assessment. The remainder of this section outlines four such problems and some of the attempts that have been made by language testers and researchers to address these problems.

The first problem is the sampling and extrapolation of results. As mentioned above, tests take samples of language and these samples are used for the purpose of making inferences about the test takers’ ability outside the test situation (Weir, 1990). Communicative testers include contexts and tasks which reflect those which the candidates will encounter in real life. However, the specificity of the contexts reduces the generalizability of the information generated. One way of obtaining a fuller sample of the candidate’s language would be to include as many tasks in the test as possible. However, as Weir (1990) points out, this conflicts with the need for efficiency. Bachman (1991) states that in order to make inferences or predictions, test developers need to

…demonstrate two kinds of correspondences: (a) that the language abilities measured by our language tests correspond in specifiable ways to the language abilities involved in nontest language use, and (b) that the characteristics of the test tasks correspond to the features of a target language use context. (p. 681)
Davies et al. (1999) argues that the adequacy of a sample can be assessed through content validity. He points out that “our inadequate understanding of language needs, of language skills and of tasks and of their relative importance and frequency permits only an approximation to a representative sample” (p.174 [original emboldening]). Davies suggests the use of other validity evidence (e.g. predictive validity) in order to support the results of content validity.

The second problem is the claim that communicative methodology is seen as a model that encourages neglect of grammatical accuracy for the sake of fluency. As a reaction against this misinterpretation of the communicative methodology, researchers (e.g. Canale and Swain, 1980; Bachman, 1990; Bachman and Palmer, 1996) insist on the inclusion of grammatical knowledge as a part of language ability. Weir (1990) points out that “linguistic competence must be an essential part of the communicative competence” (p.13) and that empirical research must be conducted to indicate how linguistic competence and communicative competence relate to a communicative ability. This kind of empirical research is provided by Bachman (1990) and Bachman and Palmer (1996).

The third problem is that of assessment. Tests of communicative ability have been generally found to be less reliable than discrete point tests. As mentioned earlier, discrete point tests tend to have high reliability because they are assessed quantitatively while communicative tests are assessed qualitatively which inevitably throws some doubt on their reliability due to the involvement of subjective judgments. Weir (1990) points out that “the holistic and qualitative assessments of productive skills, and the implications of this for test reliability, need to be taken on board” (p. 13). This problem has been recognized and steps have been taken to address it. In the case of rater reliability, it is now possible, given sufficient training of examiners and raters on the procedures and the scales used, to obtain sufficiently high rater reliability in order for test results to be valuable (McNamara, 2000; Hughes, 2003).

The fourth and the last problem discussed in this section is that of the rating scales used in communicative tests. According to Brindley (1991), although the scales are widely accepted, it is very difficult to find any empirical basis for them. He also cites a number of other drawbacks to the use of rating scales in criterion referenced tests, including Bachman’s argument that unless there are upper and lower reference or end points, criterion-referencing is not possible. These end points exist only in theory because no test taker has either zero ability or the status of a perfect speaker.
To address this problem, several ways of producing criteria to be used in proficiency communicative testing have been suggested. One of these ways is to consult expert judges such as teachers, supervisors, etc. Another possible source of opinion is the so-called native speakers (Brindley, 1991), since these are the people the test takers will encounter when using the language. The third possible source is the test takers themselves. Self-assessment using learner-defined criteria is gaining ground in classroom-based assessment (Brindley, 1991).

In conclusion, it may be reiterated that testing has progressed a long way since the pre-scientific era, with its disregard for reliability, in favor of ‘fair’ testing. It has passed through eras when reliability and objective testing were dominant to the period today when testers are more interested in how candidates are able to use their knowledge of language in a communicative situation rather than demonstrate this knowledge in isolation.

**An Eclectic Approach**

In the previous section, a number of influential ‘schools’ of language testing have been examined. The later ‘schools’ are claimed to “supersede the earlier ones on the grounds of the advances they have made in understanding the essential nature of performance (language use)” (McNamara, 2000, p. 21). It is essential to judge the most recently marked approaches in the light of what has gone before. The suggestion made by many testers and researchers is that it is better to integrate and account for, rather than sweep away, past approaches. McNamara (2000) points out that the practices associated with earlier approaches to language testing have “far from disappeared, which is why appreciating earlier work is necessary for understanding the current rather eclectic scene in language testing” (p. 21). The issue is not the preference of one approach over another on grounds of ‘superiority’. The more central issue is the integration of the relevant features of the different approaches in order to enhance the validity and reliability of language tests of different kinds and for different purposes.

The idea of eclecticism is supported by many language testers and researchers (e.g. Davies, 1978a; Canale and Swain, 1980; Heaton, 1988; Davies, 2003). Davies (1978a, cited in Assad, 1984, pp. 48-49) was one of the proponents of an eclectic approach, which he describes as the most satisfactory view of language testing, and the most useful kinds of language tests. Davies argues that.
It is possible in any case that no test can be analytical or integrative alone, that on the one hand all language ‘bits’ can be (and may need to be) contextualised; and on the other, that all language tests and discourse can be comprehended more effectively by a parts analysis (p.149).

Canale and Swain (1980) justify the use of both discrete point tests and integrative tests to measure communicative competence. The authors argue that although integrative tests must be used in the measurement of communicative competence, they need to be supplemented with discrete point tests. Canale and Swain find discrete point tests to be “useful” in their proposed communicative approach because “such tests may be more effective than integrative tests in making the learner aware of and in assessing the learner’s control of the separate components and elements of communicative competence (p. 34-35).” Heaton argues that the four approaches to testing should not be regarded as being “strictly confined to certain periods in the development of language testing”. Nor should these approaches be viewed as “always mutually exclusive”. According to Heaton, a “useful test will generally incorporate features of several of these approaches” (p.15). The limitation of a test to one approach, and the neglect of other approaches, could bring about weaknesses to the test, no matter how attractive a particular approach may appear. For Davies (2003), one approach cannot properly exist in isolation from the other. A proficiency test, according to Davies, needs to be both analytical and integrative. Emphasis on any of these approaches, at the expense of the other, “distorts the view of language provided by the test: too structural and uncontextualized at one end; too local and ungeneralizable at the other” (p.357).

This new awareness brought about the revision of large-scale tests that had originally been developed with audiolingual/structuralist principles such as the TOEFL test. In fact, the TOEFL is a good example of this journey towards eclecticism. The TOEFL test measures the ability of non-native speakers of English to use and understand North American English as it is spoken, written and heard in college and university settings. Most people who take the TOEFL test are planning to study at colleges and universities where the instruction is in English. In addition, many government agencies, scholarship programs and licensing/certification agencies use TOEFL scores to evaluate English proficiency. Daerr (n.d.) reports that almost 800,000 people take the TOEFL test every year worldwide.
Edwards (2006) observes that the TOEFL test, like any other standardized test, will continue to be “a topic of heated debate among educators throughout the United States and in many parts of the world” (p. 9). The consequences of this educational debate could have a profound effect on the development and use of the test around the world. The TOEFL test witnessed many changes since its first administration in 1963-64. These changes will be discussed below as part of the general discussion of the journey of the TOEFL test towards eclecticism.

According to the Educational Testing Service (2003), The National Council on the testing of English as a foreign language, which was formed in 1962, supported the development of the TOEFL test for use starting in 1963-64. The TOEFL program was administered, at first, by the Modern Language Association. In 1965, the College Board and Educational Testing Service (ETS) assumed joint responsibility for the program. Later on, ETS became responsible for administering the TOEFL program with oversight from the TOEFL Board. ETS has been responsible for the administration of the test since then. According to the Educational Testing Service (2003), the test originally contained five sections. As a result of extensive research, a three-section test was developed and introduced in 1976. These three sections are: listening comprehension, grammatical structure, and reading comprehension (p. 43). In July 1995, the test item format was somewhat modified but within the same three-section structure. During this stage, the TOEFL suffered from some “very serious drawbacks”. The most important, and most relevant, drawbacks are two. Firstly, the TOEFL tested neither the ability to speak English nor the ability to write it, despite the fact that these two skills are essential for anyone embarking on a course of higher studies. Traynor (1985) remarks that since the speaking and writing abilities are not tested, they are often neglected in teaching. These students become “so mesmerized by the mystical number 500 that they do not want ‘to waste time’ on skills which do not feature in the examination” (p. 44). The second drawback is related to the theory of communicative competence. According to Davies (2003), the TOEFL “continued for 40 years more or less unchanged, having no truck with the communicative revolution” (p.360).

These drawbacks have been addressed in recent years when various constituencies called for a new TOEFL test. According to ETS (2003), the new TOEFL test would (a) be more reflective of communicative competence models, (b) include more constructed-response tasks that direct measures of writing and speaking.
(c) contain tasks that integrate the language modalities tested, and (d) provide more information than the paper-based TOEFL test about the ability of international students to use English in an academic environment. Accordingly, the TOEFL Board initiated a broad effort under which language testing would evolve in the twenty first century. According to Bachman (2003), the introduction of the computer-based TOEFL test (CBT) in 1998 was the first step in this broad test improvement effort. The use of computer-based tests utilizes multimedia capabilities “to deliver a wide range of language test tasks that are potentially more authentic and interactive than those presented in a paper-and-pencil format, and that promise more efficient administration, scoring, and reporting of results” (p. 423). In July 1998, ETS introduced the computer-based TOEFL test (CBT) in many areas of the world. This move was the first critical step toward the long term goal of enhancing assessments by using electronic technology to test more complex skills. The introduction of the new TOEFL (CBT) program is meant to provide more extensive information than it has in the past about the candidates’ English proficiency.

The TOEFL (CBT) is divided into four sections, measuring language proficiency in listening, structure (grammar), reading and writing. A writing section (essay-writing) has been added as part of each test administration. This move took place as a response to institutions’ requests to include a productive measure of writing. The addition was one step toward a more communicative test. New types of questions were added to the listening and Reading sections, too. Visuals were also added to the listening section, providing a significant enhancement to that portion of the test. The Listening and Structure sections are computer-adaptive, meaning that the test is tailored to each test taker’s performance level. That is, the computer is programmed to continuously find questions of an appropriate difficulty for test takers of all performance levels.

Until 2005, the TOEFL test did not test the ability to speak English. In most environments, the ability to speak intelligibly and without undue delay is vital. As a result, many universities request incoming teaching assistants who are not native English speakers to take additional tests (such as the Test of Spoken English or some university-administered tests) to ensure their ability to communicate with their students. In September 2005, the Internet-based TOEFL test (iBT) was introduced to address this issue, and it soon found its way into most testing institutions worldwide.
Since then, the TOEFL (iBT) has progressively replaced both the computer-based (CBT) and the paper-based (PBT) tests.

The TOEFL (iBT) was first introduced in countries like the United States, Canada, France, Germany, and Italy in 2005 and was later adopted in the rest of the world in 2006. The TOEFL (iBT) consists of four sections each measuring mainly one of the basic language skills (although some tasks may require multiple skills) and focusing on language used in an academic, higher-education environment. The content of the test is authentic and the tasks directly assess the test takers’ ability to use English to communicate in academic situations. These four sections are Reading, Listening, Speaking and Writing. The most innovative features of the TOEFL (iBT) are the addition of a speaking measure and a revised writing measure consisting of two writing tasks. The speaking and writing measures include both independent and integrated tasks. Independent tasks require test takers to talk or write about their personal opinions or experiences while the integrated tasks require the use of more than one language skill (e.g. the second task in the writing subsection of the test). The rationale for including tasks that require speaking, writing and the integration of the four skills on TOEFL (iBT) is that these tasks directly measure the ability of the test takers to communicate in English in academic settings. In areas where the iBT and the CBT are not available, a paper-based (PBT) is used instead. The TOEFL (PBT) has three sections plus a 30-minute writing test, the Test of Written English (TWE).

The three different versions of the TOEFL test are used in different countries. But all versions are equally valid and reliable and are developed and administered based on the stringent ETS standards for quality and fairness (ETS, 2006).

Methods of Statistical Analysis

Item Analysis (IA)

Item analysis is an important statistical procedure in the analysis of language tests. It involves the “careful analysis of score patterns on each of the test items” and shows “how well each item is working, that is, the contribution it is making to the overall picture of candidates’ ability emerging from the test” (McNamara, 2000, p. 60). Item analysis is a useful analytical tool for test writers and test developers because it draws their attention to “items which contain problems or ambiguities in their construction which escaped notice during the construction of the test” (Baker, 1989, p. 53). Brown (2003) argues that item analysis is necessary because it helps test
writers make their test “shorter” and “more effective” by selecting only the “most effective items” and getting rid of the “ineffective items” (p. 16). This section discusses the procedures of IA that can be used to better understand the characteristics of individual test tasks or items. These procedures include classical item analysis (CIA) and item response theory (IRT).

Classical item analysis (CIA), called ‘classical’ because it is based on classical test theory (CTT), consists of calculating descriptive statistics for individual items and can serve a range of important purposes. According to Bachman (2004, p. 121), CIA can provide three types of feedback. The first is “diagnostic” feedback to test takers which is meant to inform them of their strengths and weaknesses on the individual test tasks. The second type of feedback is to teachers and course developers, which is meant to help them improve their teaching quality. The third type of feedback is the most relevant to this study. It is feedback to test writers and test developers which should help them “improve the usefulness of the test” and enable the test developer to

1 Control the characteristics of the total score distribution, specifically the level of difficulty of the test and the dispersion of test scores;
2 Increase the internal consistency reliability of the test; and
3 Diagnose why items fail to function appropriately.

Item response theory (IRT), the second procedure of IA, is a newer technique of measurement based on “the theorem that performance on test item is a reflection both of the difficulty of the item and of the ability of the test-taker” (Davies, 2003, p. 363). There are three IRT models, each of which is based on different formulas and assumptions regarding item properties, and each requires different numbers of test takers and test items for valid estimations (McNamara, 1996). These three main IRT models are: a 1-parameter IRT model, often referred to as the ‘Rasch’ model, which includes only a difficulty parameter (the ‘b’-parameter); a 2-parameter IRT model which includes a difficulty parameter and a discrimination parameter (the ‘a’-parameter); a 3-parameter IRT model which includes, in addition to parameters for difficulty and discrimination, a “pseudo chance”, or guessing parameter (the ‘c’-parameter) which is an estimate of the probability that low-ability test takers will respond to the item correctly (Bachman, 2004, pp.141-142).

The results of CIA and IRT can be used to improve the usefulness of the tests by helping test writers identify items that fail to function properly, and to select the
items which have the characteristics that are wanted. Both procedures of IA can be used with a wide variety of task type, from multiple-choice to short answers and cloze.

IRT has some advantages over CIA. One of the most important of these advantages is that the information provided by IRT is ‘sample free’, i.e. not confined to a particular sample. The analysis therefore applies to test takers who have not taken the test and to items that were not included in the test, while the information provided by CIA is always related to a sample under test (Davies, 2003). Despite this advantage of IRT, this study will use CIA since the generalizations made by the statistical analysis (chapter three) are limited to the 2005 admission test at Taiz University and to one particular sample of test takers. Furthermore, CIA still exerts a great influence on testing and measurement research. Bachman (2004) and Davies (2003) report that many testers continue to provide evidence of data quality based on the classical test theory (CTT). According to Fujita (2005), data from language tests are “commonly analyzed with this model in Japan because the statistical analyses are simple to carry out and the results are relatively easy to understand” (p.10). Furthermore, available computer software such as Microsoft Excel and SPSS provide data analysis based on CTT. Bachman (2004) also points out that in situations where a test is to be developed for use with a particular group of test takers, “classical IA provides a perfectly adequate tool for helping ensure that this test will be useful for its intended purpose” (p.139).

In the classical item analysis of norm-referenced tests, which admission tests at Taiz University are an example of, two item statistics are typically used, namely, item difficulty (or item facility (IF)) indices and item discrimination (ID) indices. The estimation of these two item analysis indices is important to know how difficult the individual items in the test are, since this will affect the difficulty of the test as a whole. It is also important to know how well the items discriminate between different groups of test takers because this will directly affect the distribution of the test scores and will also affect the overall reliability or consistency of the test scores.

**Item facility index (IF).** IF index is the proportion of students who answered a particular item correctly. It is expressed on a scale from 0 (no one answered the item correctly) to 1 (everyone answered it correctly). For example, if 90 out of 100 test takers answered a given item correctly, the proportion would be $90/100 = .90$. An IF of .90 means that 90% of the test takers answered the item correctly. This also means
that the item is considered a very easy item. On the other hand, an IF of .27 would mean that 27% of the test takers answered this item correctly, which means that this is fairly difficult item since 73% of the test takers missed it. “The closer the index is to 100% or 0% the less differential information it can provide about candidates” (Davies et al, 1999, p. 95). In other words, if all the test takers, or none of them, get an item right, no differential, or useful, information is gained about the test takers’ ability. Therefore, items which are very easy or very difficult should be removed because they do not contribute to the discriminability of the test. In the words of McNamara (2000),

If the items are too easy, then people with differing levels of ability or knowledge will all get them right, and the differences in ability or knowledge will not be revealed by the item. Similarly, if the items are too hard, then able and less able candidates alike will get them wrong, and the item won’t help us in distinguishing between them. (pp. 60-61)

An ideal IF value in a norm-referenced test (NRT) is 50% which means that the item is ‘well-centered’ (Brown, 2003), because 50 percent of the candidates would have answered correctly and the other 50 would have answered incorrectly. However, in actual practice, items rarely have an IF of exactly 50 % (Brown, 1989; Brown, 2003; McNamara, 2000). IF values, like any other statistical measures, have no absolute value but vary from one testing situation to another (Baker, 1989; Heaton, 1988). A range of IF from .30 to .70 is usually acceptable for NRT purposes (Brown, 2003; McNamara, 2000). Madsen (1983) considers an item too difficult if fewer than 30 percent get it right and explains his view by noting that “a person might get 25 percent on a four option test just by guessing” (p. 182). IF Index for multiple-choice questions is set higher than that of short answer questions because in multiple-choice tests test takers will guess some of the answers. Items which are too easy or too difficult should be either revised or eliminated. However, few easy items can be placed at the beginning of the test to give the test takers a chance to get over their nerves. Few difficult items can also be kept near the end of the test in order to distinguish between the higher and the lower achieving test takers and sometimes between the high and the higher ability level test takers.

**Item discrimination index (ID).** Item discrimination (ID) is defined by Bachman (2004) as “the extent to which the item discriminates between different groups of test takers” (p. 122). Bachman also remarks that for NRT, item
discrimination refers to “how well the item discriminates between individuals who scored high on the test as a whole and individuals who scored low on the test as a whole” (pp. 122-123). If more high-level than low-level test takers answered an item correctly, the item is distinguishing well between strong and weak test takers and is said to be a good discriminator. If the numbers are the same or if more low-level test takers answered an item correctly, then the item may need to be revised or changed. ID index can range from +1.0 (an item which discriminates perfectly, i.e. all high-level test takers responded correctly and no low-level test takers did) through 0.0 (an item which does not discriminate at all) to –1.0 (an item which discriminates in the wrong way, i.e. all low-level test takers responded correctly and no high-level test takers did) (Baker, 1989; Heaton, 1988; Brown, 1989; Hughes, 2003). ID values also vary from one testing situation to another. However, ID values are expected to exceed 0.30. Items with ID values less than 0.30 fail to discriminate strongly enough between candidates (Baker, 1989). Such items are called into question and they should be either reviewed or eliminated. McNamara (2000) points out that the usual method used to calculate ID involves “comparing performance on each item by different groups of test takers: those who have done well on the test overall, and those who have done relatively poorly” (p. 61). If this does not happen and the numbers of correct answers in the lower group exceeds that in the upper group (a negative ID value), it is inadvisable to use that item again (Heaton, 1988; Baker, 1989; McNamara, 2000).

Furthermore, if most of the test items have problems of discrimination, the information obtained from the test will be confusing. For example, some items would show that certain test takers are better while others would show that other test takers are better and therefore no clear picture of the test takers’ abilities is obtained from the test. The scores of such test items, in McNamara’s words, become “misleading, and not reliable indicators of the underlying abilities of the candidates” and the tests themselves need “considerable revision” (McNamara, 2000, p. 61).

While conducting an item analysis of a test, once the items that fall within the .30 to .70 values of IF are identified, the items among them that have the highest IDs should be further selected to be included in the revised test. “This process would help the test designer to keep only those items that are well centered and discriminate well between the high and the low scoring students” (Brown, 2003, p. 18).
It is important to mention here that CIA methods are not applicable to tasks whose marks are attributed according to descriptive rating scales such as those that can be found in the essay questions (Brown, 2003). This is because with items that can be assigned a score 1 or 0 for correct and incorrect and with short answer questions, the presence or absence of an item does not affect the performance of the other items. Each item behaves quite independently of the other items (Baker, 1989) and therefore the elimination of a problematic item from the test will not affect the behavior of the other items on the test. In the case of essay questions or subjective tests, however, there is a piece of writing that cannot be subdivided into items; there are no items to be analyzed or eliminated. Item analysis is therefore applicable only to objective tests.

The Present Study

In the construction of the proposed test, particularly in the selection of the test format, the present study will adopt an eclectic approach similar to the one discussed in the first section of the review. The content of the test will be similar to the content of the TOEFL (iBT) in the eclecticism that characterizes the testing technique and in the integration that characterizes the test tasks. The proposed test will, however, be a paper-based test. Computer-based and internet-based tests make use of computer technology and audio-visual aids which are not available in the admission testing environment where the proposed test will be administered. In the development of the proposed test, the researcher will follow the procedures of test construction proposed by Bachman and Palmer (1996).

The data obtained from the 2005 admission test, on the other hand, will be statistically analyzed using classical item analysis (CIA) since low-cost computer software used in such procedures (Microsoft Excel and SPSS) is readily available. Other reasons for the preference of this statistical procedure have been given in the section on methods of statistical analysis. The aim of the statistical analysis of the ‘multi-item’ questions is twofold: a) to identify items whose IF values fall within, or closely approximate to, the range .30-.70 and whose ID values exceed, or closely approximate to, .30; and b) to eliminate the other items which do not fall within these acceptable ranges of IFs and IDs. A reliability estimation of 0.9 or better is required in such norm-referenced tests. Alpha reliability estimation of less than 0.70 will be considered unacceptable. McNamara (2000) points out that a reliability of 0.90 means...
that “scores on the test are providing about 80% reliable information on candidates’ abilities, with about 20% attributable to randomness or error” (p. 62).