Language Proficiency Testing: A Comparative Analysis of IELTS and TOEFL

Ardeshir Geranpayeh

PhD – The University of Edinburgh – 2000

Abstract

There is a general belief that British and North American EFL proficiency tests represent radically different approaches to language test development. The North American tradition in language testing is heavily based on psychometric properties of tests such as reliability, and concurrent and predictive validity, whereas the British tradition is more focused on the specification of test content and expert judgement. Language proficiency tests, in either the American or British tradition, are designed to serve different purposes, so they may not be comparable in terms of defined purposes. Nevertheless, the term '*language proficiency*', no matter how it is defined, implies that we are referring to a monolithic concept. In the real world, test results are often used for screening purposes; the candidates' *ability* to cope with the future language medium is predicted by the proficiency criterion. If it is the case that language proficiency tests are used for similar purposes, i.e., measuring the general language ability of the candidates, comparability of such tests is a legitimate matter.

This study compares two English language proficiency test batteries: TOEFL and IELTS. The main objectives of the research were to investigate the extent to which TOEFL and IELTS are comparable in terms of: a) the operational definitions of language proficiency on which the two tests are based, b) the degree to which the two tests provide similar information concerning the abilities of the testees. Analysis of test content suggests that both batteries are based on the notion that proficiency is divisible by skill (e.g. reading) and element of language (e.g. syntax), thus we have tests of reading, writing, listening, speaking, as well as tests of grammar and vocabulary. However, the tests *differ* in their representation of the scope of skills and elements of language proficiency. The analysis also shows that the TOEFL differs from the IELTS in its method of testing. Despite these differences in test methods and scope, to a great extent both tests measure a common aspect of the subjects' language ability, therefore their internal structures are unifactorial. A *g-factor* (general language proficiency) comprises much of the total variance in both tests. Additional information provided by each test regarding the competence in reading, listening, writing, and knowledge of vocabulary and grammar, does not seem to contribute much to the total variance. A content analysis of the two tests indicates that, in fact, there are more similarities between the reading and listening comprehension sections of the two batteries than differences. This is supported by the factor analysis of the test scores.

Declaration

I declare that this thesis has been composed by myself and the work represented here is entirely my own.

Signed

Ardeshir Geranpayeh
November 2000

Acknowledgements

Very many people have helped me in the process that made this project both possible and enjoyable. I should first like to thank the late Kambeez Zolfaghari who initially asked me to do a comparative study on the IELTS - TOEFL scores 9 years ago. He had been instrumental in co-ordinating the administration of the tests in Iran, without this support the research would never have gone beyond year two.

I would like to thank Alan Davies, whose discussions in proficiency testing convinced me that the comparability study was worth pursuing more seriously. I am indebted to him for his inspiring supervision in the first 18 months of the project in Edinburgh and for agreeing to continue his supervision from the remoteness of Melbourne in the next 3 years, and especially for his support in the final year in Edinburgh. I would also like to thank Gibson Ferguson who agreed to take on a supervisory role from year three and who has given me invaluable help with the analysis of the rating instruments; his meticulous comments on various aspects of the work have made unique contributions to my research.

Sincere thanks go to Dan Robertson who, in addition to rating various tasks of the content analysis instruments, and writing computer programmes for text analysis, has been a great inspiration for my intellectual development. I have benefited tremendously from his criticisms on different aspects of my work. I am also grateful to the following people for helping me in various stages of the project; to Lyle Bachman for allowing me to use his rating instruments and for his comments on my earlier design of content analysis instrument; to John B. Carroll for providing me with the *Exploratory Factor Analysis* programme and for guiding me in the right path to understanding the underlying assumptions of such analysis; to Grant Henning for his comments on score comparability; to Liz Hamp-Lyons for her comments on the marking guidelines of the IELTS writing; to Joan Cutting for giving her valuable time rating the propositional content facets; to Peter Howell for proof reading the final draft; to my writing markers: Lee Wennerberg, Eleni Theodoridu, Ramon Palencia, Michael Jenkins, and Isabelle Gall; and finally to Hussein Farhady for introducing me to language testing.

Many thanks go to all the students who took the tests and to the teachers, tutors, and test centres who allowed their students to participate and themselves took an interest

in the project. In particular send my sincere thanks to Ladan Twiserkani who co-ordinated the administration of the tests in Kanoon centre and later collected the language background information from the student files there. In addition I should like to thank the following organisations: the Ministry of Culture and Higher Education in Tehran for sponsoring the cost of the administration of the tests and providing the *TOEFL Sample*; the EFL section of the University of Cambridge Local Examinations Syndicate for providing the *IELTS Specimen Module*; and the Institute for Applied Language Studies in Edinburgh for providing the *EPTB* test.

I would like to thank all my friends and colleagues, Szilvia Papp, Parveen Sandhu, Charlie Kemp, John Cleary, Seidu Alhassan, and Peter Howell, whose sense of humour, empathy, endless discussions, mutual help, understanding, and exchange of information made my stay in the basement of Applied Linguistics bearable.

Above of all, of course, I should like to thank my wife Shahnaz, and daughters Sarvnaz and Masumeh, not only for putting up with so much but also for their key role in cross-checking the test scores, entering them in relevant databases, and organising some 5,000 pages of test papers.

All those mentioned above have helped give this research whatever merit it may have had, but I am alone responsible for all faults.

Table of Contents

Chapter One

Chapter Two

Chapter Three

Chapter Four

Chapter Five

Chapter Six

List of Tables

Chapter Two

Chapter Three

Chapter Four

Chapter Five

List of Figures

Chapter Two

Chapter Four

Chapter Five

Chapter One

Introduction

1. Introduction

Hundred of thousands of individuals throughout the world take various English language proficiency tests each year to demonstrate their proficiency in English as a foreign language. The scores of such tests are used by different institutions for screening their candidates for a number of different purposes, such as offering employment, advancement in a career, or admission to an educational programme. In most cases the selection of candidates, to a large extent, is affected by the results of these tests. Thus, any variability in the scores of such tests might concern job opportunities or perhaps life chances of individuals; this makes the interpretation of the scores an extremely heavy responsibility.

Test scores are related to various aspects of proficiency demonstrating the candidates' language ability in different skills, e.g. writing, reading, speaking, or listening in a given language. During the last three decades numerous methods and test batteries have been developed to measure different aspects of language proficiency of non-native speakers. Depending on the level of tests' population and the purposes to which the test scores are put, the tests presumably differ from one another. Differences in methods and purposes are considered as evidence of their non-comparability. Yet, where statistical evidence is concerned, the tests are validated against one another and their results are compared to show the degree of similarity between the traits they are attempting to measure.

Academic institutions, nonetheless, are only interested in a clear cut-off point score of say 600 on TOEFL or 6.5 on IELTS as the evidence of their non-native speakers' proficiency in English to pursue a course of study. However, this raises the question of what it means to have a particular score in a test such as IELTS and whether the score can be related to an equivalent TOEFL score. In other words, we are asking whether scores in different batteries can be equated to one another.

1.1 Aims

There is a general belief that British and North American EFL proficiency tests represent radically different approaches to language test development. The North

American tradition in language testing is heavily based on psychometric properties of tests such as reliability, concurrent and predictive validity, whereas the British tradition is more focused on the specification of test content and expert judgement. Language proficiency tests, in both the American or British tradition, are designed to serve different purposes, so they may not be comparable in terms of defined purposes. Nevertheless, the term '*language proficiency*', no matter how it is defined, implies that we are referring to a single concept. In the real world, test results are often used for screening purposes; the candidates' *ability* to cope with the future language medium is predicted by the proficiency criterion. If it is the case that language proficiency tests are used for similar purposes, i.e., measuring the general language ability of the candidates, the comparability of such tests is a legitimate area for investigation.

Bachman and his colleagues, in an attempt to examine the differences / similarities between the American, and British approaches to test development, embarked on a large-scale project to compare Cambridge and ETS tests, the result of which has been published in Bachman et al. (1995). However, this study has been criticised by Alderson (1989) and Davies (1989) for the wrong choice of tests for comparison. They both suggested that it was far better to compare TOEFL with IELTS for a meaningful comparison, and this was one of the reasons for initiating the present research, the main objective of which is to compare two English language proficiency tests. The main objectives of the research were to investigate the extent to which TOEFL and IELTS are comparable in terms of:

a) the operational definitions of language proficiency on which the two tests are based.
b) the degree to which the two tests provide similar information concerning the abilities of the testees.

1.2 Scope

This study is limited in scope in the following ways.

- Only comparable sections of the batteries are analysed in the study. IELTS Speaking section will not be used in the research as it has no comparable section in TOEFL. The Listening, Structure and Written Expression, and the Reading

sections of TOEFL and the Listening, Reading, and Writing sections of IELTS are used in the analysis. The selection of the Writing section of the IELTS is to see if it has a relationship with the TOEFL Written Expression.

- The test takers in the study are selected from only one particular language background, i.e., Farsi speakers. This is due to the limitation of the research resources. However, every attempt is made to make sure that the range of abilities of the test takers matches with that of those reported in the literature for TOEFL and IELTS test takers.
- Only one sample of each test is used in the course of the study. The TOEFL sample is the one provided by the Ministry of Culture and Higher Education (MCHE) in Iran, based on the retired versions of TOEFL prior to 1993. The IELTS sample is the IELTS Specimen Module C provided by the University of Cambridge Local Examination Syndicate (UCLES) in 1992. This is one of the tests that has been used in the IELTS validational study reported in Clapham (1996).
- The comparability focuses on the analysis of test contents and test performance. The effect of language proficiency preparation courses on test performance will also be studied but is limited to the impact of TOEFL courses.

1.3 Structure

The thesis comprises six chapters. Chapter one is the introduction to the thesis and explains the aims, scope, and the structure of the study. Chapter two reviews the literature on language proficiency testing. It concerns the followings.

- The concept of language proficiency
- Differences in American and British traditions of proficiency testing
- Reviews of three language proficiency tests: TOEFL, EPTB, IELTS
- A review of comparability studies
- The relationship between validity and reliability
- The impact of test methods on test performance
- A review of a communicative framework for comparing language proficiency tests

Chapter three sets out the design of the study and explains the rationale, objectives, research questions, the selection of subjects, test administrations, marking of the tests, and the procedures of investigation.

Chapter four reports the results of content analysis of IELTS and TOEFL, and Chapter five reports the results of the analysis of test performance. The analysis includes the results of the exploratory factor analysis of the batteries, item difficulty, and the impact of test preparation on test performance.

Finally, Chapter six discusses the main findings of the research, and their implications for the questions of the research are put forward.

Chapter Two

Review of Language Proficiency Testing

2. Review Of Language Proficiency Testing

This chapter reviews the literature in language proficiency testing in the past forty years and gives an overview account of language proficiency test comparability studies. Given the shortcomings of the previous research, we will propose a communicative framework for the comparison of language proficiency tests.

2.1 On Language Proficiency

Definition

The will to define language proficiency goes back at least as far as Fries (1945), and the flourishing era of structuralist approaches to language teaching. Fries speaks of language proficiency in a reference to the goals of language learning: "*if an adult is to gain a satisfactory proficiency in a foreign language*" (1945, p. 5). Fries uses a number of terms, which are to be important in the discussions by later writers. Terms such as *mastery, competence, use* and *control* were all associated with language proficiency from Fries's time:

> "*Progress toward the satisfactory mastery of a foreign language... a satisfactory competence in the new language... a satisfying use of a foreign language ... satisfactory control of language material.*" (Fries, 1945, pp. 5-7)

Lado (1961), Fries's student, follows a more atomistic approach to proficiency and breaks it into individual skills and components on the assumption that the sum of the parts would be equal to the whole, a discussion we will shortly come to. Both Fries and Lado viewed proficiency in terms of the goal of learning. Hence, all the factors presuming to have affected learning, i.e., L1 influence, were considered to be influential on language proficiency. Assessment, according to Lado, was then based on what was known of courses students had followed prior to the test. In short, Fries and Lado had a functional view of proficiency, putting emphasis on the goals and outcomes of language learning.

Lado's contemporary influential colleague, Carroll, however, does not share this view. Although Carroll (1961[1972]), like Lado, followed the same principle of structuralist / psychometric approaches to testing and speaks of the need to specify *"kinds and levels of English language proficiencies"* (1961[1972: 315]), his approach to proficiency testing clearly differs from that of Lado's in the sense that Carroll disregards the effect(s) of what was known of the candidates' first language or learning history on test content. As Carroll points out, the point is, *"how well the examinee is functioning in the target language regardless of what his native language happens to be" (Ibid: 319)*. Carroll, furthermore, argues that proficiency is related to the future success of the learners in various learning tasks.

> *"An ideal English language proficiency test should make it possible to differentiate, to the greatest possible extent, levels of performance in those dimensions of performance which are relevant to the kinds of situations in which the examinees will find themselves after being selected on the basis of the test. The validity of the test can be established not solely on the basis of whether it appears to involve a good sample of the English language but more on the basis of whether it predicts success in the learning tasks and social situations to which the examinees will be exposed."* (Carroll, 1961[1972], p.319)

The implicit distinction between achievement and proficiency tests in Carroll's work was made explicit by Davies (1968, pp.6-7). According to Davies, an achievement test *"cannot make predictions as to pupils' future performance"* (Davies, 1977, p.46). Davies discusses proficiency in terms of dealing with the *future needs or control purposes.* He argues that " *proficiency in a language implies adequate control over language skills for an extra linguistic purpose"* (Ibid.). It appears that proficiency is equated here with *control over language skills.* Nevertheless, in Davies's view, predictability appears to be an important factor of proficiency testing. As Davies elsewhere (1990) mentions, the proficiency test:

> *"Establishes generalisations on the basis of typical syllabuses leading to entry and is more directly related to what it attempts to predict, namely, performance in the language under test on some future activity."* (Davies, 1990, pp.20-21)

Taylor (1988) equates proficiency with the *ability to make use of competence.* According to him, proficiency is a dynamic concept, having to do with process and

function. Performance then becomes "*what is done when proficiency is put to use*" (Taylor, 1988, p.166).

There is a view that does not consider language proficiency as a "*global factor*" (Richards, 1985, p.4), while according to others (Alderson, Krahnke & Stansfield, 1987, p. IV) " *proficiency is a global construct*'. Bachman (1990) in an attempt to define the term in the context of language testing maintains that it has been used to refer "*in general to knowledge, competence, or ability in the use of language, irrespective of how, where, or under what conditions it has been acquired*" (Bachman, 1990, p.16). Knowledge and ability are equated in this definition in an abstract sense of the use of language.

Finally, to summarise the discussion on language proficiency, it is worth referring to Davies, Brown, Elder, Hill, Lumley & McNamara (1999), who consider three main uses of the term in the past three decades:

a) a general type of *knowledge* of or *competence* in the use of a language, regardless of how, where or under what conditions it has been acquired.
b) *ability* to do something specific in the language.
c) *performance* as measured by a particular testing procedure.

Davies et al. do not equate knowledge and ability in their definition of proficiency. For them, knowledge is relevant in the *use of a language,* while ability is to do with *something specific in the language*. What they have additional in their definition is the inclusion of performance measured by language tests which is a more operational definition of the term as opposed to the common theoretical ones. In this regard, they even refer to *levels* of performance (superior, intermediate, novice) on performance assessments such as FSI scales. Having reviewed the literature, it can be observed that in the early 1960s and 70s proficiency was mainly used under the label of general language proficiency: a single faceted notion, whereas since 1980s language proficiency has been associated with communicative competence: a multifaceted notion. This is the discussion we will come to next.

Language Proficiency Hypotheses

It can be inferred from the above discussions that the term *language proficiency* has acquired a variety of meanings in different contexts and for various purposes. Prior to

any further discussion, it seems warranted to explain the underlying theories of these different, sometimes even contradictory, views on language proficiency. Thirty years ago Spolsky addressed us with the following question:

> *"Fundamental to the preparation of valid tests of language proficiency is a theoretical question: What does it mean to know a language?"* (Spolsky, *Sigurd, Sato, Walker & Aterburn*, 1968, p.79)

It is not possible to develop valid language tests without a method of defining what it means to know a language. The reason lies in the fact that until we have decided what we are measuring we cannot claim to have measured it.

2.1.1 Divisible Hypothesis

The question of what it means to know a language has many possible answers. There are basically three responses. One is to assume that knowledge is broken down into the individual structures- the rules and the lexical items- that make up the grammar and the lexicon of a language, so that *knowledge of these items and rules* is what needs to be measured. This is often termed in testing as *the divisible hypothesis* or *divisible competence hypothesis* (Vollmer 1981). The tenets of the hypothesis are based on psychometric theory and structural linguistics.

Psychometric theory has two main implications. Firstly, most questions are of the closed type in the sense that the testee has to choose between a limited number of responses, e.g. multiple-choice items. Secondly, a fairly elaborate system of statistical procedures has been evolved for developing and evaluating this kind of test. The first characteristic promises objective scoring and the second offers a ready-made set of methods and criteria for analysing and evaluating language tests.

Psychometric theory provides the tools for producing and developing tests. What is also needed is a basis for the content of the tests that are produced. Since the theory was first used in language testing in the late 50s, it was natural for language testers to have taken advantage of the same framework that was being used to devise the teaching programmes, namely, language description based on the work of the American structuralist linguists. The analysis used involved breaking the language system down into small bits, and then describing the ways in which these bits could be

put back together again to make stretches of speech. Structuralist description provided the analysis, which derives the criterion proficiency from the language.

The classical divisible hypothesis can be seen in Lado's book *Language Testing* (Lado, 1961). Lado, following the structuralist tradition, divides language into elements at four levels: Phonological, Lexical, Syntactic and Cultural. He, furthermore, categorises the bits of language that are to be taught or tested and defines the various ways in which the bits could be mobilised in actual language use. The result would be the "four skills': speaking, listening, reading, writing. Consequently, the criterion proficiency is composed of *elements* of language as well as *skills* that mobilise them in actual use.

Table 2.1: Components of Language Proficiency in Divisible Hypothesis

ELEMENTS \ SKILLS	Speaking	Listening	Reading	Writing
Phonology				
Syntax				
Lexis				
Culture				

To sum up, the divisible hypothesis denies a unitary hypothesis of proficiency and breaks down proficiency into linguistic elements and language skills, hoping the sum of the parts will be an indicative of the whole, i.e. the language proficiency. A typical example of a proficiency test of this kind is the Test Of English as a Foreign Language (TOEFL), which has been conducted since 1964. The test comprises three sections: Listening Comprehension, the Structure and Written expression, and Reading Comprehension and Vocabulary. The test follows a pure multiple-choice four-option format. A parallel British model is the Davies test of the receptive skills. The following table copied from Davies (1967) illustrates the framework of his proficiency test.

Table 2.2: Composition of Davies Test Battery

Test	Items	Content	Skill	Aspect
1	65	Phonemic Discrimination: words in isolation	List.	Ling.
2	25	Phonemic Discrimination: words in sentences	List.	Ling.
3	50	Intonation	List.	Ling.
4a	8	List. Comprehension: general	List.	Work sample
4b	5	List. Comprehension: specialised science	List.	Work sample
4c	5	List. Comprehension: Arts	List.	Work sample
5	196	Reading speed	Read.	Work sample
6a	49	Reading Comp.: general	Read.	Work sample
6b	50	Reading Comp.: specialised Arts	Read.	Work sample
6c	50	Reading Comp.: specialised science	Read.	Work sample
7	50	Grammar	Read.	Ling.

It is evident from the above table that the extreme '*discrete structure-point*' approach was not implemented in this test of language proficiency. Rather, following Carroll's (1961[1972, p. 318]) recommendation, it involves a combination of discrete-point tests, i.e. test1, and some integrative testing, i.e. tests 4a, 6a. Carroll suggested that the two approaches complement one another in language proficiency tests.

> *"I do not think, however, that language testing (or the specification of language proficiency) is complete without the use ...an approach requiring an integrated, facile performance on the part of the examinee.... I recommend tests in which there is less attention paid to specific structure points or lexicon than to the total communicative effect of an utterance."* (Carroll, 1961[1972, p. 318])

Carroll's remarks concerning the complementarity of integrative tests were sometimes misunderstood by some to mean that he was criticising or opposing the discrete-point approach. But that is not the case. As Carroll later (1986) pointed out:

> *"the discrete-point approach, if the 'points' are adequately sampled, is often a good way to measure the language competence on which language performance is based, but it needs to be supplemented by measures of integrative performance*".(1986, p. 124)

Despite the wide use of the divisible hypothesis in language tests, there has been sound criticism against its use for clarifying the nature of language proficiency. Spolsky, for example, has asserted that:

> *"it is clear that the definition of language proficiency will not come from psychometric theory, which is concerned with the measurement once you have decided what to measure."* (1989, p. 145)

Psychometric techniques will help us determine how dimensions are related to one another but will not in themselves help determine the dimensions.

Unitary Hypotheses

2.1.2 General Language Proficiency

As early as 1961, J B Carroll pointed out that it is not sufficient for a valid test of language proficiency to be a sample representative of the language. Rather it should show how adequately it can predict the future success of testees in coping with the future language situation. Although the term general language proficiency (GLP) was never used, it seems that passing judgement on someone's limited performance on a language proficiency test for predicting his future success in dealing with various language tasks implies a kind of transfer ability or overall proficiency of the learner.

Spolsky (1973) argues that knowledge of a language is more than a simple command over a certain amount of isolated elements. It is,

> *" a matter of having mastered these (as yet incompletely specified) rules; the ability to handle new sentences is evidence of knowing the rules that are needed to generate them."* (Spolsky, 1973, p. 173)

It appears that Spolsky equates knowledge of language with knowledge of rules, which in turn is the same as *underlying linguistic competence*. This competence, overall proficiency, operates in all different kinds of performances. The way it operates, however, is not necessarily the same for all skills. For example, the ability to read a Shakespeare play is not the same as the ability to write it. To put it in Spolsky's term, *"all that it does claim is that the same linguistic competence, the same knowledge of rules, underlies both kinds of performance"* (Ibid., p. 174).

Spolsky, elsewhere (Spolsky & Jones, 1975), modifies his definition and differentiates between overall proficiency and linguistic competence. Overall proficiency as he asserts:

> *"is something that presumably has what Alan Davies would call construct validity. In other words, it depends on a theoretical notion of knowledge of a language and the assumption that while this knowledge at a certain level can be divided up into various kinds of skills, there is something underlying the various skills, which is obviously not the same as competence. You have to allow, of course, for gross differences.... Whatever is left is overall proficiency."* (Spolsky and Jones, 1975, p. 67)

Spolsky's overall proficiency, which plays the role of principal factor in understanding as well as production, has mistakenly been taken by others as the *only* underlying factor in all linguistic behaviour. This view, however, is not shared by the author as he clearly states:

> *"I have the notion that ability to operate in a language includes a good, solid central portion (which I'll call overall proficiency) plus a number of specific areas based on experience and which will turn out to be either the skill or certain sociolinguistic situations."* (Ibid., p. 69)

2.1.3 The Unitary Competence Hypothesis

Oller and others believed that linguistic competence was the principal factor underlying all language skills. Oller and his colleagues studied the different range of the so-called discrete-point tests of grammar, vocabulary, etc., and integrative tests like dictation. The result was high correlations between these two seemingly different tests. Occasionally the correlations between discrete-point and integrative tests were higher than between discrete-point tests that were supposed to measure the same thing. The conclusion Oller reached was that all different tests tap the same unitary underlying factor, which he called proficiency. According to Oller, proficiency was indivisible, so there was no point in sampling it. Any performance would tap the unitary language proficiency of the testee; however, integrative tests - *pragmatic* in Oller's term -would be a better measure of that underlying unitary factor. Oller, using insights from cognitive psychology, proposed a model of this underlying proficiency,

which he termed *internalised expectancy grammar* (Oller, 1979). Proficiency was not just a theoretical construct; to Oller it was a force, which existed and governed all the processes of comprehending and producing utterances.

The unitary competence hypothesis (UCH) has seriously been questioned by a number of scholars on the assumption that Oller misinterpreted the results of his research. Farhady (1979) showed that there was a *disjunctive fallacy* between discrete-point and integrative testing. He demonstrated that discrete item tests could enter into equally high correlations as integrative tests, hence, the correlational evidence in favour of integrative testing was not as strong as Oller had suggested. Nevertheless, the main criticism of the unitary competence hypothesis was in the use by Oller of factor analytic evidence, and the importance of a dominant first factor reported by Oller (1979). Farhady (1983) and Carroll (1983) argue that principle component analysis used by Oller, analyses variance as well as covariance in the scores. Thus, the error variance is incorporated into the analysis, resulting in the overestimation of the first factor. The technique they propose to overcome this deficiency is principal factor analysis. Both Farhady and Carroll examine some of the data used by Oller and his colleague by means of principal factor analysis and arrive at different factors. Farhady (1983) and Vollmer and Sang (1983) also criticise Oller for not using the conventional factor rotation procedure, which could result in more psychologically meaningful results. The arguments are so strong that Oller himself admits that " *the strongest form of the unitary hypothesis was wrong*" (Oller, 1983, p. 352).

The unitary competence hypothesis generated a number of research projects on language proficiency and consequently increased our understanding of testing and analytic techniques. It was a reaction against the dominant psychometric theory of the time, but had certain outcomes. It showed the flaws of the divisible theory and introduced the importance of a general underlying factor. But what conclusion can be drawn? It seems that language proficiency is more complex a phenomenon than a single theory of testing or language acquisition can account for. There is an overall proficiency that is different from any particular language behaviour irrespective of the contexts and situations in which the performances are operationalised, but at the same time, this proficiency is task specific and varies in different skills.

Eclectic Hypotheses

By eclectic hypotheses we mean those theories of language proficiency that combine different elements of language skills and components with some forms of general language proficiency into a single unified theory.

2.1.4 General Language Proficiency + Language Skills

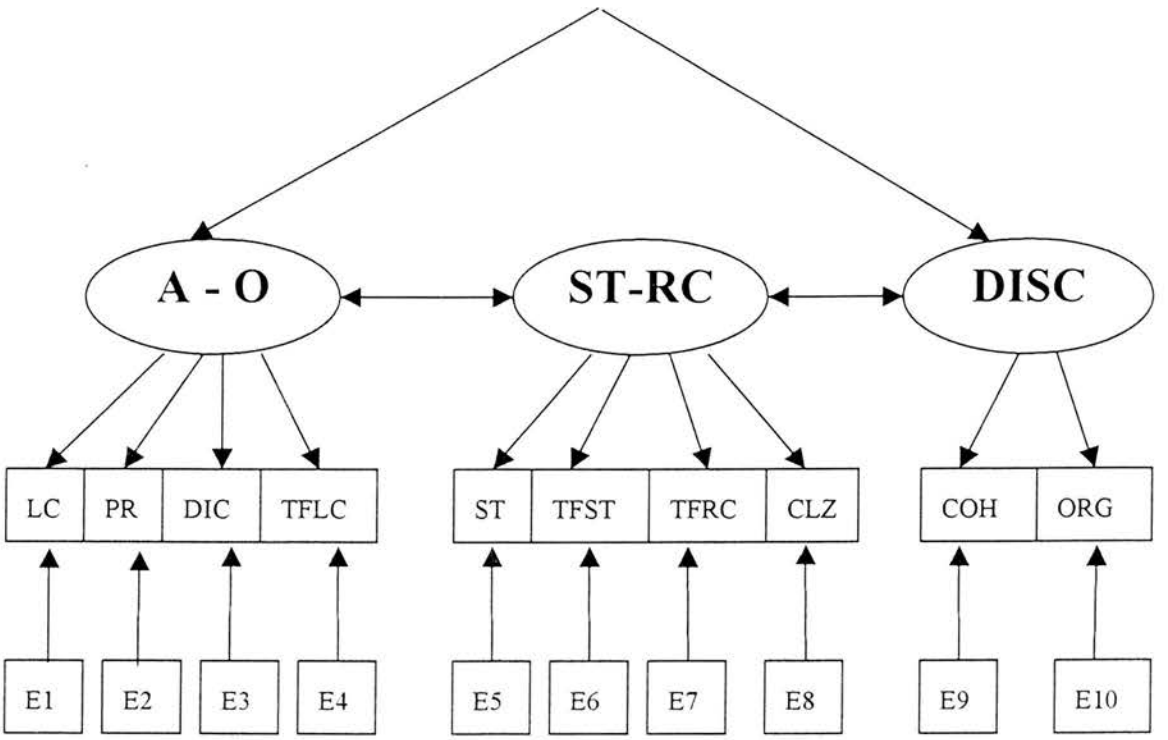
There are good reasons why skills are as important as overall proficiency. Both are important aspects of language proficiency. There is an eclectic hypothesis that accepts both the overall proficiency as well as its divisibility into skills. Carroll, based on an empirical study (1983), outlines the possibility of the existence of distinct factors underlying language tests, although acknowledging the existence of a general language proficiency factor. He maintains that:

> *"With respect to the issue of whether the data support a "unitary language ability hypothesis" or a "divisible competence hypothesis" I have always assumed that the answer is somewhere in between. That is, I have assumed that there is "general language ability" but at the same time, that language skills have some tendency to be developed and specialised to different degrees, or at different rates, so that different language skills can be separately recognised and measured."* (Carroll, 1983, p. 82)

Carroll (1983) emphasises other possibilities that might affect our understanding of language proficiency. To what extent we are concerned with L1 or L2; how we select language skills; the nature of the samples of persons; type of analysis we apply for data analysis; and the extent to which there are differentiable language skills.

The relationship between different language abilities (skills) and the general language ability has given rise to two other hypotheses regarding the nature of language competence: a) the first hypothesis, *the correlated-trait **(CT)** hypothesis*, states that *"separate traits (or factors) underlie performance on language tests and that these traits are correlated with each other,"* (Fouly, Bachman & Cziko, 1990, p. 4); b) the second hypothesis , *the higher-order **(HO)** hypothesis*, states that *"the traits underlying performance on language tests are separate and influenced by a single*

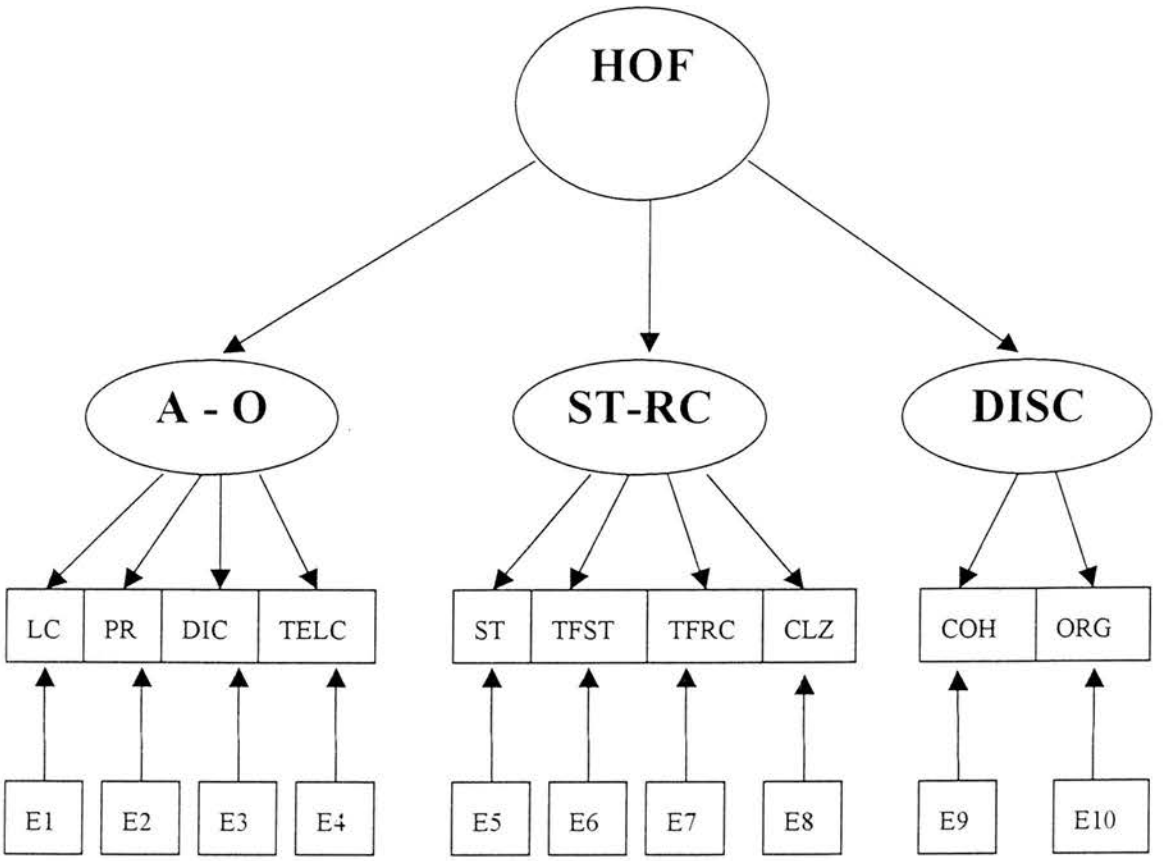
higher-order factor (HOF)" (Ibid., p. 5). Fouly et al. illustrate the fitness of their data into these two hypotheses (see Figure 2.1and Figure 2.2).



A_O=Oral-Aural ST_RC= Structure-Reading Comprehension DISC=Discourse
IEPT=Illinois English Placement Test Battery TOEFL=Test of English as a Foreign Language
OITCPE=Oral Interview Test of Communicative Proficiency in English (1983)
LC=OITCPE Listening Comprehension PR=OITCPE Pronunciation DIC=IEPT Dictation
TFLC=TOEFL Listening Comprehension ST=IEPT Structure TFST=TOEFL Structure
TFRC=TOEFL Reading Comprehension CLZ=IEPT Cloze COH=OITCPE Cohesion
ORG=OITCPE Organisation E=Errors of Measurement

Figure 2.1: Path Diagram Showing The Correlated-Trait Model

The correlated-trait model hypothesises three correlated first-order factors: An oral aural (**A-O**) construct, a construct of structure-reading comprehension ability (**ST-RC**), and discourse (**DISC**) competence. The model also hypothesises that each factor is measured by a number of observed variables, e.g. the A-O factor is measured by the four language tests **LC, PR, TFLC,** and **DIC**.



A_O=Oral-Aural ST_RC= Structure-Reading Comprehension DISC=Discourse
IEPT=Illinois English Placement Test Battery TOEFL=Test of English as a Foreign Language
OITCPE=Oral Interview Test of Communicative Proficiency in English (1983)
LC=OITCPE Listening Comprehension PR=OITCPE Pronunciation DIC=IEPT Dictation
TFLC=TOEFL Listening Comprehension ST=IEPT Structure TFST=TOEFL Structure
TFRC=TOEFL Reading Comprehension CLZ=IEPT Cloze COH=OITCPE Cohesion
ORG=OITCPE Organisation HOF=Higher Order Factor E=Errors of Measurement

Figure 2.2: Path Diagram Showing The Higher-Order Model

The higher-order model is similar to the CT model with the exception that *"the correlations among the first-order factors are accounted for by one higher-order factor, which is assumed to affect the three (uncorrelated) primary factors: A-O, ST-RC, and DISC"* (Fouly et al., 1990, p. 10). Fouly et al.'s findings support both of the models. That is, there is no significant difference between a correlated-trait model and a higher-order model. Both models would fit well with their data. The findings of their study, nevertheless, support the claim that *"in addition to differentiated language skills, there exists a general factor"* (Ibid., p. 16).

2.1.5 Communicative Language Proficiency

Current researchers have shifted towards understanding proficiency as a kind of *ability* to use language communicatively. This ability involves not only the knowledge of language but also the *capacity* for implementing that knowledge in communicative language use. The crucial question, then, is how skills and knowledge are related. Reviewing the inadequacy of early models of the skills/components model of proficiency, Bachman (1990) proposes a very complex model of language proficiency. His framework consists of language competence, strategic competence, and a psychophysiological mechanism. Figure 2.3 and Figure 2.4 illustrate Bachman's model of language proficiency.

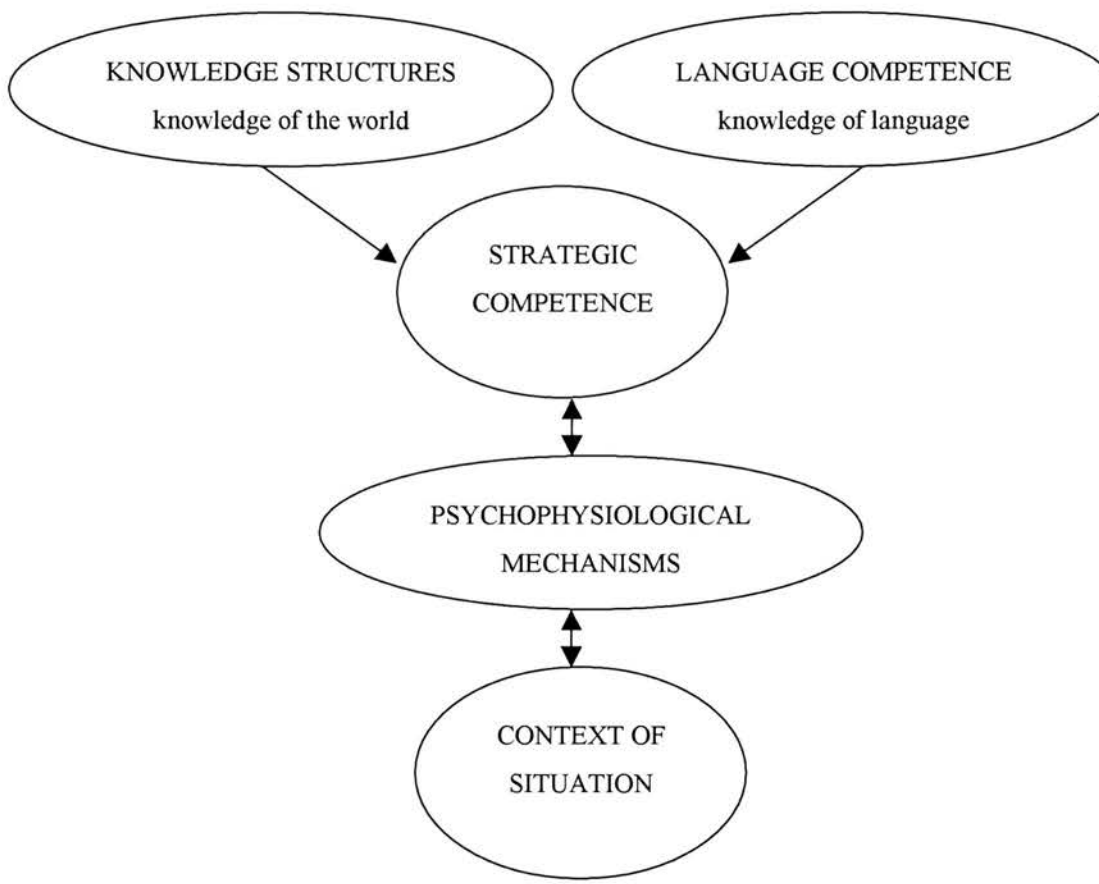


Figure 2.3: Bachman's (1990) Components Of Communicative Language Ability In Communicative Language Use

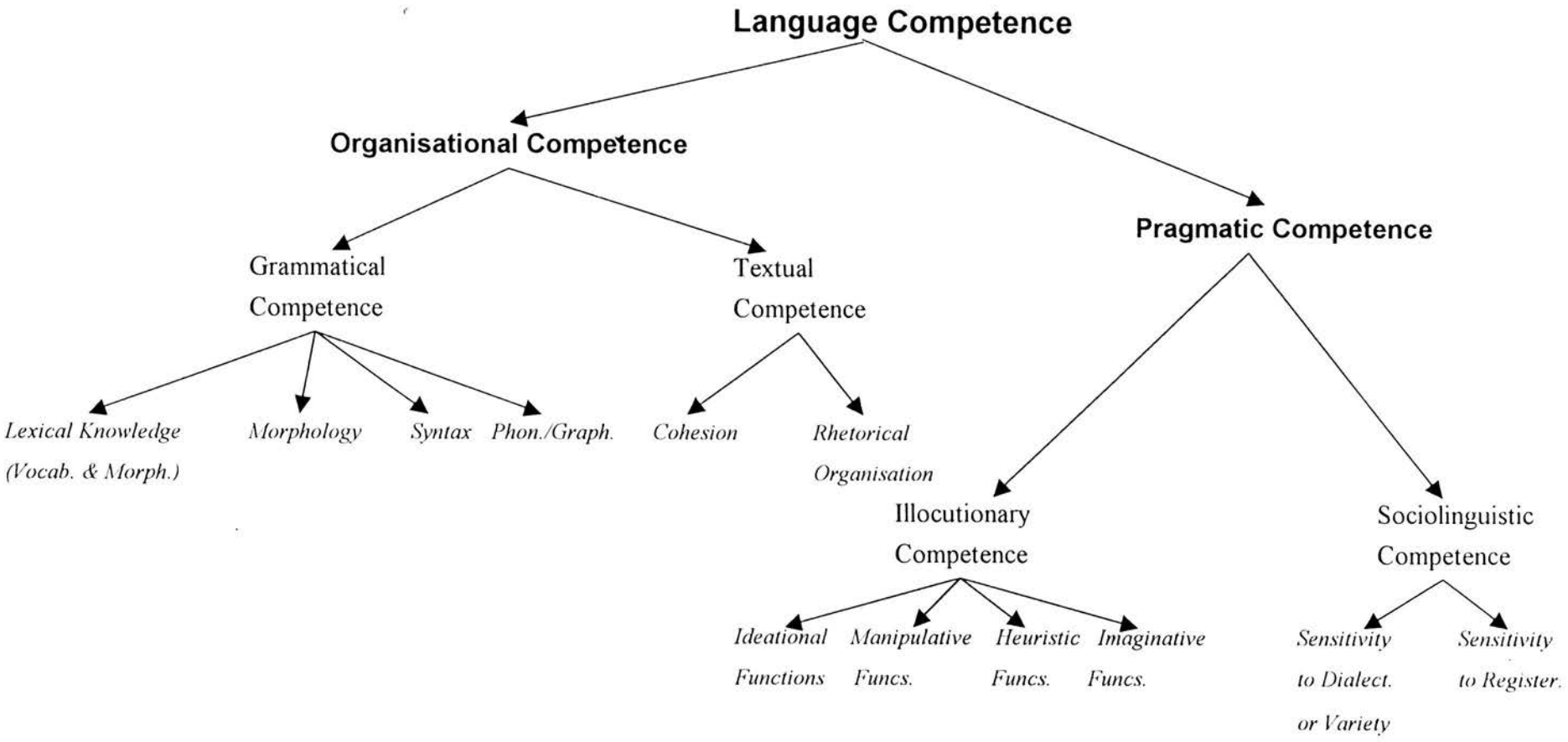


Figure 2.4: Components of Language Competence

(Adapted from Bachman, 1990)

Bachman summarises the model in this way:

> *"Language competence includes organisational competence, which consists of grammatical and textual competence, and pragmatic competence, which consists of illocutionary and sociolinguistic competence. Strategic competence is seen as the capacity that relates language competence, or knowledge of language, to the language user's knowledge structures and the features of the context in which communication takes place. Strategic competence performs assessment, planning, and execution functions in determining the most effective means of achieving a communicative goal. Psychophysiological mechanisms involved in language use characterise the channel(auditory, visual) and mode (receptive, productive) in which competence is implemented."* (Bachman, 1990, pp. 107-8)

The framework is so hypothetical that even Bachman admits that it is meant to be represented as a *"guide, a pointer, if you will, to chart directions for research and development in language testing"* (Ibid., p. 82).

Bachman has later modified his model of language ability in collaboration with his colleague Palmer (Bachman & Palmer, 1996). They define language ability essentially in Widdowson's (1983) terms as the capacity for using the knowledge of language in conjunction with the features of the language use context to create and interpret meaning. Their model includes two types of components: a) areas of language knowledge, which they would hypothesise to be unique in language use , and b) metacognitive strategies that are probably general to all mental activity. Although there are changes in the new model, the description of language knowledge is essentially the same as Bachman's (1990) discussion of language competence.

2.1.6 Preference Model

This is a model of language proficiency rooted in communicative competence theory but expressed in terms of preference linguistics. The model was first proposed by Jackendoff (1983) and later adopted by Spolsky to account as an interim solution for his theory of second language acquisition. A *preference theory* as Spolsky points out, *"is an approach that allows for a kind of valued eclecticism"* (1989, p. 146). It has the following characteristics:

1. Various conditions for language learning are not all necessary for learning to take place.
2. The knowledge and skills that are the outcome of learning are described by a number of specific conditions.
3. Each condition is to be seen as an independently measurable dimension that overlaps with, but is not directly translatable into, the others.

The theory has practical implication for language testing. It does not make any *a priori* decision on what should be included in a test. Rather it sets *possible* dimensions. The criterion for judging the content of tests is determined by pragmatic or ethical considerations. It is the tester "*who must decide how best to proceed with the maximum efficiency to obtain as accurate a measurement as possible of the minimally relevant dimensions*" (Spolsky, 1989, p. 147).

2.2 Final Remarks On Language Proficiency

> *"One cannot develop sound language tests without a method of defining what it means to know a language, for until you have decided what you are measuring, you cannot claim to have measured it."* (Spolsky, 1989, p. 140)

The same old song of *what it means to know a language* reiterates to remind us of the fact that in spite of the endless effort to demystify the concept, little advance has been made in the clarification of language proficiency. Indeed, our picture of language proficiency today is less clear than it was thirty years ago. What is it then that makes it so difficult to define? Is it a theoretical issue or a practical one? Is it really an issue? Davies (1981a) tends to take the position that the issue - in the sense of general language proficiency - is essentially a non-issue theoretically but, at the same time, its practical implications are important.

It seems that the 1990s have been an era of uncertainty and conjectures in communicative competence theories. Theoreticians do not make any strong claim; they instead prefer to give guidelines, sets of *possible* course objectives, descriptions of possible stages achievable in a program, descriptions of some possible criteria for judging success. Nevertheless, they provide no clear final answer to the need for a description of language proficiency. The characterisation of language proficiency remains unmet. Possibly, as Spolsky mentions:

> *"Part of the answer... lies in the construct of communicative competence itself. Communicative competence theories have not yet clarified the relationship between function and structure, nor provided a theoretical basis for exhaustively describing the components of language proficiency or delimiting the boundaries between them."* (Spolsky, 1989, p. 144)

Having reviewed the literature, there appears to be no consensus for a clear definition of language proficiency. Yet the quest for a theoretical framework never seems to cease. This is partly due to the complex nature of the construct as illustrated in Bachman's model (1990), but probably more to the practical needs of the testers. The validity of a test depends upon the coherence of the underlying construct and until this construct is defined, no test can claim to test what it purports to test.

Alderson and Clapham (1992) gathered numerous applied linguists' views for the construction of an international language proficiency test (ELTS Revision Project). Their survey revealed no dominant theoretical model on which they could base their test construction and construct validation. What it revealed was that:

> *"The applied linguists seemed to be content with the concept that proficiency is divisible by skill, and there are thus tests of the four macro-skills: reading , writing, listening, and speaking."* (Alderson & Clapham, 1992a, p. 164)

Alderson and Clapham were obliged to take an eclectic approach to the establishment of specifications for their test writers. Therefore, they selected those aspects of the responses they had received that they judged to be *practicable*. The result was far from being a theoretically pure model of language proficiency and the most they could claim for their underlying construct was that it did not *"appear to contradict or conflict in any serious way with what theorists and empirical research have revealed as to the nature of language proficiency"* (Ibid., p. 164).

But did they really need a pure theoretical model? It seems that one should make clear the boundary between theory and practice. Theories take a long time to be formed if they are not challenged by other competing theories. On the one hand, there are no promises of a well-formed coherent model of language proficiency in the near future, if it is at all possible. On the other hand, language testers are dealing with practical issues of assessing norms of linguistic behaviour; they cannot wait for the theories to be developed. What they can do is to find some means of operationalising those aspects of language proficiency models, which are relevant to their specific task and situation: a reflection of the Preference Model. Nevertheless, the insights coming out of test results analysis can contribute a lot to the development of a better understanding of *what it means to know a language*.

2.3 A Description Of Three Language Proficiency Tests: TOEFL, IELTS, And EPTB

In this section we will provide some descriptions of three language proficiency tests that have been used in this research. We will describe TOEFL and IELTS, which are the two major tests under examination here. Additionally, we will describe EPTB (English Proficiency Test Battery) that will later be used in the research for validational purposes.

2.3.1 TOEFL: Origin, Structure, and Statistical Features

The Test of English as a Foreign Language (TOEFL), a highly secure test, is the most widely administered, standardised , multiple-choice test of language proficiency (1963-2000). TOEFL is administered 12 times a year, in a new equated form each month, at more than 1,100 centres in 170 countries and areas and its results are used by some 2500 universities and colleges in the US., Canada and other countries for a variety of academic subject areas. According to ETS (1992a) some 1,178,193 students seeking admission to institutions in the United States or Canada took the test from July 1989 through June 1991. The test is designed to "*evaluate English language proficiency of individuals whose native language is not English, most often those wishing to study in North American universities and colleges*" (Stevenson, 1987, p. 79); it is recommended for the students at 11th grade level or above. The test is currently administered in two different versions: *paper-and-pencil* and *computer adaptive* tests. The discussion here refers to the paper and pencil version only.

The test comprises three sections (since 1976) and scores are given for both the individual sections and the total. There is no pass/fail score, however, in the Manual for Score Users, information on various reference groups and how various institutions use TOEFL scores are provided.

The TOEFL is, without a doubt, the most reliable as well as the most researched of all foreign language proficiency tests and has been under constant revision and empirical research study for the past thirty-seven years. The TOEFL Research series as of

spring 1999, consisted of 64 Research Reports and 14 Technical Reports. Over the years TOEFL has been used as a criterion for the validation of other tests. Among the most recent attempts of this kind is the Cambridge - TOEFL Comparability Study (Bachman, et al. 1995). Prior to any further generalisations about the characteristics of this test of foreign language proficiency, it is necessary to study the roots from which the test originates and its internal structure.

2.3.1.1 TOEFL: Origin

As Spolsky (1990) reports, on May 11-12, 1961 a conference was held in Washington to establish a battery test of English proficiency that could meet the needs of US colleges and universities who were considering the admission of foreign students. Carroll's keynote speech (1961[1972]) influenced most of the discussions in the two-day conference intended to bridge the gap between the theory and practice of language testing. Carroll set out the principle of language aspects (phonology or orthography, morphology, syntax, lexicon) by skill (auditory comprehension, oral production, reading, writing) involved in language proficiency. The same principle was later used by Lado (1961) in his discussion of skill / elements of language proficiency explained in 2.1.1. This framework provided a basis for testing specific items essential for reliable and valid testing. Although Carroll suggested the addition of some integrative testing in his model of language proficiency, due to the influence of psychometrists of the time, the latter approach was not recommended in the final decisions of the conference.

Based on various discussions, the conference came to the general conclusion of accepting a new programme of proficiency testing. The testing programme as Spolsky (1990) points out was supposed to be:

> *"An English proficiency test, 'an omnibus battery testing a wide range of proficiency and yielding meaningful (reliable) subscores in addition to total score'.... be aimed, at first, at the college level."* (Spolsky, 1990, p. 111)

Furthermore, the nature of the English Proficiency Test was described as having four subtests: 1) control of English structure, 2) auditory comprehension, 3) vocabulary and reading comprehension and 4) writing ability. Oral production was not to be tested while emphasis was placed on finding objective techniques for testing writing

ability. Concerning the administration and the scoring of tests, it was suggested that testing be carried out in the student's country of origin with new forms for each administration. Nevertheless, the scoring was to be done in the US. In order to develop the promising new test, an interim organisation was set up (see Spolsky, 1990).

The Test of English as a Foreign Language (TOEFL) was finally developed in 1963 by a National Council on the Testing of English as a Foreign Language comprising of over 30 organisations. Since the test was developed in the early 60s, it was heavily weighted towards then current psychometric principles. It was a reasonable decision the conference came up with, permitting the design of, as Spolsky points out:

> *" the best possible testing programme, given the state of language testing knowledge and the general intellectual atmosphere of American language teaching theory and practice at the time."* (1990, p. 114)

In 1965 the responsibility for the programme was taken jointly by Educational Testing Service (ETS) and the College Board and in 1973 a co-operative arrangement for the operation of the programme was entered into by ETS, the College Board and the Graduate Record Examinations Board.

2.3.1.2 The Structure of TOEFL

The early versions of TOEFL consisted of five areas of competence in English: I. Listening comprehension; II. English structure; III. Vocabulary; IV. Reading comprehension; and V. Writing ability. Since 1976, due to the recommendations of empirical research (Pike, 1979), TOEFL has consisted of three sections, each separately timed: Listening Comprehension, Structure and Written Expression, and Reading Comprehension and Vocabulary.

Listening comprehension (50 minutes), which measures the examinee's ability to understand English, as spoken in the US has three parts: Sentences, Dialogues, and Lecture. In the first part there are two kinds of tasks. One kind is answering a short question; the other one is understanding a short statement. The examinee should select the written option that most closely corresponds to a statement spoken once on audiotape. In the Dialogues part, the examinee hears a series of short conversations between two speakers. At the end of each conversation, a third voice asks a question

about what has been said. The best response is chosen from the four written options. In the third part, the examinee listens to several brief talks, lectures, public announcements, etc., representative of academic or student contexts in the US. Each is followed by spoken questions. The Listening Comprehension section is designed to be representative of real life situations to which the examinee is assumed to be exposed.

The Structure and Written Expression section (25 minutes) has two parts: incomplete sentences and error recognition. In the first part, a number of sentences are presented. The examinee selects the option (words or phrases) which best completes the sentence. In the second part, several sentences in which some words or phrases are underlined are presented. The examinee should identify the underlined part in each sentence that is not appropriate to the standard, formal written English.

Finally, the Reading Comprehension and Vocabulary section (45 minutes) also has two parts. In part I a word or a phrase in a sentence is underlined. The examinee selects the option which when substituted best preserves the original meaning of the underlined word or phrase. In part II several short reading passages are presented, each followed by a number of questions. The examinee selects the options best answering the questions.

2.3.1.3 TOEFL: Scoring and Reliability

Raw scores are converted to 20-80 scaled scores for each section. As for the total score, the converted scores of the three sections will be added and then multiplied by 10 divided by three, which comes to a scale of 200-800. However, in practice, section scores range from 22-67, while the total scores range from 227-677. As mentioned earlier, there is no pass / fail score and institutions require different range of scores for different subject areas. Nevertheless, it should be mentioned that there is a general tendency that students scoring below 450 are considered to be weak in English, while those scoring above 600 are considered to have an excellent mastery over the English language.

The reliability of the test has repeatedly been reported satisfactory. Stevenson (1987) reports that "*the average reliabilities for 12 forms (administered in 1981-1982) are*

0.89, 0.87, and 0.89 for the three sections, and 0.95 for the total score" (p. 80). This is well within the desirable range for this type of test.

2.3.1.4 TOEFL: Validity

Content Validity

Validity of a test, by definition, depends on the extent to which a test measures what it purports to measure. TOEFL intends to measure the English-language proficiency of non-native speakers of English who wish to study at North American universities. Hence, the content of the test should be representative of the social situations to which the examinees are expected to be exposed. The specification of such a context is not an easy task, given the wide range of TOEFL populations and target language-use situations. It seems that the traditional techniques of contrastive analysis and error analysis are not appropriate for content selection of TOEFL. Like all proficiency measures, the content validity of TOEFL depends on the degree to which experts perceive it to be valid. Stevenson points out that:

> *"TOEFL does agree that content is best specified by experts, and does rotate membership in this group often to avoid stagnation or the dominance of one view, leads to the reasonable conclusion, if not demonstration, that the content of TOEFL in general, is representative."* (1987, p. 81)

A Committee of Examiners composed of linguists and specialists in English language pedagogy is responsible for the content validity of the test. Peirce (1992) in an attempt to demystify the TOEFL Reading test explains how the content of the reading comprehension section is selected.

> *" The passages that are chosen for the reading comprehension section are expository texts that have been drawn from academic magazines, books, newspapers, and encyclopaedias; they are not written specifically for TOEFL. To preserve the original quality of the texts, test developers are discouraged from changing the author's words, although deletions are permitted. The rationale for such a policy is that TOEFL candidates should be exposed to what is called authentic language used by a variety of writers and not a customised 'TOEFL English'."* (Peirce, 1992, pp. 668-669)

The development process of the test involves both the ETS members of the test development team and individual item writers outside ETS. Item writers are private individuals outside ETS who are trained to find a variety of passages of appropriate length and to develop six or seven items based on each passage. Then the completed assignments are forwarded to a member of the test development team in ETS whose responsibility is to convert the passage and items into a publishable pre-test set. The set will then go for a review process.

> *"There are two cornerstones of the review process: first, a series of test reviews by approximately six different test development specialists; second, a pre-setting process... After the test developer is satisfied that the pre-test has been adequately prepared, the test goes for a test specialist review (TSR). The test specialist reviewer (also TSR) is a member of the TOEFL test development team; indeed, all test developers are reviewers and all reviewers are test developers. The passage and items are systematically reviewed by the TSR, who is simultaneously "taking" the test and reviewing it.... After the test has gone through the TSR stage, it goes to the TOEFL co-ordinator who examines all the items again, two editors who focus on stylistic problems in the test, and a sensitivity reviewer who seeks to eliminate any potentially offensive material in the test.... it is then returned to the Test Development department for a final review before it is published in a TOEFL test booklet."* (Ibid., pp. 672-673)

There is always a question of whether the tasks and content are representative of the situations the non-native speakers are going to face in academic settings. Since there is no definitive list of specification of different linguistic and communicative abilities necessary for given sociolinguistic situations, one has to rely on the judgements of the experts involved in the development process of the test for the degree of context validity. Bachman, Vanniarajan & Lynch (1988, p. 148) accuse TOEFL of using *"culture-specific (American) topics'*. But, is it not what the test purports to measure? Indeed, if the sole purpose of TOEFL is to challenge examinees with questions on specific points related to American English grammar, vocabulary, and usage, then TOEFL items perform this task. American English and culture are inseparable features of the typical situations TOEFL intends to simulate. Thus, the inclusion of culture-specific items in TOEFL seems legitimate. There is no particular weakness in this regard unless some other issues such as whether TOEFL might achieve its goals better with items that have greater face and content validity are discussed. This leads us to the authenticity of TOEFL items and tasks.

The issue of authenticity, although theoretically very important, has serious practical limitations in all language proficiency measures of the kind of TOEFL. The extent to which TOEFL can simulate authentic language situations that may be faced by examinees when attending North American colleges determines the academic and social naturalness of TOEFL items. The very fact that individual test items and item stimuli are very abbreviated in comparison to natural language use in actual authentic communication contexts is indicative of the degree of TOEFL distance from the criterion of authenticity. Duran, Canale, Penfield, Stansfield, & Liskin-Gasparro list "*the absence of information about a language use context, the inability of examinees to exchange language reciprocally, and the brevity of discourse-length texts*" (1985, p. 60) as the most critical factors "*inducing judged lack of authenticity*" of TOEFL items. Even item types such as mini talks / extended conversations and reading comprehension, which involve greater and richer length of discourse and are judged to have the greatest authenticity have:

> "*limited authenticity because they retain an isolated character that renders them as fragments of extended discourse drawn in an ad hoc fashion from the range of all possible social-academic situations and academic context experiences that students might encounter in college.*" (Ibid.)

But how can one remove the isolated character, which limits the authenticity of the test items? Should one, as Duran et al. suggest, revise item content specifications to strengthen the naturalness of language on TOEFL items? Or as Henning (1988) suggests, extend the length of the test items to include more illocutionary acts and include a possible wider range of social-academic situations and make the test more *authentic* but less practical? Both suggestions, although possible in principle, involve a number of practical, psychometric, and operational issues, which make them less promising in practical terms.

It seems that there is confusion here between the criterion and the test. What is our expectation of a valid language proficiency test? We want it to be representative of the situations the examinees are expected to be involved in in the future. It is important to note that the purpose of language tests is to represent a sample of real life situations; the intention is not to create the same situations. No matter how hard we try to make the test as close a sample of real life situations, we will be limited by test constraints. Time is an important factor, which restricts our options in lengthening test items. One has to make decisions about the appropriate length,

perhaps a kind of "*shibboleth decision*" (Davies, 1991). In language testing, like any other decision-making field, one has to decide to separate the sheep from the goats, the proficient from the non-proficient, the appropriate from the inappropriate. It is the great responsibility of the test constructors to take decisions about sampling, which, of course, have to be taken with great care. Sampling in proficiency testing should *resemble* the real language contexts that it is supposed to be predicting. It should not, however, replicate the real language contexts, in which case it ceases to be a test and becomes precisely what it is supposed to be predicting.

Returning to the discussion of the authenticity of TOEFL items, one might conclude that although TOEFL items represent some authentic language contexts, their authenticity is not the greatest strength of the test.

Concurrent Validity

TOEFL developed out of a desire to predict the future capabilities of the examinees to cope with various language tasks necessary for comprehending academic texts. Therefore, the aim of most TOEFL validation studies was towards establishing concurrent validity rather than construct validity. A number of criterion-related studies tend to support TOEFL validity. For instance, Hale, Stansfield, Rock, Hicks, Butter, & Oller (1989) report a correlation of 0.95 with multiple-choice cloze. TOEFL has also been shown to have moderate to high correlations with direct or integrative measures such as standard cloze test, oral interview, and essay rating. Some of the correlations between TOEFL and other standard measures reported by TOEFL Research Reports are as follows. Clark & Swinton (1980) report a correlation of 0.68 with the Foreign Service Institute (FSI) interview test. Alderman (1981) reports the correlations of TOEFL with five different tests: 0.83 with the verbal component of the Secondary Aptitude Test (SAT), 0.74 with the mathematical component of SAT, 0.82 with the Test of Standard Written English (TSWE), 0.91 with the English as a Second Language Aptitude Test (ESLAT), and 0.66 with the mathematics achievement test. Wilson (1982) reports correlations of: 0.71 with verbal scores of the Graduate Management Admission Test (GMAT), 0.70 with the quantitative component of GMAT and 0.21 with the quantitative component of GRE. The above figures support the idea that TOEFL correlates well with those instruments claiming to measure similar abilities and correlates less well with those that do not.

Construct Validity

By definition construct validity concerns "*the extent to which performance on tests is consistent with predictions that we make on the basis of a theory of abilities, or constructs*" (Bachman, 1990, p. 255). We mentioned earlier that there is a lack of consensus about the nature of language proficiency. The abilities involved in the construct of language proficiency are theoretical, yet to be defined and agreed upon. Hence, they constrain our efforts to test the extent to which we can make inferences about these hypothetical abilities on the basis of test performance. Unless we have a clear definition of the construct, we cannot claim to have measured it. TOEFL constructors seem to be very conservative in stating what construct they purport to test. For example, *the TOEFL Bulletin of Information for TOEFL / TWE and TSE, 1992-1993* (ETS, 1992b, p. 3) states that the Vocabulary and Reading Comprehension section of the test "*measures ability to understand non technical reading matter*" in standard written English. It goes on to talk about the multiple-choice format of the questions *implied, stated* or otherwise. But it never explicitly defines the construct. As Peirce (1992, p. 668) points out, "*the construct of reading that is measured in the TOEFL reading test is not made explicit in the ETS literature*". Indeed ETS cannot make it explicit as there is no promising definition in the state of the art at present. Despite the conservatism, there seems to be a general agreement in ETS that there exists a general proficiency factor, which is divisible by skills and components.

A number of factor analysis studies have been conducted in ETS to explore the underlying constructs of the TOEFL. Pike (1979) did various correlational comparisons on early (5-section) TOEFL scores of different native language groups and concluded that the original 5-section TOEFL could be reduced to a 3-section format. He found that the *listening comprehension* section was relatively independent from other objective measures and correlated well with spoken communication. Moreover, he observed that the *English Structure* section correlated highly with the spoken communication and Essay Form, the language production criteria, as well as with the *Writing Ability*. The suggestion he proposed was to combine the English Structure and the Writing Ability scores. Finally, he found that the *Vocabulary* was highly efficient and its high correlation with the *Reading Comprehension* suggested the combination of these two scores. In short, he found three clusters of measures in

TOEFL: Listening Comprehension, English Structure and Writing ability, and Reading Comprehension and Vocabulary measures.

Swinton & Powers (1980) also ran a factor analysis on the TOEFL population but came to a different conclusion. They observed some evidence that three major factors underlie performance on TOEFL. A factor underlies the listening comprehension section for all language groups. However, they found differences among the language groups in the interpretation of the other two factors. For Indo-European language groups (the Germanic and the Spanish groups),

> *"the second and third factors correspond with the TOEFL subscores (Structure + Written Expression and Reading Comprehension + Vocabulary).... For most other groups (African, Arabic, Chinese, and Japanese), the Reading Comprehension items tend to load on the same factor with Structure and Written Expression, with Vocabulary splitting off from Reading Comprehension to define a third factor. For Farsi speakers,...the least proficient of any group with respect to total TOEFL scores,... the results are more suggestive of a single listening comprehension factor and a global factor underlying performance on Sections II and III than are the structures of the other groups".* (Swinton & Powers, 1980, p. 15)

The above studies illustrate differential performance across different native language groups of the TOEFL population, which yield different factor structures accordingly. That is to say, the performance on a language test is affected by the test taker's characteristics, which may lead us to reconsider the definition of language proficiency. It should, however, be mentioned that given the world population sample of TOEFL, such individual differences will be mitigated. What can be inferred from the above studies is that TOEFL is measuring several major language proficiency areas of which *listening comprehension* is definitely one. The interrelations of different language proficiency areas, furthermore, suggest that whatever clustering one might use for different sections, TOEFL is measuring *Structure, Written Expression, Reading Comprehension, and Vocabulary of the English language.*

It might be argued that TOEFL does not reflect current theoretical developments and research in language testing. That is, it does not challenge the communicative proficiency of the examinees. Duran et al. review the different aspects of TOEFL and conclude that:

> *"The existing TOEFL, even with modifications, could not be expected to assess some important proficiency skills. For example, the test cannot test speaking ability or writing ability directly."* (Duran et al.,1985, p. 65)

Accordingly, they suggest the addition of some new sections to the test to include an extended range of communication skills such as speaking and writing. In response to such demand and current emphasis on communicative approaches, TOEFL has moved minimally towards a communicative mode and has changed only in terms of skill extension in its provision of a Test of Written English (TWE). The new version (TWE) has been included in TOEFL in only six versions of the test throughout the year and its score (1-6 scale) is reported separately. The score of the new writing section does not affect the TOEFL total score and is reported only, at this stage, for institutional information. It seems that ETS has finally agreed to include the writing test in its measurement of language proficiency, a proposal suggested as early as 1961: *"An unscored composition will be furnished to test users for whatever use they may wish to make of it"* (Spolsky, 1990, p. 112).

The reluctance of ETS to include the production section (TWE) in TOEFL for nearly thirty years was partly due to the lack of objective criteria of scoring writing, which has aroused *"causes for concern"* (Raimes, 1990), and was partly due to the increased cost of administration, which has to be reflected somehow in the examinees' fees. The latter, is going to weaken one of the great strengths of TOEFL, that is, offering *"at a modest cost a service that would otherwise be lacking to hundreds of thousands of examinees and to the co-operating institutions"* (Stevenson, 1987, p. 81).

It appears that ETS is abandoning its traditional view of separating the tests of receptive skills (TOEFL) from the productive ones (TWE and TSE). In the ongoing review of TOEFL as part of the TOEFL-2000 project (Chyn, DeVincenzi, Ross, & Webster, 1992; ETS, 1991), the two skills are brought side by side to operationalise a more comprehensive view of language proficiency as well as to concord with new developments in language testing research. Whatever the outcome may be, the merit of the present TOEFL as one of the most reliable and valid tests of English language proficiency cannot be ignored.

2.3.2 EPTB: Structure and Statistical Features

The English Proficiency Test Battery (EPTB), also known as the Davies Test, was designed to determine whether a candidate has sufficient ability in English to follow university level courses in Britain. The test was intended to function somewhat like TOEFL for candidates who wanted to study at British universities. As of 1977 four different versions of this test were constructed and validated against one another. The final form (Form D) comprised three main sections: listening, grammar (multiple-choice), and rational deletion cloze (exact word). The converted scores of these three sections were used to evaluate whether candidates had achieved a minimum level of English proficiency necessary for studying English medium in Britain. The converted scores were also used to assign candidates to two other proficiency levels; 1) requiring 4-12 weeks intensive tuition in English in Britain; and 2) requiring a minimum of 6-months full-time preliminary tuition.

Table 2.3: Reliabilities Reported for Davies Test

Form A 1967

	Content	Reliability
1	Phonetic Discrimination: words in isolation	0.91
2	Phonetic Discrimination: words in sentences	0.85
3	Intonation	0.75
4a	Listening Comprehension: general	0.51
4b	Listening Comprehension: specialised science	
4c	Listening Comprehension: specialised arts	
5	Reading Speed	0.97
6a	Reading Comprehension: general	0.94
6b	Reading Comprehension: specialised arts	0.88
6c	Reading Comprehension: specialised science	0.93
7	Grammar	0.89

EPTB is one of the rare tests in Britain, which has undergone empirical research with published statistics. Davies (1984, p. 60) reports the following reliability estimates for the test: 0.79 for Listening Comprehension, 0.91 for Modified Cloze, 0.82 for Grammar, and 0.92 for Reading Speed. With the exception of the Listening Comprehension section, these figures show lower reliabilities than those reported for

the earlier version of the test by Davies (1967); the difference, however, might not be significant .

It is evident that it is a test of the receptive skills. There are supplementary test components (essay and interview), but since their scores are not included in determining the proficiency levels of the candidates, no statistics are reported for them.

Concerning content validity, Davies believes that the test reflects the view that "*what is required in terms of English language by all overseas students is a minimum proficiency*" (1984, p. 57). The test is concentrated on assessing listening and reading comprehension. The content specification of Version D of the test is illustrated in Table 2.4.

Table 2.4: Content Specification of EPTB: Version D

	Items	Description	Skill
1	25	Interpretation of sentence stress	LC
2	25	Recognition of appropriate responses in discourse	LC
3	17	Identification of written notes correctly summarising points made in an interview	LC
4	50	Modified cloze passage: with the first letter supplied	RC
5	50	Multiple-choice Grammar	RC

A factor analysis of the first version suggested "*the existence of three factors, tentatively identified as Segmental Listening, Textual Listening and a more general Reading Comprehension*" (Davies, 1967, p.169). Due to the length constraints, the Reading Comprehension section was dropped in later versions, hence, it is not clear what factors might have emerged in later versions of the test had new factor analysis been conducted. Nevertheless, as we will discuss in 5.3.1.3 in more details, a factor analysis of Version C reveals similar results finding three factors associated with *language redundancy (cloze method), listening*, and *phonemic discrimination*.

The minimum cut-off proficiency level score rose from 36 to 44 over 10 years. Davies (1984) justifies this raise on the basis of the inevitable regression to the mean. It could well be due to the preparation for the test that the candidates were able to score higher on the test than their predecessors with the same level of English.

In spite of the limited scope of the test, it seems that EPTB functioned well over a period of 15 years (1964-1980) for the purpose for which it was designed. It is believed that some 5000 students took the test each year before it was replaced by the new ELTS test. Aside from the outmoded format of the test, there does not seem to be a sound justification for abandoning the test altogether, especially when one observes that its longer, more expensive successor, ELTS, proved to be less successful in terms of validity and reliability (Criper & Davies, 1988).

2.3.3 IELTS: Origin, Structure, and Statistical Features

The International English Language Testing Service (IELTS), formerly known as ELTS, was the immediate successor of EPTB for determining whether student's ability in English would meet the demands of a course of study in Britain and Australia. The early versions of the test (ELTS, 1980-1988) comprised 6 subject specific areas in addition to a general section. It was widely welcomed during 1980-1989 by British universities as it claimed to be a test of English for Specific Purposes (ESP). The test reflects the ideas of communicative language teaching and is probably the first standardised communicative language test administered over a large population across the world. Some 37,455 non-native speakers of English, according to Criper & Davies (1988), are reported to have taken the ELTS test between 1981-1985. With the introduction of IELTS in 1989, the candidature has increased steadily since then to some 78,898 in 1998 from over 200 nations (UCLES, 1999). IELTS, like its predecessor, is peculiar among British tests for its published statistics concerning the validity and reliability of the test.

The test comprises four sections and scores are reported for both the individual sections and the total in terms of proficiency band scales (1-9). Each band scale is associated with a language proficiency level ranging from the non-user (1) to the expert user (9).

2.3.3.1 IELTS: Origin

The idea of developing a new measurement that conforms to the shift in theoretical orientation towards communicative teaching in Britain goes back to 1975 when the

British Council approached the University of Cambridge Local Examinations Syndicate (UCLES) to replace EPTB with a new test. As Criper & Davies report, the first meeting of the ELTS Test Development Committee was held in Cambridge in 1976. The actual development of the test took four years and its first implementation was in early 1980 known as ELTS (English Language Testing Service).

ELTS was meant to be an ESP test; nevertheless, the final form of the test included an additional general section. The test followed the Munby (1978) communicative syllabus design. Carroll (1978) guided the test specifications on the basis of needs analysis. The analysis suggested a number of specific tests for different subject areas. However, in practice large compromises and reductions were made limiting the specific areas to six (General Academic Source, the Life Sciences Source, the Medicine Source, the Physical Sciences Source, the Social Sciences Source and the Technology Source). These specific areas were later reduced to three (Business Studies & Social Science, Life & Medical Sciences, and Physical Science & Technology) when IELTS was introduced in 1989. The whole specificity aspect of the test was revised later based on Clapham's works (1993, 1996) and the IELTS has been reduced to only one Academic module, in addition to the General Training module, since 1995-6.

2.3.3.2 The structure of IELTS

IELTS consists of two sections: General (G) and Modular (M). The general section consists of a listening test and a speaking test intended to test the oral skills. The Modular section is intended to test the written skills: reading and writing. The modules are limited to two forms: Academic - for academic audiences, and GT - for non-academic general training purposes. The total time allotted for the test is 2 hours and 45 minutes.

The listening part (30 minutes) consists of 40 test items of various types (e.g., multiple-choice, gap-filling, and true-false items), accompanied by a tape in four sections: 1) choosing from diagrams; 2) listening to an interview; 3) replying to questions; and 4) listening to a seminar. The speaking test (10-15 minutes) is in the form of a face-to-face interview, which is audiotaped. According to Ingram & Wylie (1997) the interview has five phases: *Introduction*, *Extended Discourse*, *Elicitation*, *Speculation and Attitudes*, and *Conclusion*. The marking of the speaking is done with reference to a single scale, which contains global band descriptions on a nine-point scale.

The overall format of the reading modules is the same. The readings contain texts taken from books, journals, reports, etc., related to a specific subject area and involve testees in the study skills necessary for academic and non-academic studies. There are altogether 40 test items (60 minutes) in each module spread over three sections.

The Writing test has two tasks in the case of each module (Academic and General Training). The first task is strictly limited to the information available in the text. The second task requires the testee to bring in his / her own experience and views on the basis of the reading text. Both tasks require the testees to write short paragraphs of 150 and 250 words in the space of 60 minutes (see UCLES, 1999, p.7).

2.3.3.3 IELTS: Scoring and Reliability

Scoring

Unlike TOEFL scoring, IELTS raw scores are converted into proficiency band levels. Weir (1987) has pointed out that,

> *"Administrators in the receiving institutions normally wish to know only whether the evidence supports admission or counsels against it, that is, they want a single overall score with a clearly defined cut-off point. Those providing remedial language tuition usually require a more comprehensive profile."* (1987, p. 29)

IELTS satisfies both these demands by providing an overall score as well as profile scores for candidates, which range between the non-user (1) and the expert user (9). Criper & Davies, in their examination of the ELTS band levels, take the view that,

> *"Although we have been apprised of the procedures for score conversion we remain unclear about the rationale behind the conversion, i.e. where adjustments between scores and between the means if various sub-tests were made and how they were arrived at in the first instance."* (1988, p. 4)

The1-9 scale seems to be Carroll's (1978) suggestion. Carroll (1980) proposed three 1-9 scales for assessing General Ability, Interview, and Academic Writing scales.

Davies[1] recalls that the conversion of the ELTS raw scores into the band scales was first done on the writing scores. Then, on the basis of the writing scale, the other scores were also converted likewise. There was a criticism that since the marking of the writing was done subjectively, the reliability of the overall band score hence achieved called for examination. In the light of this, the band scales were revised in the IELTS revision project (Clapham & Alderson, 1997). Alderson (1997) reports that profile marking is used for the two tasks in the Writing section. Task 1 scripts are rated for task fulfilment, coherence and cohesion and sentence structure, whereby they are assigned three band scores. These band scores are then totalled, divided by three and rounded to the nearest whole number to provide the Band Score for Task 1. A similar process is followed for Task 2 scripts but the criteria this time are: communicative quality; argument, ideas and evidence; word choice, form and spelling; and sentence structure. The final Band Score for Writing is arrived at using a Conversion Grid, weighting the two tasks differently. *'No Writing Band Scale exists to describe the Final Writing Band Score. Instead, the scores are reported on the Test Report form using the overall scale'* (Alderson, 1997, p. 97). See Appendix 5 for the Writing profile band descriptions.

The Speaking test is rated with reference to a single scale containing global band descriptions (1-9). No profile marking is used for the Speaking test. With respect to testing reading and listening, which are scored objectively, the raw scores are converted into Band Scale Scores (1-9) using statistical programmes. The profile scores on each test are totalled, divided by four to form the final overall score, which assigns the candidates to one of the 9 IELTS proficiency band scores. See appendix 5 for IELTS Band Scores descriptions.

Reliability

There are no reliability figures reported for the writing and speaking sections of IELTS as of 1999 (UCLES, 1999), as they are rated by one marker only. However, there are some acceptable figures for the reliability of the reading and listening sections. Alderson (1993), for example, reports acceptable reliability figures for the IELTS trial test. Aside from the variations in the size of the trial population in different modules (not all students took every test in the battery), the reliabilities reported are acceptable. Nevertheless, that of Module GT is questionable.

[1] Personal communication (1999).

Table 2.5: Reliabilities Reported for IELTS Trial Test (Alderson, 1993)

Tests	G1	MA	MB	MC	MGT	G2
Reliability	0.86	0.90	0.91	0.88	0.79	0.87

G1= Grammar Test; MA= Science and Technology Reading Test; MB= Life Science Reading Test; MC= Arts and Social Sciences Reading Test; MGT= Non-academic Reading Test; G2= Listening Test.

Alderson (1993) also reports the results of the reliabilities for the total test battery of listening, grammar, and reading tests ranging between 0.80-0.97, and that of the battery without the grammar test ranging between 0.76-0.96. Although the reliability of the total test battery declines in the absence of the grammar test, "*this decline is relatively unimportant, with the arguable exception of MGT, the General Training Model*" (Alderson, 1993, p. 215). The implication was that the grammar section should be dropped in the actual IELTS test. No reliability is reported for the total band score. Clapham (1996) reports exactly the above figures for only the reading section of the IELTS trial tests. Perhaps both Alderson and Clapham refer to the same trial test.

UCLES (1999) report similar reliability figures for the live IELTS listening and reading material during 1998/9. The figures in Table 2.6 illustrate that the constructors of IELTS have achieved satisfactory reliability figures for the sections of the test that can be scored objectively, i.e., multiple choice, and dichotomously scored items (true-false and gap-filling): reading and listening sections. Nevertheless, they have failed to achieve acceptable reliability figures for the productive sections of the test, which are marked subjectively: writing and speaking. The unreliability of the productive sections relate to the fact that the raters only mark both sections once. UCLES believe that '*the quality of the Writing and Speaking Modules is assured through training, certification and monitoring of* examiners' (UCLES, 1999, p.18). It is true that UCLES train the raters who mark the writing and speaking sections and randomly check some of the marked papers and interviews to ensure that the raters conform to the guidelines set out for them in the trainings but the final mark reported for the candidates will not be affected by this. The best one can say about the reliabilities of the writing and speaking sections of the IELTS is that no reliability can be reported for them. Until UCLES change their system of marking and publish the reliability figures of the latter sections, it is hard to place trust in the reliability of these sections.

Table 2.6: Reliability of IELTS Live Test Material during 1998/9

Modules	Alpha
Listening Version 13	0.90
Listening Version 14	0.89
Listening Version 15	0.89
Listening Version 16	0.89
Listening Version 17	0.91
Listening Version 18	0.91
Academic Reading Version 13	0.85
Academic Reading Version 14	0.84
Academic Reading Version 15	0.86
Academic Reading Version 16	0.82
Academic Reading Version 17	0.82
Academic Reading Version 18	0.85
General Training Reading Version 8	0.90
General Training Reading Version 9	0.91
General Training Reading Version 10	0.84
General Training Reading Version 11	0.88

2.3.3.4 IELTS: Validity

Concurrent Validity

Criper & Davies (1988) have shown that there was a considerable overlap between ELTS and EPTB (0.81) and between ELTS and ELBA[2] (0.77). They have also shown that different sections of the ELTS best correlated with the listening sub-test of

[2] The English Language Battery (ELBA) is a language proficiency test designed by Elizabeth Ingram in the 1960s on behalf of the University of Edinburgh on structuralist principles. See Davies, et al, 1999.

EPTB and the reading subsection of ELBA (Ibid.). This is not surprising as G1, G2, and M1, which were dominant in the ELTS overall score were all related to reading and listening skills. It seems that the modular section of ELTS, which was meant to be subject specific, did not provide much information about the language proficiency of the learners other than what was already provided by the general section of the test.

Geranpayeh (1994) reports relatively high correlations between IELTS total band level and TOEFL total score for two groups of subjects (0.83 and 0.67). This might suggest that the two tests provide similar information about some aspects of the test takers' language ability.

Mok et al., (1998) in an attempt to match the scales used in the ACCESS[3] and the IELTS have compared the results of these two tests using complex Item Response Modelling. The subjects did not all sit for the two tests; instead, some of the candidates who sat for either of the tests also sat for a third common test, which linked the results of the IELTS and ACCESS. This method of selecting subjects is known as *common-person equating* within the item response modelling (IRM) framework. The linking test chosen was the ASLPR[4]. The sample was 2,093 with 502 in ACCESS, 759 in the ASLPR, 477 in IELTS (Academic), and 355 in IELTS (General Training). The number of subjects in each linking group who took the ASLPR and one of the other tests was as follows: 6 in ACCESS, 25 in IELTS (Academic), and 1 in IELTS (General Training). Using many-facet Rasch Models, the four scales were matched onto a common external scale. The results reported by Mok et al., indicate that there is a match between most of the scales in ACCESS and IELTS. In the case of IELTS, levels 4.5 to 8 in IELTS (General Training) match levels 2 to 5 in ACCESS. Therefore, it allows one to establish the equivalence between most of the scales in IELTS (General Training) to those in ACCESS. However this equivalence of scales cannot be established in the same way for IELTS (Academic) and ACCESS. Only levels 5 to 6.5 in IELTS (Academic) could be matched to levels 3 to 5 in ACCESS.

As can be seen from the equivalence of proficiency levels in the two tests, proficiency levels at extreme ends of the scales on each test (i.e., level 1, 2, 3, and 9 in IELTS and

[3] ACCESS is the English language examination of the Department of Immigration and Multicultural Affairs (DIMA) of the Australian Federal Government to assess migration applicants' English language skills for migration purposes. It has six levels.

[4] ASLPR is the Australian Second Language Proficiency Ratings with twelve levels, ranging from zero to native-like ability. The ratings have subscales for speaking, listening, reading, and writing.

1 and 6 in ACCESS) could not be equated. There are two possible reasons. Firstly, there are very few people who score the highest level of proficiency in either of the two tests and therefore, there are not sufficient data for a reliable link between the two tests at such levels. Secondly, the discriminatory power of proficiency tests such as IELTS and ACCESS is not very good for lower language ability candidates, i.e., those scoring below 4 in IELTS. Nevertheless, it appears that IELTS can relatively assess the language proficiency of the candidates at levels 5 to 6.5 in much the same way as ACCESS. Levels 5 to 6.5 of IELTS band scores currently cover 65% of all the IELTS candidature according to the figures published by UCLES (1999).

Construct Validity

Criper & Davies have argued that ELTS was based upon the view that language proficiency is divisible rather than unitary and that it is divisible on three dimensions:

> *"Firstly, it divides proficiency in the skills dimension, having separate tests of reading, listening, writing, and speaking ... Secondly, it divides proficiency into 'general' and 'study' proficiency, having a test of 'study skills' distinct from the tests of the four skills referred to above ... Thirdly, it divides proficiency on the subject dimension, providing options in the form of 'modules'."* (1988, pp. 9-10)

A factor analysis, reported in the ELTS Validation project (Criper & Davies, 1988) suggested the dominance of a first (General) factor on which ELTS G1 Reading and G2 Listening loaded highest, followed by a second (Reading) factor on which EPTB Grammar and ELBA Reading loaded highest, and a third (Listening) factor on which EPTB and ELBA Listening sections loaded highest (Ibid., pp. 100-102). This is in accordance with the dominance of the G section in the overall band score in ELTS. It was concluded that despite the intention of the designers of ELTS to create a multi-factorial test, the internal structure of ELTS was in favour of a unifactorial one.

This was not to exclude the importance of different aspects of proficiency that ELTS was measuring, but it was to say that general proficiency was a better predictor of the ELTS overall score. Indeed, as Criper & Davies asserted:

> *"ELTS does appear to be measuring some aspects of proficiency that are not touched by EPTB or ELBA, though not perhaps as much or the same kind - given the dominance in ELTS of G1 and G2 - that was intended by the original construction."* (Ibid., p. 112)

The large overlap with the two traditional, discrete-point type tests of EPTB and ELBA might well have been due to the dominance of G1 and G2. It might also have been due to the mostly similar discrete-point sub-tests of G1 and G2, which concentrated on sentential understanding.

A factor analysis of IELTS trial test results also reveals the emergence of a first dominant (general) factor, followed by a second (writing) factor. Alderson (1993) reports that *"in general, an analysis of reading, grammar, and listening yielded only one common factor. The addition of writing occasionally gave rise to a second factor."* (p. 213)

Since the Interview was not included in the test analysis, nor any other external criteria, it is difficult to predict what factors might have emerged had they been included in the analysis. The only statistics available in Alderson's report are the correlations between the two reading tests (PST & BSS) from the new IELTS test and their counterpart in the old ELTS test. The purpose of comparison *"was to enable the calculation of band scores for the new test (test scores are not reported raw, but in bands of scores, which are simple transformations of raw scores)"* (Alderson & Clapham, 1997, p. 42). There were significant variations in the relationship between the new IELTS reading tests and the old ELTS reading tests: the PST reading modules correlated 0.39 while those of BSS reading modules correlated 0.76. The differences were justified on the assumption that the new IELTS test was an improvement on the old test and that the readings were not directly parallel to each other in content or topic.

Moderate correlations reported in the IELTS trial study between different modules support the ESP aspect of the trial test. IELTS did look and function like an ESP test. The test appears to have been favoured more by its face validity than any other objective criteria. Due to the lack of published data, it is difficult to observe the extent to which the test measures what it purports to test. Nevertheless, the factor analysis of the trial study does give evidence for the unifactorial structure of the test. As we will illustrate in Chapter Five (5.3.1.1), a factor analysis of the IELTS specimen module also supports the dominance of a primary factor in IELTS.

IELTS, like its predecessor ELTS, seems to have been based on a notion that proficiency is divisible by skill and as Alderson & Clapham (1992a, p. 164) have put it, *"there are thus tests of the four macro-skills: reading, writing, listening, and speaking."*

Predictive Validity

There have been a number of research studies which have attempted to investigate the predictive validity of IELTS and its predecessor ELTS. It should be borne in mind that the predictive validity of English proficiency tests is not great, perhaps not more than 0.30. This is due to the fact that language plays a limited role in academic success. As Criper & Davies maintain:

> *"...once the minimum threshold of adequate proficiency has been reached. Thereafter it is individual non-linguistic characteristics, both cognitive and affective, that determine success."* (1988, p. 113)

Criper & Davies (1988) report a correlation of 0.30 between language proficiency as measured by ELTS and academic success based on a non-representative sample of 720 students.

One of the first attempts in establishing the IELTS predictive validity was carried out by Gibson & Rusek (1992) in South Australia. A sample of 63 students entering one of the South Australian universities, who sat IELTS between December 1989 and February 1991, was selected. The students' scores on IELTS were compared with their academic progress. Gibson & Rusek (1992) concluded that IELTS scores did not predict subsequent academic success. Although, due to a major flaw in the design of the study with regard to the measure of success[5], the researchers failed to observe any meaningful relationship between the IELTS scores and academic success, their study contributed to the discussion about the difficulties inherent in such a study.

Fiocco (1992) also failed to find any relationship between IELTS global scores and the semester-weighted academic results of 61 students at Curtin University in Western Australia. She reports a negligible correlation coefficient of 0.063 for the relationship between the two.

[5] The measure for academic success was simply permission to proceed to second semester. All the students deemed to be successful as they were allowed to proceed to second semester regardless of their failure in some of the units.

The predictive validity of IELTS was investigated by Elder (1993) with a small sample of (32) overseas students studying at a number of academic institutions in Melbourne. The research findings show a correlation of 0.35 between global IELTS scores and first semester course progress ratings, and a correlation of 0.40 between the listening subtest and first semester academic ratings. The correlations drop significantly for second semester.

Ferguson & White (1998) investigated the predictive validity of the IELTS in a small-scale study at the University of Edinburgh. Although the number of students taking part in the study was small (28), the research design was comprehensive; the subjects took the IELTS test at the beginning and at the end of the year. In addition, the subjects and their supervisors were interviewed four times each over the period of one year. The results indicated a positive, albeit weak, correlation (0.39-0.46) between the IELTS scores as a measure of language proficiency and academic outcome.

Cotton & Conrow (1998) report rather different figures for the correlation between the IELTS scores of a small sample of 45 students studying at the University of Tasmania and the students' academic performance (Grade Point Averages) in the first and second semester. They found no positive correlations overall.

Table 2.7: Correlation Coefficients Of IELTS Global And Subtest Scores And Students' Academic Results (University of Tasmania)

IELTS Scores	Academic Results	Semester 1 Results	Semester 2 Results
Global	-0.24	-0.62	-0.47
Reading	0.42	0.09	0.17
Writing	0.11	-0.03	0.05
Listening	-0.19	-0.58	-0.56
Speaking	-0.55	-0.41	-0.32

Academic Results= Total 1996 academic results (n=26) *Semester 1 Results*= First semester results (n=17) *Semester 2 Results*= Semester two results (n=17)

As can be seen from the Table 2.7, with the exception of the IELTS reading subtest correlation with the total 1996 academic results, the rest of the IELTS sections have poor predictive validity with regards to students' academic performance at Tasmania University. Further analysis of the results showed that,

> *"Three (12%)students who achieved IELTS scores of 7- did very poorly in their examinations, whilst two (7%) students who achieved scores of 5.5, obtained good Grade Point Averages. With a small sample, such a distribution is likely to affect the correlations between IELTS scores and academic performance, and indicates the existence of other factors influencing academic outcomes." (Cotton & Conrow, 1998, p. 94)*

Cotton & Conrow, by and large, report a similar pattern of correlation between IELTS scores and staff ratings of students' academic achievement. The pattern of correlations for the IELTS reading and writing is, nevertheless, positive. The correlations between the IELTS reading and writing subtest scores with staff ratings of students' academic achievement is moderately positive (0.36 and 0.34), and even more positive with students' self estimates of academic performance in second semester (0.46 and 0.39).

Finally, Hill, Storch & Lynch (1999) have investigated the usefulness of IELTS as a predictor of readiness for the Australian academic context. A sample of 35 international students from 17 different first languages who had completed their courseworks was selected. The results of the study show that there was a moderately strong relationship (r=0.540) between English language proficiency as measured by IELTS and student academic achievement as measured by the average of first semester grades at university. Regressing IELTS scores on Grade Average, however, revealed that the model was weak in its predictive ability (R^2 = 0.291).

2.4 An Overview Of Language Proficiency Test Comparability Studies

During the last three decades numerous test batteries have been developed to assess the language proficiency of non-native speakers of English. Along with these batteries and with the advancement of language testing theory, various test methods have also been developed and employed in the construction of language proficiency tests. The batteries may differ from one another in two respects: *a)* aspects of language proficiency being assessed (oral, written, etc), and *b)* the scope and the purposes for which the tests are designed. Nevertheless, the batteries share an important feature: assessing the language proficiency of the candidates taking the tests. The raw scores obtained on each test are often converted to norms so that the interpretation of the scores and their association with different language ability levels become possible. In some cases, IELTS for instance, the test scores are converted into band scales assigning the candidates to several language ability groups ranging from non-user (1) to expert user (9).

Clearly, language proficiency tests are distinguished from one another in one or more respects. Yet where statistical evidence is concerned, the tests are validated against one another and their results are compared to show the degree of similarity between the traits they are measuring. The validity of the tests, that is, the degree to which the tests serve the purposes for which they are designed depends, to a great extent, on the results of their comparisons with other tests. The higher the correlation between the scores of two tests the greater is the evidence of their similarity; the lower the correlation between the two, on the other hand, the greater is the evidence of their differentiation.

2.4.1 History of Comparability Studies

The history of comparability studies is perhaps as old as test validation studies and is not exclusive to the study of language proficiency batteries. However, the literature indicates that language proficiency test comparability studies are often limited in scope and scale to the examination of one aspect of language proficiency only (see Buel, 1993; Geranpayeh, 1994; Gillespie, 1990; Ioanidou, 1990; and Irvine, 1990). The only exception is the Cambridge-TOEFL Comparability Study (CTCS), which

examines the tests under comparison from different aspects. The limited scope of comparability studies is due partly to the lack of support from administrative organisations responsible for the construction of language proficiency tests, which reflects politics in testing, and partly to the arguments against comparability studies of language proficiency tests. The opponents argue that language proficiency tests are not comparable because they are designed to serve different purposes. For example, as they argued against the Cambridge-TOEFL Comparability Study, TOEFL is designed to assess the language proficiency of candidates wishing to enter North American universities, while Cambridge tests are designed to assess that of candidates wishing to enter British academic institutions. The above mentioned tests have different audiences, test methods, and perhaps populations; hence, they are not comparable. The critics usually refer, among other things, to the moderate correlations between the scores on such tests as evidence of the non-comparability of the tests.

But moderate correlations can be due to various factors. For instance, no two language tasks are identical, so we cannot expect that even method-wise-similar tests correlate highly. Farhady (1979) has shown that basing our differentiating judgement on correlational studies is a "*disjunctive fallacy*". Tests using similar methods might also correlate moderately. The most one can infer in such cases is that the methods with which one measures a construct might influence the measures of the construct, in which case a comparability study becomes a good means for exploring such influence as the study accounts for the method variance.

Moreover, the effect of test methods on the measures of a construct should be minimised in language proficiency batteries; otherwise, the validity of the tests would be questionable. The validity and reliability of language proficiency tests rely on how effectively and consistently they measure language abilities rather than on how differently they measure language traits. To elaborate, it should be borne in mind that in dealing with the interpretation of language proficiency test results, one is interested to know how well a test battery assesses the abilities of the testees in coping with various language tasks. Since language tasks vary from one context to another, reliable and valid language proficiency tests should sample tasks (texts), which share a general framework replicable in a number of contexts. The degree of replicability indicates the generality of the test results. It follows that differences in test methods should not affect the measures of a construct in a meaningful way. That is not to

ignore the effect of test methods but to look at the effect as only one probable source of error variance.

2.4.2 Purposes of Comparability Studies

Language proficiency comparability studies may serve two purposes. They can increase our understanding of the nature of language proficiency by providing information concerning the underlying constructs of the tests. They can also promote the construction and development of future language proficiency tests by allowing us to examine the descriptive or explanatory adequacy of various theoretical and operational frameworks across test batteries, which, in turn, enable the researchers to lessen the gap between the theory and practice of language testing. A good example is the trialling of Bachman's (1990) two Content Analysis rating instruments during the CTCS: the Communicative Language Ability (CLA) and the Test Method Facets (TMF).

2.4.3 Cambridge-TOEFL Comparability Study

The Cambridge-TOEFL Comparability Study has been the most comprehensive attempt at looking at the comparability of language proficiency tests. Bachman and his colleagues have done a large-scale qualitative as well as quantitative comparability study between Cambridge and ETS tests. They have examined the EFL proficiency test batteries developed by Cambridge (FCE) and ETS (TOEFL, TSE, and TWE). The first issue they addressed was the abilities measured by the two tests. Bachman, Vanniarajan, and Lynch (1988) compared the contents of a version of TOEFL with that of the papers from FCE to find out the theory of language proficiency on which these tests are based. They arrived at a general conclusion that:

> *"the ETS tests represent language test development driven largely by measurement theory, while the Cambridge tests represent language test development guided primarily by applied linguistic theory."* (Bachman et al., 1988, p. 142)

Their analysis, furthermore, illustrated that Cambridge tests represent more illocutionary acts and perhaps should tap communicative competence in a more meaningful way.

2.4.3.1 Analysis of Latent Traits in the Two Batteries

In another study, Bachman, Davidson, & Foulkes (1993) examined the performance of a large sample who took both ETS tests and FCE Papers. The exploratory factor analysis of the study suggests that two factors underlie both sets of tests. The analysis of FCE papers supports the hypothesis of a higher-order factor [6]. That means, all the FCE papers 1,2, and 3 (Reading Comprehension, Composition, Use of English) loaded high on a first primary factor, whereas Papers 4 and 5 (Listening Comprehension and Oral Interview) loaded high on the second primary factor. It is suggested that:

> *"the FCE Papers all tend to measure a common component of the subjects' English language ability, with two specific ability factors, 'reading, structure and writing' and 'speaking and listening', being identified."* (Bachman et al., 1993, p. 36)

The exploratory factor analysis for ETS tests also supports the hypothesis of a higher-order general factor. Moreover, the TOEFL listening section and the SPEAK (Interview) ratings loaded most heavily on the first primary factor, whereas TOEFL sections 2 and 3 and TEW (equivalent to TWE) loaded on the second. The results suggest that:

> *"the ETS tests also tend to measure a common component of the subjects' English language ability, with specific factors associated with listening and speaking on the one hand, and reading ,structure and writing on the other being identified."* (Ibid., p. 37)

The similarities in the factor structures of the two sets of test scores would appear to reflect similarities in the abilities of the subjects in the study. It can be inferred that *"these two sets of tests measure these abilities in much the same way"* (Ibid.).

Moreover, despite the similarity of factor structures, with higher-order general factors accounting for much of the common variances in the two test batteries, the tests behave differently with respect to first-order factors. Only 10% of the common variance in FCE papers is accounted for by first-order factors, compared to 26.6% of

[6] See Figure 2.1and Figure 2.2 in 2.1.4 for the discussion of higher-order factors

the common variance in the ETS tests accounted for by the two first-order factors. This may suggest that *"the ETS tests provide relatively more information about specific language abilities than the FCE papers do"* (Ibid.), although both batteries appear to measure a single language ability.

Bachman et al. (1995) combine the two battery tests so that they can study across battery factor structures. The exploratory factor analysis once again supported the dominance of higher-order factor followed by four first-order factors. It is suggested that all these tests measure, to a considerable degree, the same common aspect of language proficiency, a general factor, whatever name it may be given. In addition to this general or common ability, there is a component appearing to be associated with Speaking, followed by two other components, *'ETS written test factor'* and *'FCE written test factor'* and finally, a listening factor which shares the least variance.

2.4.3.2 Merits of CTCS

The Cambridge-TOEFL Comparability Study (CTCS) has strong as well as weak points. The main merit of this study is the trialling of Bachman's theoretical content analysis instruments of the components of *communicative language ability* and the *facets of test methods*. The CTCS provides the means for the examination of the operationality of this model of content analysis in such a way that future similar studies may be carried out with greater accuracy. The revised improved model of content analysis instruments can now be of great help to researchers interested in examining the content analysis of language proficiency tests. The CTCS does not directly address the issue of content validity, yet the instruments used and revised in the course of this study provide a useful means for exploration into the content validity and in a deeper sense the validation of language tests. The latter has created on-going research in the University of Cambridge Local Examination Syndicate (UCLES) on the validation of Cambridge proficiency tests. There are many interesting findings in CTCS as well; for example, that tests as diverse as Cambridge and ETS tests not only tap the same aspect of the subjects' language proficiency but also that they do so in much the same way.

The Cambridge-TOEFL Comparability Study, like any other study, has its own limitations. Davies (1989) believes that the central flaw in the study is the wrong choice of comparison. *"The FCE/CPE tests are not sensibly compared with TOEFL.*

Far better to compare TOEFL with ELTS in a meaningful comparison" (p. 6). This remark seems to be very sound. ETS tests (TOEFL, SPEAK, and TWE) do not form a single battery, and their scores do not contribute to an overall score. They are in fact tests for different purposes, which test some aspects of general language proficiency without any specific syllabus. Cambridge tests (FCE/CPE), on the other hand, form a single battery and their scores add up to a single score. They are different from ETS tests in that they are actually examinations based on specific syllabuses. The subjects taking Cambridge tests will almost always participate in preparation courses prior to the examination. Therefore, Cambridge tests best fall into the category of achievement rather than proficiency tests.

The Cambridge-TOEFL Comparability Study provides a good framework for the comparison of language tests. The empirical evidence gathered in this study supports the idea that tests of language proficiency, despite their differences in methods of testing, tap virtually the same aspect of the language proficiency of the test takers. This allows us to compare other language proficiency tests with legitimacy. The CTCS' results can help us to see the effect of test methods on different language tasks as well as on different language abilities. This may resolve many of our problems in defining what we mean by *a 'communicative test'*. The content analysis instruments developed in the course of CTCS provide a good framework for investigating the relationship between test content and test performance. This framework might speed bridging the gap between the theory and practice of language testing.

Finally, should the results suggest that different tests tend to measure almost the same aspects of the subjects' language proficiency, as it was the case in CTCS, justifiable score comparisons can be made across the tests in a meaningful way. This may have practical advantages for both test takers, as they may not have to sit for several tests, and for academic institutions willing to offer admissions to overseas candidates.

2.5 How to Judge Tests: Validity Question

In this section we will examine the concept of test validation and the effect of test methods on test content and consequently on test scores. We will argue that test validation, in particular construct validation, is the most important aspect of comparability studies. One cannot compare language proficiency tests meaningfully unless they demonstrate that the tests measure similar aspects of language proficiency. To achieve that one has to first illustrate that the contents of the tests are comparable in some ways. Once content comparability is observed, it is then possible to investigate how similar / different the tests measure different language abilities of the candidates taking them. Evidence gathered in the content comparability and the abilities that are being measured by language tests is often referred to in the literature as the evidence of test validity. However, it has to be borne in mind that validity is not an absolute feature of a test and can vary along a number of dimensions such as different groups of subjects or uses of test scores.

2.5.1 Concept of Test Validation

One of the primary concerns in test construction is demonstrating that the interpretation of test scores is *sound* and *relevant*. The question one usually comes across in test validation is whether the test serves the purposes for which one intends to use it. A test, which may be excellent in many ways, becomes worthless if wrongly interpreted or used. So test users, as Cronbach (1990, p. 150) puts it, must ask, "*How valid is this test for the decision to be made*". For example, a highly valid test of grammar may not be valid if used to assess the vocabulary of the test takers. A test, which is designed to assess the language proficiency of adult ESL learners, may be invalid if used for assessing ESL at elementary school. Validity, then, is relative and can vary along a number of dimensions.

Traditionally, validity has been classified into different types. The 1954 *Technical Recommendations for Psychological Tests and Diagnostic Techniques* (APA, 1954), the 1966 and 1974 *Standards for Educational and Psychological Tests and Manuals* (APA, 1966, &1974), and the 1999 *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 1999) have all broken

validity into three or four distinct but interrelated types: content, criterion, and construct validity. Messick (1989a, p. 17) recapitulates the traditional definitions as:

> *Content validity is based on professional judgements about the relevance of the test content of a particular behavioural domain of interest and about the representativeness with which item or task content covers that domain.*

> *Criterion-related validity is based on the degree of empirical relationship, usually in terms of correlations or regressions, between the test scores and criterion scores.*

> *Construct validity is based on an integration of any evidence that bears on the interpretation or meaning of the test scores.*

It is very difficult to separate these three and to associate a test score with any of these categories. Interpretation of the results is an on-going process, with every bit of information contributing to its formation. Although content validity is an essential part of any validational study, one cannot judge the validity of a test based on the mere content analysis of the test. The analysis of content is only one, but important, stage in the process of validation. Determination of content validity is a matter of expert judgement. A test is valid to the extent that the experts believe it to be representative of the target syllabus (content). Content validity alone cannot provide us with a definitive interpretation as the judges often do not know what precisely a test item is purporting to test. This confusion is due, partly, to the lack of precise and agreed definitions as to what a construct is - as already discussed in 2.1 on the construct of language proficiency - and partly to the disagreement of how to operationalise theoretical definitions into concrete test items. Alderson (1990a) concludes from a series of studies (Alderson 1990a, 1990b; Alderson & Lukmani, 1989) that judges are unable to agree as to what an item is testing. What are we trying to measure? If we do not know what we are measuring, we cannot claim to have measured it. The same old argument appears to lead us to a vicious circle. It is here that construct validity comes to our rescue by building on an interpretation of what we are trying to measure.

In recent years there has been an emerging consensus among measurement scholars about the *"centrality of construct validity to the evaluation of any assessment-based*

interpretation" (Moss, 1992). Anastasi (1990), Cronbach (1988, 1989, 1990), Messick (1975, 1980, 1989a, 1989b, 1995) and Moss (1994) have all stressed the centrality of construct validity in the validation procedure. Construct validity has become such an important part of the validity studies that the latest edition of the *Standards for Educational and Psychological Testing* (American Educational Research Association et al. 1999, p. 174) considers it *"redundant with validity"*. Construct validity is actually a further stage after content analysis in the validation of a test. Construct validity starts with a definition and, based on the results of a test, tries to build on a persuasive interpretation. In doing so, it requires the incorporation of any relevant information from the results of content analysis, criterion validation, or any other available sources of information.

Validation is viewed here as a *unitary concept* (Messick, 1989a) though it is *multifaceted*. It is unitary in that every bit of information contributes to the formation of a single interpretation concerning the construct under investigation. It is multifaceted because it has different dimensions from the examining of the areas of content including the study of the elements of language ability and test method facets to the criterion with which one intends to measure the test, and finally arriving at an interpretation of what the construct may actually represent. Validation then is a *"fluid, creative process"* (Cronbach, 1990, p. 178), which starts with a definition and moves toward developing an interpretation for persuading others of its soundness. It has to be revised when inadequacies are recognised in due research. This complexity, as Cronbach points out, *"means that validation cannot be reduced to rules, and that no interpretation is the final word, established for all time"* (Ibid.).

Construct validation, as such, seems to be the most comprehensive and complex part of validation as it starts with a definition and tries to tap the underlying construct(s) of a test and match it (them) with the claim of the test. There is confusion here as to what the definition of a construct is. As we have already argued, there is a lack of agreement among applied linguists as to what language proficiency is. One can neither claim to measure language proficiency when it is not defined nor reasonably define this construct until pertinent observations are made. The argument might yet lead us into another vicious circle. That is precisely why Cronbach invites us to look at the issue as a process, which is open to continuous investigation. The more observations we make, the more likely we come to a persuasive interpretation. There is no final word. The process has to be continued as more evidence is gathered continually. On the basis of new evidence, one has to revisit the present

interpretations and re-examine their adequacy for describing the traits meant to be measured by language tests.

A word of caution. Another form of validity evidence, which has not received much attention by researchers in the past, is based on the social consequences of test interpretation and use. Messick (1995) in trying to draw our attention to this important but forgotten aspect of test validity maintains that:

> *"It is ironic that validity theory has paid little attention over the years to the consequential basis of test validity, because validation practice has long invoked such notions as the functional worth of the testing- that is, a concern over how well the test does the job for which it is used.... And to appraise how well a test does its job, one must inquire whether the potential and the actual social consequences of test interpretation and use are not only supportive of the intended testing purposes, but also at the same time consistent with social values."* (*Messick*, 1995, p. 744)

Consequential validity is in fact another aspect of construct validity, which deals with the positive or negative evidence and rationales for evaluating the consequences, deliberate or inadvertent, of score interpretation and use. Messick's call for the examination of the consequences of test use in validity studies has been taken seriously by a number of language testing professionals in recent years. For example, Stansfield (1993), Davidson et al. (1997), Davies (1997a), Hamp-Lyons (1997), Norton (1997), and Wall (1997) have discussed the importance of the consequences of language test use. It has received so much attention in language testing that a special issue of *Language Testing*, guest edited by Alan Davies (1997b) has been devoted to this topic. Despite the importance of consequential validity, it will not be examined in this research because score interpretation is not within the scope of our study.[7]

Reliability / Validity Relationship

Another fundamental concern in the design of language proficiency tests is to identify the potential sources of error variance in a given measure of a trait and to control the effect of such factors on that measure. The potential sources of measurement error are any factors other than the ability being measured that affect the test scores. For

[7] See Messick, 1989a, pp. 58-92 for a full description of the consequential basis of test interpretation and use.

example, in a language test, factors such as test-wiseness, motivation, health, and stress can affect test takers' performance but are not associated with the language ability being measured. Thus, their influence on test performance is undesirable and is considered as the source of unreliability. Other potential sources of unreliability could be related to test method facets. Factors such as length of the test, number of items, genre, test takers' preparedness, and all related to examinees' characteristics can affect test reliability. Since the potential sources of unreliability are endless, it is important to control as many factors as possible that may contribute to error variance. By controlling such factors, one minimises the measurement error and hence maximises test reliability.

Reliability, by definition (Anastasi, 1990, p. 109), refers to the consistency of scores obtained by the same persons when re-examined with the same test on different occasions, or with different sets of equivalent items, or under other variable conditions. Ideally, establishing reliability requires several administrations of a single test or parallel forms of a test on a population sample. Nevertheless, in practice most of the tests of interest are administered only once, which can explain why we rely so much on single administration methods such as Cronbach alpha, KR20, and KR21[8] that assess internal consistency. All the above methods are based on *classical true score* measurement theory, which assumes that an observed score on a test is a composite of a *true score* that reflects an individual's level of ability, and an *error score* that is due to factors other than the ability being measured. The lower the error score is, the closer our observed score will be to the true score, hence the more reliable is our measurement tool, bearing in mind that the purpose of a test is to achieve reliable observed scores, i.e. as close as possible to true scores.

There is an alternative model of test theory, namely, *generalisability theory*, which is grounded in the framework of factorial design and analysis of variance. According to generalisability theory, reliability is a function of the circumstances under which the test is developed, administered and interpreted. Reliability in this model is a matter of generalisability and,

[8] See Bachman, 1990, pp. 160-235, amongst others, for a full description of reliability coefficient formulas.

> *" the extent to which we can generalise from a given score is a function of how we define the universe of measures. And the way we define a given universe of measures will depend upon the universe of generalisation- the decisions or inferences we expect to make on the basis of the test results. "* (Bachman, 1990, p. 188)

No matter what method of reliability estimate we adopt, establishing that an observed score on a test is reliable is an essential preliminary step in most[9] validity studies. That is, a reasonably high reliability is needed for most kinds of validity, but does not guarantee validity. Reliability is concerned with how consistently and accurately we measure a construct and how likely it is that we get the same result each time we measure it, while validity is to do with whether we are measuring the right thing for the purposes of our test. For example, do the skills measured by a language proficiency test correspond to the skills needed for competent performance in, let us say, business English in an English speaking country?

The general rule about the relationship between reliability and validity is that reliability is a necessary but not sufficient condition for validity and that a test can be reliable and not valid, but before a test can be valid it must first be reliable. This is based on the principles of construct-related and criterion-related validity. Content validity is not validity in the same sense as these two. For a test to be content valid it must accurately sample the content that it is supposed to be assessed. Theoretically, a test could have perfect content validity but zero reliability. Think of a test, which consists of items that cover the content perfectly, but all the items are so bad that getting one right is a matter of chance. Or consider a test covering appropriate content that is so easy that all students get every item correct or so difficult that students get every item wrong. Admittedly, this is unlikely to happen in practice, but content validity is certainly not tied to reliability in the same way that other measures of validity are.

There is an additional paradox about the relationship between content-related validity and reliability. In many cases the better the content validity is, the lower is the reliability. Although reliability is easiest to explain in terms of repeatability and equivalence of performance across parallel forms, in practice, however, as we have already explained above, most of the tests of interest are administered only once. We must, therefore, rely on single administration methods that assess internal consistency (coefficient alpha, KR20, KR21). If the content to be covered were broad, a content

[9] In the case of content validity, high reliability is not a necessary condition as we will shortly come to.

valid test would be less reliable than a test that lacked content validity because it only assessed part of the content. Consider a communicative language proficiency diagnostic test used for placement purposes that is supposed to cover grammar, vocabulary, reading comprehension (both general and specific), writing, speaking, and listening comprehension. A test that covered all of this content is likely to be less reliable (have lower item intercorrelations) than a test that was not content valid because it only covered grammar in a multiple-choice format.

Finally, it is important to remember the tenets of the *"attenuation paradox"* in reliability theory regarding the number and type of items and test score reliability. The gist of the attenuation paradox is that if test items are selected that are too homogeneous, reliability may increase but at the expense of validity. That is, the resulting test may measure a very narrow construct that has limited relations with a limited set of other narrow variables. In contrast, a test with items that are less homogeneous but which are from the same construct domain may have more moderate reliability, but increased validity. Of course, a test that is a hotchpotch of items from different construct domains would more or less be useless. Thus, there are times when a decrement in reliability, through the selection of less homogeneous items within a construct domain, can produce more valid measures.

2.5.2 Test Method Effects

The characteristics of test methods, which influence test performance, have long been studied by many researchers in language testing. Research has shown that test performance varies as a function of an individual's language ability and the characteristics of test methods. Some test takers, for example, might perform better in the context of a laboratory speaking to a microphone than they would in front of a panel of judges in an oral interview. Some test takers might find it easier to choose responses from among alternatives in a multiple-choice test of vocabulary than to complete an open-ended cloze format of a similar test. Completion of isolated sentences as opposed to completion of blanks in a text, live versus recorded speech, aural in contrast to written tests are but few examples of how test methods may vary. These characteristics of test methods may, in turn, influence the test performance, casting doubt on the reliability and validity of language tests. Controlling these characteristics, thus, becomes an important issue in the theory and practice of language testing.

The study of test methods dates back to 1959 when Campbell & Fiske (1959) illustrated that method variance might influence the measures of a construct. They argued that a hypothetical large correlation between two traits of say, A and B, and no correlation between traits A and C might be a function of method variance common to the measures A and B and not to C, if the measures A and B are obtained by one method and that of C by another method. To control the method effect, they proposed a Multitrait-Multimethod (MTMM) design for validating tests. The main focus of the MTMM design is to separate trait and method factors. It recognises that "*any test score is a function of both the trait it intends to measure and of the method by which it is measured*" (Bachman & Palmer, 1979, p. 54). Therefore, the method involved in measuring might become as important as the trait it is intended to measure. Table 2.8 illustrates how methods and traits interact in a hypothetical Multitrait-Multimethod matrix.

Table 2.8: Hypothetical Multitrait-Multimethod Matrix

(From Campbell & Fiske, 1959:82)

		Method 1			Method 2			Method 3		
	Traits	A_1	B_1	C_1	A_2	B_2	C_2	A_3	B_3	C_3
Method 1	A_1	(.89)								
	B_1	.51	(.89)							
	C_1	.38	.37	(.76)						
Method 2	A_2	**.57**	.22	.09	(.93)					
	B_2	.22	**.57**	.10	.68	(.94)				
	C_2	.11	.11	**.46**	.59	.58	(.84)			
Method 3	A_3	**.56**	.22	.11	**.67**	.42	.33	(.94)		
	B_3	.23	**.58**	.12	.43	**.66**	.34	.67	(.92)	
	C_3	.11	.11	**.45**	.34	.32	**.58**	.58	.60	(.85)

Letters A, B, C refer to traits, subscripts 1,2,3 to methods. Validity coefficients (monotrait-heteromethod) are the three diagonal sets of boldface numbers; reliability coefficients (monotrait-monomethod) are the numbers in parentheses along principal diagonal. Solid triangles enclose heterotrait-monomethod correlations; broken triangles enclose heterotrait-heteromethod correlations.

According to MTMM design, to observe the validity of a test, that is, to see whether the test is measuring what it purports to test, the application of more than one method seems necessary. If independent methods testing the same construct do tend to correlate highly, it is concluded that convergent validity is achieved. On the other hand, to achieve discriminant validity, i.e., to show that there are independent traits irrespective of the methods applied, introduction of more than one trait in the analysis is necessary. Low correlation between different traits indicates that they are really different from one another and hence discriminant validity is achieved.

As it stands, independence of methods is an important issue in validity as well as reliability studies. Convergence of *independent methods* claiming to test similar constructs is a proof of the validity of a test. However, in the case of reliability, convergence of similar *methods* is indicative of the reliability of the test. Since independence is a matter of degree, it may be concluded that reliability and validity can be considered to be on a continuum, depending on the degree of independence of test methods. To put it in other words,

> *"Reliability is the agreement between two efforts to measure the same trait through maximally similar methods. Validity is represented in the agreement between two attempts to measure the same trait through maximally different methods."* (Campbell & Fiske, 1959, p. 83)

The MTMM design of Campbell and Fiske was influential for those interested to know whether the techniques the testers use distort the results that they obtain. Bachman & Palmer (1981), for example, used a complex MTMM research design to investigate the comparative influences of two traits (Speaking and Reading) and three methods (Interview, Translation and Self-rating). They found that scores from self-ratings loaded consistently more highly on method factors than on specific trait factors, and that translation and interview measures of reading loaded more heavily on method than on trait factors. The researchers obtained similar results in another study. Bachman & Palmer (1982) found that scores from both self-ratings and oral interviews consistently loaded more heavily on test method factors than on specific trait factors, while the scores from the multiple-choice and writing tests were least affected by method factors.

A number of other studies have also examined the effect of test methods on test performance. Alderson (1978) studied the effect of different deletion procedures on

cloze tests and showed that different deletion rates on a given test as well as different text difficulties of the same test influenced the results obtained. Bachman (1982) illustrated that fixed-ratio or rational deletion procedures affected the results obtained. Chapelle & Abraham (1990) also studied four different methods of cloze testing on a given text. Their study describes results obtained from altering trait and method facets of the cloze procedure while holding text and student ability constant. The findings suggested that different methods had striking effect on the difficulty level of the texts, predicting the fixed-ratio as the most difficult and the multiple-choice as the easiest of these methods.

Other studies have also shown that multiple-choice format is the most transparent method for learners. Shohamy (1984), for instance, found that multiple-choice tests of reading were easier than open-ended tests for L2 learners. Lewcowicz (1983), as Skehan (1989) reports, mentions similar results concerning multiple-choice format as opposed to open-answer format, concluding that different methods seemed to be measuring somewhat different things. The latter conclusion, nevertheless, seems to question the validity of language tests. Whether different methods measure different things or measure the same thing in a different way, the influence of test methods on test scores cannot be ignored.

2.5.2.1 Characteristics of test methods

What are the characteristics of test methods? The facets of test methods can be viewed from different perspectives. There have been a number of descriptions of test method facets over the years. Prior to the 1980s psychologists referred to *stimulus* and *response* when describing the characteristics of language testing methods. Carroll (1968), for example, has discussed four general types of language test tasks required in individual language test items in terms of differences in their stimulus-response characteristics. Discussing '*test modalities*', Clark (1972) uses the term *stimulus* to describe any written or spoken material presented to the student in a test situation. He associates *response* with *"any physical activity on the part of the student in reaction to the stimulus materials"* (p. 27), and classifies them into two types: *free and multiple-choice*. Since 1980s with the shift in theoretical fashion towards communicative testing, there has been a shift in testing methods as well. Weir (1983), for example, proposes a framework for describing a '*communicative test event*'. This framework comprises three sections: *general descriptive parameters of*

communication, *dynamic communication characteristics*, and *task dimensions*. The first section includes activities, setting, and dialect, while the second includes realistic context, relevant information gap, and normal time constraints. Finally, task dimensions refer to the amount of communication involved, functional range, and referential range.

2.5.2.2 Bachman's Facets Of Test Methods

The most comprehensive framework for studying the facets of test methods has been proposed by Bachman (1990, p. 119). His framework comprises five main categories: facets of the testing environment, facets of the test rubric, facets of the input, facets of the expected response, and relationship between input and response. Figure 2.5 illustrates Bachman's categories of test method facet. The large number of dimensions along which test methods vary in this framework are reflections of the variety of testing techniques that are used in language tests, and the ways in which these techniques vary.

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 organisation
- 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, specialisation)
  - Degree of contextualisation (embedded/reduced)
  - Distribution of new information (compact/diffuse)
  - Type of information (concrete/abstract, positive/ negative, factual/counter-factual)
  - Topic
  - Genre
- Organisational characteristics
  - Grammar
  - Cohesion
  - Rhetorical organisation
- Pragmatic characteristics
  - Illocutionary force
  - Sociolinguistic characteristics

4 FACETS OF THE EXPECTED RESPONSE
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, specialisation)
  - Degree of contextualisation (embedded/reduced)
  - Distribution of new information (compact/diffuse)
  - Type of information (concrete/abstract, positive/ negative, factual/counter-factual)
  - Topic
  - Genre
- Organisational characteristics
  - Grammar
  - Cohesion
  - Rhetorical organisation
- Pragmatic characteristics
  - Illocutionary force
  - Sociolinguistic characteristics

Restrictions on response
- Channel
- Format
- Organisational characteristics
- Prepositional and illocutionary characteristics
- Time or length of response

5 RELATIONSHIP BETWEEN INPUT AND RESPONSE

Reciprocal

Nonreciprocal

Adaptive

Figure 2.5: Bachman's (1990:119) Facets Of Test Methods

Facets of the testing environment are the characteristics of the test relating to the conditions under which the test is administered and might affect the performance of

test takers in any given test. They include familiarity of the place and equipment used in administering the test, the personnel involved, the time of testing, and physical conditions under which the test is administered. The facets of the test rubric relate to the specification of the way in which test takers are expected to proceed in taking the test. They are test organisation, time allocation, and instructions. The facets of the input and the expected response are the characteristics of the input and expected response, which might affect the performance on language tests. *Input*, as Bachman (1990) defines it, "*consists of the information contained in a given test task, to which the test taker is expected to respond*" (p. 125). It is viewed from two dimensions: the *format* of the input such as channel, mode, form, vehicle and language of presentation, together with the identification of the problem and the degree of speededness to read the input; and the *nature of language* referring to the length, propositional content, organisational characteristics, and pragmatic characteristics. The facets of the expected response are slightly more complex as there are restrictions on the response. These restrictions apply to the channel, format, organisational characteristics, propositional and illocutionary characteristics, and time or length of response. There are of course some factors such as language ability to be measured in expected response, which are beyond the test developer's control and thus cannot be incorporated in the facets of the test methods.

Finally, *the relationship between input and response* is the last category in Bachman's model, which must be taken into account when examining the effect of test methods on the performance on language tests. The relationship can be *reciprocal* as in an interview where the test taker can get feedback from the examiner, or *nonreciprocal* where there is no feedback as in the test of receptive skills, or it can be *adaptive*, that is, the input is influenced by the response but without the feedback of the sort of nonreciprocal one, i.e., computer adaptive testing.

Bachman's framework has been used for examining various dimensions, or facets of test methods in a large-scale study, namely, Cambridge - TOEFL Comparability Study (Bachman, Davidson, Ryan & Choi, 1995). The latter study concludes that different methods as diverse as Cambridge and ETS test batteries not only tap, to a large degree, similar abilities of the subjects in the sample concerned but also measure these abilities in much the same way.

2.6 A Communicative Framework For The Comparison Of Language Proficiency Tests

In this section we will propose a framework for the comparison of language proficiency tests. The framework is based on the examination of content comparability of language tests and has two main instruments: test method facets and communicative language ability components. This framework is a modified version of the one used in the CTCS; the modification being made partly on the basis of ongoing research in the UCLES and partly on the limitation of research resources. The proposed framework will allow us to compare 34 different aspects of language proficiency tests; 21 facets of test methods and 13 components of communicative language ability. The detailed operational definitions of the facets and components will be explained later in 3.7.

It is worth mentioning here that test methods characteristics are limited to those aspects of test methods that are related to the *nature of language* in Bachman's (1990) *facets of the input*. Other facets such as the *testing environment* will be discussed separately in 3.5.

2.6.1 Test Method Characteristics

Test methods are restricted to those features of the *input* (test items), which directly influence the test content. They comprise *length, propositional content,* and *organisational characteristics. Length* has usually an effect on other characteristics of test methods and may eventually influence the performance. Longer text, for instance, results in more grammatical components (content words, clauses, embeddings, passives, and cohesive devices) and a heavier load of information for the test taker to process, which could contribute to task difficulty. Examining this variable, thus, becomes an important issue when comparing test contents.

2.6.1.1 Propositional Content

Propositional content is described as the characteristics of the information in the context and in the discourse: it includes *degree of contextualisation, distribution of information, type of information, topic* and *genre*. Bachman (1990) also includes *vocabulary* in his category of propositional content but it was excluded from the framework here because one of the tests (IELTS) lacked an independent vocabulary section.

Degree of Contextualisation

Contextual information can be described as the familiar or known information within a context that is relevant to the information expressed in the discourse. Contextual information or *context embeddedness* (Cummins, 1983) may be supported by a variety of linguistic and paralinguistic cues in the context. The discourse is said to be contextualised if the information expressed in the context is known to the reader whether by the immediate physical context, by the information in the input language or by the progression of the discourse itself and the language used is said to be context-embedded. If, however, the discourse is full of new information, the language used may be called context-reduced. The degree of contextualisation can therefore be expressed in terms of the ratio of contextual information to new information in the discourse. The greater the ratio of contextual information to new information, the higher the degree of contextualisation will be.

Although contextualisation can greatly influence the propositional content of a test item, it is itself subject to variations depending on the world knowledge of the testees. On the one hand, the more familiar and relevant contextual information is in the input, the more likely that test taker will be able to respond to the propositional content of the discourse. On the other hand, the test taker's ability to interpret and respond to the propositional content of the discourse depends not only on the amount of context in the input but also on the amount of world knowledge (schemata) she is able to activate. As the amount of this presupposed knowledge varies from one test taker to another, the degree of contextualisation of the input also varies and might favour one particular test taker over others. That is why in designing general language

proficiency tests utmost care should be taken to avoid writing texts of a highly technical nature that might favour any particular group of test takers.

Distribution of New Information

The distribution of new information that must be processed so that a test taker can successfully complete a given test task can make a task more or less difficult. Discourse can be *compact* or *diffuse* (Bachman, 1990) depending upon the distribution of new information over time or space. For example, the discourse of a speeded reading comprehension may be called compact because the distribution of new information over time and space is relatively short; there are too many questions to be answered in a very short time. If, however, the distribution of new information over time or space is relatively long, the discourse may be called diffuse, as it is the case of a listening comprehension where the test taker is required to summarise a lengthy lecture.

Highly compact or diffuse discourse may be quite difficult to process and demanding of the test taker's competence. For instance, in a speeded listening comprehension, the test taker has little time to consider her answers before the next question begins. She is sometimes called beyond her comfort and may have very little opportunity to negotiate meaning, or may have to negotiate meaning very quickly. Such highly compact discourse is very demanding of her competence. Equally demanding can be a diffuse discourse where the test taker is required to summarise a lengthy lecture in a relatively short time. In doing that, the test taker has to constantly process the information and keep it in her memory for further recall.

It may be concluded that the difficulty or the ease of a test task, among other things, depend on the compactness or diffuseness of the new information within the discourse.

Type of Information

The type of information in a test task can be classified along two dimensions: concrete / abstract, negative / positive. Abstract information is here referred to as the kind of information whose mode of representation is primarily symbolic or linguistic, whereas the concrete information mode of representation can take other forms as well. Most mathematical concepts, for instance, are considered to be abstract as there are usually

no ways to represent them other than using the symbolic language of formulae. Other instances of abstract information can be moral concepts such as equality and humbleness, which are even more difficult to represent, and are almost always associated with a short story or a parable. In contrast, concrete concepts such as *Bentley* can be represented in different modes: *visually* by showing a picture of a Rolls Royce, *linguistically* by means of an expression "a four-wheeled aristocratic automobile", and *auditorily* by listening to the sound of a Rolls Royce engine. It is commonly believed that input with abstract information is more demanding of the test taker's competence than one with concrete information.

Type of information can also be classified according to the degree to which it is negative. If there is no negative element in the information, it is considered as positive, as in "Mr Smith went to school". Negative information can vary in its negation both on the level and the number of negative markers used. For example, both sentences "I didn't ask her to go to the school" and "I asked her not to go to the school", carry only one negative element but they differ in the level of negation: the former carries the negation in its matrix sentence while the latter includes the negation in its embedded sentence. It is also possible to have negative elements both in the matrix and in the embedded sentence at the same time, as in "I didn't intend not to complete the job". Increasing the amount of negative information in the input can make the task more difficult to process.

Topic

Generally speaking *topic* refers to what a given piece of discourse is about. Test items are usually written with the view that the subject matter is interesting and relevant to test takers. In most cases, test writers try to avoid subjects, which might presuppose a certain background knowledge on the part of the testees, instead, they tend to use neutral topics to minimise the potential differences in test takers' background knowledge. There is a dilemma here for the test writers as how to write test items whose subjects are neutral, yet interesting and relevant to the testees. On the one hand, a neutral and broad topic might be quite unengaging to test takers and question the relevance of the test items to the population tested. On the other hand, familiar topics may seem to be interesting and engaging the test takers, but they may also favour a particular group of testees over others and hence be a major source of test bias. The solution seems to be the selection of a variety of topics, which would lessen the test bias and increase the engagement of the testees.

Genre

Genre is here referred to as the types of language tests that have some identifiable patterns (i.e., multiple-choice, essay, cloze and dictation). These patterns can activate certain expectations in test takers, hence facilitating the task of test taking for those individuals familiar with the patterns, while making it more difficult for those unfamiliar with them. Additionally, it can be argued that if the language of an input is characteristics of a particular genre, tasks depending upon the interpretation of that genre would be relatively difficult for those test takers unfamiliar with it. For example, writing a business letter applying for a job in a firm could be quite difficult for an individual who has never worked in an office before, whereas completing the same task could be relatively easy for someone who has had work experience. Writing a business letter requires an understanding of the office environment and the expectations associated with it, which presume that the test taker is already familiar with the task. The purpose of such tasks could well be testing how competent the test taker is in responding to the input, i.e. writing the letter. This is especially true in the case of ESP testing where the language test is going to be used as the predictor of how well the testee will cope with the tasks in a particular field. Since in communicative testing variability of tasks plays an important role in the formation of the test battery, care must be taken to ensure that tasks associated with particular genres are familiar to the test takers.

2.6.1.2 Organisational Characteristics

Bachman (1990) in discussing communicative language ability mentions that organisational competence is related to the abilities *"controlling the formal structure of language for producing or recognising grammatically correct sentences, comprehending their propositional content, and ordering them to form texts"* (p. 87). Organisational characteristics can likewise be described as those features of the discourse, which relate to the formal organisation of language. This formal organisation is of three types: grammar, cohesion and rhetorical organisation.

It is implicit from the term '*organisation*' that the length of the language sample is a determining factor in the amount of incorporation of organisational characteristics

required for a successful interpretation of the language. As a general rule, the longer the language sample, the greater the need to incorporate organisational characteristics to make it interpretable. It follows that inputs with short single sentence items involve very little organisation, whereas inputs with long reading passages or lectures may involve the full range of organisational characteristics. For example, in a simple vocabulary-matching item, there is very little need for the incorporation of the organisational characteristics. In contrast, writing a short summary of a long lecture will involve the full range of the organisational characteristics from different aspects of grammar and cohesion to the rhetorical structure of the language used. The *length* of a piece of language, then, may often be associated with the involvement of the organisational characteristics. This is especially true when the input requires a constructed response, as is the case in a reading comprehension where the test taker must choose the *best* response from among several answers; the input material in such a case is generally accurately organised in terms of grammar, cohesion, and rhetorical organisation.

However, that is not to say that inputs, which require a selected response, cannot test organisational characteristics. An example would be an input, which involves the recognition of grammatical errors in a sentence where several words or phrases are underlined and the test taker is required to identify the part that makes the sentence ungrammatical.

From what has been explained above, one may conclude that organisational characteristics are important features of any communicative testing. They have to be carefully examined in content analysis of language tests.

2.6.2 Language Ability Components

As mentioned earlier, communicative language ability in Bachman's (1990) framework includes three components: language competence, strategic competence and psychophysiological mechanisms. The latter deals with neurological and physiological processes of language use and is, therefore, beyond the scope of this study. Language ability components are restricted here to language competence and strategic competence.

2.6.2.1 Language Competence

Language competence comprises a wide range of linguistic and paralinguistic features which are utilised in an effective communication through language. These features vary from the ability to control the formal structure of language, such as, producing correct grammatical sentences, to the abilities to perform language functions in the contexts of communication.

Munby (1978), Canale & Swain (1980), Hymes (1982), Bachman & Palmer (1982), Canale (1983), Allen et al. (1983) and Bachman (1990) have all attempted to describe a theoretical framework for specifying the components of language competence in a language use situation. In this research, language competence, following Bachman's taxonomy, is divided into two higher order components: organisational competence and pragmatic competence.

Organisational Competence

Organisational competence relates to the abilities that control the formal organisation of language. The abilities to produce and comprehend grammatical sentences, understand their propositional content, and sequence them to form meaningful texts are all manifestations of this competence. Organisational competence is considered to be operative at two different levels: sentence and text. The competencies, which are involved at each level, are called grammatical and textual competence, respectively.

Grammatical Competence

Grammatical competence refers to the abilities involved in language *usage* (Widdowson 1978), which, as Bachman puts it,

> *"govern the choice of words to express specific significations, their forms, their arrangement in utterances to express propositions, and their physical realisations, either as sounds or as written symbols."* (Bachman, 1990, p. 87)

These abilities include knowledge of vocabulary, morphology, syntax, and phonology / graphology. Each ability is relatively independent. For example, lexical competence

demonstrates test taker's choice of appropriate words; morphological knowledge demonstrates the affixing of inflectional morphemes; syntactic knowledge demonstrates the proper order of words in composing a sentence; and finally the phonological competence illustrates how accurate the representation of the language is according to phonological rules. Grammatical knowledge is, then, restricted to that competence which is involved in producing and recognising sentences.

Textual Competence

Textual competence is similar to grammatical knowledge in that it is involved in controlling the formation and recognition of pieces of language, only that it is usually operative at a higher level than a sentence. Textual competence is the ability to structure independent sentences and utterances to form a text - a larger unit of language - according to the conventions of cohesion and rhetorical organisation. Cohesion refers to explicitly marked relationships across clauses within the same sentence or across sentences. This explicit marking may be in the form of lexical connectors or of specific grammatical patterns that provide appropriate topicalisation. Types of cohesion include reference, substitution, ellipses, conjunction and lexical cohesion (Halliday and Hassan 1976). Rhetorical organisation is generally related to the overall effect of a text on a language user. That is, how the overall conceptual structure can impress the reader and influences them for or against an opinion or a certain course of action. There are numerous conventions of rhetorical organisation. Some of them are formally taught in expository writing classes such as narration, description, comparison, and classification. Other conventions are either too complex to teach or too difficult to understand. Conventions of rhetorical organisation are limited here to only common methods of development: narration, description, comparison, classification, argumentation, and process analysis.

Pragmatic Competence

All the characteristics studied so far are related to the ways in which linguistic signals are recognised and how they are used to refer to persons, objects, etc. Equally important in any communicative language use is the ability to produce and understand sentences or utterances that are *appropriate* to the *context* in which they occur. This ability, which subsumes the things one needs to know in order to communicate effectively in different social settings, is called pragmatic competence. Bachman maintains that pragmatics is concerned with,

> *"the relationships between utterances and the acts or functions that speakers (or writers) intend to perform through these utterances, which can be called the* illocutionary force *of utterances, and the characteristics of the content of language use that determine the* appropriateness *of utterances."* (1990, p. 89)

Pragmatic competence, then, comprises two abilities: the ability to understand the pragmatic conventions for performing acceptable language functions or *illocutionary competence* and the ability to understand the sociolinguistic conventions for performing language functions appropriate in a given context - *sociolinguistic competence*.

Illocutionary Competence

Illocutionary competence pertains to the ability to use language functionally, or to perform speech acts, or language functions. These functions are numerous but can be grouped into four macro functions: *Ideational, Manipulative, Heuristic,* and *Imaginative,* each of which will be described briefly below. Language functions can be performed with varying degrees of directness, ranging, for example, from very direct forms such as "I request you to close the door," to less direct forms such as "I wonder if someone could close the door," and "It's really cold here." Two factors can generally be considered to determine the level of ability required to interpret the realisation (utterance, sentence, text) of a given function: 1) the amount of information in and complexity of the function and 2) the degree of directness or indirectness with which it is expressed. Thus "basic" realisations of functions would be those that express simple functions and which realise the functions quite directly, while "advanced" realisations are those that express complex functions or that state the functions very indirectly.

Ideational Function

The ideational function is perhaps the most pervasive function by which we communicate information - ideas or feelings. To put it in other words, it is a function by which we express meaning in terms of our experience of the real world (Halliday, 1973, p. 20). For example, language is used ideationally to present knowledge in textbooks or scholarly articles.

Manipulative Function

The primary purpose of the manipulative function is to affect the world around us. That is, the use of language to get various things done. The realisation of manipulative functions can include various other functions. For example, we may use language to get things done by forming or uttering suggestions, requests, or commands. Such use of language can be called *instrumental* function. Other types of manipulative functions are *regulatory* and *interactional* functions. The former refers to the use of language to control the behaviour of others, or to manipulate persons in the environment while the latter refers to the use of language to form, maintain or change interpersonal relationships.

Heuristic Function

The heuristic function is by far the most common function in the act of teaching, learning, and problem solving where language is used to extend our knowledge of the world around us. An exemplification of this function in teaching a language could be the use of utterances such as *"Mary goes to school every day"* not to convey information about a fact (ideational function) but rather to illustrate a grammatical point (subject-verb agreement).

Imaginative Function

The imaginative function pertains to those aspects of language use, which enable us to create our own environment for humorous or aesthetic purposes. Some of the examples are telling/listening to jokes, creating/interpreting figures of speech, and reading literary works for enjoyment. The purpose of all the above activities is to appreciate the way in which the language itself is used.

Although all the above four functions are distinct instances of language use, they, by and large, fulfil several functions simultaneously.

Sociolinguistic Competence

Sociolinguistic competence refers to that aspect of pragmatic competence that enables us to see the appropriateness of a language function to a particular social context. While illocutionary competence enables us to perform different functions and

understand the illocutionary force underlying such functions, the appropriateness of the use of functions to individual social context is influenced by a variety of sociolinguistic characteristics of language use which relate to such matters as the social attitudes to language, standard / non-standard forms of language, and social varieties and levels of language. The two most influential sociolinguistic factors that are generally linked to the context in which language use take place are dialect and register.

Dialect

There are dialects in virtually every language depending upon the geographical location of the language users or the social class with which the users are associated. These variations in language use can be classified into formal / informal, black / standard, or Southern / Northern dialects, each of which has distinct language functions relevant to their social context. For example, the use of an informal greeting in black American English dialect among black peers is considered as appropriate language use, whereas the use of the same language in a different social context such as that in a high class golf club peers may be considered quite inappropriate. Understanding different language dialects or language varieties, hence, is an essential part of the sociolinguistic competence.

Register

Register is here referred to as a variety of language defined according to its use in social situations. Examples include, formal, intimate, and casual registers.

2.6.2.2 Strategic Competence

Strategic competence as described by Bachman is *"the capacity that relates language competence ... to the language user's knowledge structures and to the features of the context in which communication takes place"* (1990, p. 107). It performs three functions of assessment, planning, and execution. In CTCS, strategic competence referred to the mental capacity that enables language users to implement the components of language competence listed under 2.6.2.1 in contextualised communicative language use. That is, strategic competence enables language users to relate the features of the language use context to the intended illocutionary force, to

relate an utterance form to the features of the context, to interpret its illocutionary force appropriately. Strategic competence includes the skills associated with test preparation such as how to read the questions relating to a text or apportioning a suitable time to each sub-section of the test. It seems to be operationally more transparent for us to associate it with test wiseness. We will therefore, refer to strategic competence as that capacity of the language user's competence that controls the skills associated with test wiseness.

Chapter Three

Design of the Study

3. Design of the Study

This section sets out the design of the study and explains the rationale, objectives, research questions, the selection of subjects, test administrations, marking of the tests, and the procedures of investigation.

3.1 Rationale

There is a general belief that British and North American EFL proficiency tests represent radically different approaches to language test development. The North American tradition in language testing is heavily based on psychometric properties of tests such as reliability, concurrent and predictive validity. Objectivity of scoring and generalisability of the results play a dominant role in the development of test methods in this tradition. For example, multiple-choice items are often used in testing receptive skills to gain desired internal consistency, even if the test is expected to measure communicative competence, as is the case in *Functional Testing* (Farhady, 1980). In the British tradition, however, the emphasis is on the specification of test content and expert judgement. Whereas in this tradition less attention is paid to reliability (degree of generalisability of the results), content and face validity win the major concerns of test designers. It may follow that British tests enjoy more variability in their formats and include various communicative activities.

Language proficiency tests, in both the American and British tradition, are designed to serve different purposes, so they may not be comparable in terms of defined purposes. However, when we use the term '*language proficiency*', no matter how it is defined, we are referring to a single concept. It follows that tests of language proficiency may all tap the same underlying construct, albeit how imperfectly they may do so. In the real world, test results are often used for screening purposes; the candidates' *ability* to cope with the future language medium is predicted by the proficiency criterion. As we have already mentioned in chapter two, the components of most proficiency tests tend to load heavily on similar factors. Although differences in results do exist in different tests indicating the effect of test methods and test takers' characteristics, the large similarities among them suggest that not only do these tests measure, to a large degree, the same underlying construct, but also they may do it in much the same way.

If it is the case that language proficiency tests are used for similar purposes, i.e., measuring the general ability of the candidates in coping with the future language medium, comparability of such tests should be a legitimate area for investigation.

Comparability studies provide us with an excellent opportunity to examine various aspects of language proficiency tests. By comparing different test batteries one may increase the degree of generalisation that one may draw from studying any single test battery. Moreover, critical observation of test takers' performance on various tests can help us further our understanding of the relationship(s) between the test methods and other relevant factors of test score variance such as language ability and test takers' characteristics, enabling us to progress in our attempt to define the concept of language proficiency. Additionally, comparability studies can give us insight into a better understanding of the underlying constructs of the tests which in turn might promote our understanding of the concept of language proficiency. The information hence gathered might help us examine the validity of construct interpretations. For example, in the case of the Cambridge-TOEFL Comparability Study (CTCS), the finding that instruments as diverse as those examined in the study tap virtually the same sets of language abilities might influence the future structure of language proficiency tests.

3.2 Objectives

As we have already discussed in 2.4.3.2, the Cambridge-TOEFL Comparability Study final report was criticised, among other things, by Davies (1989) for the wrong choice of tests for comparison:

> *"the CTCS was centrally flawed because it made the wrong comparison. The FCE/CPE tests are not sensibly compared with TOEFL. Far better to compare TOEFL with ELTS in a meaningful comparison."* (Davies, 1989, p. 6)

The same view seems to be shared by Alderson (1989) when he comments on one of the important findings of the study- *the finding that the two test batteries show remarkable similarities in their factor structures and therefore can be said to measure a single language ability*- and maintains that:

> *"I am led to the conclusion that the wrong test was used in the comparison. What would certainly throw more light on this interesting matter would be to conduct further research that included the IELTS test, which is intended to be directly relevant to the same population as the TOEFL and which claims to measure relevant and different abilities. If a comparison of TOEFL / SPEAK / TEW / CPE / IELTS (with reliable test scores) continues to reveal only one general factor, then we will indeed have made an interesting finding."* (Alderson, 1989, p. 8)

It was due to the above calls for concern that the present study was formed. This study compares two English language proficiency test batteries: TOEFL and IELTS. TOEFL is perhaps the most reliable and researched test of English language proficiency. A satisfactory score on TOEFL is one of the prerequisites for entering most academic institutions in North America, Britain, and Australia for non-native speakers of English. The test is based on psychometric techniques of language testing resembling a typical North American test. IELTS, originally meant to be an ESP test, has been welcomed by most British and Australian universities as it claims to sample real academic challenges, which the candidates are to face in their field of study. It is based on a communicative approach to language testing and has achieved high face

validity as well as producing acceptable reliability figures. Both tests are tests of language proficiency and are thus not based on any specific syllabus.

The main objectives of the research are to investigate the extent to which TOEFL and IELTS are comparable in terms of:

- the operational definitions of language proficiency on which the two tests are based. This involves examining the test method facets in the two batteries as well as performing statistical analyses on the test scores of a group of subjects on these two batteries.
- the degree to which the two tests provide similar information concerning the *abilities* of the testees. This involves investigation into the construct validity of the two tests. The answer to this question will be derived from both the qualitative analysis of the contents of the tests as well as the quantitative analysis of candidates' performance on the two tests. The former is achieved through examining the components of communicative language ability in each test. The latter is achieved by comparing the patterns of relationships of test results across the two batteries.

3.3 Research Questions

To examine the above objectives, firstly, one needs to investigate the content comparability of the two tests, which involves comparing the facets of test methods and the components of communicative language ability across the two tests. Secondly, one needs to study the performance of a group of subjects on these two tests and analyse their test scores for any meaningful relationship that might emerge across the two tests. Finally, it is important to study the impact of test preparation on test performance. Investigating all these objectives requires empirical research questions. Therefore, it was decided to form different sets of questions for examining test method facets, communicative language ability components and the impact of test preparation. In each set we may have one or more main questions (**1, 2, ...**) related to a class of components or facets and a few minor questions (**2.1, 2.2, ...**) related to the specific components or facets concerned in a category. Each question (**Q**) will be followed by a null hypothesis ($\mathbf{H_0}$) and an alternative hypothesis ($\mathbf{H_1}$).

The general descriptions of test method facets and communicative language ability components have already been explained in 2.6. Their operational definitions will be explained shortly in 3.7.1.1 and 3.7.1.2. We will first examine the questions related to test method facets. The following research questions, null hypotheses, and alternative hypotheses were formed for examining the facets of test methods.

3.3.1 Questions Related to Test Method Facets

Twenty one different facets of test method have been defined so far in this research which are classified under four general headings of length, propositional content, organisational characteristics, and relationship of item to passage, all of which will be operationally defined in 3.7.1.1. Thus we can have four empirical research questions examining these broad categories of test methods. Since propositional content and organisational characteristics have various sub-components, we need to formulate further research questions to investigate their different aspects. Thereby, we will have four major research questions and eight minor questions relating to test method facets.

Length

Length has been measured here as the total number of words in a passage.

Q 1: Are the reading passages in the two batteries significantly different in terms of their length?

H_0 1: The reading passages in the two batteries are not significantly different in terms of their length.

H_1 1: The reading passages in the two batteries are significantly different in terms of their length.

Propositional Content

Propositional content relates to the characteristics of the information in the context and in the discourse and comprises degree of contextualisation, distribution of new information, type of information, and topic.

Q 2: Are the reading passages in the two batteries significantly different in terms of their propositional content?

H_0 2: The reading passages in the two batteries are not significantly different in terms of their propositional content.

H_1 2: The reading passages in the two batteries are significantly different in terms of their propositional content.

Organisational Characteristics

Organisational characteristics are those features of the discourse, which relate to the formal organisation of language.

Q 3: Are the reading passages in the two batteries significantly different in terms of their organisational characteristics?

H_0 3: The reading passages in the two batteries are not significantly different in terms of their organisational characteristics.

H_1 3: The reading passages in the two batteries are significantly different in terms of their organisational characteristics.

There are two main dimensions of formal organisation: grammar and cohesion. Grammar will be assessed in terms of syntactic complexity, lexical density, and text difficulty while cohesion will be assessed in terms of number of cohesive markers in the passages. Question 3 can therefore be extended into further minor questions.

Q 3.1: Are the reading passages in the two batteries significantly different in terms of their syntactic complexity?

H_0 3.1: There is no significant difference in the syntactic complexity of the reading passages in the two batteries.

H_1 3.1: There is a significant difference in the syntactic complexity of the reading passages in the two batteries.

Q 3.2: Are the reading passages in the two batteries significantly different in terms of their lexical density?

H_0 3.2: There is no significant difference in the lexical density of the reading passages in the two batteries.

H_1 3.2: There is a significant difference in the lexical density of the reading passages in the two batteries.

Q 3.3: Are the reading passages in the two batteries significantly different in terms of their text difficulty?

H_0 3.3: There is no significant difference in the text difficulty of the reading passages in the two batteries.

H_1 3.3: There is a significant difference in the text difficulty of the reading passages in the two batteries.

Q 3.4: Are the reading passages in the two batteries significantly different in terms of their cohesive markers?

H_0 3.4: There is no significant difference in the number of cohesive markers in the reading passages of the two batteries.

H_1 3.4: There is a significant difference in the number of cohesive markers in the reading passages of the two batteries.

Relationship of Item to Passage

Reading comprehension questions can be rated for the relationship they have to the reading passages.

Q 4: Are the relationships of the test items to the reading and listening passages significantly different in the two batteries?

H_0 4: There is no significant difference in the relationships of the test items to the reading and listening passages in the two batteries.

H_1 4: There is a significant difference in the relationships of the test items to the reading and listening passages in the two batteries.

3.3.2 Questions Related to the Components of Communicative Language Ability

As explained in 2.6.2 communicative language ability components are restricted here to language competence and strategic competence. Language competence comprises a wide range of linguistic and paralinguistic features which are utilised in an effective communication through language. It has two major components: organisational competence and pragmatic competence, all of which will be operationally defined in 3.7.1.1. Like the facets of test method, components of communicative language ability do have some sub-components. That means, in addition to three broad empirical research questions, we will have as many as eight minor questions related to communicative language ability components.

Organisational Competence

Organisational competence relates to the abilities that control the formal organisation of language. It involves two other components of grammatical and textual competence, which are operative at two different levels of sentence and text.

Q 5: Do the test items in the two batteries involve the same degree of organisational competence[1] for successful completion of a given task?

> ***H_0 5:*** *There is no significant difference in the degree of organisational competence required for successful completion of the test items in the two batteries.*
>
> ***H_1 5:*** *There is a significant difference in the degree of organisational competence required for successful completion of the test items in the two batteries.*

Grammatical competence includes the knowledge of vocabulary, morphology, syntax, and phonology / graphology. Textual competence includes the ability to structure independent sentences and utterances to form a text according to the conventions of cohesion and rhetorical organisation. To investigate the comparability of the above sub-components, Question 5 could be broken down into the followings:

Q 5.1: Do the test items in the two batteries require the same degree of lexical knowledge for successful completion of a given task?

> ***H_0 5.1:****There is no significant difference in the degree of lexical knowledge required for successful completion of test items in the two batteries.*
>
> ***H_1 5.1:*** *There is a significant difference in the degree of lexical knowledge required for successful completion of test items in the two batteries.*

Q 5.2: Do the test items in the two batteries require the same degree of morphological knowledge for a successful completion of a given task?

[1] As operationally defined in 3.7.1.2.

H_0 5.2: There is no significant difference in the degree of morphological knowledge required for successful completion of test items in the two batteries.

H_1 5.2: There is a significant difference in the degree of morphological knowledge required for successful completion of test items in the two batteries.

Q 5.3: Do the test items in the two batteries require the same degree of syntactic knowledge for successful completion of a given task?

H_0 5.3: There is no significant difference in the degree of syntactic knowledge required for successful completion of test items in the two batteries.

H_1 5.3: There is a significant difference in the degree of syntactic knowledge required for successful completion of test items in the two batteries.

Q 5.4: Do the test items in the two batteries require the same degree of phonological / graphological knowledge for successful completion of a given task?

H_0 5.4: There is no significant difference in the degree of phonological / graphological knowledge required for successful completion of test items in the two batteries.

H_1 5.4: There is a significant difference in the degree of phonological / graphological knowledge required for successful completion of test items in the two batteries.

Q 5.5: Do the test items in the two batteries require the same degree of knowledge of cohesive relations for successful completion of a given task?

H_0 5.5: There is no significant difference in the degree of knowledge of cohesive relations required for successful completion of test items in the two batteries.

H_1 5.5: There is a significant difference in the degree of knowledge of cohesive relations required for successful completion of test items in the two batteries.

Q 5.6: Do the test items in the two batteries require the same degree of knowledge of rhetorical organisation features for successful completion of a given task?

H_0 5.6: There is no significant difference in the degree of knowledge of rhetorical organisation required for successful completion of test items in the two batteries.

H_1 5.6: There is a significant difference in the degree of knowledge of rhetorical organisation required for successful completion of test items in the two batteries.

Pragmatic Competence

Pragmatic competence is the ability to produce and understand sentences or utterances, which are appropriate to the context in which they occur. This ability subsumes the things one needs to know in order to communicate effectively in different social settings.

Q 6: Do the test items in the two batteries involve the same degree of pragmatic competence for successful completion of a given task?

H_0 6: There is no significant difference in the degree of pragmatic competence involved for successful completion of test items in the two batteries.

H_1 6: There is a significant difference in the degree of pragmatic competence involved for successful completion of test items in the two batteries.

Pragmatic competence comprises two abilities: illocutionary competence which is the ability to understand the pragmatic conventions for performing acceptable language functions and the sociolinguistic competence which is the ability to understand the sociolinguistic conventions for performing language functions appropriate in a given context. To investigate these abilities the following two questions were formed.

Q 6.1: Do the test items in the two batteries involve the same degree of illocutionary competence for successful completion of a given task?

H_0 6.1: There is no significant difference in the degree of illocutionary competence involved for successful completion of test items in the two batteries.

H_1 6.1: There is a significant difference in the degree of illocutionary competence involved for successful completion of test items in the two batteries.

Q 6.2: Do the test items in the two batteries involve the same degree of sociolinguistic competence for successful completion of a given task?

H_0 6.2: There is no significant difference in the degree of sociolinguistic competence involved for successful completion of test items in the two batteries.

H_1 6.2: There is a significant difference in the degree of sociolinguistic competence involved for successful completion of test items in the two batteries.

Strategic Competence

Strategic competence enables language users to relate the features of the language use context to the intended illocutionary force and to relate an utterance form to the features of the context to interpret its illocutionary force appropriately. Strategic competence seems to be operationally more transparent when it is associated with test wiseness.

Q 7: Do the test items in the two batteries involve the same degree of strategic competence for successful completion of a given task?

H_0 7: There is no significant difference in the degree of strategic competence involved for successful completion of test items in the two batteries.

H_1 7: There is a significant difference in the degree of strategic competence involved for successful completion of test items in the two batteries.

3.3.3 Questions Related To Test Performance

Subjects' scores on the two batteries will be analysed to see if test performance supports the interpretations of the content analysis findings. Two specific research questions are addressed in test performance analysis. Firstly, it is important to investigate whether the grading of text difficulty in content analysis is supported by the subjects' scores on the tests. That is, whether the subjects perceive the difficulty of the test items in the two batteries differently.

Q8: *Are the test items in TOEFL and IELTS significantly different in terms of their perceived item difficulty?*

> ***H_0 8:*** *The test items in the two batteries are not significantly different in terms of their item difficulty.*
>
> ***H_1 8:*** *The test items in the two batteries are significantly different in terms of their item difficulty.*

Secondly, we need to investigate if test preparation has any impact on the performance of the subjects on these tests. Various researchers have discussed the effect of test preparation on test performance; see Messick (1980), Alderson & Wall (1993), and Bachman et al. (1995). Since some of the subjects in this research were participating in a TOEFL preparation course, it was necessary to investigate the effect of such courses on the performance of test takers on these tests. Hence the following question was formed.

Q 9: *Does test preparation have a significant effect on the performance of subjects on TOEFL and IELTS?*

> ***H_0 9:*** *Test preparation has no significant effect on the performance of subjects on TOEFL and IELTS.*
>
> ***H_1 9:*** *Test preparation has a significant effect on the performance of subjects on TOEFL and IELTS.*

As it stands, we have nine major research questions and twelve minor ones in this research.

3.4 Subjects

The subjects were selected from two different English language institutes in Tehran, Iran. The first group consisted of 65 graduate students and university lecturers from various academic backgrounds who had received scholarships to pursue their studies in an English speaking country leading to the PhD degree. They were attending an intensive TOEFL preparation course run by the Ministry of Culture and Higher Education (MCHE), henceforth referred to as the M group. The second group comprised 90 high school and undergraduate students who had completed the FCE courses in Kanoon Zaban Iran (former Iran-America Society), hereafter referred to as the K group. The K group were generally at above intermediate level according to Kanoon's officials. The M group subjects were assessed to be at pre-intermediate level of language proficiency by the MCHE. There is not much information about the M group language background other than that gathered by the questionnaire distributed among them after the administration of the tests[2]. While group M subjects practised only TOEFL preparation materials, which emphasises multiple-choice items, group K subjects, additionally, did some writing assignments required for the FCE courses.

Both groups can be categorised under the wide range of typical TOEFL test takers (undergraduates and postgraduates seek admission to universities in English speaking countries). Since IELTS "*is recognised as a language requirement for entry to academic courses by institutes of further and higher education in the United Kingdom, Australia, ...*" (UCLES, 1999, p. 5), the K and M groups can also be categorised under the typical IELTS test takers.

Group K subjects all volunteered to take the proficiency tests in this research. All of them were familiar with TOEFL but very few knew about IELTS and almost none had heard about EPTB. They were told that IELTS and EPTB were experimental tests administered for research purposes by the University of Cambridge Local Examination Syndicate and the University of Edinburgh, respectively[3]. As an incentive to participate in all three tests, group K subjects were told that they would

[2] See Appendix 1 for the original English version of the questionnaire.
[3] This was because the IELTS specimen module was provided by the EFL section of the UCLES and the EPTB by the Institute of Applied Language Studies of the University of Edinburgh.

receive an official TOEFL result certificate signed by Kanoon Zaban, which might be taken into consideration when applying for future employment with Kanoon as an English tutor. Convincing group M subjects to participate in all three tests was not an easy task. TOEFL was given to them as part of their course with a week's notice. Since the majority of the M subjects were applying to British and Australian universities, it was argued that sitting for a mock IELTS would help them familiarise themselves with the test in case they needed to sit for IELTS. Some, however, did not take it seriously, did not come to the exam session, and if they did, did not complete all the tasks. Hence the total number of subjects taking the IELTS dropped slightly. There was no reaction against EPTB as it was given to them as practice for listening comprehension and grammar exercises.

Information on test takers' current age, sex, current status, qualification, academic field, and the proficiency preparation courses was obtained from responses to the background questionnaire[4] given to them at the end of the TOEFL test. Table 3.1 and Table 3.2 present the descriptive statistics for the test takers' characteristics. Their median age was 25 with the youngest test taker being 16 and the oldest 41 years old. The proportion of male / female was different in the two test centres but in total there were roughly equal numbers of female and male participants; slightly over half 55.6 % were female. The number of participants in the two test centres was also roughly equal with 51.9 % in the K centre and 48.1 % in the M centre. Most test takers had participated in proficiency preparation courses prior to taking the tests. TOEFL preparation course seemed to be the most popular one with 42 % of the test takers indicating that they were participating in one. All the M centre participants were attending a TOEFL preparation course at the time. 16 % participated in FCE, 4.9 % in IELTS, and 2.5 % in CPE courses. Of 130 who responded to the question about the qualification they held, 46.9 % indicated that they were either in high school or had already obtained their diploma, 20 % had completed their first degree and a further 33.1 % were postgraduates intending to pursue their studies for a PhD degree. The latter group were mostly those in the M centre.

[4] See Appendix 1.

Table 3.1: Test Takers' Characteristics: Age, Sex, Prep. Course

Current Age					
Mean		Median		Minimum	Maximum
25.93		25.00		16	41
Sex and Test Centre					
			Test Centre		Total
			Kanoon	MCHE	
Sex	Female	Number	69	21	90
		% of total	42.6 %	13.0 %	55.6 %
	Male	Number	15	57	72
		% of total	9.3 %	35.2 %	44.4 %
Total		Number	84	78	162
		% of total	51.9 %	48.1 %	100.0 %
Preparation course					
	TOEFL		FCE	CPE	IELTS
% YES	42.0 %		16.0 %	2.5 %	4.9 %

120 responded to the question related to the topic of their course of which 20 % were in Humanities, 35 % in Science, 30.8 % in Engineering, and a 14.2 % Medicine. The subjects were asked to comment on their current status. Of the 132 who responded to this question, 11.4 % indicated that they were still in high school, 31.4 % were undergraduate students (part-time and full-time college), 14.4 % were in a language institute or an English course, and 43.2 % were not studying at the time (these held Master degrees).

There are not many published statistics about IELTS test takers. The available statistics only report the percentile rankings by band score for the whole IELTS population. According to UCLES (1999), during the 1998 administrations of IELTS (Academic), 20% of the candidates achieved band score 5 or less, 32% achieved band scores 5-6, and 41% achieved band scores 6-7.5. Due to the lack of information about Iranian test takers, we cannot compare our sample with those who actually sit for IELTS.

There are some statistics reported in Wilson (1982) about the characteristics of TOEFL examinees but they refer to data gathered between 1977 to 1979. The

latest information (ETS, 1999) about the TOEFL test takers does not cover all the characteristics that Wilson (1982) report does. One would expect a change in test takers' characteristics over the years. For example, Wilson reports an average of 23.85 for the age group of the undergraduate-level degree and graduate-level planners as compared to 25.93 in the present study.

Table 3.2: Test Takers' Characteristics: Academic Status

			Qualification				
			High school	Diploma	B.Sc./BA	Master	Total
Sex	Female	Number	9	41	11	12	73
		% of total	6.9 %	31.5 %	8.5 %	9.2 %	56.2 %
	Male	Number	4	7	15	31	57
		% of total	3.1 %	5.4 %	11.5 %	23.8 %	43.8 %
Total		Number	13	48	26	43	130*
		% of total	10.0 %	36.9 %	20.0 %	33.1 %	100.0 %

			Academic field				
			Humanities	Science	Engineering	Medicine	Total
Test center	Kanoon	Number	16	28	9	7	60
		% of total	13.3 %	23.3 %	7.5 %	5.8 %	50.0 %
	MCHE	Number	8	14	28	10	60
		% of total	6.7 %	11.7 %	23.3 %	8.3 %	50.0 %
Total		Number	24	42	37	17	120*
		% of total	20.0 %	35.0 %	30.8 %	14.2 %	100.0 %

			Current status					
			High-school student	Part-time college	Full-time college	Language student	Non-student	Total
Test center	Kanoon	Number	13	3	30	16	9	71
		% of total	9.8 %	2.3 %	22.7 %	12.1 %	6.8 %	53.8 %
	MCHE	Number	2	2	6	3	48	61
		% of total	1.5 %	1.5 %	4.5 %	2.3 %	36.4 %	46.2 %
Total		Number	15	5	36	19	57	132*
		% of total	11.4 %	3.8 %	27.3 %	14.4 %	43.2 %	100.0 %

* Some subjects did not answer all the questions.

The main difference in test takers' characteristics is the change in the number of male participants over the years. Wilson reports that 72 % of the TOEFL participants were male while they comprise 51% of the population in the ETS (1999) report. Bachman et al. (1995) also report a smaller percentage of male participants (41 %) in their study. It appears that the number of female TOEFL test takers has risen over the years. In our sample 44.4% of the population were male.

Another aspect on which our sample can be compared with the actual TOEFL population is the test scores. TOEFL Test and Score Manual Supplement (ETS, 1994, p.7) reports the average TOEFL scores for the 9393 Iranians who sat for the test between July 1992 and June 1994. Table 3.3 compares the mean scores of our subjects on the TOEFL sample with the ones reported by the official TOEFL.

Table 3.3: Comparison of TOEFL Scores

TOEFL Sample	Number of examinees	Listening Comprehension	Structure and Written Expression	Reading Comprehension	Total Score Mean
TOEFL scores for Iranian test takers (1992-1994)*	9393	51	52	50	511
TOEFL scores for this sample (Aug 1994)	127	48	52	50	501

*It refers to the average scores of the Iranian test takers on TOEFL from July 1992 through June 1994

Although the number of examinees in our sample was significantly lower than the actual TOEFL, the scores reported are so close that it allows us to claim that our sample was not different from the Iranian candidates who took the actual TOEFL from July1992 through June 1994[5].

Moreover, the comparison of our sample's mean scores on TOEFL with the average scores of the total TOEFL population is not much different with respect to sex. Table 3.4 illustrates the comparison of our sample's scores on TOEFL with the actual total TOEFL population who sat for the test during the same period. The actual TOEFL

[5] It is worth mentioning that the TOEFL sample in this research was administered in August 1994.

scores were extracted from the Tables Seven and Eight of the TOEFL Test and Score Manual Supplement (ETS, 1994, p. 4).

Table 3.4: Comparison of Official TOEFL Scores with the Sample with Respect to SEX

SEX	TOEFL Sample	Number of examinees	Listening Comprehension	Structure and Written Expression	Reading Comprehension	Total Score Mean
Male	TOEFL scores for Iranian test takers (1992-1994)*	722,247	51.9	52.8	52.3	523
	TOEFL scores for this sample (Aug 1994)	53	49.7	52.6	50.5	510
Female	TOEFL scores for Iranian test takers (1992-1994)*	595,383	52.3	52.2	51.1	519
	TOEFL scores for this sample (Aug 1994)	69	48.3	54.5	50	509

*It refers to the average scores of all the candidates who took the TOEFL from July 1992 through June 1994 and responded to sex group membership on answer sheets.

It can be observed from the above that the scores of the test takers on the TOEFL sample is close enough to the scores of the average TOEFL male and female test takers to allow us to say that the two populations, despite the significant difference in the number of their examinees, are comparable. That is, the test takers in this study can be considered as typical TOEFL test takers.

3.5 Test Administrations

The two groups of subjects were tested separately in the two test centres: the Language Unit of the MCHE and Kanoon Test Centre. Since there were three test samples, six test administrations had to be planned during a period between July 16th and August 5th 1994. It was decided to administer the tests in both centres in the following order; 1. IELTS, 2. EPTB, 3. TOEFL. The facets of the testing environment will be explained in 3.5.2 below. The Language Unit of the MCHE personnel were involved in the administration of the tests in both centres. They were the same trained personnel who had helped the official administrations of TOEFL and IELTS in Tehran since 1992, hence, were quite competent in their job. While test administration in the MCHE centre was organised and performed entirely by the same personnel, the task in Kanoon was aided by Kanoon officials. The researcher himself acted as the chief testing officer in both centres.

3.5.1 Test Samples

IELTS

Two sample tests of IELTS and TOEFL were chosen for this research. The IELTS sample was chosen from Module C of the Specimen Materials of IELTS provided by UCLES and the British Council in 1992[6]. Based on O'Neill, Steffen and Broch (1994), who found that scientists were not disadvantaged if they read social science texts, but that social science and humanities students were disadvantaged if they read scientific texts, it was decided to choose Module C of the specimen materials of IELTS; Module C was intended for non-native students of English who were planning to study Business Studies and Social Sciences. Clapham (1996, p. 121) has shown that Module C texts did not disadvantage non-social science students. It consists of four subtests: Listening, Speaking, Reading, and Writing. Since the Speaking part of IELTS had no counterpart in TOEFL, it was excluded from the analysis. The Listening part consisted of 39 questions: 4 multiple-choice, 9 True/False, and 26 completion items all of which were based on audiotape material. The Listening sub-

[6] See Appendix 2 for the IELTS sample used in the administration.

test (30 minutes) was divided into four sections, testing general listening comprehension. The Reading sub-test (55 minutes) contained 36 questions: 4 four-option multiple-choice and 32 various completion items based on three different passages. It was meant to test the general reading ability of the testees in coping with Business Studies and Social Sciences texts. The Writing part consisted of two tasks. Task 1 (15 minutes) required the testees to describe a process based on the information illustrated in a diagram. Task 2 (30 minutes) asked the testees to write an essay based on a topic relevant to one of the reading passages.

TOEFL

The TOEFL sample was provided by the Iranian Ministry of Culture and Higher Education[7]. It was based on retired versions of TOEFL prior to 1994 and consisted of three sections: Listening comprehension, Structure and Written Expression, and Vocabulary and Reading comprehension all in a four-option multiple-choice format. The Listening comprehension (50 questions) was divided into three parts. Part A (20 items) required the testees to listen to short statements on an audio-tape and decide which option in the test booklet was closest in meaning to what they heard. Part B (15 items) required the testees to listen to the questions at the end of each conversation on the tape and select the best answer to the question from the test booklet. Part C (15 items) was similar to Part B but with longer conversations. The Structure and Written Expression section (40 items, 25 minutes) was divided into two parts testing the ability to recognise language that is appropriate for standard written English. Section three (45 minutes) was divided into two parts: Vocabulary and Reading comprehension. The Vocabulary section (30 items) required the testees to choose the word or phrase which best kept the meaning of the underlined word. The Reading comprehension part contained 30 questions based on five different passages.

EPTB

In addition to TOEFL and IELTS, a version of the EPTB test was also selected for assisting validating IELTS and TOEFL samples. Short Version Form C of EPTB was obtained from the Institute of Applied Language Studies (IALS) in Edinburgh[8]. Form C was prepared by Alan Davies and Alan Moller in 1973 and comprised three main

[7] See Appendix 3 for TOEFL sample used.
[8] See Appendix 4 for the EPTB test used.

sections: listening, reading and grammar. The listening section contained two tests. Test 1 (58 items, five-option multiple-choice) was the recognition of word stress. The testees would hear three words and they had to decide which words sounded the same. Test 2 (44 items, true-false) was based on 24 short conversations and tested the recognition of appropriate responses in discourse. The reading section, test 3 (49 items), contained two modified cloze passages with the first letter supplied. Test 4 (47 three-option multiple-choice items) was testing grammar.

3.5.2 Facets of the Testing Environment

Facets of the testing environment are those environmental conditions, which might affect the performance of the test takers. They include the familiarity of the place and equipment, time of testing, and physical conditions of the testing room. The researcher rated these facets for each site on a three-band scale. Since the subjects took the tests in the same institute where they were studying, they were quite familiar with the place. All three tests were pen and paper tests and demanded no training with any equipment for the subjects. The administrative personnel involved were also familiar to the subjects. In the M centre, the personnel of MCHE were actually the same who ran the TOEFL preparation courses. In the K centre, the participation of local personnel from Kanoon in the administration of the tests helped to ease any anxiety, which might have arisen as the result of the presence of the MCHE personnel. Therefore, it can be concluded that the familiarity of the place, equipment, and personnel was very similar in the two test centres. Table 3.5 shows the ratings on this facet.

Table 3.5: Familiarity Rating

	K	M
Place	2	2
Equipment	2	2
Personnel	2	2

K= Kanoon Zaban Test Centre M=MCHE Test Centre

Scale: Unfamiliar Familiar

0 1 2

Due to administrative problems, the tests were administered at different times in the centres. All the tests in Kanoon centre were administered during morning, while the

MCHE ones were done in the afternoon. It could be argued that afternoon sessions might slightly disadvantage some of the test takers as it affected the temperature of the exam rooms. The rating on this facet is given in Table 3.6.

Table 3.6: Time of Testing Rating

	K	M
Time	2	1

Scale: Not conducive to good test performance 0 1 Very conducive to good test performance 2

The physical conditions of the rooms where the tests were performed were slightly different. It was a hot summer in Tehran when the temperature could reach as high as 38° Celsius. Both test centres enjoyed central air-conditioning, however, there were occasional air-conditioning failures in the M centre during one of the administrations which affected the temperature in the room to some extent. The seating, lighting, and sound equipment of the two centres were of acceptable standard. The K centre was in general better equipped for testing large groups of subjects and enjoyed an internal audio system fitted in the ceiling. The M centre did not have any fitted audio system and audio players were used instead at the time of listening comprehension tests. The rating of the physical conditions is illustrated in Table 3.7.

Table 3.7: Physical Conditions Rating

	K	M
Weather	1	1
Room	2	1
Temperature in Room	2	1
Seating	2	2
Lighting	2	2
Audio facilities	2	0

Scale: Bad 0 OK 1 Good 2

3.6 Marking of the Tests

The three tests were marked according to the *scoring instructions' manuals* for each test.

EPTB

The 17-page answer sheet for this test battery was marked according to *Scoring Instructions for English Proficiency Test Battery Short Version Form C 1973* prepared by Alan Davies and Alan Moller[9]. The keys to the tests contained the correct responses for each part of the battery. No alternative responses were allowable in any test. Once the tests were marked, the number of *Right* (**R**) responses for each part was written in the appropriate scoring box provided in the covering page of the answer sheet (Table 3.8). Then, the raw scores were converted into the standardised scores (**S**) already calculated for each part. Each answer sheet was marked by two raters.

Table 3.8: Scoring Box for EPTB

Test	R	S
1		
2		
3		
4		
Total	XX	

R= raw score S= standardised score

TOEFL

The TOEFL sample was administered with a multiple-choice answer sheet. The answer sheets were hand-scored based on the TOEFL answer key provided by the MCHE[10]. A correct answer was assigned a score of 1 while an incorrect answer received a score of zero. The number of correct answers for each section of the test

[9] See Appendix 4 for the EPTB key and instructions to mark the test.
[10] See Appendix 3 for the TOEFL sample key and score conversion table.

was written above each section. Then, raw score subtotals and totals were converted to TOEFL standard scores via a conversion table provided with the scoring key. The answer sheets were checked by three markers for accuracy.

IELTS

The listening comprehension and reading comprehension sections of the IELTS Specimen Module C (1992) were hand-scored based on the answer keys available in the *Specimen Materials Handbook for the International English Language Testing System 1992*, provided by the EFL Section of the University of Cambridge Local Examinations Syndicate[11]. Each correct answer equalled one mark. Raw score subtotals for each section were recorded on the cover of answer sheets. The general guidelines for acceptable responses were as follows. For questions where the answer was a single letter, answers with more than one letter would be marked wrong. For questions where one or more words were necessary, spelling mistakes did not matter as long as the meaning was still clear. The only exception to this was Question 7 in the Listening Test, where spelling was important.

The *Writing* section of IELTS is marked subjectively by trained examiners. The examiners assign bands to the candidates by comparing the candidates' writing with criteria presented in the *Band Descriptions*. The writing is marked by one examiner only. Occasionally, UCLES randomly selects some of the marked papers to be rated by a second marker to ensure that all raters- especially those who are new- mark the papers according to the guidelines. Since this marking is done after the results have been sent out, the rating of the second marker does not affect the final official IELTS score. The researcher decided to follow the same procedure for rating the writing assignments. Five professional EFL teachers (three native and two non-native speakers of English) who were doing an MSc option in Language Testing at the Department of Applied Linguistics , University of Edinburgh volunteered to mark the papers.

Instructions for marking the scripts were obtained from UCLES together with some of the test materials used for training the examiners. The researcher planned three one-hour training sessions for the markers. In the first session Band Descriptors[12]

[11] See Appendix 2 for the IELTS answer keys and answer sheets.
[12] See Appendix 5 for the Writing Band Descriptors.

were distributed among the participants and they were asked to comment on the descriptors. Once it was agreed that there was no confusion about the descriptors, two sample scripts, which had previously been marked by IELTS official examiners, were given to the trainees for marking. They represented two different band levels. The trainees were then asked to discuss the discrepancy between their own markings and those of the IELTS examiners. At the end of the session, seven more script samples were given to the trainees to be marked for the next session.

The second session was one week later during which the trainees discussed in detail their markings against those of the IELTS markers. Each had a chance to explain why they felt a piece of script represented the particular band level to which they had assigned it. As the discussion went along, the markers seemed to agree more and more on assigning similar band levels. Finally, they were given two actual exam papers to be marked according to the IELTS guidelines. Each paper consisted of two questions (tasks). For question 1, the accuracy of the information was more important than that of the question 2, while communicative value of the writing and the correctness of the English used was also important. For question 2, the markers were asked to judge the communicative quality of the writing, the effectiveness with which the arguments are presented, the logical structure of the presentation and the accuracy and appropriateness of the language used. Table 3.9 illustrates how each question was marked.

Table 3.9: Question Sub-Scales Rating For The IELTS Writing Section

Questions	Sub-scales	Band	
Question 1	Task fulfilment		
	Coherence and cohesion		
	Sentence structure		
			Global Band
Question 2	Communicative quality		
	Arguments, ideas & evidence		
	Word choice, form & spelling		
	Sentence structure		
			Global Band
			Final Band

In the third session the trainees discussed their ratings of the two writing assignments. Four out of five judges seemed to perfectly agree on assigning similar band levels to the scripts. Even the one judge, who differed in his ratings from the rest, had only disagreed in one band level, which is perfectly acceptable according to the IELTS standards. The training hence concluded and the markers were given approximately 26 papers each, 16 from the K group and 10 from the M group. Some of the test takers did not attempt the *writing* section. In addition to marking instructions, the markers were also given sample rated pieces of scripts for each band level and a conversion table for assigning the final band[13]. All the scripts were marked and returned within a three-week period.

3.7 Procedures of Investigation

Since construct comparability is the most important aspect of the study, gathering evidence for construct validation became the major task of the research. The evidence was gathered mainly with respect to two criteria: test content and test performance.

3.7.1 Methods of Content Analysis

There were limited resources available to the researcher. For example, the applied linguists (the expert judges who helped rating some of the components of communicative language ability) could only offer a limited amount of their time for rating the research instruments, and the research budget did not allow room for using incentives to employ further qualified staff to rate the instruments. Hence, content analysis was carried out mainly on the reading passages and their corresponding items.

The Rating Instruments

The rating instruments used were based on Bachman's (1990) characteristics of *test methods* (TM) and components of *communicative language ability* (CLA). The rating scales are modified versions of the ones used in Bachman et al. (1995) and Bachman et al. (1996) studies and reflected ongoing modifications based on their

[13] See Appendix 5 for the sample rated writings, marking instructions, and conversion writing band tables.

actual use in research. Although original scales produced satisfactory inter-rater reliability figures in Bachman et al.'s (1995,1996) studies, they could not be used in exactly the same format in this research. Bachman and his colleagues designed the scales to be used by *expert judges* for rating test method and communicative language ability. The collective judgement of the judges was the prime criterion in this regard; that was not the case here.

There were problems in this study with the use of expert judges. Firstly, the limited scope of the research and availability of resources did not allow the researcher to call upon the expertise of the judges to rate *all* the characteristics of test method and components of communicative language ability; it required too much of their time. Secondly, except for Bachman et al.'s (1995, 1996) studies, most similar researches (Alderson 1990a, 1990b; Alderson and Lukmani 1989) had shown poor inter-rater reliability as far as the content relevance of the tests was concerned. Judges usually disagreed as what an item was testing or how an ability was being measured by test items. The high reliability figures reported by Bachman and his colleagues could well be due to the fact that the judges had worked together as part of a research team for a period of at least three years. The judges would have probably shared the same frame of reference after such a long time of co-operation. Moreover, the rating scales were developed and modified by the same judges; expecting completely different judges to agree on the basis of the same rating scales was not warranted.

Clapham (1996) has reported that Bachman's rating instruments (TM & CLA) were problematic for her judges.

> *"None of the three raters were confident about their assessments. They felt that although some facets were unambiguous and straightforward to answer, ..., they were still worried by others. They all said that they would not expect to give the same ratings another time as they felt their internal rules for assessing the facets kept changing."* (Clapham, 1996, p. 150)

Finally, it seemed that some of the ratings could easily be achieved objectively by the researcher without asking the judges to carry them out. For example, counting the number of times passive constructions or cohesive markers occur in a passage.

In view of the above, it was decided to limit the number of communicative language ability components to 13 and the test method characteristics to 21. See 2.6 for a full description of the communicative framework used for comparing the language proficiency tests. It was also decided to use the judgements of expert judges on only those characteristics that could not be graded on the basis of pure objective linguistic knowledge. For instance, the length of the passage was rated on a count of content words and clauses in a paragraph: pure linguistic knowledge. This was done by using a simple count command on a computer. Whereas the degree of contextualisation was rated by the expert judges; it required the subjective judgements of the raters.

3.7.1.1 Analysis of Test Method Facets

Test method characteristics are limited to those aspects of test methods that are related to the *nature of language* in Bachman's (1990) *facets of the input*, which have been explained in 2.6.1. Other facets such as those of the *testing environment* have already been discussed separately in 3.5.2.

Length

Length of the passages was determined by counting the number of words, each word being defined as any string of letters separated by two spaces on each side. This was done by converting the original reading passages into pure text files (ASCII characters), removing any CONTROL characters, and running a UNIX command to count the number of words. The researcher, intentionally, avoided using commercial word-processor *word count* commands, which are very common in research today, because such commands most often treat the characters differently from what linguists consider to be a character; spaces are often treated by word processors as characters as well. Using the Medical Research Council (MRC) database[14], content words in each passage were also computed for further Type / Token analysis.

[14] MRC (Medical Research Council) is a database of over 100000 English words for First Language Acquisition on a UNIX environment. The database lists a number of linguistic features for each entry.

Propositional Content

As discussed in 2.6.1.1, propositional content relates to the characteristics of the information in the context and in the discourse and includes degree of contextualisation, distribution of information, type of information, and topic.

It was decided to call upon the judgement of expert judges to rate the propositional content of the reading passages of IELTS and TOEFL. The judges were three highly qualified applied linguists who had had long experience of EFL teaching. One of the judges was involved in writing reading items for IELTS, in addition to a research project he was conducting for UCLES on the predictive validity of IELTS band scale. A 3-scale band (0 , 1 , 2) was used to rate the components of the propositional content.

The degree of contextualisation was assessed with respect to two criteria: cultural content, and specific topical content. The judges were asked to rate this facet considering the relative proportion of new to contextual information. New information was defined as that which is not known to the test taker and cannot be predicted from the context whereas contextual information was defined as that which is developed in the passage itself. Thus, a passage was not at all contextualised if there was a lot of information in the passage that was not explained through definition, example, paraphrase, etc. The passage was highly contextualised if there was either not much new information, or if the new information was explained. The rating on this facet can be summarised as in Table 3.10.

Table 3.10: Degree of Contextualisation Rating

	Not at all contextualised		Highly contextualised
With respect to cultural content (CTXCULT)	0	1	2
With respect to specific Topical content (CTXSPEC)	0	1	2
	←		→

Distribution of new information was rated in relation to the compactness or diffuseness of the discourse. The judges were asked to rate the discourse of a given passage as highly compact if the distribution of new information was over a relatively

short space or time, and rate it highly diffuse if the distribution of new information in the discourse was over a relatively long space or time. Table 3.11 illustrates how this facet was rated.

Table 3.11: Distribution of New Information Rating

	Highly compact		Highly diffuse
Distribution of new information (C / D)	0	1	2
	←——————→		

Type of information in a test task was classified along two dimensions: concrete / abstract, and negative / positive. The judges were asked to rate this facet in terms of the relative degree of abstractness or negativeness of the passage. What was of concern here was the information contained in the text, and not how the test taker was expected to process that information. For the negative (NEG) rating, the raters were expected to consider only explicitly marked negatives, recognising that negatives might be explicitly marked in English in a variety of ways. Table 3.12 illustrates how type of information was recognised along abstract/concrete and negative/positive continuums.

Table 3.12: Type of Information Rating

	Rating		
Type of Information	0	1	2
ABSTRACT	Abstract	←——→	Concrete
NEG	Negative	←——→	Positive

Topic deals with the topic of the text. Since all the reading passages in the study were judged to be academic, the judges were asked to observe whether the texts were biased towards a specific discipline or they were just general academic: a two point scale was used for this facet.

Table 3.13: Topic Rating

	Discipline specific	General academic
Specialised topic (TOPSPEC)	1	2

The judges were asked to go over all the eight reading passages (5 TOEFL and 3 IELTS) and mark their ratings of the propositional content of the texts on a response sheet (Appendix 6). Since the same judges were also involved in determining the text difficulty of the reading passages, they were interviewed sometime later for their

comments on the rating instruments, the result of which will be explained later in 4.1.1.

Organisational Characteristics

Organisational characteristics, as discussed in 2.6.1.2, are those features of the discourse, which relate to the formal organisation of language. The two main components of formal organisation are grammar and cohesion. Both these facets were rated objectively using linguistic analysis.

Grammar is a complex facet, which incorporates several different components. In interpreting how the formal structure of the texts is laid out so that a meaningful grammatical comparison may be possible across the texts, it is hard to pinpoint which aspect of grammar should be examined. Therefore, it was necessary to look at different components of this facet to see which ones could provide meaningful comparisons of the two tests. To achieve that end, it was decided to rate the grammar of each piece of discourse on the basis of a comparison of the followings: number of sentences, clauses, embeddings, pre-modifiers, lexical density, voice, and text difficulty.

The number of *sentences* per paragraph was counted for all the reading passages. Then , the number of *clauses* was counted. Each clause was defined as a unit of post-modification, which has a finite verb in it. All relative clauses as well as reduced forms were considered as clauses. Coordinations as well as subordinations were also considered as clauses as long as they had predicatives.

Embeddings were defined as grammatical structures, which include relative clauses, and noun phrase complement structures. Both full forms and reduced forms of embeddings were counted for each reading passage. In addition to embeddings, noun phrases were examined to count the number of *pre-modifiers*, which preceded them.

There are two approaches to measuring *lexical density*. The first approach is the traditional approach discussed by Halliday (1966), which is the ratio of the number of lexical items to total words (token) in a passage. Ure (1971), amongst others, has used this approach in investigating lexical density and register differentiation. The second approach, which has been proposed by Halliday (1985), measures lexical density in terms of the number of lexical items as a ratio of the number of clauses.

Lexical items, often called content words, are members of an open system [15] that may consist of more than one word: for example, *turn on, stand up.* Since the advantages of one method over the other are not uncontroversial, it was decided to use both methods of measuring lexical density to see which one provided more information but report only one of them in the final analysis.

An arbitrary categorisation of passives was made to count for *voice*. Something was counted as passive only if it was in a clause while it contained at least the auxiliary verb form and participle as remnants.

Measuring *text difficulty* was one of the most difficult tasks in this research. Difficulty of a text depends on what the reader is asked to do with it. In everyday usage, difficulty is a relation between an action and an actor who performs that action, not a quality that is intrinsic or inherent to an object. So when we say, for example, *'That's a difficult mountain'*, what we mean is, in effect, *'That's a difficult mountain for me (or whoever) to climb'*. In other words, although the form of words in sentence *'That's a difficult mountain'* suggests that we are ascribing an inherent quality to the mountain, it is clear that a successful interpretation of the sentence requires us to fill in the defaults for the *'actor'* and *'action'* slot in the situation.

Furthermore, in order to assess whether a particular action is difficult for a particular actor to perform, we need some criterion of what counts as success. Thus it would be easy for us to read a sonnet by Shakespeare aloud to an audience, but it would be difficult for our neighbour, who is eighty years old, nearly blind and nervous of public speaking. We say it is easy for us to read the sonnet, but what counts *as 'reading a sonnet'* here? Would we read it as well as a professional actor, for example? Clearly not, so the difficulty of a task must depend not only on who is performing it, but also on how we judge that it has been satisfactorily accomplished.

In short, the call for assessing the text difficulty is based on the assumption that there is such a thing as intrinsic difficulty in a text, which can be assessed regardless of the reader's purpose. It is clear from the remarks above that this assumption cannot be totally justified. Against this background, it seems warranted to assess text difficulty from two different perspectives: assuming that there is such an intrinsic value as text

[15] As opposed to grammatical items which belong to a closed system. Examples of grammatical items are, *him, me, you, it, us, them, and one.*

difficulty, and that how difficult a text can be for a particular audience. The former can be predicted using a readability formula, while the latter can be determined by the judgement of the experts. We will first discuss the use of readability formulas to predict the difficulty of a text.

There are various readability formulas with which one can predict the readability of a text. Some of them measure the readability in terms of the complexity of the words used in a passage based on some word lists, i.e., the Dale-Chall (1948), whereas others measure the readability on the basis of word length; examples are Flesch Reading Ease, Gunning, and Fry. All readability formulas seem to produce satisfactory reliability indices in predicting the readability of reading passages. However, Flesch Reading Ease was used to predict the readability of texts in this research, as it seems to be favoured the most by researchers[16]. The Flesch formula used here is the revised version of the 1948 formula reported in Klare[17] (1984).

There are criticisms against using Flesch formula or indeed any other readability formulas in predicting the readability of texts for L2 readers. Standal (1987) classifies the criticisms of readability measures into three broad categories: text based, reader based, and text-reader-interaction based. He argues, firstly, that factors counted and measured, i.e., word/sentence lengths are not the only factors contributing to difficulty. Secondly, he argues that *"measures of reader skill level are not sufficiently sophisticated to allow precise matching of reader to text"* (Standal, 1987, p. 126). Finally, the argument is that *"text and reader attributes that are not known can (and do) interact in ways that render readability information invalid"* (Ibid.). Carrell (1987), while reiterating the same arguments, emphasises that ignoring reader variables such as background knowledge is the main weakness of such formulas. Readability formulas do not take into account the interactive nature of the reading process, which is the interaction of the reader with the text.

There is an important criticism of readability, which is related to the way(s) these formulas are correlated with readers' ability. The predictability of the formulas is determined by comparing a wide range of texts and readers' ability in L1. As Carrell (1987) reports *"the predictability of the formulas - that is, the high correlations –*

[16] See Klare, 1984, Carrell, 1987, and Standal, 1987.

[17] Flesch Reading Ease formula:
GL = .39 (words / sentence) + 11.8 (syllables / word) –15.59
Where GL = grade level

drops dramatically when more limited ranges of reader abilities, text subjects and numbers of test passages are considered" (p. 25). In the case of language proficiency testing, although large population of test takers are considered as potential readers, none will be from the L1 population for whom the formulas were originally developed.

To sum up the argument, it can be concluded that the readability formulas, despite all their deficiencies, can still be used in a meaningful way in ESL to assess the textual features if they are used as some kind of *predictors* of readability rather than *measures* of readability, as Klare (1984) suggests. The distinction between predicting the readability of text and measuring the readability of text can be illustrated by the example that we can *predict* the readability of a text for a certain audience with a particular level of language ability by use of formulas, but we can *measure* it only when a real person has read a given passage. Flesch Reading Ease was used here as a predictor of what readers would find easy/difficult. The actual difficulty would depend on the performance of the readers on test items.

Flesch Reading Ease (FRE) was used to measure the readability of texts as an intrinsic value. Using the MRC[18] database, the researcher wrote a Perl programme[19] to calculate the FRE of the reading passages. FRE is based on the calculation of the number of words and their syllables within a paragraph assuming that the lower the value obtained is, the more difficult the text is for comprehending.

To predict the difficulty or ease of a text for potential test takers of IELTS and TOEFL, the reading passages were given to expert judges. The three judges, who rated the propositional content, were asked to rate the difficulty of the passages on a 1-5 scale: 1 being 'not difficult' and 5 as being 'very difficult'. They were first asked to rate the difficulty level of each passage on the 1-5 scale and then rank the passages in terms of their difficulty: 1 being the 'easiest' and 8 as being the 'most difficult'. The judges were later interviewed for their comments on the scale. The interviews were recorded for further analysis. See chapter 4, section 4.1.1. for results.

Cohesion was rated purely on the basis of linguistic descriptions of the texts. Non-explicit lexical cohesion was not counted, as it required subjective judgement of some

[18] Amongst other things, MRC features include, pronunciation, number of syllables and characters necessary to calculate FRE.
[19] Credits to Dan Robertson for helping me write the programme.

form, which meant the use of the expert judges that were not available for this exercise. The facet of cohesion was subdivided into five sub-facets: *reference* (REF), *substitution* (SUB), *adversatives* (ADV), *causals* (CAU), and *temporals* (TEM). Following Halliday & Hassan (1976), explicit cohesive markers of the above types were counted for each passage as the evidence of their cohesive comparability. Cohesion in Hallidayan approach refers to those surface-structure features of a text, which link different parts of sentences or larger units of discourse. Some examples of cohesive markers are given in Halliday & Hassan, 1976, pp. 242-243.[20]

Relationship of Item to Passage

Reading comprehension and listening comprehension questions were rated for the relationship they had to the reading passages. Based on the results of the expert judges' earlier judgements of text difficulty and the propositional content,[21] and the successful experience of the training of the raters for marking the writing section of IELTS, it was decided to train the judges before attempting to rate any more tasks. One of the judges who was involved in the rating of the previous exercises (see 4.1) volunteered to rate the relationship of item to passage in addition to the researcher. The judge and the researcher had discussions about different aspects of the research for over two years, in particular about the degree of the complexity of the reading passages. We felt that we shared a lot in common about the way we perceived the rating instrument. The researcher and the judge had several meetings to study the instrument and discuss the judge's criticisms about the practicality of such a framework. Once the limitations of the scope of the ratings were agreed upon, it was decided to carry on with the ratings of this last facet. The instructions asked the judges to rate the relationship of the items to passage on a 1-5 scale. The description for each level was read as follows:

5=Requires test taker to relate information in passage to the real world
4=Item relates to the entire passage, and requires an understanding of the entire passage
3=Relates to several specific parts of the passage, or requires test taker to relate one part of the passage to several others

[20] See Appendix 7 for the examples of cohesive markers.
[21] See 4.1.1 and 4.1.2 for the details of the analyses.

2=Relate to a specific part of the passage, and requires only localised understanding of that part

1=No relationship to the passage; items can be answered without reference to the passage, or relationship of item to passage is not clear

This facet was rated for all the reading comprehension questions as well as the listening comprehension items. Table 3.14 summarises the complete list of the test method facets used for examining text comparability across IELTS and TOEFL.

Table 3.14: Facets of the Test Methods

Facet	Description	Code
Length	No. of words	WDS
Propositional content		
Degree of contextualisation	Culture sensitive	CTXCULT
	Special topic	CTXSPEC
Distribution of new information	Compact / Diffuse	C / D
Type of information	Abstract / Concrete	ABSTRACT
	Negative / Positive	NEG
Topic	Special / General	TOPSPEC
Organisational characteristics		
Grammar	No. of sentences per paragraph	SENT
	No. of clauses	CLAU
	No. of embeddings	EMBED
	2 or more pre-modifiers	PREMOD
	Lexical density (Halliday 1966)	LEXDEN-A
	Lexical density (Halliday 1985)	LEXDEN-B
	Passives	VOICE
	Text difficulty (FRE)	DIF-FRE
	Text difficulty (subjective)	DIF-RATE
Cohesion	Reference	REF
	Substitution	SUB
	Adversatives	ADV
	Casuals	CAU
	Temporals	TMP
Relationship of item to passage	Relationship of item to passage	RTP

3.7.1.2 Analysis of Language Ability Components

Communicative language ability components, derived from Bachman (1990) model, are the ones explained in 2.6.2. They include, *lexicon, morphology, syntax, phonology/graphology, cohesion, rhetorical organisation, ideational, manipulative, heuristic, imaginative, dialect, register,* and *strategic competence.* A brief description is given below for each component.

Lexicon is related to the lexical competence of the language user and assesses their choice of appropriate words. Morphology or morphological knowledge consists of the affixing of inflectional morphemes. Syntax involves the proper order of words in composing a sentence. And phonology / graphology relates to the ability that illustrates how accurate the representation of the language is according to phonological and orthographic rules. All the above four mentioned abilities relate to the grammatical competence, which is involved in producing and recognising sentences.

Cohesion and rhetorical organisation are part of the textual competence involved in controlling the formation and recognition of pieces of language at a higher level than a sentence. Cohesion refers here to explicitly marked relationships across clauses in the form of lexical connectors or of specific grammatical patterns that provide topicalisation. Rhetorical organisation is related to the overall effect of a text on a language user and is restricted here only to the formal aspect of conventional organisations used in expository writing classes: narration, description, comparison, and classification.

Ideational, manipulative, heuristic, and imaginative functions pertain to illocutionary competence explained in 2.6.2.1 as the ability to use language functionally. They can briefly be defined as follows: ideational function is a means by which we express meaning in terms of our experience of the real world; manipulative function is to use language to get various things done; heuristic function is where language is used to extend our knowledge of the world around us; and imaginative function relates to those aspects of language use which enable us to create our own environment for humorous or aesthetic purposes.

Dialect and register reflect the sociolinguistic competence of the language users. Dialect refers to the varieties of language dialects depending upon the geographical location of the language users or the social class with which the users are associated, i.e., black / standard, and British / American / Australian. Whereas register is referred to as a variety of language defined according to its use in social situations, i.e., formal, intimate, and casual.

Although strategic competence is not part of communicative language ability, it will be defined here as associated with a single aspect of the user's competence, i.e., test

wiseness. See 2.6.2.1 for the detailed descriptions of all the communicative language ability components.

The instrument used for assessing communicative language ability consisted entirely of rating scales, which "*attempted to capture both the degree to which components of CLA were involved in the successful completion of a given task, and the approximate level of ability required*" (Bachman et al., 1995, p. 102). Ratings of communicative language ability were made on the basis of:

a) the extent to which the judges felt the ability was required for the successful completion of the task and
b) the general '*level*' of that ability required, according to the following scale:

Table 3.15: Scale Used For Rating Communicative Language Ability Components

Not Required	Somewhat Involved	Critical Basic	Critical Intermediate	Critical Advanced
0	1	2	3	4

The instructions for using the scale went as follows: If you feel the ability is not required for successful completion of the task, write "0", if the ability may be involved, but is not critical to successful completion of the task, write "1", if the ability is critical to successful completion of the task, and at a basic level, write "2", if critical, but intermediate level, write "3", and if critical and advanced level, write "4". The rating instrument combined three types of information, the perceived degree of involvement of the ability or abilities in a given test task (0,1,2), the level of ability required (2,3,4), and the information about what abilities might be involved, but not critically, in completing the task (1).

Once again in addition to the researcher, the judge who participated in the rating of *the relationship of item to passage* facet volunteered to rate the test items on the basis of the above scale. The judges were asked to rate all the reading comprehension and listening comprehension items of both IELTS and TOEFL. The judges had joint meetings to discuss the rating instruments in the course of which some of the components were slightly redefined operationally. For instance, it was agreed that strategic competence be equated with test-wiseness and nothing else. Each judge was provided with instructions how to rate the components and a full description of what

they meant. For further details and the checklists used for rating communicative language ability see Appendix 8.

3.7.2 Analysis of Test Performance

If the results of content analysis of the tests reveal that there are similarities in the kind of abilities the two batteries measure, it is then necessary to investigate whether patterns of performance support such interpretation. To achieve this, patterns of correlations within each of the two test batteries and across the two are compared for possible similarities / differences in their latent traits.

In the discussion of test method effects (2.5.2) mention was made that one needs to use a Multitrait-Multimethod (MTMM) design - a minimum of three different test batteries - to separate trait and method factors which influence test results. Following the same principle we decided to employ the EPTB test, in addition to TOEFL and IELTS, so that we could meaningfully examine the underlying constructs of the tests under study. The above three proficiency batteries claim to measure different aspects of language ability of test takers with apparently independent test methods. We would examine the correlational matrices of test scores on these three batteries. If the result of the analysis shows that similar constructs on different tests do tend to correlate highly, it is the evidence of convergent validity. That is to say that the batteries provide similar information about some aspects of the examinees' language ability. If, however, the correlation between similar traits is found to be low, discriminant validity is achieved, i.e., the traits are independent irrespective of the method applied and provide different / additional information about the language abilities of the test-takers.

It has to be borne in mind that a full Multitrait-Multimethod approach cannot be adopted in this research, as the three tests under examination (IELTS, TOEFL, and EPTB) do not all have identical sections. That is, they do not all attempt to tap similar underlying traits. The focus of the Multitrait-Multimethod approach in this study is on the comparison of listening and reading sections of the tests and possibly on the grammar subsections.

Using exploratory factor analysis[22] the correlational matrices of test scores on these three tests will be examined to explore possible identifiable patterns within the interrelationships among the different sections of the tests. We will look into four different intercorrelation matrices between: a) the raw scores for the three sections of IELTS; b) the raw scores for the three TOEFL sub-tests; c) the raw scores for the four EPTB tests; and d) all ten of these measures. This should allow us to compare the internal structure(s) of these tests and to see if the identifiable patterns support the interpretations made in the content analysis section. It should also allow us to observe what aspect(s) of language proficiency the tests are aimed at tapping. See 5.3 for the details of factor analysis procedures and the results obtained.

Various other statistical analyses such as t tests and multiple regression analyses are carried out to find out it the test takers perform equally well on both tests or whether test takers' characteristics are in any way influencing the test results. See Chapter 5, sections 5.5 and 5.6 for the results of the analysis.

[22] Exploratory factor analysis is used for examining the correlational matrices because the batteries studied are seemingly based on different operational interpretations of language proficiency and therefore their traits' similarities / differences need to be explored without any presumption of the underlying constructs. The application of confirmatory factor analysis is judged to be inappropriate in such circumstances. See Gorsuch, 1983 for more details of the application of factor analysis methods.

Chapter Four

Content Analysis of TOEFL and IELTS:

Results and Discussions

4. Content Analysis of TOEFL and IELTS: Results and Discussions

We pointed out in section 3.2 that the purpose of carrying out content analysis in this research was to find out whether the contents of the two batteries are comparable in terms of the components of communicative language ability and the facets associated with test method. This chapter reports and discusses the results obtained from the analysis of test content and addresses the questions raised in 3.3.1 and 3.3.2, the answers to which would enable us examine the construct comparability of the two language proficiency batteries under examination: TOEFL and IELTS. We will first begin by reporting the results of the analysis of test method facets.

4.1 Analysis of Test Method Facets

In 3.7.1.1 we discussed the facets of test methods associated with the test input, which Bachman defines *as "the information contained in a given test task, to which the test taker is expected to respond"* (1990, p. 125). This section reports the results obtained from the analysis of test method facets and will provide answers to questions Q1 to Q4 raised in 3.3.1. Facets of the test methods were categorised under four general headings: length, propositional content, organisational characteristics, and relationship of item to passage. Each general heading was associated with a main research question and some minor questions related to sub-sections of each heading. Since the most problematic task carried out by the judges in this research was to do with text difficulty, a sub-section of the Grammar component (organisational characteristics), we begin our discussion of test method analysis by reporting the results of the analysis of this facet. We will then report the ratings of the components of the propositional content and discuss the results of the ratings of all the other organisational characteristics and the relationship of item to passage.

4.1.1 Text Difficulty Results

This section seeks to find answer to question 3.3.

Q 3.3: *Are the reading passages in the two batteries significantly different in terms of their text difficulty?*

It has already been argued that in order to establish the comparability of the two tests, one has to demonstrate that, among other things, the two tests are comparable in terms of their text difficulty. Text difficulty can be assessed in two main ways by using: (I) a readability formula, (II) the subjective judgement of difficulty by experts. The former is based on the assumption that there is such an intrinsic value as text difficulty, which can be measured directly from the text, whereas the latter looks at how difficult a text can be for a particular audience. Let us first discuss the application of the Flesch Reading Ease (FRE) formula for measuring the readability of the reading passages. Table 4.1 lists the Flesch Reading Ease for each of the 8 reading passages in the two batteries concerned: TOEFL and IELTS[1].

Table 4.1: Readability Indices for the Reading Passages

	TOEFL					IELTS		
	TR1	TR2	TR3	TR4	TR5	IR1	IR2	IR3
Flesch Reading Ease	54.1	44.5	36.9	63.3	45.3	40.8	45.8	39.5

TR= TOEFL Reading IR= IELTS Reading

The lower the Flesch reading index, the more difficult the passage is deemed to be. According to Flesch[2] (1974, p. 177), a Reading Ease Score of 30-50 describes a difficult text suitable for college students, and a score of 60-70 a standard text suitable for students in 7th or 8th grades. As can be seen, with the exception of TOEFL reading 4, all the reading passages are measured on a scale of 30-60 interpreted as *fairly difficult* to *difficult,* suitable for high school/college students. To test whether the

[1] Clapham (1996, p. 92) reports slightly different FRE values for the IELTS texts but the pattern of difficulty ranking is the same as the one reported here. The difference in the FRE values reported in the two studies could be because of the use of different application software in the two studies. Clapham has used an early version of Word for Windows for the calculation of FRE, whereas we have used a Perl programme to calculate the Flesch Ease.

[2] Flesch has proposed the following descriptions for interpreting the Reading Ease:

Description of Style	Reading Ease Score	Estimated School Grades
Very Easy	90-100	4th grade
Easy	80-90	5th grade
Fairly Easy	70-80	6th grade
Standard	60-70	7th or 8th grade
Fairly Difficult	50-60	some high school
Difficult	30-50	high school/ some college
Very difficult	0-30	college graduate

difference in text difficulty of the reading passages in the two batteries is significant, a Mann-Whitney test was performed. Table 4.2 shows that the p-value obtained for Z (2-tailed p, Asymptotic Significance) is greater than 0.05, indicating that there is no significant difference between the reading passages in the two batteries in terms of their text difficulty as measured by Flesch Reading Ease. This confirms our null-hypothesis **H_0 3.3**[3]. Therefore, the passages in the two proficiency tests seem to be approximately of the same readability level with respect to the Flesch formula.

Table 4.2: Comparison of Mean Facet: FRE

TOEFL / IELTS Passages

	FRE rating
Mann-Whitney U	5.000
Wicoxon W	11.000
Z	-.745
Asymptotic Sig. (2-tailed)	.456
Exact Sig. [2*(1-tailed Sig.)]	.571[a]

a= Not corrected for ties FRE= Flesch Reading Ease

Table 4.3, nevertheless, indicates that TOEFL reading passages varied much more in their readability distributions than those of the IELTS. The relatively large variance (102.67) across TOEFL texts is indicative of the fact that these texts were aimed at different levels of language abilities, whereas the relative small variance (11.06) of IELTS readability distribution suggests that IELTS texts were prepared for more or less the same level of language ability. The comparison of means (48 and 42 for TOEFL and IELTS, respectively)[4] show that IELTS texts were on average more difficult than the TOEFL ones, although the difference is not statistically significant as set out in Table 4.2.

Table 4.3: Readability Distributions in the Two Tests Based on FRE

	Mean	Std. Deviation	Variance
TOEFL Readings	48.82	10.13	102.67
IELTS Reading	42.03	3.33	11.06

FRE= Flesch Reading Ease

[3] **H_0 3.3:** There is no significant difference in the text difficulty of the reading passages in the two batteries.

[4] The lower Flesch mean is suggestive of a more difficult text.

Figure 4.1 illustrates how the reading passages vary in their degree of readability, with TOEFL passages having both the easiest (Reading 4) and the most difficult (Reading 3) of the texts.

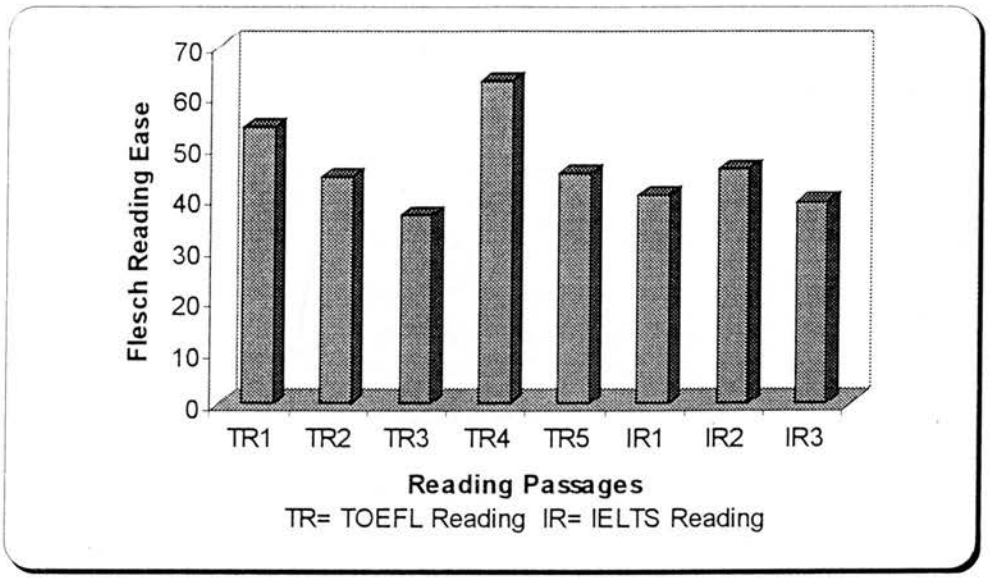


Figure 4.1: Readability Comparison Across TOEFL and IELTS

In addition to Flesch estimates, the reading passages were given to three expert judges to determine their difficulty level[5]. The judges first ranked the passages in terms of difficulty; 1 being the *easiest* text and 8 being the *most difficult* one. Then, they rated text difficulty of the texts on a 1-5 scale; 1 as being *not difficult* and 5 as *very difficult* for typical IELTS/TOEFL test taker. The results of their ranking and their difficulty ratings are shown in Table 4.4 and Table 4.5 respectively.

As can be seen from Table 4.4, the judges did not agree about the difficulty ranking of the texts. Since the number of rating categories was small, it was not possible to calculate a standard reliability index for the raters. However, it was possible to check the agreement between the raters using Kappa statistics. The following Kappa figures were obtained for their agreements: Rater1-Rater2= 0143, Rater1-Rater3= -0.143, and Rater2-Rater3= -0.143, none of which was significant at 0.05 level. The only passages that the judges could relatively agree on their difficulty level were IELTS reading 2 and 3; the judges agreed that they were the most difficult of all. The judges

[5] For the details of the instructions given to the judges see section 3.7.1.1

almost always disagreed among themselves about the difficulty ranking of other passages.

Table 4.4: Reading Difficulty Ranking

	TOEFL					IELTS		
	TR1	TR2	TR3	TR4	TR5	IR1	IR2	IR3
Rater 1	5	4	3	1	6	2	7	8
Rater 2	1	2	7	4	6	3	5	8
Rater 3	3	1	5	2	4	8	6	7

TR=TOEFL Reading IR=IELTS Reading

Ratings: 1=easiest 8=most difficult

Looking at Table 4.5, one can see that how the judges had difficulty assigning difficulty grades to each reading passage.

Table 4.5: Reading Difficulty Grading

	TOEFL					IELTS		
	TR1	TR2	TR3	TR4	TR5	IR1	IR2	IR3
Rater 1	3	2	2	2	3	2	4	4
Rater 2	2	2	5	4	4	3	4	5
Rater 3	2	1	3	1	2	5	4	5

TR=TOEFL Reading IR=IELTS Reading

Ratings: 1=not difficult 5=very difficult

It was not possible to compute Kappa statistics for the agreement between the raters on reading difficulty grading, as Kappa statistics require a symmetric 2-way table in which the values of the first variable match the values of the second variable; Rater 1 avoided the use of difficulty levels 1 and 5, and Rater 2 avoided that of level 1. One viable option for raters' agreement was the use of Pearson correlation. The Pearson correlation statistics for the raters (Rater1-Rater2 = 0.305, Rater1-Rater3 = 0.466, and Rater2-Rater3 = 0.412) do not show much better agreement among the judges. Overall inter-rater reliability for difficulty grading was r_{tt} = 0.59, using Z transformations as a correction[6]. The judges had clearly had different concepts of

[6] The inter-rater reliability using Z transformation is obtained through $r_{tt} = nr_{AB} / 1 + (n-1)r_{AB}$ formula, where r_{AB} is the average correlation between the raters and n is the number of raters.

reading difficulty, which called for a follow-up interview to find out why they rated the texts so differently.

Rater 1 indicated in her interview that she rated the texts based on first impression; had she had more time, she would have possibly come to a different conclusion. She justified her first impression ratings on the basis that most test takers do not have time to negotiate during the examination time, hence it is important how the texts appeal to one's mind on first impression. In grading the texts, she looked primarily at how a text was accessible to the average reader. For instance, she thought TOEFL reading 4 (Isadora Duncan) was very accessible to the average reader partly because the dance topic was a universal one - which is familiar to all mankind irrespective of their individual cultures - and partly due to the relatively short sentences and simple structures used in the text. Rater 3 gave a similar explanation about the same text. Rater 1 believed that IELTS reading 1 (Quality Circles) was equally easy, a view shared by Rater 2, firstly because the visual layout helped the reader follow the descriptions. Secondly, the text was very descriptive using similar words and recycling them throughout the passage, and finally it did not contain a great amount of information in spite of its relative length. However, the text could have been very difficult for some readers if they had little background in business promotions, where special language is used. This is precisely why rater 3 considered the text as the most difficult of all. As Rater 1 argues, the task of ranking the difficulty order of texts could be quite misleading if the reader's background was ignored. She elaborated on this point saying that it was difficult to imagine an *average IELTS* test taker as there was very little published evidence of who they were. In the absence of such descriptions, one tended to judge the degree of difficulty of a text based on texts themselves, assuming that the readers were familiar with the background knowledge presupposed in the texts.

Rater 1 ranked IELTS reading 2 (The Purpose of Continuing Education) and IELTS reading 3 (Access to Higher Education) as the most difficult of all because she thought both texts had a lot of memory load and forced the readers to really understand the texts. IELTS reading 2, in particular, was wordy and abstract with no clear reference to facts, which could make the text quite boring for some readers. Longer texts, in Rater 1's opinion, could cause difficulty problems in examinations. With the same line of argument, TOEFL texts looked easier overall to Rater 1, because they were shorter and less argumentative.

In deciding to allocate difficulty level (Table 4.5), Rater 1 put the universe of the texts and the universe of the test takers side by side. In doing that she thought none of the texts were easy enough to be allocated level 1; neither could she allocate level 5 to any text as she believed the testees could cope with the tests. Therefore, the difficulty level of the texts varied between 2-4 for this rater. Background knowledge played a part in her judgement. Overall, Rater 1 thought that TOEFL difficulty was more associated with vocabulary and less with understanding, whereas in IELTS the difficulty was more to do with understanding.

Rater 3 seemed to have focused on different aspects of difficulty. He believed that ranking the two tests was misleading as there was a big gap between the difficulty of the two tests: IELTS texts were more difficult overall. IELTS reading 1 was reckoned to be the most difficult text of all because it was commercial and it was not clear for whom it was written. Furthermore, he commented that there was a fair amount of *culture-specific* references, *register-specific* jargon, *complex* sentences and *anaphoric* pronouns that could potentially contribute to the difficulty of the text. In his judgement, IELTS reading 3 was equally difficult because of its complex structure and culture-specific references. Unless one were familiar with the references, the texts would have been very difficult to comprehend. IELTS reading 2 was perhaps less complex than the other IELTS texts but definitely more difficult than any of the TOEFL ones. As far as the sign-posting of the titles and pictures (layout) were concerned, he believed that they could help the understanding in IELTS reading 2 and IELTS reading 3 but not in IELTS reading 1. Rater 3, overall, thought that the difficulty of IELTS readings was not intrinsic in the texts but rather in the tasks related to them and in the amount of culture specific references.

In the case of TOEFL, he believed that three factors contributed to their difficulty: *lexis* (technical, lexical difficulty and abstractness of lexical items), *sentence structure* (length and anaphoric references), and *rhetorical structures*. He considered TOEFL texts as context independent where the meaning was clear from the texts and not required to have the knowledge of the context. But in IELTS, the context from which they were derived was important. This was also reflected in the instructions of the tests. Rater 3 commented that TOEFL instructions were very clear but related to the familiarity of the testees; in IELTS, that was not the case and one had to read it carefully. Tasks also varied a lot in IELTS; therefore, the candidates had to read them carefully and needed to be familiar with them. Finally, IELTS tasks were generally more authentic, which made them intrinsically more difficult.

Unlike Rater 1 and Rater 3, who looked the difficulty from different perspectives, Rater 2 had a more eclectic approach and accommodated both perspectives in his approach. This is reflected in his grading of the texts (see Table 4.4 and Table 4.5). Perhaps the fact that he was familiar with the rating instrument and had helped the researcher to rephrase some of the descriptors of the rating instrument helped him to have a clear idea of what to look for in grading the texts.

So far we have discussed the raters' differences in what they perceived to contribute to the difficulty / ease of the texts. This discussion cannot be closed without reporting what the raters agreed that contributed to the difficulty / ease of each text. The following comments are extracts from the judges' follow-up interviews and indicate what the raters *did* agree about each text.

TEXT	*JUDGES' COMMENTS*
TOEFL Reading 1: "Sea Anemones"	*It is a straightforward factual description with short sentences and some subordination, but rhetorically is not particularly complex. The major difficulty, if there is one, is likely to be the unfamiliarity with technical vocabulary: "stem, tentacles, hydras, coelenterates, wharf pilings"; there are no difficulties of register, style or tone.*
TOEFL Reading 2: "Steamships"	*It is a little more difficult than Reading 1, largely because of the final sentence. The difficulty here is to establish a referent for the subject pronoun "it"; if the reader does not make the connection with the phrase "an enormous sum for the time" in the previous paragraph, the final sentence is difficult to make sense of. There are fewer specialised lexical items in this text than in Reading 1: "entrenched, public debt, per capita", mostly relating to economic history. There are more cultural references here, but not such as to cause any difficulty.*
TOEFL Reading 3: "The Archaeological Record"	*The sentence structure and rhetorical complexity of this text seem to render it more difficult than any of the other TOEFL texts. A point of possible difficulty is to get a clear idea of the topic. The topics of the other TOEFL readings are readily identifiable, while the topic of this reading (roughly, what constitutes data for the archaeologist) is more abstract and more complex. To the extent that the reader does not grasp the topic, understanding will be impeded, because unless the topic is understood, it will be difficult to integrate the parts of the text into a whole, i.e. to see the relevance of each constituent part of the text to the other parts and to the whole text. There are some tricky sentence constructions using comparisons: "a source of history, not just..., ... in their own right, ... not mere, Just as much as ..., an archaeologist studies ..." etc. Additionally, there are some tricky, idiomatic, or specialised vocabulary: "humble, in their own right, mere, in so far as, ephemeral, scraps, peat bogs".*
TOEFL Reading 4: "Isadora Duncan"	*There are relatively short sentences with, for the most part, simple structure, clear sign posting of rhetorical structure ("First.... Her second contribution ... etc."), easily identifiable topic, and clear line of argument.*

	Possible difficulty in lexis could relate to: "exacting, dread, swaying, codified".
TOEFL Reading 5: "Plate Tectonics"	*It is a straightforward descriptive writing, in short sentences, and with relatively simple structure. There are some possible difficulties in establishing discourse coherence through resolution of anaphoric reference, particularly in the last paragraph: "the theory of plate tectonics ... The hypothesis; the geomagnetic field ... the field ... Reversal of the field; the rift, etc.*
IELTS Reading 1: "Quality Circles"	*This text can present real difficulties of comprehension. In the first place, it takes careful reading and some familiarity with the type of document to place this in context. If the reader cannot do this, much of the document will be very difficult to understand. For example, it is not clear who "we" (paragraph 6) refers to – presumably the IDS Public Sector Unit – but we have no idea who they are. Neither is it clear who the intended audience is – this appears to be the introduction to a publication issued by the IDS Public Sector Unit, but we have no way of knowing what kind of publication this is. It is referred to as a "Study", but this is not exactly transparent. There is a reference to "Appendix 1", but we do not see this.* *It is also dense with cultural references and idiomatic vocabulary which will be familiar only to people with some knowledge of Britain, and British business in particular, e.g., "NHS, Further Education, public sector, flavour of the month, catch on, Black and Decker, Jaguar, Tioxide, Honeywell Control Systems, London Life Association, Rolls-Royce Aero Division". Secondly, there is a good deal of business and management jargon: " grading and salary structures, pay and conditions, service organisations, trade unions", etc.* *Finally, the discourse is dense, with many examples of anaphoric reference, presenting difficulties in text processing. Much of the terminology is abstract or opaque. In spite of all the above characteristics, all the judges agreed that the text would not cause major comprehension problems if the reader was familiar with the context. That is why two of the judges ranked the text as one of the easiest readings while the third one considered it as the most difficult of all. It very much depends on what background knowledge the reader has.*
IELTS Reading 2: "The Purpose of Continuing Education"	*This text is somewhat simpler than the text about Quality Circles as it is apparently easier to contextualise. The structure is clearly sign-posted ("... at least eight purposes... The first... The third... The fourth... etc."). The conversational style makes for easy reading: "What are the purposes of continuing career education?" etc. There are some technical or specialised lexis: "patterns of resource allocation, peer review, self-appraisal, adjunct, fully-fledged". Short paragraphs aid comprehension.* *One particular difficulty with this text is the high degree of abstraction and depersonalisation. Thus, instead of "it is important that you should maintain freshness of outlook", we get "maintaining freshness of outlook on one's work ... is also important". There is also a lot of nominalisation, with some heavy pre-nominal modification: "career-oriented continuing education".*
IELTS Reading 3: "Access to Higher Education"	*All the raters agreed that this text is the most difficult of the texts. It is relatively easy to place in context, and it is, for the most part, a straightforward recital of facts and statistics relating to the subject matter.*

	The subject matter is easy to identify from the title, apart from anything else. The first two or three paragraphs are straightforward and do not present much difficulty, but the second section ("The Future") is more difficult, requiring some sustained concentration. Some culturally specific references may cause difficulty: "BTEC, SCOTVEC, TVEI, YTS, the Robbins principle," etc. *There is also considerable use of jargon: "wastage rates, on a par with, projections, differential demand" etc. The syntactic structure is in places quite complex: "Of potentially greater impact, however, is the assumption... etc.". Finally, there are some difficult anaphoric references to resolve: "a significant increase... This will depend", etc.*

Discussion of Text Difficulty

Two different approaches were used for investigating the text difficulty of the reading passages of TOEFL and IELTS. The first approach employed the Flesch Reading Ease formula to predict the difficulty of the texts. The comparison of Flesch indices across the two batteries indicates that there is no significant difference in the text difficulty of the reading passages in the two batteries. The second method involved the judgement of three experts in determining the difficulty of the texts. It appears that the judges could not agree either on the readability of the texts or on what contributed to the difficulty of the passages. The follow-up interviews of the judges traced the main source of the disagreement among the judges to the definition of the typical test taker. Although one of the judges - Rater 2, who was an IELTS item writer- was confident about who the test takers were, the other two judges did not share that view and were not sure who the test takers were. The judges' uncertainty about the audience of these tests made grading the difficulty of the texts unreliable.

Clapham (1996) reports a similar problem with her judges' uncertainty about the expected test takers for IELTS. She points out that "*the most important problem, and certainly the most enduring, related to those facets where the assessment had to be made in relation to the expected test taker*" (Clapham, 1996, p. 149). She believes that at the time she carried out her research, "*there was no such thing as a typical IELTS test taker*" (Ibid, pp. 149-150).

It is clear from the remarks above that the operational definition of the typical IELTS and TOEFL test takers needs revisiting. The typical test takers for these two tests were defined as those non-native speakers of English who seek to pursue their studies in English speaking countries and whose English language competence varies on a large scale from pre-intermediate level to advanced near-native speaker. The wide

range of test takers that fall into this vast category confused the judges in determining whether a text was easy / difficult for the average test taker. The judges had problems defining such an average test taker and believed that it was difficult to think of an average test taker, given the wide range of test takers sitting for these tests. But isn't this the case with all other general language proficiency tests?

As argued in Chapter 2, proficiency tests are not based on any syllabus and provide information about the general language competence of the testees in coping with the future language medium. These tests are targeted at a large population of testees with variable language abilities and schemata, which makes it difficult to judge the ease / difficulty of a text for such a wide population. Since the language abilities of the testees taking the proficiency tests varies significantly on a large scale, the schemata of such test takers can not be incorporated into the criteria for the selection of the passages by the test developers. The common practice as explained in 2.3.1.4 is to select texts that have been drawn from academic magazines, books, newspapers, encyclopaedias, etc., and then write appropriate tasks that are believed to be the representative of the situations the non-native speakers are going to face in their future settings. Since there is no definitive list of specification of different linguistic and communicative abilities necessary for given sociolinguistic situations, one has to rely on the judgements of the experts involved in the development process of the test for the degree of context validity. The difficulty of the texts and items will usually be assessed in the pre-test stage using statistical analyses. Those texts and items that are judged to be of appropriate difficulty[7] will be used in the final version of the actual test.

One may conclude from the above discussions that text difficulty is far too complex a facet to be measured by any single scale of assessment. On the one hand, text difficulty is associated with inherent text qualities such as complexity of vocabulary and syntax, degree of contextualisation, distribution of new information, and abstractness of the topic. These qualities contribute to the ease or difficulty of a text. On the other hand, text difficulty is related to the proficiency level of the target reader for whom the texts have been written. Unless the target audience is known, it is not possible to decide if a text is easy/difficult for them.

[7] The appropriateness of the difficulty will be assessed statistically using various item analysis techniques.

To recapitulate we can say that determining the difficulty of a text in a language proficiency test is a complex and lengthy procedure which starts with the subjective judgement of the expert item writers about the appropriateness of a particular text or item for a specific test population. This judgement will later have to be verified by test performance at the pre-test stage; only the texts and tasks that meet certain statistical features will be selected for the inclusion in the proficiency test. In this research we needed to establish ways of comparing the text difficulty of the passages in the two different proficiency batteries, thus we employed two methods of comparing text difficulty prior to analysing the results of the administration of the tests. One method, which was based on the subjective judgement of three expert judges, did not produce satisfactory reliability figures for determining the text difficulty of the passages; this is in accordance with the bulk of previous research[8] findings that the judges could not agree as to what an item was testing. The other method, using Flesch Readability Ease, produced a reliable estimate of text difficulty, confirming that the reading passages in the two batteries were not significantly different in terms of their text difficulty. Mention was made that this evidence should be considered as a *predictor* not a *measure* of text difficulty, which can only be assessed when the tests are administered.

4.1.2 Propositional Content Results

This part is mainly related to the investigation into the propositional content comparability of TOEFL and IELTS, and reports the analysis of the propositional content of the passages in the two batteries. We have explained in 3.7.1.1 that the propositional content relates to the characteristics of the information in the context and in the discourse and includes degree of contextualisation, distribution of information, type of information, and topic. The main research question here is:

Q 2: *Are the reading passages in the two batteries significantly different in terms of their propositional content?*

The three expert judges who rated the text difficulty of the reading passages were asked to rate the *degree of contextualisation, distribution of new information, type of information, and topic* of the reading passages according to the criteria explained in

[8] See 4.1.3 for a discussion of the previous research on this issue.

Chapter 3 (Tables 3.10-3.13). We will first report the results of their ratings for each subsection of the propositional content and then discuss their implications for the research question.

Prior to reporting the results, it seems warranted to explain the statistic used for measuring the agreement among the raters. Like the rating of text difficulty, a standard reliability index cannot be calculated for the raters, as the number of rating categories for propositional content facets is small. Bachman, Davidson and Milanovic (1996) have suggested the use of a measuring scale, which they call RAP (Rater Agreement Proportion). The RAP measures the proportion of rater agreement on each facet/item, where there are five raters. The RAP is 1.0 (5/5) if all five raters agree, .8 if four do, .2 if two do, and 0 if no agreement. The RAP is easy to conceptualise but it has one disadvantage: it cannot take account of extreme ratings. For example, if two raters gave a 2 and three a 1, the RAP would be the same as two gave a 2 and three a 0. Clapham (1996) suggests a similar method of calculating the rater agreement for three raters and taking account of extreme ratings.

> *"If all three raters agree, the RAP figure is 1, and if two agree with a difference of 1 between the ratings, then the RAP is .67 (2/3). If, however, two agree, but the third is two points away from the others, then the RAP is .33. If no one agrees the RAP is .0. This, therefore, could be called a Weighted RAP or WRAP."* (Clapham, 1996, p.151)

Following Clapham (1996), we have used the WRAP statistic to determine the agreement among the raters in the discussions to follow[9].

Contextualisation Results

The judges were asked to rate the contextualisation of the passages considering the relative proportion of *new* to *contextual* information. New information was defined as that which is not known to the test taker and cannot be predicted from the context, whereas contextual information was defined as that which is developed in the passage itself. The degree of contextualisation was assessed with respect to two criteria: cultural content, and specific topical content.

[9] This is because there were three raters in this research.

Table 4.6 demonstrates how the judges rated the reading passages for the degree of their contextualisation with respect to cultural content.

Table 4.6: Degree of Contextualisation With Respect To Cultural Content

	TOEFL					IELTS		
	TR1	TR2	TR3	TR4	TR5	IR1	IR2	IR3
Rater 1	-	1	-	1	-	0	0	1
Rater 2	2	1	0	0	1	0	1	1
Rater 3	0	1	1	1	0	2	2	2
WRAP	0	1	0	.67	0	.33	0	.67

TR=TOEFL Reading IR=IELTS Reading

0 = not at all contextualised 2 = highly contextualised

With the exception of TOEFL Reading 2, the judges disagreed among themselves on the degree of contextualisation of the passages with respect to their cultural content. The mean WRAP for the agreement of the raters was only 0.33. A careful observation indicates that the judges had difficulty on deciding about the cultural content of the IELTS reading passages, which is a reflection of their view expressed earlier about the subjects' schemata. Rater 3 thought that IELTS texts were highly sensitive to British culture while the other two did not think this was the case as they assumed that the texts were supposed to represent British academic texts, and thereby ignored the influence of British culture on the contextualisation of the passages. This is an indication of the judges' misinterpretation of the instructions.

Table 4.7: Degree of Contextualisation With Respect To Specific Topical Content

	TOEFL					IELTS		
	TR1	TR2	TR3	TR4	TR5	IR1	IR2	IR3
Rater 1	2	1	2	2	2	2	1	2
Rater 2	2	2	2	2	2	2	1	1
Rater 3	0	1	2	1	0	2	2	2
WRAP	.33	.67	1	.67	.33	1	.67	.67

TR=TOEFL Reading IR=IELTS Reading

0 = not at all contextualised 2 = highly contextualised

Looking at Table 4.7 one can observe that the judges had also their disagreement on deciding about the degree of contextualisation with respect to specific topical content. However, the mean WRAP for the agreement of the raters improved significantly in this exercise to 0.67. Despite their disagreement on most reading passages, the judges had agreed on the degree of contextualisation of TOEFL Reading 3 and IELTS Reading 1, considering them as highly contextualised with respect to specific topical content. This, again, is a reflection of how judges perceived the difficulty of the texts as expressed in their interviews. TOEFL Reading 3, as has already been explained, owes much of its difficulty to the topic and in that respect is quite different from the other TOEFL texts. That is why all the judges considered the passage as *highly contextualised* (2) with respect to topical content. IELTS Reading 1 is similar in some ways to TOEFL Reading 3 in that unless the subjects understand or are familiar with the topic, it can impede their comprehension. The topic requires the readers to have some knowledge of Britain, and British business in particular.

Distribution of New Information Results

This facet was rated in relation to the compactness or diffuseness of the discourse, assuming that highly compact or diffuse discourse may be quite difficult to process and demanding of the test taker's competence. The judges were asked to rate the discourse of a given passage as highly compact[10] if the distribution of new information was over a relatively short space or time, and rate it highly diffuse if the distribution of new information in the discourse was over a relatively long space or time. Table 4.8 shows how the judges rated the distribution of new information in the texts.

Table 4.8: Distribution of New Information Rating

	TOEFL					IELTS		
	TR1	TR2	TR3	TR4	TR5	IR1	IR2	IR3
Rater 1	0	-	-	2	-	2	-	1
Rater 2	2	2	1	2	1	1	1	0
Rater 3	0	0	1	1	1	2	2	2
WRAP	.33	0	.67	.67	.67	.67	0	0

TR=TOEFL Reading IR=IELTS Reading

0 = Highly compact 2 = Highly diffused

[10] See Chapter three, 3.7.1.1 for the description of Compact/Diffuse.

This was the most difficult task for the judges. The mean WRAP for the agreement of the raters was only 0.38. Rater 2 perceived the concept of new information in a completely different way from the other two judges. Although the judges did not find the instructions for rating this facet ambiguous at first, the follow up interviews revealed that the judges had serious doubts about what they were asked to rate in this task. Compactness and diffuseness confused the judges as how to rate the distribution of new information. The judges felt that these terms were not useful in ascribing new information to a context. Compact or diffused did not seem to be as transparent as they were expected to be.

Type of Information Results

Type of information in a test task was classified along two dimensions: concrete / abstract, and negative / positive. The judges were asked to rate this facet in terms of the relative degree of abstractness or negativeness of the passage. What was of concern here was the information contained in the text, and not how the test taker was expected to process that information.

Table 4.9 demonstrates how type of information was rated with respect to abstractness of the passages. Since abstractness was rated in relation to the information contained in the passages, not how the test takers were expected to process that information, the judges had less difficulty in rating this facet.

Table 4.9: Type of Information: Abstract

	TOEFL					IELTS		
	TR1	TR2	TR3	TR4	TR5	IR1	IR2	IR3
Rater 1	2	2	1	1	2	1	0	1
Rater 2	2	2	1	1	1	1	1	0
Rater 3	2	2	1	2	1	0	1	1
WRAP	1	1	1	.67	.67	.67	.67	.67

TR=TOEFL Reading IR=IELTS Reading

0=Abstract ◄——————► 2=Concrete

The mean WRAP for the agreement of the raters was relatively high 0.79. It is interesting to note that none of the judges rated TOEFL texts as abstract, while they

rated at least one of the IELTS texts as abstract. Nevertheless, they failed to agree which IELTS passage was abstract; each rated a different passage as abstract. This needed more attention, hence, the researcher decided to exclude the IELTS texts from the analysis and look at the correlation figures computed based on TOEFL texts. The mean WRAP obtained increased to an acceptable 0.87 figure. This might suggest that the judges had more problems deciding on the abstractness of IELTS texts than they did for the TOEFL texts

For the negative (NEG) rating, the raters were expected to consider only explicitly marked negatives, recognising that negatives might be explicitly marked in English in a variety of ways. Although negativeness was expected to be one of the easiest tasks for the raters, the mean WRAP for the agreement of the raters 0.67 is not very promising. Table 4.10 illustrates how type of information was rated in terms of the relative degree of negativeness of the passages.

The judges were asked to consider only explicitly marked negatives and in doing that one would expect very little discrepancy in their ratings. The correlation results prove otherwise. Since negatives might be explicitly marked in English in a variety of ways, the judges could not agree what explicitly marked negatives were in different passages. Further review of the judges' follow-up interviews indicated that they had difficulty deciding about IELTS passages. It was decided to exclude the IELTS passages from the analysis and see if there was still disagreement among the judges. The mean WRAP for the agreement of the raters increased significantly to a 0.80 when IELTS texts were excluded from the analysis.

Table 4.10: Type of Information: Negative

	TOEFL					IELTS		
	TR1	TR2	TR3	TR4	TR5	IR1	IR2	IR3
Rater 1	2	2	1	2	2	2	1	2
Rater 2	2	2	0	1	1	2	2	1
Rater 3	2	2	1	1	1	1	2	0
WRAP	1	1	.67	.67	.67	.67	.67	0

TR=TOEFL Reading IR=IELTS Reading

0=Negative ◄——————► 2=Positive

It may be that some IELTS text features discouraged the judges from carefully examining this facet. Perhaps, the length or the information load of the IELTS passages was a factor; as we will shortly discuss in 4.1.4, TOEFL texts were much shorter. Irrespective of the possible cause, the relatively high raters' agreement (WRAP=0.80) on TOEFL texts seems to indicate that TOEFL texts were much more transparent in terms of their negativeness / positiveness for our judges.

Topic Specificity Results

Topic refers to what a given piece of discourse is about. The judges were asked to rate the specificity of the reading passages with respect to their topics. Since all the reading passages in the study were judged to be academic, the judges were asked to observe whether the texts were biased towards a specific discipline or they were just general academic. Table 4.11 shows the outcome of this task.

Table 4.11: Topic Rating

	TOEFL					IELTS		
	TR1	TR2	TR3	TR4	TR5	IR1	IR2	IR3
Rater 1	1	1	2	1	1	1	2	2
Rater 2	2	2	1	2	1	2	2	2
Rater 3	1	2	1	2	2	2	2	1
WRAP	.67	.67	.67	.67	.67	.67	1	.67

TR=TOEFL Reading IR=IELTS Reading

1=Discipline specific 2=General academic

Despite a moderate agreement among the raters (Mean WRAP=0.71), there was a substantial degree of disagreement between the judges. What was assumed as biased towards a specific discipline for one judge did not necessarily mean the same thing for other judges. Instruction for rating this facet was intended to be self-explanatory and none of the judges indicated otherwise in their follow-up interviews. It seems that we are caught here in the familiar argument as what '*specific*' means. What is the boundary of specificity and how are we going to decide when to consider a text as biased towards a subject discipline? There is no easy answer for this. Clapham (1996) reports similar problems when she tried to rate IELTS texts with respect to their subject specificity. Clapham argues that it is difficult to know in advance how specific a passage will be.

> *"My own intuitions about the texts' specificity proved to be only partly correct, and I had to adjust the 'general' and 'specific' labels that I had assigned to the reading passages once I had studied the results of repeated measures analyses of variance, the analysis of bias, and the students' comments on the familiarity of the reading passage subject areas. The fact that these passages were selected by experienced EAP teachers, and checked by members of the IELTS project committee, suggests that the test constructors were not aware that test specificity might pose a problem."* (Clapham, 1996, pp. 198-199)

What is more revealing in this research is the fact that two of the judges (Rater 1 and Rater 2) have both had a long history in ESP teaching and testing and have published some joint papers together, as well as working as colleagues in the same department; none of which appears to have brought their concepts of specificity any closer. As it stands, it seems that specificity means different things to different people and a reliable measure of specificity is hard to achieve.

4.1.3 Discussion of Propositional Content Ratings By Judges

In comparing the propositional content of the tests, it was decided to use the subjective evaluation of the expert judges. Previous research (Alderson, 1990a) has shown that in evaluating what an item is testing, the judges usually disagreed. However, Bachman and his colleagues (1993, 1995, 1996) report very satisfactory agreement figures for using the judgement of expert raters in their research based on a rating instrument proposed by Bachman (1990). Using Bachman's modified rating instrument (see 2.6 for the details) three expert applied linguists rated the propositional content of the reading passages in this research. The purpose of the ratings was to see the degree of the comparability of the propositional content of the two batteries; an attempt to find an answer to question 2:

Q 2: *Are the reading passages in the two batteries significantly different in terms of their propositional content?*

The results of the ratings suggest that although the judges found the rating instructions relatively self-explanatory and had very few problems in following them, they disagreed on what constituted the propositional content of the passages. The

source of the majority of disagreement among the judges can be traced back to the background knowledge of the test takers for whom the texts were prepared. The judges felt that it was difficult to rate some of the facets of the propositional content if the intended audience was not known to them. Since proficiency tests are targeted at a wide population of testees with variable language abilities, it is difficult to comment on the schemata of the test takers. If the background knowledge of the intended audience is not taken into consideration, the rating of the facets of the propositional content will be subject to the raters' interpretation of what they believe an intended audience should be. This is obviously a source of error variance, which affects the reliability of their ratings.

The rating instrument used in assessing the propositional content did not produce satisfactory reliability figures though it seemed to be logically valid for assessing a subjective rating of this nature. The instrument has been carefully developed over the years by a team of expert judges to be used in similar situations. The instrument produced acceptable reliability figures in the original study (Bachman et al., 1995), but it failed to produce acceptable reliability figures in a different study (Clapham, 1996). One possible explanation for this failure in the present study, and possibly in Clapham's (1996) study, could be the fact that the expert judges in both studies did not develop the rating instrument themselves and hence did not have much chance to discuss it together. On the other hand, in the case of Bachman et al. (1995) study, the judges not only developed and modified the instrument but also worked together as a team on the same project for a very long period of time, during which they possibly shared the same frame of reference. The judges in this research did not enjoy that company and relied on their own intuition of what they perceived from the instructions.

One may argue then that in order to improve the reliability of the ratings, the judges should have spent several meetings in discussing the instrument and in improving it so that they would have all shared the same frame of reference. There are two problems associated with this suggestion. Firstly, although this is ideal from the research point of view, it is impractical in the real world with the limited resources most researchers have at their disposal. Secondly, if we claim that a rating instrument is carefully worded and is worthy of being used in similar situations, we imply that there is no need for further major modifications. In other words, we suggest that the instrument can be used by other expert judges to produce reliable ratings in similar situations. In this research the judges were provided with a trialled rating instrument designed to

assess the propositional content of language proficiency tests. They were briefed about the underlying assumptions of the instrument, and in most cases, they were supplied with samples, which had been rated by other expert judges. Yet the reliability of the judges' ratings was not satisfactory. It is probable that the reliability figures could have been improved had the judges spent more time on the instrument and had several joint discussions. Had that been the case, it would have possibly resulted in the modification of the entire instrument and the production of some totally new instrument. The question that remains is whether the new instrument, if used by some different judges, would produce sufficient reliability figures in similar future exercises.

It seems warranted to clarify an important distinction between what we understand about the reliability of a rating exercise and the usefulness of the rating instrument. Reliability is an issue related to the rating exercise not to the instrument used in that exercise. A rating exercise is reliable to the extent that the judges' ratings are consistent. It is possible to improve the reliability of an exercise by asking the judges to work together for several sessions and try to compromise among themselves some of their differences about the application of the rating criteria in the exercise; this may involve the modification of some of the components of the rating instrument used in the exercise. But it is not appropriate to think of the reliability of the instrument itself. The instrument is a tool made up of some criteria that has to be *valid* for the purposes for which it is designed. That is, the judges using the rating instrument should agree on its usefulness in measuring the facets under investigation. Once the initial agreement on the usefulness of the instrument is achieved, it is then possible to observe how reliable the actual rating exercise is. The follow-up interviews indicated that the judges in this research had their disagreements on the usefulness of some of the facets of the propositional content instrument, i.e., compactness / diffuseness of new information and contextualisation. This could have been a factor affecting their ratings.

In an attempt to investigate the usefulness of the propositional content instrument, the researcher and Rater 3 decided to work together and rate the propositional content of the passages once more. They met a few times to discuss the instrument, however, due to the serious reservations that Rater 3 had about the propositional content instrument, they could not find common ground on which they could base their judgements. Their interpretations of the descriptors for contextualisation with respect to cultural content, distribution of new information (compactness / diffuseness), and

topic specificity were so different that did not leave much room to compromise. However, they agreed on their interpretation of contextualisation with respect to specific topical content and type of information (abstract / concrete, negative / positive). Since overall there was more disagreement than agreement between the judges about the usefulness of the propositional content instrument, it was decided to abandon the exercise. It is worth mentioning that the researcher and Rater 3 could agree on the interpretation of the communicative language ability instrument and the reliability of their ratings is satisfactory. This will be discussed in 4.1.6 and 4.2.

It is clear from the remarks above that a valid rating instrument, which has produced reliable ratings in the past, may not be valid if used by different judges on different samples. The propositional content rating instrument used here was part of the two rating instruments used by Bachman and his colleagues in their research, which resulted in satisfactory reliability figures. The failure of the same instrument to produce reliable estimates of the propositional content ratings of the passages here could be due to two reasons. Firstly, the descriptions of the criteria of the ratings were not as transparent as Bachman and his colleagues would have hoped them to be. Neither the three expert judges, nor Rater 3 and the researcher, could agree on their interpretations of the criteria[11]. This indicates that the validity of the propositional content rating instrument was questionable.

Secondly, the three-point scales (0-1-2) used for the rating instrument and the number of readings (8) used were too small a sample to produce high correlation figures. A slight discrepancy between the judgement of any two judges would skew the correlation. This is especially true when, for instance, we observe that in the case of the ratings of the contextualisation with respect to specific topical content and type of information (abstractness and negativeness) where the judges agreed on the usefulness of the criteria, we still fail to observe satisfactory inter-rater reliability. A careful observation of the ratings in Tables 7, 9, and 10, demonstrates how close the judges were in their ratings. Figure 4.2, for instance, visualises this closeness of the judges' views about the degree of the abstractness of the passages in the two batteries.

[11] The three expert judges did not meet to discuss the instrument while the researcher and Rater 3 had a few joint sessions to discuss it.

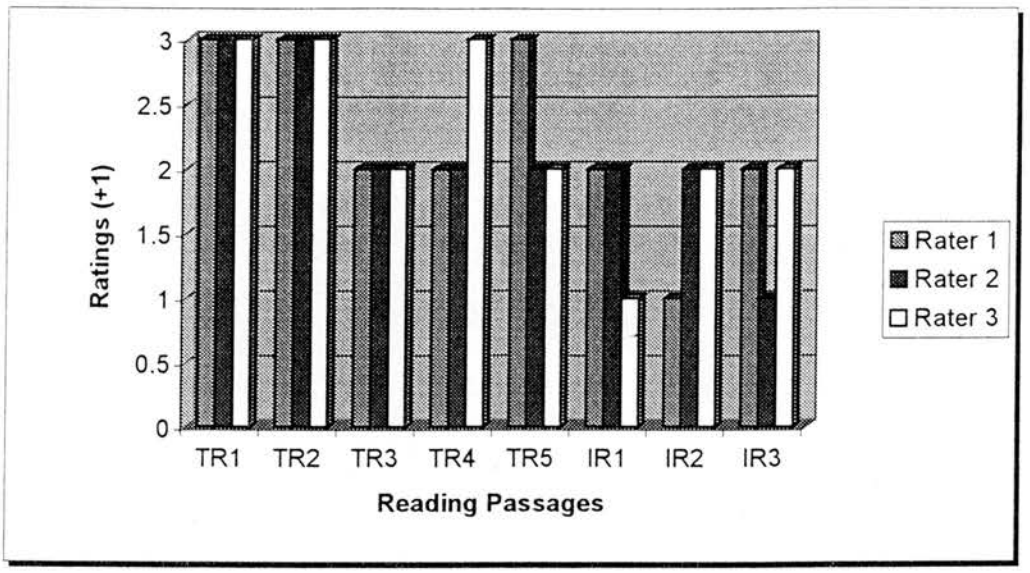


Figure 4.2: Ratings of Type of Information: Abstract

It appears that at least on these three ratings we would have achieved a better reliability figure had there been more items to rate. It is worth noting that even in the case of the ratings of Bachman et al. (1995) study, the reliability of the ratings for the propositional content of the reading passages is not reported separately; what is reported is the reliability of the whole 48 facets of test methods and components of communicative language ability on all the individual test items (over 250 items). It is possible that the reliability of their ratings on these facets could also have been not as high as the one reported for the whole items. Our sample of the propositional content rating instrument was just too small a sample to allow it gain any statistical significance. As we will see shortly in the discussions of the communicative language ability ratings, the reliability of the ratings improves significantly when more items are included in the sample to rate.

Although the above-mentioned problems did not allow us to examine Question 2 of the research reliably, it shed light on the complexity of the propositional content of texts and how the subjective ratings of this sort could be pursued. The propositional content analysis exercise reported above served as a pilot study, giving us insights on modifying the ratings of the rest of the facets of test methods and the components of communicative language ability. The two important outcomes of this exercise were: the need for the training of the judges, and for an increase of the number of items to

be rated. These two shortcomings will be dealt with in the subjective ratings to be discussed in 4.1.6 and 4.2.

4.1.4 Results of the Analysis of Length Across The Batteries

Question one of the research dealt with the comparability of the passages in terms of their length.

Q 1: Are the reading passages in the two batteries significantly different in terms of their length?

We have already argued that length usually has an effect on other characteristics of test methods. The longer a text is, the more probable it is that the grammatical components (content words, clauses, embeddings, passives, and cohesive devices) are abundant too (See Henning, 1988). This brings in a heavier load of information for the test taker to process, which in the end could contribute to task difficulty. Examining this variable is believed to be one of the factors that need to be accounted for in comparing test content. Measuring length was achieved by counting the total number of words for each reading passage. Additionally, the content words as well as the total vocabulary were counted for *type / token* ratio[12]. Figure 4.3 demonstrates that IELTS texts were much longer than the TOEFL ones.

[12] See Table 4.13 for the results.

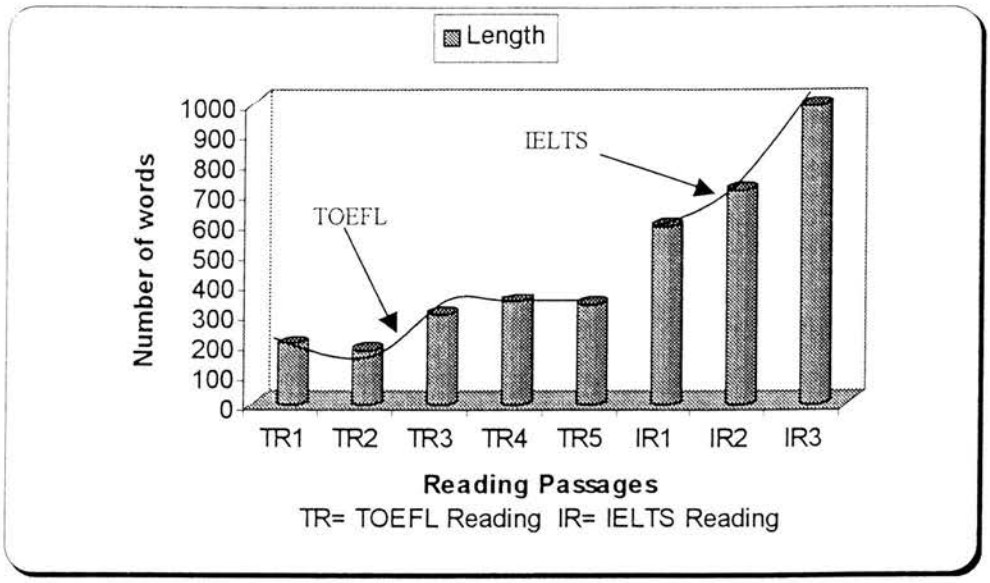


Figure 4.3: Comparison of Length: TOEFL and IELTS

Table 4.12 sets out the relevant means, median and standard deviations, indicating that the length of the reading passages in the two battery tests varied widely; IELTS passages are on average 2.5 times longer than the TOEFL ones.

Table 4.12: Comparison of Mean Facet For Length

Total No. of Words	No. of Passages	Mean	Median	Standard Deviation
TOEFL	5	273.20	303	76.66
IELTS	3	764.67	713	204.46

A Chi-Square test (Chi-Square=235.296, df=1, $p<0.000$) confirms that the difference between the length of the passages in the two batteries is statistically significant. This allows us to reject the null-hypothesis $\mathbf{H_0}$ **1**, which holds: *The reading passages in the two batteries are not significantly different in terms of their length.* Thus, the alternative hypothesis $\mathbf{H_1}$ **1** seems to be in order in this case: *The reading passages in the two batteries are significantly different in terms of their length.* This shows that the two tests are not comparable in terms of the length of their reading passages.

We have mentioned that length of a test could potentially affect other facets of test methods and in some cases add an unnecessary burden on the test taker's memory to

process a piece of discourse. That is perhaps on the assumption that the longer the text is, the more likely that there is an increase in the number of sentences, content words, clauses, modifiers, and paragraphs. However, as will be shown in the discussion of Grammar shortly, there is no evidence that the length had any significant effect on the grammatical complexity of the texts here, making them more or less accessible to the readers.

4.1.5 Results of The Analysis of Organisational Characteristics

It is implicit from the term '*organisation*' that the length of the language sample is an important factor in the amount of incorporation of organisational characteristics required for a successful interpretation of the language. Since it was found that IELTS passages were significantly longer than their TOEFL counterpart, the examination of the organisational characteristics of the passages should shed light on the possible effect of length on such features. The main research question here is,

Q 3: *Are the reading passages in the two batteries significantly different in terms of their organisational characteristics?*

Organisational characteristics are defined as those features of the discourse, which relate to the formal organisation of language. They are divided into Grammar and Cohesion components. The reading passages of each battery were analysed for the comparability of their organisational characteristics[13].

4.1.5.1 Results of the Analysis of Grammar Facet

In 3.7.1.1 we discussed the complexity of Grammar facet and the difficulty of deciding which aspect of grammar to examine when interpreting how the formal structure of the texts is laid out, so that a meaningful grammatical comparison may be possible across the texts. It was decided to arbitrarily associate Grammar with syntactic complexity, lexical density, and text difficulty. Text difficulty has already been discussed and its results have been reported in 4.1.1. We should now discuss the

[13] See section 3.7.1.1 for the details of how these components were measured.

other two aspects: syntactic complexity, and lexical density. The two research questions being addressed here are:

Q 3.1: *Are the reading passages in the two batteries significantly different in terms of their syntactic complexity?* And,

Q 3.2: *Are the reading passages in the two batteries significantly different in terms of their lexical density?*

Syntactic complexity was basically defined in terms of sentence complexity and voice. Sentence complexity was assessed with regard to the ratios of Word per Sentence, and Clause per Sentence, while voice was measured in terms of the proportion of Passives in a text. Lexical density was rated in terms of the ratios of Character per Word and Type / Token, as well as the traditional Lexical Density measurement (ratio of Content Words to Total Words).

A simple count was first used for each of the following in the reading passages of the two batteries: content words, total words, total vocabulary, clauses, embeddings, two or more pre-modifiers in NP, sentences, paragraphs and characters. Then, various relevant ratios were obtained. Table 4.13 presents the details of the Grammar rating for the reading comprehension passages.

Table 4.13: Organisational Characteristics: Grammar Facets

Reading Comprehension Passages

	TOEFL					IELTS		
Categories	R1	R2	R3	R4	R5	R1	R2	R3
A: COUNT								
CW	105	84	159	167	165	339	342	475
TOKEN	204	179	303	344	336	591	713	990
TYPE	127	107	185	180	167	312	346	371
CLAU	25	14	34	44	36	65	74	78
EMBED	11	7	13	19	14	23	31	44
PREMOD	4	4	4	4	4	18	5	25
Sentences	11	7	14	22	15	26	35	25
Paragraphs	1	1	2	6	3	10	10	8
Characters	984	885	1585	1657	1709	3145	3513	5172
B: RATIOS								
TYPETOK	0.62	0.65	0.61	0.52	0.50	0.53	0.49	0.38
LEXDEN	0.52	0.47	0.53	0.49	0.49	0.57	0.49	0.48
CHARCWD	4.82	4.94	5.23	4.82	5.09	5.32	4.93	5.22
SENT	11	7	7	3.66	5	2.6	3.5	3.13
CLSENT	2.27	2	2.43	2	2.4	2.5	2.11	3.12
WDSENT	9.54	25.57	21.64	15.64	22.4	22.73	20.37	39.6
VOICE	18	33	21	18	53	26	26	12

CW= # content words TOKEN= # total words TYPE= # total vocabulary
CLAU= #of clauses EMBED= # of embeddings PREMOD= # 2/more pre-modifiers in NP
TYPETOK= TYPE/TOKEN LEXDEN= lexical density (CW/TOKEN)
CHARCWD= Character/word SENT= # of sentences per paragraph
CLSENT= Clause/Sentence WDSENT= Words/Sentence VOICE= % passives

On the one hand, section A of Table 4.13 might lead one to the initial conclusion that IELTS passages, on average, have more content words, clauses, embeddings, pre-modifiers, sentences, and paragraphs than the TOEFL ones. This is only a reflection of their longer texts; the longer the text is, the more likely that these factors are abundant too. On the other hand, section B seems to illustrate that the complexity of the sentence structures and that of the vocabulary across the passages are not very different. In order to test the significance of the difference of sentence complexity, voice, and lexical density across the batteries, a t-test was conducted on these measures. Additionally, a Levene's test of equality of means and variances was applied as the samples did differ in their size. We also included Flesch Reading Ease into the analysis as it was considered to be one of the Grammar components. Table 4.14 presents the comparison for the equality of means and variances for the Grammar facets.

Table 4.14: Comparison for the Equality of Means and Variances: Grammar Facets in the Reading Passages of TOEFL and IELTS

Facet	Variance assumption	Levene's Test for Equality of Variances		t-test for Equality of Means		
		F	Sig.	t	df	**Sig. (2-tailed)**
Type/Token ratio	Equal variances assumed	.035	.857	2.200	6	.070
	Equal variances not assumed			2.089	3.711	.110
Lexical density	Equal variances assumed	5.220	.062	-.642	6	.544
	Equal variances not assumed			-.522	2.462	.645
Character / word	Equal variances assumed	.027	.876	-1.292	6	.244
	Equal variances not assumed			-1.241	3.837	.285
Word / sentence	Equal variances assumed	1.713	.239	-1.494	6	.186
	Equal variances not assumed			-1.298	2.891	.288
Clause/sentence	Equal variances assumed	2.687	.152	-1.445	6	.199
	Equal variances not assumed			-1.164	2.417	.347
% Passive Sentences	Equal variances assumed	1.272	.302	.760	6	.476
	Equal variances not assumed			.890	5.999	.408
Flesch Reading Ease	Equal variances assumed	3.223	.123	1.094	6	.316
	Equal variances not assumed			1.379	5.228	.224

The *t* values reported for word / sentence and clause /sentence, which were the criteria for sentence complexity, are not significant at $p<0.05$ level. Neither are the values for the proportion of passive sentences (voice), suggesting that the reading passages in the two batteries did not differ significantly in terms of their voice or sentence complexity. Since these two facets are part of the syntactic complexity facet, it can be concluded that the reading passages in the two batteries did not differ significantly in terms of their syntactic complexity. Therefore, we are obliged to retain the null-hypothesis of no significance with respect to syntactic complexity.

H_0 3.1: There is no significant difference in the syntactic complexity of the reading passages in the two batteries.

The *t* values reported for type/token, character/word, and lexical density are not significant at $p<0.05$ level either, suggesting that the reading passages in the two

batteries did not differ significantly in terms of their lexical density. Hence, the null-hypothesis **H_0 3.2** is retained.

H_0 3.2: There is no significant difference in the lexical density of the reading passages in the two batteries.

The *t* values reported for Flesch Reading Ease are not significant at $p<0.05$ level. This yet again confirms that the reading passages in the two batteries did not differ significantly in terms of their readability, a finding earlier reported in 4.1.1.

We started the discussion about the grammar component of test method facets by dividing it into three facets: syntactic complexity, text difficulty and voice. We then analysed TOEFL and IELTS passages and concluded that they did not differ significantly in terms of the above three facets. In doing our analysis we used a number of other linguistic features associated with grammar such as the number of two or more pre-modifiers in NP, embeddings, etc. In this section we will report the results of the correlational analysis of such features and discuss how various grammar components correlated. This is done to identify any potential factor, which might have affected our results on the grammar facet.

Table 4.15 sets out the correlation matrices for most of the components of the grammar facet. Significant correlations are highlighted. A careful observation of this table reveals that VOICE and LEXICAL DENSITY have no significant correlations with any of the other grammar facets. That is, changes in these two facets will not have meaningful changes on the other components of the grammar facet.

Table 4.15: Correlation Matrices (Pearson) for Grammar Facets: Reading Passages (IELTS & TOEFL)

	CHARWD	WDSENT	TYPETOK	CLSENT	SENTPARG	EMBED	PREMOD	LEXDEN	VOICE	FRE	CW	CLAU
CHARWD	1.000											
WDSENT	0.624	1.000										
TYPETOK	-0.348	-0.637	1.000									
CLSENT	**0.711***	**0.731***	-0.675	1.000								
SENTPARG	-0.439	-0.545	**0.732***	-0.276	1.000							
EMBED	0.350	0.656	**-0.906****	0.677	-0.669	1.000						
PREMOD	0.656	**0.750***	-0.697	**0.849***	-0.534	**0.778***	1.000					
LEXDEN	0.544	-0.246	0.158	0.162	-0.047	-0.089	0.237	1.000				
VOICE	0.037	-0.079	0.132	-0.273	-0.026	-0.437	-0.395	-0.095	1.000			
FRE	**-0.831***	-0.637	0.138	-0.575	0.166	-0.241	-0.441	-0.276	-0.125	1.000		
CW	0.536	0.680	**-0.861****	**0.714***	-0.722	**0.964****	**0.850****	0.121	-0.353	-0.403	1.000	
CLAU	0.417	0.494	**-0.848****	**0.548**	**-0.769***	**0.937****	**0.704**	0.159	-0.334	-0.249	**0.961****	1.000

CHARWD= character/word WDSENT= word/sentences TYPETOK= type/token CLSENT= clause/sentence SENTPARG= sentence/paragraph

EMBED= embeddings PREMOD= # 2/more pre-modifiers in NP LEXDEN= lexical density VOICE= % passives FRE= Flesch Reading Ease

CW= # content words CLAU= # clauses

* Correlation is significant at the 0.05 level (2-tailed).

** Correlation is significant at the 0.01 level (2-tailed).

On the contrary, the number of content words (CW) seems to correlate highly with most of the other components. The greatest association, as one would suspect, is between the number of content words and the number of embeddings (0.964). Since the number of embeddings is also highly correlated with the number of clauses (0.937), it is not surprising that content words also correlate highly with the number of clauses (0.961). The more embeddings result in more clauses and consequently in more content words. Figure 4.4 illustrates the co-incidence of these three facets.

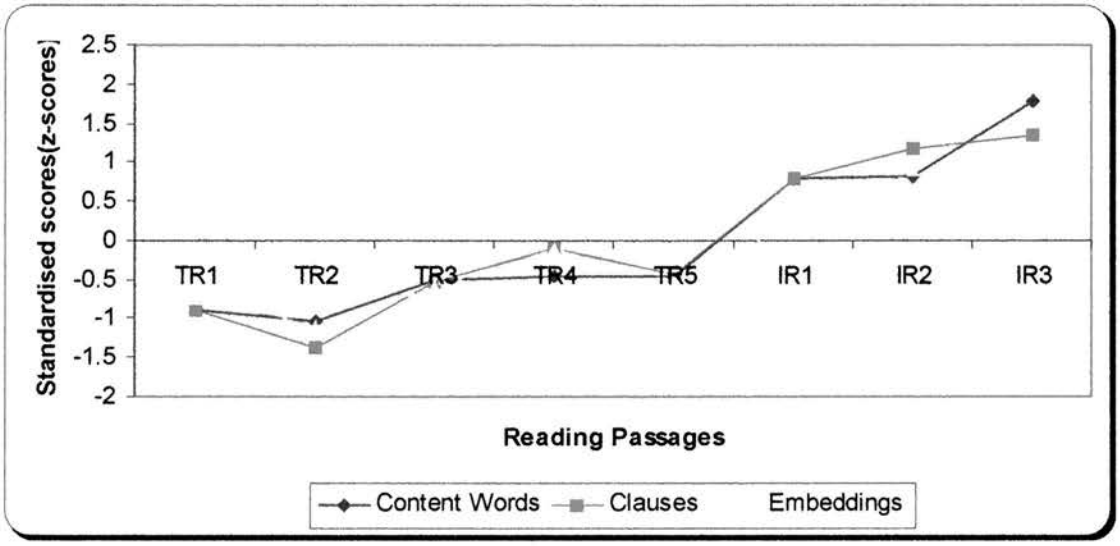


Figure 4.4: Correlations Between Grammar Facets: Content words, Clauses, Embeddings

Figure 4.5 illustrates an equally important association between the number of content words and the number of two or more pre-modifiers in NP (0.850). This is not surprising as modifiers are almost always classified under content words. Although the correlation between the number of pre-modifiers and the ratio of clause per sentence is significant (0.849), its correlation with the number of clauses (0.704) is not sufficient to be significant at 0.05 level. That is, an increase in the number of pre-modifiers in NP does not necessarily increase the number of clauses. Pre-modifiers also correlate highly with the ratio of word per sentence (0.750), but this is perhaps due to the high correlation that the pre-modifiers have with the number of sentences (0.931).

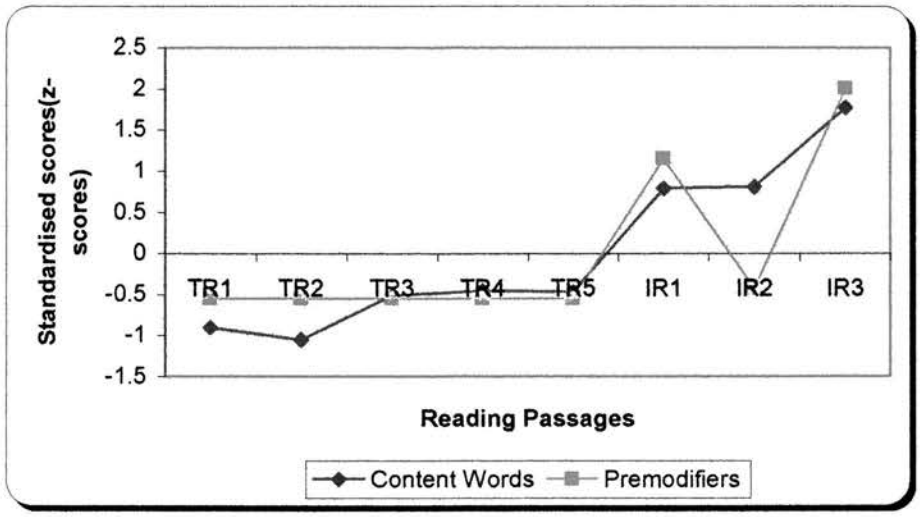


Figure 4.5: Correlations Between Grammar Facets: Content Words, Pre-Modifiers in NP

Finally, there is a high negative correlation between the Flesch Reading Ease index and the number of characters per word (-0.831). That means, the higher the rate of the characters per word (longer words) is, the lower Flesch index (more difficult text) becomes. This is a reflection on how Flesch Reading Ease is computed[14].

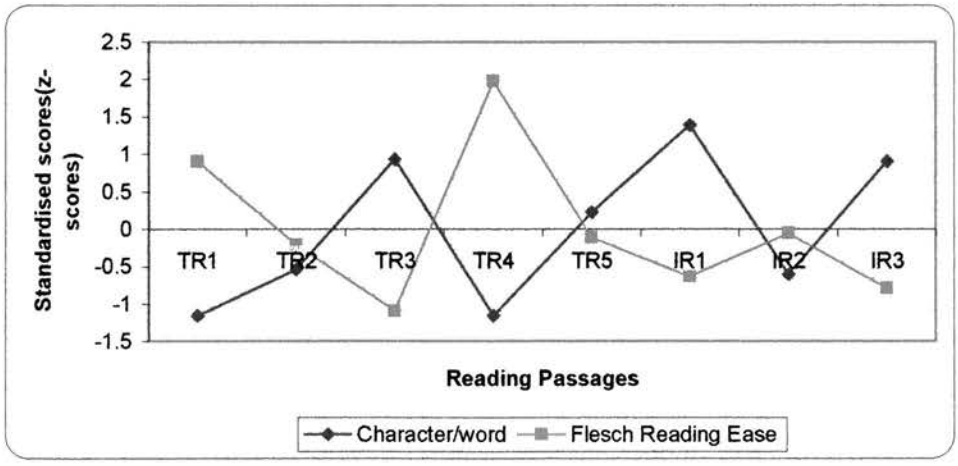


Figure 4.6: Reverse Relationship Between Flesch Index And Word Length

[14] Flesch Reading Ease is computed using the following formula.
FRE= 0.39(words/sentence) + 11.8(syllables/word) – 15.59

4.1.5.2 Results of the Analysis of Cohesive Markers

This section reports the results related to research question 3.4.

Q 3.4: Are the reading passages in the two batteries significantly different in terms of their cohesive markers?

Cohesion refers to those surface-structure features of a text, which link different parts of sentences or larger units of discourse. This facet was subdivided into six sub-facets: *Reference, Substitution, Additive, Adversatives, Causals,* and *Temporals.* Explicit cohesive markers of the above types were counted for each passage as evidence of their cohesive comparability[15]. Table 4.16 presents the number of cohesive markers found in each reading passage.

Table 4.16: Organisational Characteristics: Cohesion Facet

Reading Comprehension Passages

	TOEFL					IELTS		
	TR1	TR2	TR3	TR4	TR5	IR1	IR2	IR3
REFERENCE	4	0	1	8	1	1	5	2
SUBSTITUTION & ELLIPSES	1	2	4	0	4	2	8	4
ADDITIVE	2	1	4	1	0	12	17	9
ADVERSATIVE	1	1	2	2	2	2	4	2
CAUSAL	0	0	0	1	2	0	5	3
TEMPORAL	2	1	1	6	1	3	10	4

TR = TOEFL Reading IR = IELTS Reading

Table 4.17 sets out cohesive markers' mean and standard deviation values across the batteries.

[15] The counting was based on the definitions given in Halliday & Hassan 1976: pp. 242-3 (Appendix 7). See section 3.7.1.1 for how the cohesive markers were counted.

Table 4.17: Cohesion Facet Means Across The Batteries

Cohesion facet	Reading tests	N	Mean	Standard Deviation
Reference	TOEFL	5	2.80	3.27
	IELTS	3	2.67	2.08
Substitution & ellipses	TOEFL	5	2.20	1.79
	IELTS	3	4.67	3.06
Additive	TOEFL	5	1.60	1.52
	IELTS	3	12.67	4.04
Adversative	TOEFL	5	1.60	.55
	IELTS	3	2.67	1.15
Causal	TOEFL	5	.60	.89
	IELTS	3	2.67	2.52
Temporal	TOEFL	5	2.20	2.17
	IELTS	3	5.67	3.79

To test whether the mean values of the reading passages in the two batteries differed significantly in terms of their cohesive markers, parametric tests for the equality of means and variances were applied. Table 4.18 illustrates how the reading passages in the two batteries differed with respect to Cohesion.

Table 4.18: Comparison for the Equality of Means and Variances: Cohesion Facets in the Reading Passages of TOEFL and IELTS

Cohesion Facet	Variance assumption	Levene's Test for Equality of Variances		t-test for Equality of Means		
		F	Sig.	t	df	**Sig. (2-tailed)**
Reference	Equal variances assumed	.992	.358	.062	6	.952
	Equal variances not assumed			.070	5.872	.946
Substitution & ellipses	Equal variances assumed	1.095	.336	-1.475	6	.191
	Equal variances not assumed			-1.274	2.847	.297
Additive	Equal variances assumed	3.334	.118	-5.737	6	**.001**
	Equal variances not assumed			-4.554	2.344	**.033**
Adversative	Equal variances assumed	5.463	.058	-1.819	6	.119
	Equal variances not assumed			-1.502	2.553	.245
Causal	Equal variances assumed	3.318	.118	-1.740	6	.132
	Equal variances not assumed			-1.371	2.308	.288
Temporal	Equal variances assumed	1.938	.213	-1.688	6	.142
	Equal variances not assumed			-1.450	2.810	.249

With the exception of the means for the Additive category, the rest of the p-values for *t* are all greater than 0.05, rejecting the hypothesis that there is a significant difference between the reading passages in the two batteries in terms of the cohesive facets. This supports the null-hypothesis 3.4.

$H_0$3.4: *There is no significant difference in the number of cohesive markers in the reading passages of the two batteries.*

Because IELTS texts were significantly longer than the TOEFL ones, one would have expected to find significantly more cohesive markers in IELTS texts; the longer the text, the more frequent the use of cohesive markers to make it coherent. However, the results reported in Table 4.18 demonstrate otherwise. Only additives were more abundant in IELTS passages. There was no significant difference in the number of References, Substitutions and Ellipses, Adversatives, Causals, and Temporals across the reading passages.

At the outset of the organisational characteristics discussion mention was made that the length of a passage could be a determining factor in the amount of incorporation of organisational characteristics required for a successful interpretation of the language. The results of the analysis of various facets of organisational characteristics suggest that such an assumption is not warranted by the evidence presented here.

Having analysed all the facets related to the organisational characteristics, we can now address question 3 of the research:

Q 3: *Are the reading passages in the two batteries significantly different in terms of their organisational characteristics?*

The results of the analysis of grammar facets (syntactic complexity, lexical density, text difficulty) and cohesive facets support the hypothesis that the reading passages in the two batteries did not differ significantly in terms of these features. That lends support to null-hypothesis 3. That is,

$H_0$3: *The reading passages in the two batteries are not significantly different in terms of their organisational characteristics.*

4.1.6 Results of the Analysis of the Relationship of Item To Passage

The last facet related to Test Methods is the *relationship of item to passage.* This section addresses question four of the research.

Q 4: Are the relationships of the test items to the reading and listening passages significantly different in the two batteries?

All the reading and listening comprehension items in the two batteries were analysed to find out what kind of relationship they had with the passages to which the items referred. One of the judges who took part in the subjective judgement of the propositional content of the passages volunteered to rate the test items, in addition to the researcher. Based on the failure of the propositional content instrument to produce satisfactory reliability agreement, it was decided to train the judges. Since the judges were expected to rate both this facet and the components of communicative language ability, they arranged several sessions to discuss the instrument. Once the raters were relatively satisfied that they could follow the instructions, they embarked on rating all the reading and listening items in the two batteries. The instructions asked the judges to rate the relationship of the items to passage on a 1-5 scale.[16] The ratings were entered manually in the instrument sheets provided with each sub-test of the batteries.

There were altogether 155 test items to be rated. Figure 4.7 and Figure 4.8 illustrate the ratings of the judges on the reading comprehension items of IELTS and TOEFL, respectively. These figures indicate how close the raters were in their ratings. The correlation between the ratings of the two judges for the reading comprehension items was 0.88, and their inter-rater reliability was 0.82.

[16] The rating instrument read as follows:

5= Requires test taker to relate information in passage to the real world.

4= Item relates to the entire passage, and requires an understanding of the entire passage.

3= Relates to several specific parts of the passage, or requires test taker to relate one part of the passage to several others.

2= Relate to a specific part of the passage, and requires only localised understanding of that part.

1= No relationship to the passage; items can be answered without reference to the passage, or relationship of item to passage is not clear.

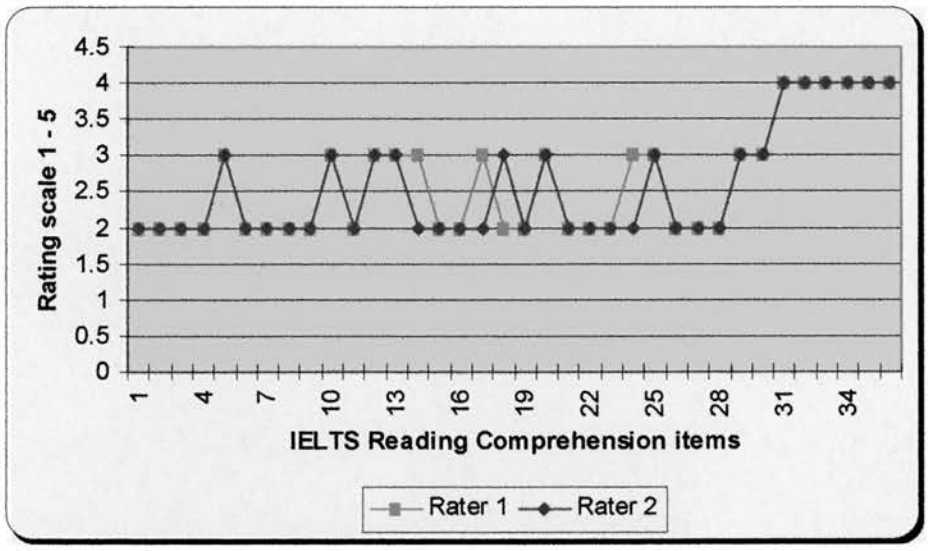


Figure 4.7: Relationship of Item To Passage Ratings: IELTS Reading Comprehension Items

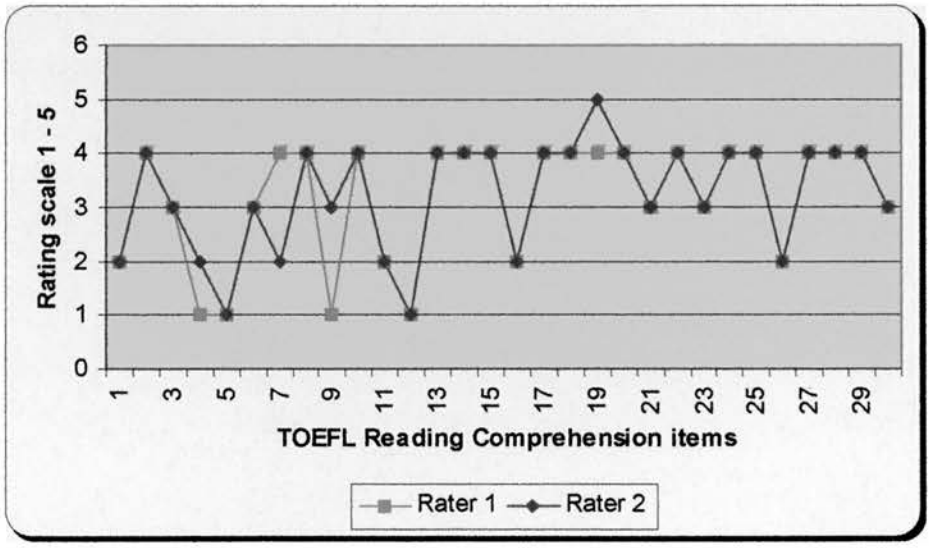


Figure 4.8: Relationship of Item To Passage Ratings: TOEFL Reading Comprehension Items

Figure 4.9 and Figure 4.10 illustrate the ratings of the judges on the listening comprehension items of IELTS and TOEFL, respectively. Again the figures show a

close relationship between the ratings of the judges. The correlation between the ratings of the two judges for the listening comprehension items was 0.73, and their inter-rater reliability was 0.75.

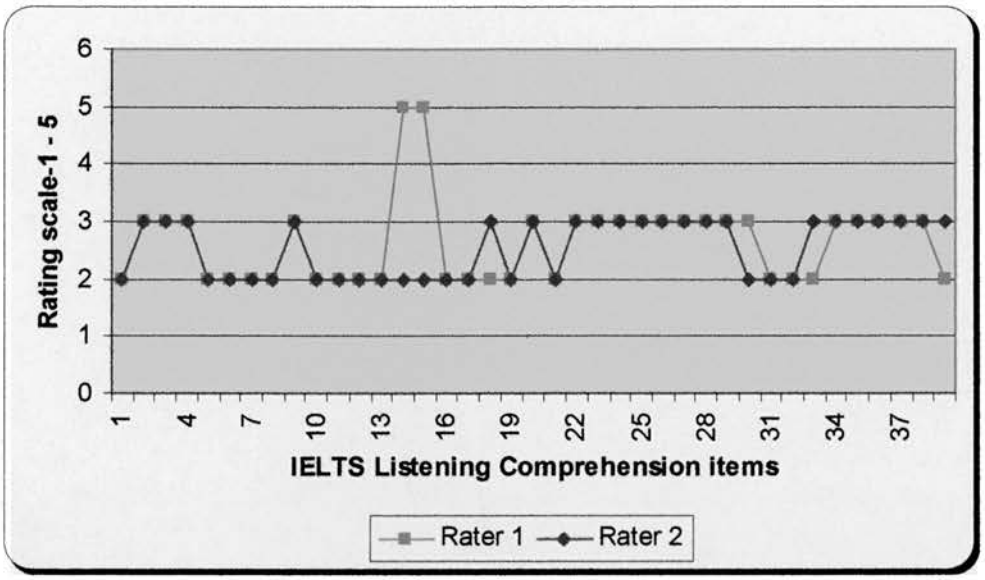


Figure 4.9: Relationship of Item To Passage Ratings: IELTS Listening Comprehension Items

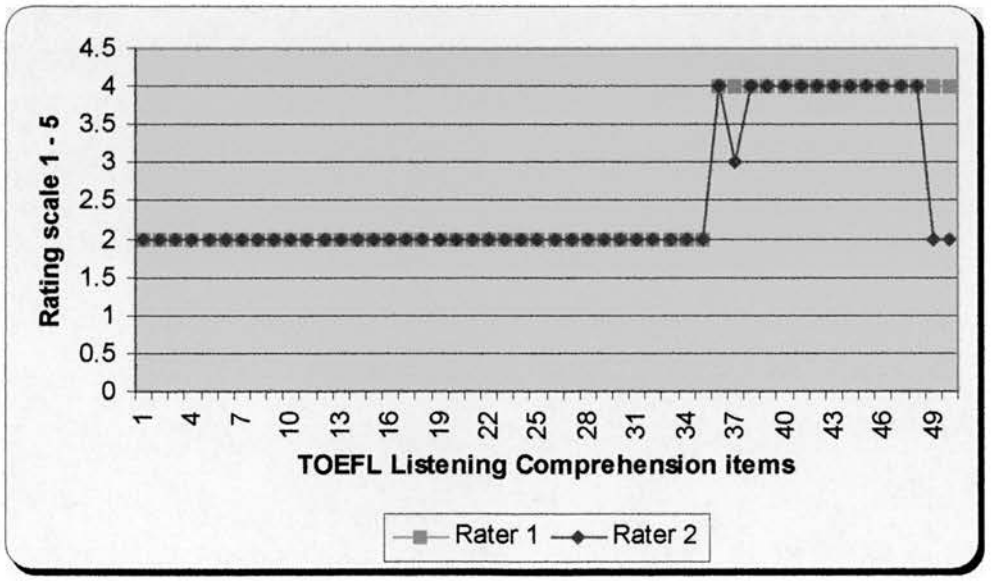
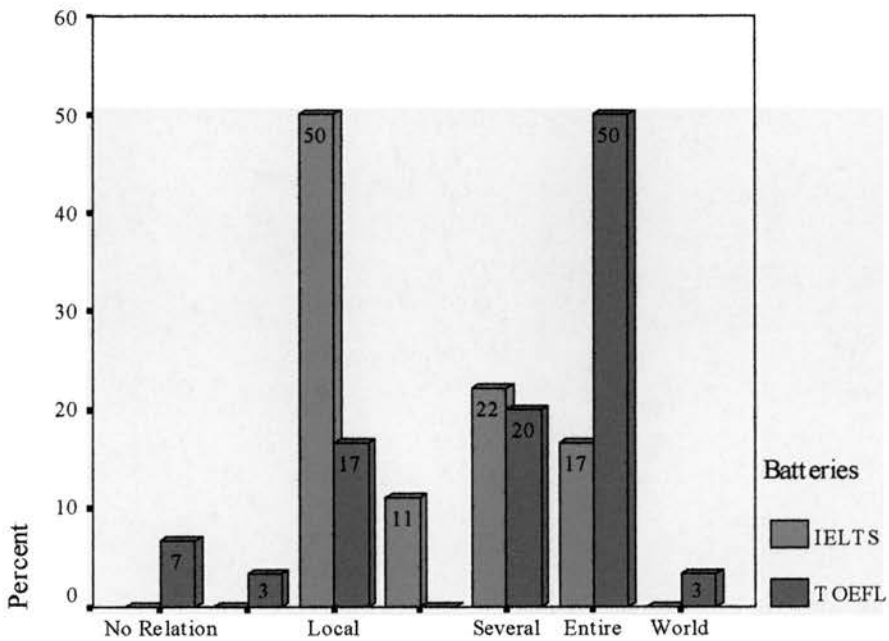


Figure 4.10: Relationship of Item To Passage Ratings: TOEFL Listening Comprehension Items

The correlation between the ratings of the two judges for the reading and listening comprehension items combined was 0.81 for the facet of relationship of item to test passage. The overall inter-rater reliability for this facet was r_{tt} = 0.79, using z-transformation. This is just on the threshold of acceptability. That means, we can rely on these ratings with relative confidence.

In order to compare the ratings for items across the two test batteries, the average ratings of the two raters across all the items in a given test was used.
Figure 4.11 illustrates how reading comprehension items in the two batteries were rated for their relationship to the passages.



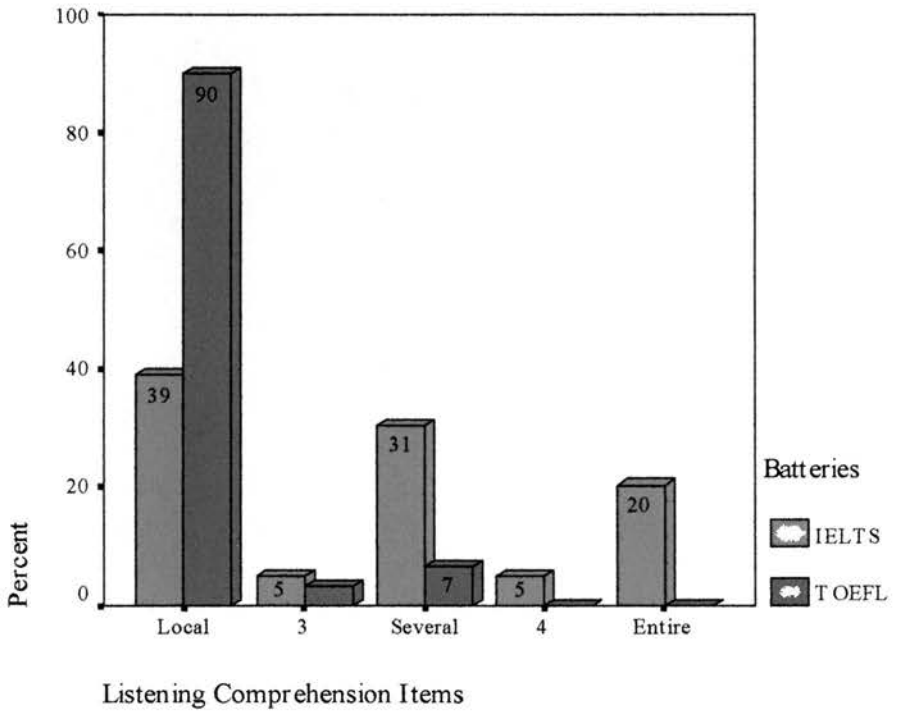
No Relation = can be answered without reference to the passage, **Local** = relates to specific part of the passage, **Several** = relates to several parts of the passage, **Entire** = relates to the entire passage, **World** = relates information in passage to the real world.

Figure 4.11: Relationship of Item To Passage: Reading Mean Facet Ratings

The patterns of relationships show some differences in the ways the items are related to the passages in each battery. In the case of IELTS, fifty percent of reading comprehension items relate to specific part of the passage, between twenty to thirty

percent relates to several parts of the passage, and seventeen percent relate to the entire passage. There are no items in IELTS that relate to the world knowledge, nor are there any items that can be answered without reference to the passage. However, this is slightly different in TOEFL. Seven percent of the items can be answered without reference to the passage and three percent relate the information in the passage to the real world. While fifty percent of TOEFL reading items relate to the entire passage, seventeen percent relate to a specific part of the passage and twenty percent relate to several parts of the passage.

The patterns of relationships are different in listening comprehension items. Figure 4.12 demonstrates how listening comprehension items in the two batteries were rated for their relationship to the passages.



No Relation = can be answered without reference to the passage, **Local** = relates to specific part of the passage, **Several** = relates to several parts of the passage, **Entire** = relates to the entire passage, **World** = relates information in passage to the real world.

Figure 4.12: Relationship Of Item To Passage: Listening Mean Facet Ratings

There are no listening items in either TOEFL or IELTS that can be answered without reference to the passage. Nor are there items that relate the information in the passage to the world. Whereas ninety percent of TOEFL listening items relate to a specific part of the passage and ten percent relate to several parts of the passage, forty percent of the IELTS listening items relate to a specific part of the passage, forty percent relate to several parts of the passage and twenty percent relate to the entire passage. It appears that, by and large, TOEFL listening items can be answered by relating to a specific part of the passage. While IELTS listening items, in addition to local information, are equally related to several parts of the passage and in twenty percent of the items to the entire passage.

In order to make a meaningful comparison between the two batteries one has to examine the mean ratings of the judges on this facet. Table 4.19 sets out the mean facet ratings for the relationship of item to passage for both the listening and reading comprehension items across the batteries.

Table 4.19: Comparison of Mean Facet Ratings For The Relationship of Item To Passage Across The Two Batteries

Batteries		READING			LISTENING		
		# Items	Mean	Std Dev.	#Items	Mean	Std Dev.
	IELTS	36	2.64	0.76	39	2.64	0.74
	TOEFL	30	3.17	1.12	50	2.60	0.93
		Difference in Mean	0.53			0.04	

It was difficult to decide how to interpret the meaningfulness of a difference in mean ratings as the facet was rated on a five-point scale and the amount of variation in ratings differed across the tests. In order to set a criterion for interpreting a difference as meaningful, it was decided to follow Bachman et al.'s (1995, p. 105) approach and interpret any difference between the mean ratings for the TOEFL and the IELTS items on a given facet as *meaningful* if that difference was greater than the standard deviations of the ratings for that facet on either of the two tests. Based on this criterion[17], there is no meaningful difference between the mean ratings for this facet on reading and listening items in the two batteries. This lends support to the null-hypothesis 4.

[17] As can be seen from the last raw of the above table, differences in mean ratings are not greater than the standard deviations of the ratings for the facet.

H_0 4: There is no significant difference in the relationships of the test items to the passages in the two batteries.

4.1.7 Summary Discussion of Test Method Facets Comparison

The purpose of analysing the facets of test methods was to find answers to questions 1, 2, 3, 3.1, 3.2, 3.3, 3.4, and 4.

Q 1: Are the reading passages in the two batteries significantly different in terms of their length?

Q 2: Are the reading passages in the two batteries significantly different in terms of their propositional content?

Q 3: Are the reading passages in the two batteries significantly different in terms of their organisational characteristics?

Q 3.1: Are the reading passages in the two batteries significantly different in terms of their syntactic complexity?

Q 3.2: Are the reading passages in the two batteries significantly different in terms of their lexical density?

Q 3.3: Are the reading passages in the two batteries significantly different in terms of their text difficulty?

Q 3.4: Are the reading passages in the two batteries significantly different in terms of their cohesive markers?

Q 4: Are the relationships of the test items to the reading and listening passages significantly different in the two batteries?

It was not possible to come to a definitive answer for question 2 as the judges disagreed on what constituted the propositional content of the passages. The disagreement was twofold. The sample to be rated (eight reading passages) was just too small to produce acceptable reliability indices. The second aspect was related to the judges' interpretation of the rating instructions. The follow-up interviews showed that three factors affected the judges' decision. Firstly, some of the instructions for rating propositional content facets, despite their apparent self-explanatory descriptions, proved to be highly controversial for the judges. For example, reference to the *compactness* and *diffuseness* of the new information confused the judges as to what to look for in the passages. Secondly, the judges had difficulty deciding who the

audience of the tests were, hence could not decide how suitable the texts were in terms of their ease/difficulty, abstractness/concreteness, and compactness/diffuseness for a particular audience. Finally, the judges did not meet to discuss the rating instrument; therefore they used their own interpretation where they judged the instructions to be ambiguous.

In attempting to re-rate the propositional content facets, the researcher and one of the judges also failed to agree on the interpretation of some of the descriptions of the facets. They concluded that this rating instrument was just unworkable for the judges here. Based on the smallness of the sample and unworkability of some of facets of the propositional content instrument, it was decided not to use the results of the propositional content analysis as they failed to produce satisfactory reliability figures. Although question two remained unanswered, the analysis of the propositional content shed light on the conduct of the subjective ratings for the rest of the facets to be rated.

The descriptions and methodology for rating and measuring the remaining facets of test methods proved to be useful. As for question one, it was found that the IELTS texts were significantly longer than the TOEFL texts. Hence, the reading passages in the two batteries were not comparable in terms of their length. Nevertheless, it was argued that although length could have had an impact on the overall performance of test takers on the tests, it did not have significant impact on the sentence complexity, text difficulty, voice, lexical density, or any other facet of test methods that were examined here.

The results of the analyses of syntactic complexity, lexical density, text difficulty, and cohesive markers indicated that the reading passages in the two batteries did not differ significantly in terms of these organisational characteristics. In other words, with the exception of one cohesion facet (Additive), TOEFL and IELTS passages were comparable in terms of the facets of organisational characteristics: Grammar (Syntactic Complexity, Voice, Lexical Density, Text Difficulty), and Cohesion (Reference, Substitution and Ellipses, Adversatives, Causals, and Temporals).

Finally, a detailed analysis of the relationship of item to passage facet supported the hypothesis that the relationships of the test items to the passages in the two batteries did not differ significantly. The two batteries appear to be comparable in this respect.

4.2 Analysis of Communicative Language Abilities

This section reports the results obtained from the analysis of communicative language ability facets explained in section 3.7.1.2. Since a single rating instrument was used for rating all the facets of communicative language ability, we will first briefly explain the facets and their corresponding questions, and then report the results of the analysis.

Communicative language ability is divided into language competence and strategic competence. Language competence comprises a wide range of linguistic and paralinguistic features which are utilised in effective communication through language. The two major components of language competence are organisational competence and pragmatic competence. Strategic competence is associated with test wiseness in this research. Thus, there are three major questions addressed in this section:

Q 5: Do the test items in the two batteries involve the same degree of organisational competence for successful completion of a given task?

Q 6: Do the test items in the two batteries involve the same degree of pragmatic competence for successful completion of a given task?

Q 7: Do the test items in the two batteries involve the same degree of strategic competence for successful completion of a given task?

There are eight other research questions related to the components of language competence, which will be addressed in this section. The first six are associated with the abilities that control the formal organisation of language, organisational competence, and involve two other components of grammatical and textual competence, operative at two different levels of sentence and text. They are related to the facets of: *Lexicon, Morphology, Syntax, Phonology/Graphology, Rhetorical Organisation,* and *Cohesion.*

Q 5.1: Do the test items in the two batteries require the same degree of lexical knowledge for successful completion of a given task?

Q 5.2: Do the test items in the two batteries require the same degree of morphological knowledge for successful completion of a given task?

Q 5.3: Do the test items in the two batteries require the same degree of syntactic knowledge for successful completion of a given task?

Q 5.4: Do the test items in the two batteries require the same degree of phonological / graphological knowledge for successful completion of a given task?

Q 5.5: Do the test items in the two batteries require the same degree of knowledge of cohesion for successful completion of a given task?

Q 5.6: Do the test items in the two batteries require the same degree of knowledge of rhetorical organisation features for successful completion of a given task?

The last two minor research questions are related to pragmatic competence; the ability to produce and understand sentences or utterances, which are appropriate to the context in which they occur. Pragmatic competence comprises two abilities: the ability to understand the pragmatic conventions for performing acceptable language functions or illocutionary competence, and the ability to understand the sociolinguistic conventions for performing language functions appropriate in a given context - sociolinguistic competence. Thus we examine questions,

Q 6.1: Do the test items in the two batteries involve the same degree of illocutionary competence for successful completion of a given task?

And,

Q 6.2: Do the test items in the two batteries involve the same degree of sociolinguistic competence for successful completion of a given task?

Finally, the tests should be compared on the amount of the involvement *of Strategic competence* in a successful completion of a given task. Strategic competence is associated here with test wiseness. The question to be investigated is,

Q 7: Do the test items in the two batteries involve the same degree of strategic competence for successful completion of a given task?

Ratings of communicative language ability facets were made on the basis of:

a) the extent to which the judges felt the ability was required for the successful completion of the task, and
b) the general level of that ability required.

With the exception of strategic competence, all the other facets were rated on five-point scales (zero to 4)[18]. Strategic competence was rated on three-point scales (zero to 2)[19].

The two judges who rated the *Relation of item to passage* with satisfactory reliability figures in 4.1.6 were asked to do the ratings of the communicative language ability components. They were expected to go over all the 155 reading and listening comprehension items and judge what abilities they were testing on the basis of the instruction given to them[20]. The two judges had discussed the rating instrument in detail and had been co-operating in the research for a period of over three years. Hence, they were quite familiar with the instrument.

4.2.1 Results of the Ratings: Communicative Language Ability

The ratings of the two judges on communicative language ability facets correlated in a range of 0.70 to 1.0, with an average of 0.90. This is a relatively high correlation. The lowest correlation was for the facet of phonology/graphology and the highest correlation was for the *Imaginative* and *Register*. The perfect correlation (1.0) for two of the facets was due to the fact that the judges believed that they were not involved for the successful completion of any of the items. To find out how reliable the ratings were, inter-rater reliability was computed using Fisher Z-transformation. This was to correct the distortion inherent in using Pearson correlation for ordinal data. Table 4.20 and Figure 4.13 show that the inter-rater reliability varied between 0.73 to 1.0, with an average reliability of 0.86 across the facets. This figure is quite satisfactory, indicating that the ratings of the communicative language ability components were of acceptable reliability.

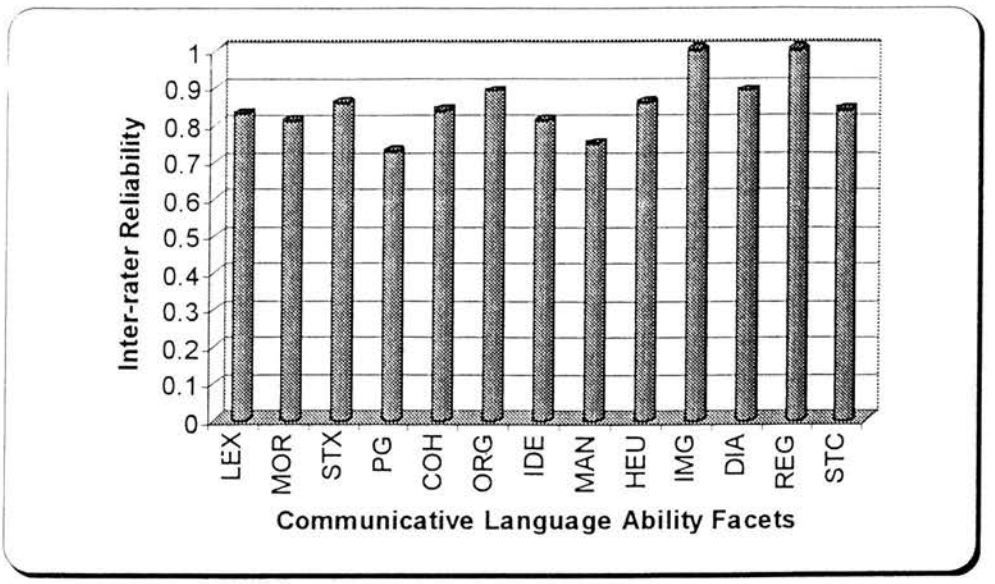
[18] The five-point scale for rating communicative language ability components:

Not Required	Somewhat Involved	Critical Basic	Critical Intermediate	Critical Advanced
0	1	2	3	4

[19] The three-point scale for rating strategic competence:

	Very much		Not at all
Degree to which engaged	2	1	0

[20] See Appendix 8 for the details of the instructions.



LEX= Lexicon MOR= Morphology STX= Syntax PG= Phonology/Graphology
COH= Cohesion ORG= Rhetorical Organisation IDE= Ideation MAN=Manipulative
HEU= Heuristic IMG= Imaginative DIA= Dialect REG= Register
STC= Strategic Competence

Figure 4.13: Inter-Rater Reliabilities: Facets Of Communicative Language Ability

Table 4.20: Correlation and Reliability Estimates For The Ratings Of Communicative Language Ability Facets

	LEXICON	MORPHOLOGY	SYNTAX	Phonology/Graph	COHESION	ORGANISATION	IDEATION	MANIPULATIVE	HEURISTIC	IMAGINATIVE	DIALOGUE	REGISTER	STRATEGIC COMPETENCE
r_{AB}	0.90	0.86	0.95	0.70	0.91	0.98	0.85	0.74	0.95	1	0.98	1	0.91
r_{tt}	0.83	0.81	0.86	0.73	0.84	0.89	0.81	0.75	0.86	1	0.89	1	0.84

r_{AB} = correlations r_{tt} = inter-rater reliability using Fisher Z-transformation

4.2.1.1 Reading Comprehension Results (CLA)

Like the relationship of item to passage facet, the averaged ratings of the two raters across all the items in a given test were used. Table 4.21 sets out communicative language ability mean facet ratings for the reading comprehension items across the batteries.

Table 4.21: Comparison of Mean Facet Rating For CLA: Reading Items (TOEFL and IELTS)

	TOEFL			IELTS		
	# Items	Mean	Std Dev	# Items	Mean	Std Dev
LEXICON	30	3.03	1.16	36	1.99	.94
MORPHOLOGY	30	.00	.00	36	.00	.00
SYNTAX	30	.00	.00	36	.18	.43
PHONOLOGY / GRAPHOLOGY	30	2.00	.00	36	2.00	.00
COHESION	30	.13	.73	36	.75	.97
ORGANISATION	30	2.28	1.72	36	.67	1.04
IDEATION	30	.00	.00	36	.28	.61
MANIPULATIVE	30	.00	.00	36	.04	.25
HEURISTIC	30	.00	.00	36	.00	.00
IMAGINATIVE	30	.00	.00	36	.01	.08
DIALECT	30	.00	.00	36	.00	.00
REGISTER	30	.00	.00	36	.01	.08
STRATEGIC	30	1.90	.28	36	1.76	.59

CLA= Communicative language ability (0-4) Scale

It can be inferred from Table 4.21 that there are relatively more similarities than perceived differences in the average ratings of the communicative language ability facets on reading comprehension items. It appears that the raters did not consider several abilities such as knowledge of morphology and syntax, illocutionary competence (ideation, manipulative, heuristic, and imaginative functions), as well as sensitivity to differences in dialect and register to be involved for the successful completion of the reading items in either of the two reading tests. Although the mean rating of knowledge of lexicon (3.03 for TOEFL and 1.99 for IELTS) suggests that overall this ability was perceived to be involved at basic to intermediate level in both

reading tests. Figure 4.14 illustrates that this faculty was considered to be required at a more critical advanced level in TOEFL. In other words, TOEFL reading items did require this competence for successful completion of a task to a great extent. That could mean that a number of TOEFL reading items were testing the knowledge of lexicon. This is not surprising as half of TOEFL reading items were actually testing vocabulary[21].

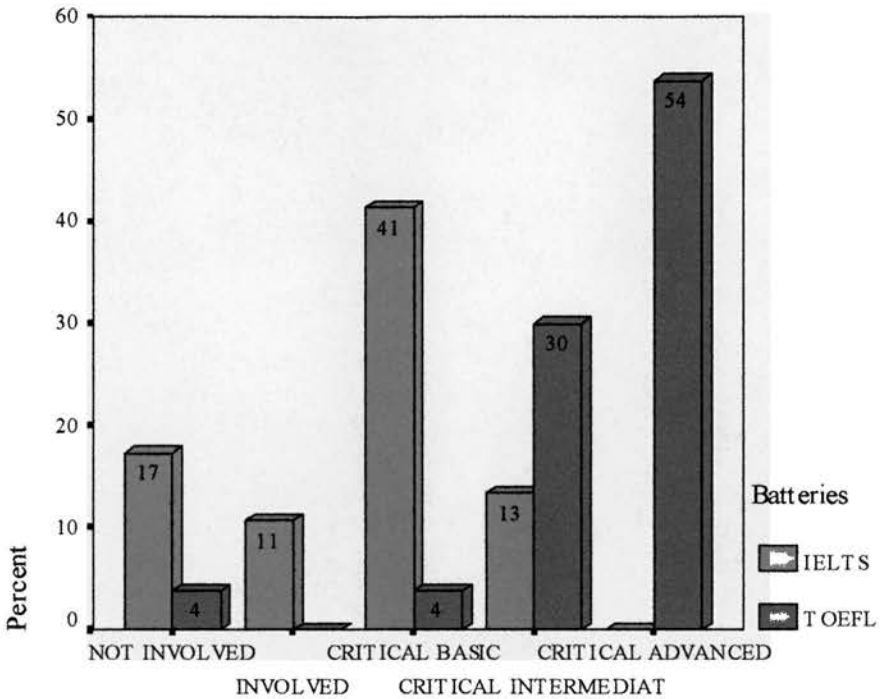


Figure 4.14: Mean Facet Rating On Lexicon: Reading Comprehension Items

The raters perceived knowledge of cohesion to be *somehow involved*[22] in both tests, though it was of less significance in TOEFL (0.13) than in IELTS (0.75). They also perceived knowledge of rhetorical organisation features to be *somehow involved* in IELTS (0.67), whereas it was required in TOEFL reading items at a basic level (2.28). Table 4.21 also indicates that knowledge of phonology / graphology (2) was required at a critical level for both reading tests; therefore, there was no difference between the two reading tests in that respect. Finally, Figure 4.15 demonstrates that the raters perceived strategic competence to be very much involved in a successful completion

[21] See Appendix 3

[22] We use the term *somehow involved* as the values obtained are between 0-1. Zero means no involvement, whereas one means somewhat involved. See 4.2.1, Footnote 18 for the descriptions of the scale.

of any reading item in the two tests (1.90 for TOEFL and 1.76 for IELTS). They judged over ninety percent of the items to be sensitive to test wiseness. This brings up the importance of test preparation for success on these tests; an issue to be dealt with in chapter five, section 5.6.

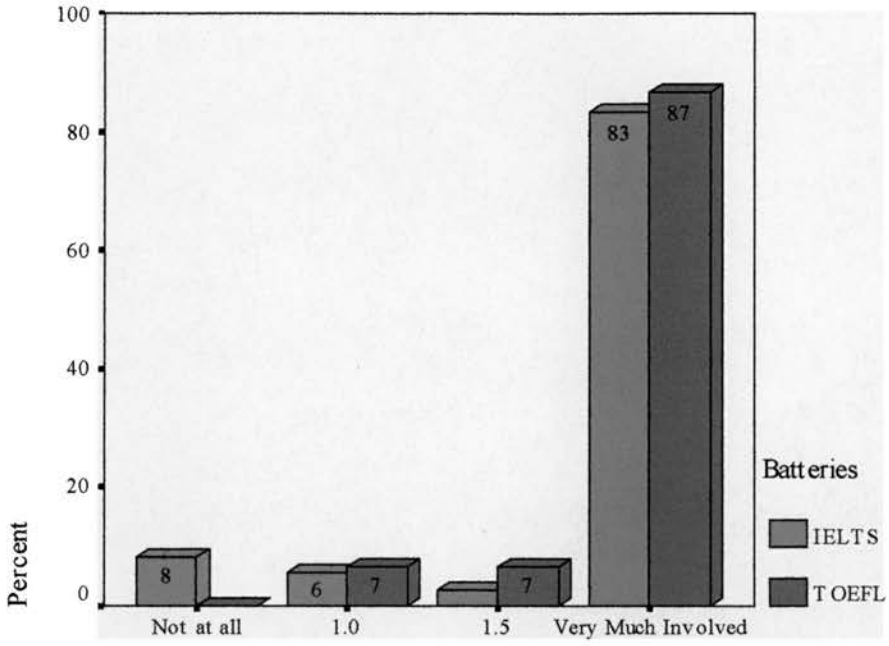


Figure 4.15: Mean Facet Rating: Strategic Competence

4.2.1.2 Listening Comprehension Results (CLA)

Table 4.22 sets out the communicative language ability mean facet ratings for the listening comprehension items across the batteries. Again, one can infer that there are relatively more similarities than perceived differences in the average ratings of the communicative language ability components. The values for morphology, syntax, cohesion, rhetorical organisation and competence in heuristic and imaginative functions, and sensitivity to dialect and register indicate that knowledge of these facets was not required for the successful completion of any of the IELTS listening comprehension items and the majority of TOEFL listening items. However, the judges perceived that knowledge of syntax (0.68), cohesion (0.76), and heuristic function (0.62) were sometimes *'somehow involved'* in TOEFL as the values approximate to 1, which means some involvement of the facet. The judges also

perceived that knowledge of ideational (1.34 for TOEFL and 1.56 for IELTS) and manipulative (0.9 for TOEFL and 0.76 for IELTS) functions were '*somehow involved*' in the successful completion of listening items in the two tests.

Table 4.22: Comparison of Mean Facet Rating For CLA: Listening Items (TOEFL and IELTS)

	TOEFL			IELTS		
	# Items	Mean	Std Dev	# Items	Mean	Std Dev
LEXICON	50	3.65	.55	39	1.46	.99
MORPHOLOGY	50	.27	.95	39	.00	.00
SYNTAX	50	.68	1.42	39	.00	.00
PHONOLOGY / GRAPHOLOGY	50	2.39	1.29	39	1.00	.00
COHESION	50	.76	1.50	39	.00	.00
ORGANISATION	50	.00	.00	39	.00	.00
IDEATION	50	1.34	1.56	39	1.56	.70
MANIPULATIVE	50	.90	1.46	39	.76	.69
HEURISTIC	50	.62	1.30	39	.00	.00
IMAGINATIVE	50	.00	.00	39	.00	.00
DIALECT	50	.00	.00	39	.38	.96
REGISTER	50	.00	.00	39	.00	.00
STRATEGIC COMPETENCE	50	2.00	.00	39	2.00	.00

CLA= Communicative language ability

Figure 4.16 displays the comparison of the judges' ratings for the involvement of the lexical knowledge for a successful completion of a listening task across the batteries. The raters perceived that the level of lexical knowledge required for the successful completion of a listening item was at a much higher level for TOEFL than for IELTS. In the case of TOEFL, the knowledge of lexicon required was close to critical advanced level (3.65), whereas in IELTS it was judged to be somewhere between *somewhat involved* and critical basic (1.46)[23].

[23] The five-point scale for rating communicative language ability components:

Not Required	Somewhat Involved	Critical Basic	Critical Intermediate	Critical Advanced
0	1	2	3	4

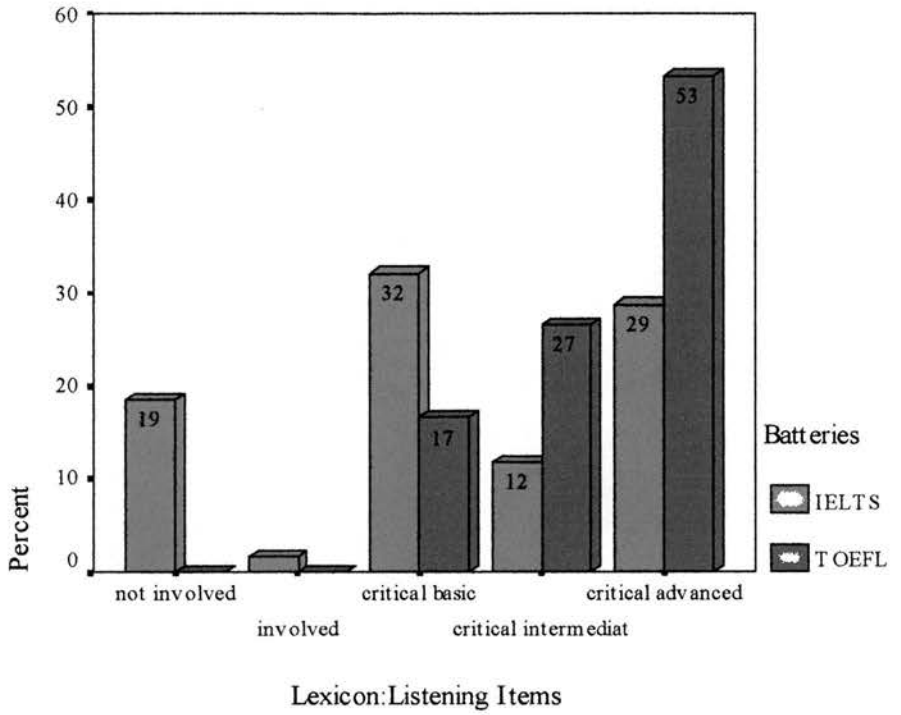


Figure 4.16: Mean Facet Rating On Lexicon: Listening Comprehension Items

It appears that the batteries also differed in the degree of the involvement of knowledge of phonology / graphology in the successful completion of a task. Figure 4.17 illustrates the difference in patterns for the phonology/graphology competence across the listening items of the two batteries. The knowledge of phonology / graphology was only *somewhat involved* in IELTS (1) listening items, while it was slightly above the critical basic level for TOEFL listening items (2.39).

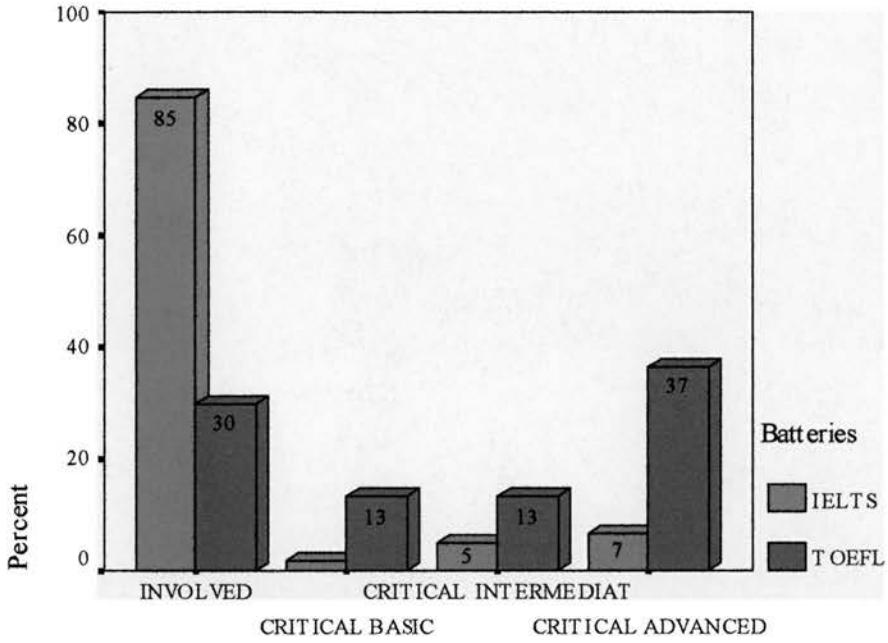


Figure 4.17: Mean Facet Rating On Phonology/Graphology: Listening Comprehension Items

Finally, the judges believed that strategic competence (2) was very much involved in all the listening items across the batteries. It is important to reiterate here that although strategic competence covers a wide range of abilities, it was only associated with test wiseness in this research.

4.2.2 Discussion of Ratings On Communicative Language Ability

The amount of variation in ratings on communicative language ability facets differed across tests and facets so that a mean difference of, say, 0.5 could not be interpreted as equally meaningful for all facets. It was decided to adopt the same criterion as was used in analysing the relationship of item to passage facet for interpreting the meaningful differences. That is, to follow Bachman et al.'s (1995, p. 105) approach and interpret any difference between the mean ratings for the TOEFL and the IELTS items on a given facet as *meaningful* if that difference was greater than the standard deviations of the ratings for that facet on either of the two tests.

Following the above criterion, all the 26 possible comparisons between the facets of communicative language ability of the reading and listening comprehension items across the two batteries were analysed for meaningful differences. Table 4.23 presents the meaningful comparisons across the facets and batteries.

Table 4.23: Meaningful Differences In CLA Ratings

Subtests	Facet	Difference magnitude	Difference direction
Listening items	Lexicon	2.19	TOEFL>IELTS
Listening items	Phon /Graph	1.39	TOEFL>IELTS

CLA= Communicative language ability

Out of 26 possible pair comparisons of communicative language ability facets, only two proved to have meaningful differences in the two batteries: lexicon and phonology/graphology in listening items. The TOEFL listening items were judged to involve significantly more knowledge of lexicon for a successful completion of a given task than the IELTS listening items. In other words, the knowledge of lexicon was judged to be a determining factor in successfully completing the TOEFL listening items; this was not the case for the IELTS listening items. However, the involvement of lexical knowledge was not judged to be a determining factor in completing the reading comprehension items. This has a dual implication for question 5.1 of the research.

In question 5.1 we wanted to know whether the test items in the two batteries required the same degree of lexical knowledge for successful completion of a given task. The results of the communicative language ability analysis suggest that, on the one hand, the two batteries are not significantly different with regards to the involvement of this facet for completing reading comprehension items. On the other hand, the meaningful differences reported in Table 4.23, suggest that the batteries are significantly different with regards to the involvement of this facet for completing the listening comprehension items.

To observe the effect of lexical knowledge for successful completion of a given task in both reading and listening items, it was decided to combine the reading and listening comprehension items and then analyse the mean facet ratings for the lexical

knowledge across the batteries. Table 4.24 sets out the mean facet rating for lexical knowledge for the combined items.

Table 4.24: Comparison of Mean Facet Rating For Lexicon: Reading & Listening Items Combined

	TOEFL			IELTS		
	# Items	Mean	Std Dev	# Items	Mean	Std Dev
LEXICON	80	3.42	0.88	75	1.71	0.99
	Difference magnitude			Difference direction		
Meaningful difference	1.71*			TOEFL > IELTS		

* Meaningful

The difference magnitude between the two means in the batteries is meaningful according to the criterion set. TOEFL items were judged to require significantly more lexical knowledge for a successful completion of a given task than the IELTS items. This lends support to alternative-hypothesis 5.1.

H_1 5.1: There is a significant difference in the degree of lexical knowledge required for successful completion of test items in the two batteries.

To put it in different words, the two batteries may not be similar in terms of the involvement of the lexicon facet in the completion of a given task.

Phonological / graphological knowledge was also rated to be significantly more influential in successfully completing a listening item in TOEFL than in IELTS. Although all the listening items in the two batteries required some degree of the involvement of this competence, the two tests differed significantly in the degree of the involvement of Phonological / graphological knowledge in completing their tasks. Since the involvement of phonological/graphological knowledge was not judged to be a determining factor in completing the reading comprehension items, it was decided to combine the reading and listening items and then analyse the mean facet ratings for the phonological / graphological knowledge across the batteries. Table 4.25 sets out the mean facet rating for phonological/graphological knowledge for the combined listening and reading comprehension items.

Table 4.25: Comparison of Mean Facet Rating For Phonology/Graphology: Reading & Listening Items Combined

	TOEFL			IELTS		
	# Items	Mean	Std Dev	# Items	Mean	Std Dev
Phonology/graphology	80	2.24	1.04	75	1.48	0.50
	Difference magnitude			Difference direction		
No Meaningful difference	0.76			TOEFL > IELTS		

The difference magnitude between the two means in the batteries is not meaningful according to the criterion set. The implication of this finding for research question 5.4 is that the two batteries do not appear to be significantly different in terms of the involvement of phonological/graphological knowledge in a successful completion of a given task. This lends support to null-hypothesis ***H_0 5.4.***

H_0 5.4: *There is no significant difference in the degree of phonological / graphological knowledge required for successful completion of test items in the two batteries.*

The results presented in Table 4.23 demonstrate that apart from lexicon and phonology/graphology in listening items, no other difference between the means of the facets in the two batteries appears to be meaningful. This allows us to reject the rest of the alternative-hypotheses raised in this section and lends support to the null-hypotheses about the communicative language ability components. Hence, the following conclusions may be drawn for the rest of the facets of communicative language ability.

- For question 5.2 about the difference in the degree of morphological knowledge across the batteries, the null-hypothesis ***H_0 5.2*** holds: *There is no significant difference in the degree of morphological knowledge required for successful completion of test items in the two batteries.*
- For question 5.3 about the difference in the degree of syntactic knowledge across the batteries, the null-hypothesis ***H_0 5.3*** holds: *There is no significant difference in the degree of syntactic knowledge required for successful completion of test items in the two batteries.*

- For question 5.5 about the difference in the degree of knowledge of cohesive relations across the batteries, the null-hypothesis ***H_0 5.5*** holds: *There is no significant difference in the degree of knowledge of cohesive relations required for successful completion of test items in the two batteries.*
- For question 5.6 about the difference in the degree of knowledge of rhetorical organisation features across the batteries, the null-hypothesis ***H_0 5.6*** holds: *There is no significant difference in the degree of knowledge of rhetorical organisation required for successful completion of test items in the two batteries.*
- For question 6.1 about the difference in the degree of illocutionary competence across the batteries - all the forms of the illocutionary competence (ideational, manipulative, imaginative, and heuristic) appear to be involved[24] in the same degree for completing a given task in the batteries, the null-hypothesis ***H_0 6.1*** holds: *There is no significant difference in the degree of illocutionary competence involved for successful completion of test items in the two batteries.*
- For question 6.2 about the difference in the degree of sociolinguistic competence across the batteries - both forms of the sociolinguistic competence (sensitivity to register and dialect) appear to be involved[25] in the same degree for completing a given task in the batteries, the null-hypothesis ***H_0 6.2*** holds: *There is no significant difference in the degree of sociolinguistic competence involved for successful completion of test items in the two batteries.*
- Finally, for question 7 about the difference in the degree of strategic competence across the batteries, the null-hypothesis ***H_0 7*** holds: *There is no significant difference in the degree of strategic competence involved for successful completion of test items in the two batteries.*

There remain two other questions for investigation: Question 5, organisational competence and Question 6, pragmatic competence. Organisational competence comprises the following facets: *lexicon, morphology, syntax, phonology/graphology, cohesion, and rhetorical organisation*. In investigating research questions 5.1-5.6, we have shown that almost all the facets of organisational competence, apart from lexicon, have been judged to be required in the same degree for the completion of a given task across the two batteries. Hence, there are more similarities than differences in the average ratings of organisational competence. That lends support to the null-hypothesis ***H_0 5:*** *There is no significant difference in the degree of*

[24] To be exact, illocutionary competence in most cases appears not to be involved.
[25] Again it appears that this facet is not involved for the successful completion of a task.

organisational competence required for successful completion of the test items in the two batteries.

Finally, it was found that both of the components of pragmatic competence, that is, sociolinguistic competence and illocutionary competence, were involved in the same degree for the completion of a task in the two batteries. Therefore, it can be inferred that the two batteries did not differ significantly in the involvement of pragmatic competence.

This supports the null- hypothesis H_0 *6: There is no significant difference in the degree of pragmatic competence involved for successful completion of test items in the two batteries.*

Communicative language ability facets were rated with respect to 155 TOEFL and IELTS listening and reading comprehension items with satisfactory reliability figures. Thirteen different communicative language ability facets were rated for their degree of involvement in completing a given task. The findings show that despite some obvious differences in the two tests such as the length of the passages and the heavy reliance on lexical knowledge in TOEFL, there were more similarities than perceived differences between the two batteries on the facets of communicative language ability. Out of twenty six possible pair comparisons of communicative language ability facets of the listening and reading comprehension items, only two proved to have meaningful differences: lexical and phonological/graphological knowledge in listening items. In terms of the combined scores, only lexical knowledge proved to have meaningful difference across the batteries. This may suggest that TOEFL and IELTS listening and reading comprehension items are comparable in terms of their communicative language ability facets.

4.3 General Discussion Of Content Analysis

To assist the content analysis of TOEFL and IELTS samples, two rating instruments were developed for the measurement of the facets of test methods and communicative language ability. The assessment of the communicative language ability facets was entirely based on a subjective rating instrument, which produced acceptable reliability figures. The rating instrument used in the analysis of test methods had two components: objective and subjective. The former was based on pure linguistic

analysis of texts such as counting the number of words and phrases in a particular syntactic category, or counting the percentage of passive constructions in a passage. The latter employed subjective ratings of expert judges on a number of facets. The objective scoring was used in analysing length and organisational characteristics of the batteries. The subjective scoring was used in assessing the propositional content of the batteries and the relationship of item to passage. The propositional content ratings did not produce satisfactory reliability figures, whereas the relationship of item to passage ratings did produce acceptable reliability. It was concluded that in order to improve the reliability of the subjective ratings, one needs to take two actions: train the judges and increase the number of items for rating.

The results of various analyses of test method facets indicated that the two batteries differed only in two respects: length and lexicon facets. IELTS reading passages were significantly longer than their TOEFL counterpart; the average IELTS reading passage was over two and half times as long as the average TOEFL passage. This finding is similar to what Bachman et al. (1995) report for the comparison of FCE with TOEFL: *" the FCE reading passages were considerably longer on average than the TOEFL passages"* (p. 121). This is, perhaps, a reflection of the communicative theory of language proficiency on which IELTS and FCE tests are based, where the emphasis is on task-based items, which eventually require longer texts to accommodate various test tasks. However, we have already argued that the length of a passage, despite its potential impact on other facets, did not seem to have affected any other facets of test methods in this research.

The second meaningful difference was related to TOEFL reading and listening items, which were judged to require a much higher level of lexical knowledge for the successful completion of a given task. The knowledge of vocabulary for TOEFL items on average was judged to be critical at an intermediate to advanced level (3.42). Whereas in the case of IELTS this knowledge was critical at a basic level (1.71). See Table 4.24. This finding is also similar to what Bachman et al. (1995) report about the significance of vocabulary knowledge in successfully completing a given TOEFL reading and listening task. Heavy emphasis on lexical knowledge is one of the features, which differentiates TOEFL items from the IELTS.

It can be concluded from the above that the two batteries may not be comparable in terms of their length of passages and the amount of lexical knowledge that is required for completing a given listening and reading task. However, the results of the analysis

of the rest of the facets indicate that the two batteries are not significantly different in terms of the following facets: *syntactic complexity, lexical density, text difficulty, cohesive markers, relationship of item to passage, competence in morphology, syntax, phonology/graphology, cohesion, rhetorical organisation, illocutionary competence (ideation, manipulative, heuristic, imaginative), sensitivity to dialect* and *register*, and *strategic competence*. This tends to support the suggestion that the two batteries have significantly more similarities than differences.

The communicative framework used in the content analysis appears to have strong as well as weak points. Most of the facets examined provided some information about the similarities / differences of the batteries. However, not all of them seem to be useful in practice. There were problems with the rating instrument used in the comparison of the propositional content discussed in 4.1.2. Firstly, there was a validity problem that the judges did not find the instructions for this instrument as transparent as Bachman et al. (1995, p. 122) would have hoped them to be. The judges found some of the descriptions of the facets unclear and confusing. For example, the distinction of the distribution of new information through compactness or diffuseness of the load of information was very confusing for the judges. The judges also had problems with the descriptions of contextualisation with respect to cultural content and topic specificity. Each judge interpreted these facets differently. The judges' follow-up interviews indicated that they had serious reservations as to the use of this instrument for rating the facets. If the usefulness of an instrument is in question, one cannot rely on the results achieved using the instrument.

Secondly, there were problems with the reliability of the exercise using the instrument. Two caveats were in order here. The 3-point scale (0-1-2) was too narrow and the number of items (8 passages) was too small a sample to produce an acceptable reliability figure. A slight discrepancy in the judgement of the raters would have skewed the results. The above problems eventually led to abandoning the propositional content instrument in the course of the research.

While the propositional content instrument did not seem to be very useful for the comparison of the two batteries, the communicative language ability and the relationship of item to passage instruments, which were also based on subjective ratings, did prove to be useful. There is obviously a rater training factor involved here. The judges who rated the communicative language ability components and the relationship of item to passage had worked together for a long time which may have

helped them to interpret the instructions in much the same way. Nevertheless, the same judges could not agree on the ratings of the propositional content. That indicates that training of the judges is only one important factor in a subjective rating exercise. What is even more important is the validity of the instrument for the purposes for which one intends to use it. The two instruments used for the ratings of the relationship of item to passage and the components of communicative language ability had transparent instructions, which the judges could agree on and follow. Besides, the two caveats regarding the propositional content instruments were not relevant in the latter exercise. The communicative language ability and relationship of item to passage instruments were both based on five-point scales (0-4, and 1-5) and the number of items (155) to be rated was large enough to produce a satisfactory reliability figure. Even in the case of strategic competence, where a three-point scale (0-2) was used, the number of items to be rated (155) was sufficient for producing a satisfactory reliability coefficient.

This latter point may shed some light on the usefulness of the subjective rating instruments used in our analysis. It is possible that some of the judges' difficulties in using the propositional content instrument could have been caused because they were expected to apply it in rating the features of reading comprehension passages. As we have already mentioned in 4.1.3, previous research has shown that judges usually have difficulty in determining what an item is testing with regard to a reading comprehension passage. In the case of the propositional content instrument, the judges were only asked to rate the contents of the passages, not the items, with respect to the criteria given to them. Separating the passages from the test items might have caused some of the confusion for the raters. The limitation of resources did not allow us to ask the judges to apply the instrument to the reading items, as was the case in rating the communicative language ability instrument. Had we had the resources, the same instrument could have produced a better reliability had it been applied to the reading items.

The results of the communicative language ability ratings suggest that the rating instrument can only be used reliably if the raters are very comprehensibly trained in a very long process, which might involve the modification of the rating instrument and bringing in different interpretations of what each facet may mean to the raters. For example, strategic competence was equated with test wiseness for the raters in this exercise, while the original definition proposed by Bachman (1990) included a much wider scope. The fact that the rating instrument used in the rating of the propositional

content of the tests reported in 4.1.2 failed to produce satisfactory reliability figures could be indicative of the importance of the raters' training.

This raises the question of the trade off between the reliability of an exercise and the validity of the instrument used in the exercise. On the one hand, achieving satisfactory reliability for an exercise is generally time consuming and expensive and in most cases, is only possible through objective measurements, as was the case in assessing most of the test method facets. Most objective measurements tend to focus on one aspect of the trait under investigation and are thus questionable for the validity of what they try to assess. On the other hand, valid instruments will usually take into consideration some kind of subjective measurements, which are difficult to replicate in similar situations, hence are prone to throw into question the reliability of the exercise. The rating instrument used in the ratings of the communicative language ability facets seems to have produced some kind of valid measurement of what the items were expected to measure with relative confidence.

The results of the communicative language ability ratings show that the two batteries have similarities along a number of dimensions, despite their apparent differences, in particular the longer length of IELTS reading passages. We are aware that some of the similarities could have been related to the sensitivity of the rating instrument to the components of communicative language ability. For example, the zero rating of illocutionary competence facets and sensitivity to dialect and register facets meant zero variances across the tests and a case of prefect correlation, which has increased the reliability of the ratings. This also means that these facets were not tested in the batteries. In other words, some of the similarities between the two batteries relate to what these tests are not measuring.

A more viable explanation for the similarities between the two tests is the fact that only similar sections of the two batteries were selected for analysis: listening and reading items. We have shown that the two tests have a number of similarities in the ways they measure listening and reading abilities of the test takers. However, one should bear in mind that IELTS includes Speaking, Writing, Reading, and Listening sections, while TOEFL comprises Listening, Reading, and Structure sections. The two productive sections of IELTS were not analysed for their content for three reasons. Firstly, there were no TOEFL counterparts for these sections. Secondly, there were reliability problems about the ratings of these two productive skills, in particular in relation to the Speaking section. And thirdly, the rating instrument used

in this research seemed to be more relevant to the assessment of the receptive skills. Although the result of content analysis shows much more similarity across the tests, there are major differences between the two tests in the variety of language skills they are measuring.

Chapter Five

Analysis of Test Performance

5. Analysis of Test Performance

This chapter reports the results of the analysis of test performance. The first four sections of the chapter (5.1, 5.2, 5.3, and 5.4) report and discuss the findings of factor analysis techniques to see if patterns of performance support the findings of the content analysis section in Chapter 4. The remaining sections of the chapter report and discuss issues related to questions 8 and 9 of the research regarding item difficulty and test preparation impact.

We have argued in Chapter 3 that if the results of content analysis of the tests reveal that there are similarities in the kind of abilities the two batteries measure, it is then necessary to investigate whether patterns of performance support such interpretation. We have already shown that the two tests have more similarities than perceived differences in the facets of test methods and communicative language ability components. It is now necessary to investigate whether patterns of correlations within each of the two test batteries and across the two are comparable. To achieve that, the correlational matrices of test scores on TOEFL, IELTS, and EPTB will be examined using exploratory factor analysis. The addition of EPTB to the analysis is due to the view that we need to follow a multitrait-multimethod design for investigating convergent and discriminant validity. It also helps to explore possible identifiable patterns within the interrelationships among the different sections of the tests.

The main focus of this section is to investigate whether the analysis of test scores supports the findings of the content analysis that the batteries do not differ significantly in terms of the degree of the involvement of various facets in the completion of the test tasks.

5.1 Reliability of the Test Batteries Measured

In the review of reliability in Chapter 2, section 2.5.1, we mentioned that a fundamental concern in the design of language proficiency tests is to identify the potential sources of error variance in a given measure of a trait and to control the effect of such factors on that measure. There are various potential sources of measurement error that arise from the effect of any factors, other than the ability being measured, which affect the test scores. For example, in a language test, we would like to control the effect of non-language factors such as test-wiseness, motivation, and health, on test takers' performance as they are sources of unreliability. Since the potential sources of unreliability are endless, it is important to control as many factors as possible that may contribute to error variance. By controlling such factors, one minimises the measurement error and hence maximises test reliability.

Prior to any further analysis it is essential to demonstrate that the tests used in this research had adequate reliability so that analysis can be done with more confidence. Using the ITEMAN conventional item analysis programme from Assessment Systems Corporation (1993), the coefficient α reliability estimates were calculated for each battery. Table 5.1 reports the reliability estimates for the tests examined in this research.

Table 5.1: Reliability Estimates

Scale	K	N	α	Norm
IELTS LC	39	131	0.85	NA
IELTS RC	35	127	0.81	0.85
TOEFL LC	50	127	0.87	0.90
TOEFL ST	40	127	0.85	0.86
TEOFL RC	60	127	0.83	0.90
EPTB T1	58	134	0.81	NA
EPTB T2	44	134	0.67	NA
EPTB T3	49	134	0.76	NA
EPTB T4	47	134	0.78	NA

K = No. of items N = No. of examinees NA = Not available

The last column of Table 5.1 reports the published reliability figures for each sub-test. Clapham (1996) has reported reliability estimates for the reading comprehension texts of IELTS based on 634 examinees. The IELTS texts used in the Clapham study are exactly the same texts used in this research; hence, the reliability estimate achieved here can be considered with relative confidence. The IELTS writing sub-test was excluded from the reliability analysis as the marking was done subjectively. Although every attempt has been made to follow UCLES instructions for training the judges, the inter-rater reliability could not be carried out because each paper was rated by only one rater, as is the norm with the IELTS markings.

In the case of TOEFL, the norm refers to values obtained from ETS (1987), for examinees tested in the US and Canada between December 1984 and February 1986. With the exception of TOEFL RC, these estimates are as high as the values reported by Bachman et al. (1995) for the reliability of the institutional TOEFL (TOEFL LC = 0.889, TOEFL ST = 0.834, TOEFL RC = 0.874) based on 1467 candidates. Cronbach (1990) considers the range of scores as one of the factors affecting the reliability coefficient; "*the reliability coefficient is higher in a wide-range group*" (Cronbach, 1990, p.206). Since only 127 candidates took the TOEFL sample in our research compared to over 1,300,000 TOEFL candidates reported in Chapter Three, Table 3.4, the reliability estimates based on our small sample are well within the acceptable threshold level for reliability of a test.

There are no published data for EPTB Short Version Form C used in this study. The available published data refer to Form A and Form D reported by Davies (1967 and 1984). The EPTB battery changed significantly from one version to another in terms of listening comprehension items, and it is therefore very difficult to say what the reliability figure would have been for Form C. The reliability estimates reported for Test1, however, were usually very high (0.85 and 0.91) for Form A, while they were not that high for Test 2 (only 0.51 and 0.75 for Form A and 0.79 for Form D). In the case of Grammar items (Test 4), the α reported were 0.89 and 0.82 for Form A and Form D respectively[1]. The figures calculated in this research also follow a similar pattern, with Test 1 having the most reliable estimate and Test 2 the least reliable one[2]. Figure 5.1 illustrates the distribution of reliability figures in all the tests concerned here.

[1] See Davies 1967 and 1984.

[2] The researcher has reservations concerning the reliability of cloze test as it violates the underlying assumption of independence of items on which the traditional item analysis is based.

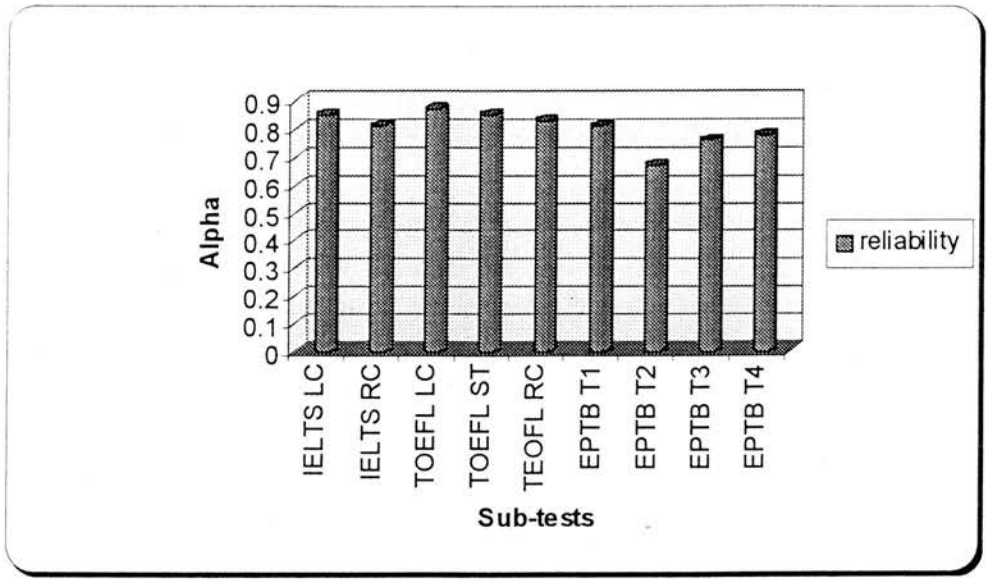


Figure 5.1: Comparison of Reliability Estimates Across The Three Tests

As can be seen from the above, with the exception of EPTB Test 2, all the sub-tests studied in this research achieved reliability similar to other studies (e.g., Davies, 1984, ETS, 1987, Alderson, 1993, Bachman et al., 1995, and Clapham, 1996). This allows us to carry out factor analysis with relative confidence. Other relevant item statistics are reported in Appendix 9.

5.2 Validity of the Abilities Measured

In 4.2.2, we have shown that content analysis of language abilities tested in TOEFL and IELTS supports the hypothesis that the test items in the two batteries had more similarities than differences in terms of the assessment of communicative language ability components and the facets of test methods. It was necessary to determine whether the patterns of performance in the two batteries also supported such interpretation. To achieve this, patterns of correlations within each of the two test batteries and across the two were examined using factor analysis techniques. Following the *Multi-Trait-Multi-Method* approach discussed in Chapter 2, the scores of examinees on EPTB tests were also included in the analysis to allow easier interpretation of the factor loadings and to allow the examination of divergent as well as convergent validities across the batteries.

Exploratory factor analysis was used for examining the correlational matrices for two reasons. Firstly, as has already been argued, there is no consensus among applied linguists about the definition of language proficiency and hence confirmatory factor analysis was not a suitable method for our inquiry. Secondly, the batteries studied are seemingly based on different operational interpretations of language proficiency and therefore their traits' similarities / differences needed to be explored without any presumptions.

A word of caution. Despite all the advantages of factor analysis techniques for investigating the underlying structure of the batteries studied, the researcher is aware of the limitations of factor analysis. It has been argued (Gorsuch, 1983, among others) that factor analysis only investigates the intercorrelations among various variables and reduces the number of variables to some common factors on which the majority of variables load. Factor analysis by itself cannot determine what the underlying constructs may be called. It is the job of the experts in each field to look at the relationships between the factors and decide subjectively what they may be called. Since the judgements are subjective, they might be interpreted differently by different researchers. To overcome this potential problem, the researcher has decided to interpret the factors on the basis of simplicity and interpretability criteria, to be explained shortly.

Factor Analysis Procedures

Four different correlation matrices were used in the final analysis:

- Intercorrelations among the raw scores for the three IELTS sections.
- Intercorrelations among the raw scores for the three TOEFL sub-tests.
- Intercorrelations among the raw scores for the four EPTB tests.
- Intercorrelations among all ten of these measures.

All the correlational matrices are presented in Table 5.10. To examine the appropriateness of the common factor model, the matrix of correlations among the various test scores to be analysed was examined in two ways: testing that the determinant of the matrix was positive and non-zero, and that KMO and Bartlett's[3] test of sphericity were above acceptable limits. All four of the above correlation matrices satisfied these criteria. Extraction of factors was processed using principal axis factoring with squared multiple correlations on the diagonal as initial communality estimates. The initial decision about the appropriate number of factors to be extracted was based on scree plots that were plotted using the eigenvalues obtained from the initial extraction. Additionally, a Montanelli-Humphreys (1976) parallel analysis criterion on the number of salient roots was obtained to check the number of factors. Then, the specified principal axes were extracted for each correlation matrix depending on the number of factors decided upon for each matrix. The principal axes thus extracted were rotated to four factor solutions:

- An orthogonal solution with the normal varimax procedure;
- An orthogonal solution with the weighted varimax procedure;
- An oblique solution with the direct oblimin procedure; and
- An oblique solution with Tucker and Finkbeiner DAPPER[4] (Direct Artificial Personal Probability Function Rotation) two-sided case.

---

[3] Bartlett (1950) has presented a chi-square test of the significance of a correlation matrix with unities as diagonal elements that is widely recommended. The equation is:
$X^2 = - \left(n - 1 - [2_v + 5] / 6 \right) \mathrm{Log}_e | R_{vv} |$
Where $| R_{vv} |$ is the determinant of the correlation matrix. The degrees of freedom are:
$df = v \ (v - 1) / 2$

[4] This rotation procedure was originally presented in ETS Research Report RR-81-58 and has been found to be quite successful in producing satisfactory simple structure solutions. It is performed using as the starting position whatever transformation matrix is currently in memory. For example, the starting position can be a Varimax rotation. Experience indicates that the best solutions are obtained by using the "two-sided" case, that is, allowing salient loadings sometimes to have negative signs. See discussion in Tucker & Finkbeiner's (1981) report.

Following Bachman et al.'s (1995) approach, it was decided to base the final determination of the number of factors and the best solution on the basis of simplicity and interpretability:

> *Simplicity was evaluated by examining the patterns of loadings for the orthogonal and oblique solutions and the scatter plots of loadings on the rotated axes. Interpretability was evaluated with reference to the nature of the tasks and abilities thought to be operationalised in the different measures.* (Bachman et al., 1995, p. 65)

Correlational analyses and part of the factor analysis were performed with SPSS for Windows Version 8.02 (SPSS, 1998) and exploratory factor analysis was performed with Turbo-Basic programmes written for the IBM-PC by John B. Carroll (Carroll, 1999).

5.3 Results of Factor Analysis

This section reports the results of factor analysis and is divided into two parts. The first section elaborates on the results obtained from the analysis of each test battery and their trait structures. The second section reports the results from the factor analysis across all the batteries

5.3.1 Within Test Battery Factor Solutions

We will first report the results of the exploratory factor analysis of IELTS and then report the results of TOEFL and EPTB.

5.3.1.1 Exploratory Factor Analysis of IELTS

The results of exploratory factor analysis are given in Table 5.2 and Table 5.3. Initial parallel analysis and scree plots suggested two factors underlay IELTS test scores. As can be seen from the factor correlations in Table 5.2, oblique solution (DAPPER two-sided case) was the best solution. That is, it maximised the simplicity and interpretability of the factors underlying IELTS test scores. In the DAPPER solution for IELTS, listening comprehension and reading comprehension sections loaded most

heavily on the first factor, while the writing section loaded most heavily on the second. This suggests that the first factor can be associated with the receptive skills factor and the second with the productive skill factor. This is in line with the earlier findings of Alderson (1993) on the trial study of IELTS: *"In general, an analysis of reading, grammar, and listening yielded only one common factor. The addition of writing occasionally gave rise to a second factor"* (p. 213).

Table 5.2: Exploratory Factor Analysis of IELTS Raw Scores (A)

Variable	Communalities	
	Initial	Extraction
IELTS LC	0.45266	0.66863
IELTS RC	0.39250	0.57402
IELTS WR	0.30597	0.46842

Factor	Eigenvalue	% of Variance	Cumulative %
1	2.06066	68.7	68.7
2	0.56180	18.7	87.4
3	0.37755	12.6	100.0

DAPPER rotated reference vector (factor) matrix

	Factor 1	Factor 2
IELTS LC	**0.458**	0.389
IELTS RC	**0.743**	-0.001
IELTS WR	0.001	**0.649**

Factor correlation matrix

	Factor 1	Factor 2
Factor 1	1.000	
Factor 2	0.931	1.000

Moreover, the factor correlation matrix shows that the two factors were highly correlated (0.931). That is, they were highly oblique, which called for the investigation of the possible higher-order factor. Using Carroll's (1999) HIGHERFACTOR programme a Schmidt–Leiman[5] transformation to orthogonal primary factors with a second-order general factor was performed on the factor solution. Table 5.3 illustratres that all the IELTS sections loaded most heavily on the

[5] Schmidt & Leiman (1957) proposed a procedure for the transformation of factors to orthogonal primary factors, the result of which is a matrix, P_{vo}, which has v variables as rows and the "orthogonalised factors" as columns. The highest-order factors are listed first and the primaries are listed last. The elements are the correlations of each variable with the part of that factor that is orthogonal to all factors at a higher order. All of the variance predictable from a higher level of analysis has been partialed out of the correlation between a variable and a factor. The derivation may be generalised to any number of higher-order factors. Carroll's programme will go up to order 5, although it would be unusual to require more than order 3.

general factor with listening comprehension loading the highest. The loadings on the first-order factors follow the pattern observed on DAPPER solution.

Table 5.3: Exploratory factor analysis of IELTS raw scores (B)

Orthogonalised Factor Matrix With Second-Order General Factor For IELTS Sections

	General factor	F1	F2	Communality
IELTS RC	**0.716**	**0.194**	0.000	0.551
IELTS LC	**0.817**	**0.120**	0.102	0.692
IELTS WR	**0.627**	0.000	**0.170**	0.422
SMSQ	1.574	0.052	0.039	1.666
% Variance	52.46	1.73	1.3	55.49

N= 134 Salient loadings on each factor are highlighted

As can be seen from the above table, IELTS latent trait is very much unifactorial with all the sections loading heavily on a general factor, which accounted for more than half (52.46%) of the total variance in the tests. IELTS listening and reading comprehension load, to a lesser degree, on the first first-order factor and the writing on the second, indicating the specific skills they are measuring. Although the loadings on the first-order factors are salient, they comprise such a small portion of the total variance (1.7 % and 1.3 % for F1 and F2, respectively) that is hard to justify that they provide any extra information other than the general ability to cope with the language.

This general ability or the *g-factor* is a common factor that all the IELTS sections tend to tap. This *g-factor* can be interpreted as general language proficiency. One should caution here, however, that this general language proficiency might not necessarily be the same general aspect of language proficiency found in other language tests using different groups of subjects. It is a common aspect of language proficiency shared by the subjects in this research as measured by IELTS subtests. To summarise, the exploratory factor analysis suggests that all the IELTS sections tend to measure a single language ability, with specific abilities being measured by reading and listening comprehension and writing. The high loadings of all the IELTS sections on the general factor indicate the dominance of a single general language ability in the latent trait of the battery. The less important but salient loadings of IELTS reading and listening sections on the first first-order factor and that of writing section on the

second first-order factor are indicative of the specific abilities these sections tend to measure. Figure 5.2 visualises the findings.

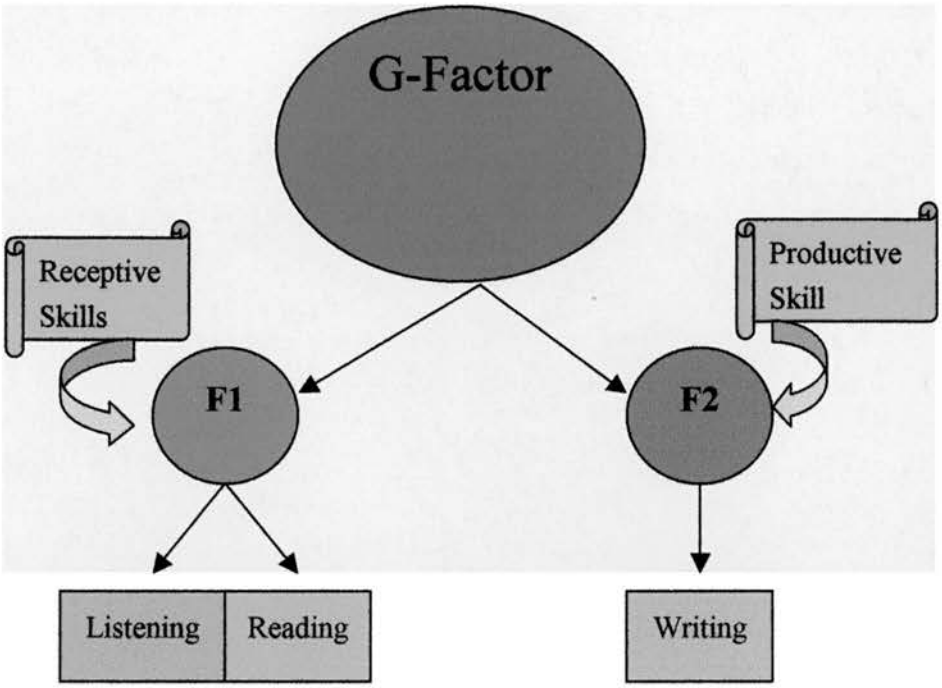


Figure 5.2: IELTS Latent Traits

5.3.1.2 Exploratory Factor Analysis of TOEFL

The results of exploratory factor analysis for TOEFL tests are given in Table 5.4 and Table 5.5. Initial parallel analysis and scree plots also suggested two factors underlay TOEFL test scores. As can be seen from the factor correlations in Table 5.4, oblique solution (DAPPER two-sided case) was the best solution. That is, it maximised the simplicity and interpretability of the factors underlying TOEFL test scores. In the DAPPER solution for TOEFL, reading comprehension and structure tests loaded most heavily on the first factor, while the listening section loaded most heavily on the second. The loading of the structure test on the first factor was very dominant (0.974) suggesting that the first factor could be interpreted as the structure and reading comprehension factor and that the second as the listening comprehension factor. One of the reasons that listening and reading comprehension in TOEFL, contrary to the IELTS case, do not load on the same factor could be due to the absence of a productive test in the analysis.

Table 5.4: Exploratory Factor Analysis of TOEFL Raw Scores (A)

Variable	Communalities	
	Initial	Extraction
TOEFL LC	0.37366	0.49944
TOEFL ST	0.58283	0.73535
TOEFL RC	0.60056	0.76461

Factor	Eigenvalue	% of Variance	Cumulative %
1	2.26400	75.5	75.5
2	0.48582	16.2	91.7
3	0.25018	8.3	100.0

DAPPER rotated reference vector (factor) matrix

	Factor 1	Factor 2
TOEFL LC	0.083	**0.589**
TOEFL ST	**0.974**	-0.124
TOEFL RC	**0.699**	0.184

Factor correlation matrix

	Factor 1	Factor 2
Factor 1	1.000	
Factor 2	0.976	1.000

Additionally, as was the case with IELTS, the factor correlation matrix of TOEFL scores shows that the two factors were even more highly correlated (0.976), suggesting that they were highly oblique, and calling for the investigation of the possible higher-order factor. A Schmidt–Leiman transformation to orthogonal primary factors with a second-order general factor was performed on the factor solution. Table 5.5 illustratres that all the TOEFL tests loaded most heavily on the general factor with the reading comprehension loading the highest. The loadings on the first-order factors follow the pattern observed on DAPPER solution.

Table 5.5: Exploratory Factor Analysis of TOEFL Raw Scores (B)

Orthogonalised Factor Matrix With Second-Order General Factor for TOEFL

	General factor	F1	F2	Communality
TOEFL ST	**0.840**	**0.151**	-0.019	0.729
TOEFL RC	**0.873**	**0.108**	0.029	0.774
TOEFL LC	**0.664**	0.013	**0.091**	0.449
SMSQ	1.908	0.035	0.010	1.952
% Variance	63.6	1.16	.33	65.09

N=126 Salient loadings on each factor are highlighted

As can be seen from Table 5.5, TOEFL latent trait is yet more unifactorial with all the sections loading heavily on a general factor, which accounted for a large proportion (63.6%) of the total variance in the tests. TOEFL structure and reading load, to a lesser degree, on the first first-order factor and the listening on the second, indicating the specific skills / component they are measuring. Although the loadings on the first-order factors are salient[6], they comprise such an insignificant portion of the total variance (1.16 % and 0.33% for F1 and F2, respectively) that it is hard to justify the view that they provide any additional information other than the general language ability. All the TOEFL sections tend to tap a common aspect of language proficiency shared by these subjects. The dominance of the general factor in TOEFL factor matrix seems to support the idea that TOEFL is a test of general language proficiency[7]. This could mean reducing the number of test sections.
Figure 5.3 visualises this interpretation.

[6] Salient, usually associated with *Significant*, has been used in the literature of factor analysis on an intuitive basis to identify high loadings. Technically, a salient loading is one that is sufficiently high to assume that a relationship exists between the variable and the factor. What is a salient level for one factor may not be a salient level for another factor. Carroll's programme computes the salient loading using Montanelli-Humphreys parallel analysis criterion (Montanelli & Humphreys, 1976, 341-348) for the number of salient roots.

[7] Although the researcher would have much preferred to use the term *ability* over *proficiency*, the latter is used here to conform to the terminology dominant in the TOEFL literature. Ability is equated here with proficiency.

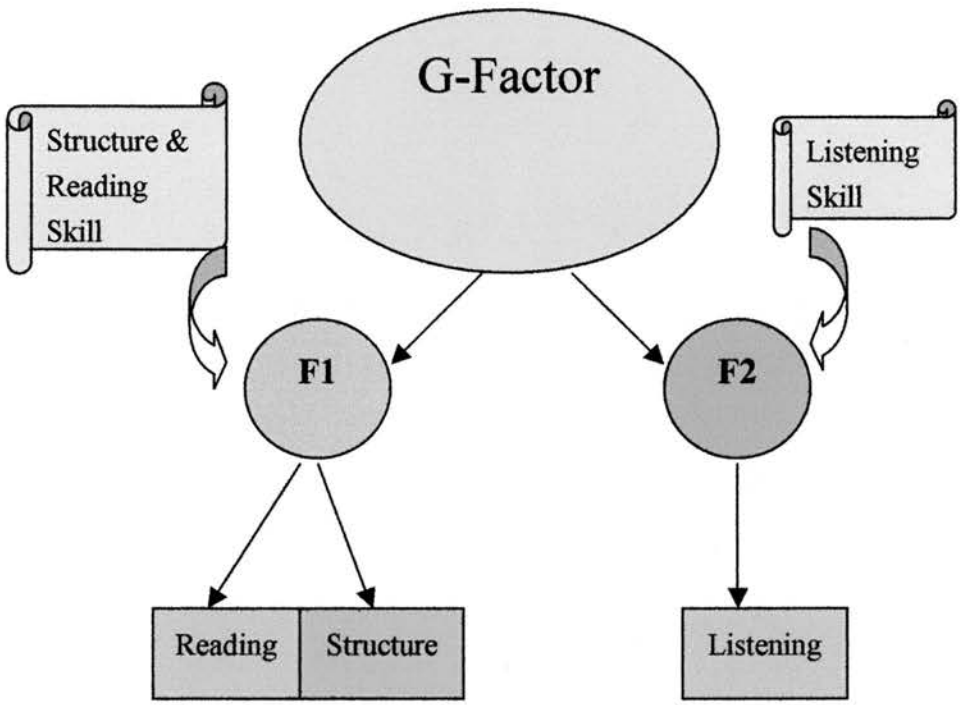


Figure 5.3: TOEFL Latent Traits

It is apparent that TOEFL latent trait is very much like that of IELTS but with one difference: the listening and the reading load on two different factors in TOEFL, whereas they load on the same factor in IELTS. A possible explanation could be that the TOEFL only tests the receptive skills (listening, structure, and reading); therefore, it is probable that in the process of the factor analysis of TOEFL the programme associated the inter-correlation of items within the subtests to two different factors, distinguishing the reading comprehension from the listening one. In IELTS, a second dimension, the writing that is a productive skill, is tested in addition to the receptive skills. In the presence of this new dimension and the paucity of variables, the reading and the listening comprehension items tended to load on a single factor. Had there been more variables (test sections), the factor loadings might have been different, as we will shortly see in the case of the analysis of factors across the batteries in 5.3.2.

5.3.1.3 Exploratory Factor Analysis of EPTB

Exploratory factor analysis of EPTB was not intended in the original research plan but since it was included in the factor analysis across batteries, it seemed worthwhile to examine its internal structure.

The results of exploratory factor analysis for EPTB tests are given in Table 5.6 and Table 5.7. Initial parallel analysis and scree plots suggested three factors underlay EPTB test scores. It can be seen from the factor correlations in Table 5.6, as was the case with the other two batteries, that oblique solution (DAPPER two-sided case) was the best solution. It maximised the simplicity and interpretability of the factors underlying EPTB test scores. In the DAPPER solution for EPTB, Test 3 loaded most heavily on the first factor, Test 2 on the second, Test 1 on the third, while Test 4 acting in a hybrid manner loaded highly on both the first and the second factor, although it was more salient on the latter. DAPPER solution for EPTB suggests that the first factor can be associated with language redundancy (cloze method), the second with the listening, and the third with the phonemic discrimination factor. Test 4 is the grammar component of the EPTB.

Table 5.6: Exploratory Factor Analysis of EPTB Raw Scores (A)

Variable	Communalities	
	Initial	Extraction
EPTB T1	0.09692	0.23146
EPTB T2	0.28611	0.51396
EPTB T3	0.29879	0.50698
EPTB T4	0.38805	0.66551

Factor	Eigenvalue	% of Variance	Cumulative %
1	2.08625	52.2	52.2
2	0.85207	21.3	73.5
3	0.63397	15.8	89.3
4	0.42771	10.7	100.0

DAPPER rotated reference vector (factor) matrix

	Factor 1	Factor 2	Factor 3
EPTB T1	0.000	0.000	**0.487**
EPTB T2	0.000	**0.712**	0.000
EPTB T3	**0.704**	0.000	0.000
EPTB T4	**0.593**	0.643	-0.459

Factor correlation matrix

	Factor 1	Factor 2	Factor 3
Factor 1	1.000		
Factor 2	0.716	1.000	
Factor 3	0.664	0.801	1.000

EPTB T1=EPTB Test 1, EPTB T2=EPTB Test 2, EPTB T3=EPTB Test 3, EPTB T4=EPTB Test 4

The factor correlation matrix of EPTB scores shows that the three factors were highly correlated (0.716, 0.664, and 0.801), calling for the investigation of the possible higher-order factor. A Schmidt–Leiman transformation to orthogonal primary factors with a second-order general factor was performed on the factor solution. Table 5.7 shows that all the EPTB tests loaded most heavily on the general factor but not as heavily as the other two batteries did. It is worth noting that EPTB was developed at the time when discrete-point testing was dominant in language test construction. Although a cloze section was also included in this battery, the designers of the test deliberately used many items to measure specific language elements (e.g., phonology, intonation, and stress). The assumption was that the language could be broken down into elements and skills and that the sum of the parts would be equal to language proficiency: the divisibility hypothesis. That means, each section of the test

should measure firstly a specific language ability and secondly, a general ability of the testee to cope with the language as a whole.

The loadings on the first-order factors follow the pattern observed on DAPPER solution with the exception that Test 4, the grammar component, now loads more saliently with Test 3 – the cloze- on the third factor. The dominance of the general factor is less vivid in the EPTB case, accounting for only 33.6% of the total variance in the tests. First-order factors seem to comprise a good percentage of the total variance with 3.2, 2.9, and 8.6 percentage of the total variance. This is a reflection of the divisibility hypothesis operationalised in this battery.

Table 5.7: Exploratory Factor Analysis Of EPTB Raw Scores (B)

Orthogonalised Factor Matrix With Second-Order General Factor for EPTB

	General factor	F1	F2	F3	Communality
EPTB T2	**0.662**	**0.263**	0.000	0.000	0.507
EPTB T1	**0.420**	0.000	**0.247**	0.000	0.237
EPTB T3	**0.543**	0.000	0.000	**0.449**	0.496
EPTB T4	**0.659**	0.238	-0.233	**0.378**	0.689
SMSQ	1.343	0.126	0.115	0.344	1.928
% Variance	33.575	3.15	2.875	8.6	48.2

EPTB T1=EPTB Test 1, EPTB T2=EPTB Test 2, EPTB T3=EPTB Test 3, EPTB T4=EPTB Test 4
N=134 Salient loadings on each factor are highlighted

5.3.2 Across Test Battery Factor Solutions

To examine how the scores from the two batteries correlated, the intercorrelations among IELTS raw scores and TOEFL raw scores were analysed in the presence of EPTB raw scores, using exploratory factor analysis explained above. Examination of scree tests suggested three factors should be extracted, while the parallel analysis criterion indicated up to seven factors. Therefore, it was decided to examine three, four, five, six, and seven principal axes factor solutions. As in the previous exercise, the factors were rotated to orthogonal and oblique solutions. Varimax orthogonal solutions created hybrid factors, which were very difficult to interpret, whereas oblique solutions were much simpler and could be interpreted in a much more meaningful way. The best solution was a six-factor oblique DAPPER-two sided case, which maximised the interpretability. Table 5.8 and Table 5.9 report the results of the exploratory factor analysis across the batteries.

Table 5.8: Exploratory Factor Analysis Across Batteries (A)

Variable	Communalities	
	Initial	Extraction
IELTS LC	0.69408	0.86075
IELTS RC	0.57582	0.64114
IELTS WR	0.42346	0.51162
TOEFL LC	0.72020	0.89338
TOEFL ST	0.62020	0.77169
TOEFL RC	0.60657	0.72404
EPTB Test 1	0.16860	0.38192
EPTB Test 2	0.51272	0.70840
EPTB Test 3	0.37740	0.58909
EPTB Test 4	0.60202	0.75608

Factor	Eigenvalue	% of Variance	Cumulative %
1	5.29314	52.9	52.9
2	0.95505	9.6	62.5
3	0.91434	9.1	71.6
4	0.73089	7.3	78.9
5	0.59483	5.9	84.9
6	0.42519	4.3	89.1
7	0.33828	3.4	92.5
8	0.32110	3.2	95.7
9	0.25232	2.5	98.3
10	0.17487	1.7	100.0

DAPPER rotated reference vector (factor) matrix

	F1	F2	F3	F4	F5	F6
IELTS LC	0.001	-0.007	0.557	0.035	**0.543**	0.048
IELTS RC	**0.691**	-0.008	-0.004	0.150	0.021	0.008
IELTS WR	-0.001	0.016	**0.726**	0.011	0.002	-0.047
TOEFL LC	0.630	0.014	0.001	-0.025	**0.472**	-0.055
TOEFL ST	**0.922**	-0.513	0.210	0.046	-0.015	0.037
TOEFL RC	**1.108**	-0.286	0.002	-0.014	-0.170	-0.004
EPTB T1	0.004	0.003	0.004	-0.018	0.004	**0.819**
EPTB T2	0.001	**0.900**	0.001	0.022	-0.008	0.022
EPTB T3	0.004	0.005	0.006	**0.970**	-0.000	-0.004
EPTB T4	0.243	-0.012	**0.688**	-0.054	0.150	-0.221

Factor correlation matrix

	F1	F2	F3	F4	F5	F6
F1	1.000					
F2	0.695	1.000				
F3	0.747	0.534	1.000			
F4	0.513	0.427	0.613	1.000		
F5	0.499	0.572	0.248	0.234	1.000	
F6	0.374	0.280	0.355	0.225	0.243	1.000

The DAPPER solution suggests that factor one can be identified as the *reading comprehension and Structure* factor exclusive to TOEFL and IELTS, where TOEFL

reading comprehension, structure and written expression as well as IELTS reading comprehension load most heavily. Factors two and six may be the listening and the phonemic discrimination factors associated with EPTB, where EPTB Test 2 and Test 1 load quite uniquely (0.900 and 0.819 respectively). Factor three can be the *writing ability* factor, where the IELTS writing section and the EPTB Test 4 - which is a grammar test - load saliently. Factor four might be the language redundancy factor, where EPTB cloze Test 3 loads highest (0.970). The fact that the only other variable loading on this factor is the IELTS reading comprehension, which has a gap-filling task, is another indication for its association with cloze. Finally factor four can be a listening factor associated with TOEFL and IELTS; only TOEFL and IELTS listening comprehension tests load saliently on this factor.

Table 5.9: Exploratory Factor Analysis Across Batteries (B)

Orthogonalised Factor Matrix With Second-Order General Factor Across Batteries

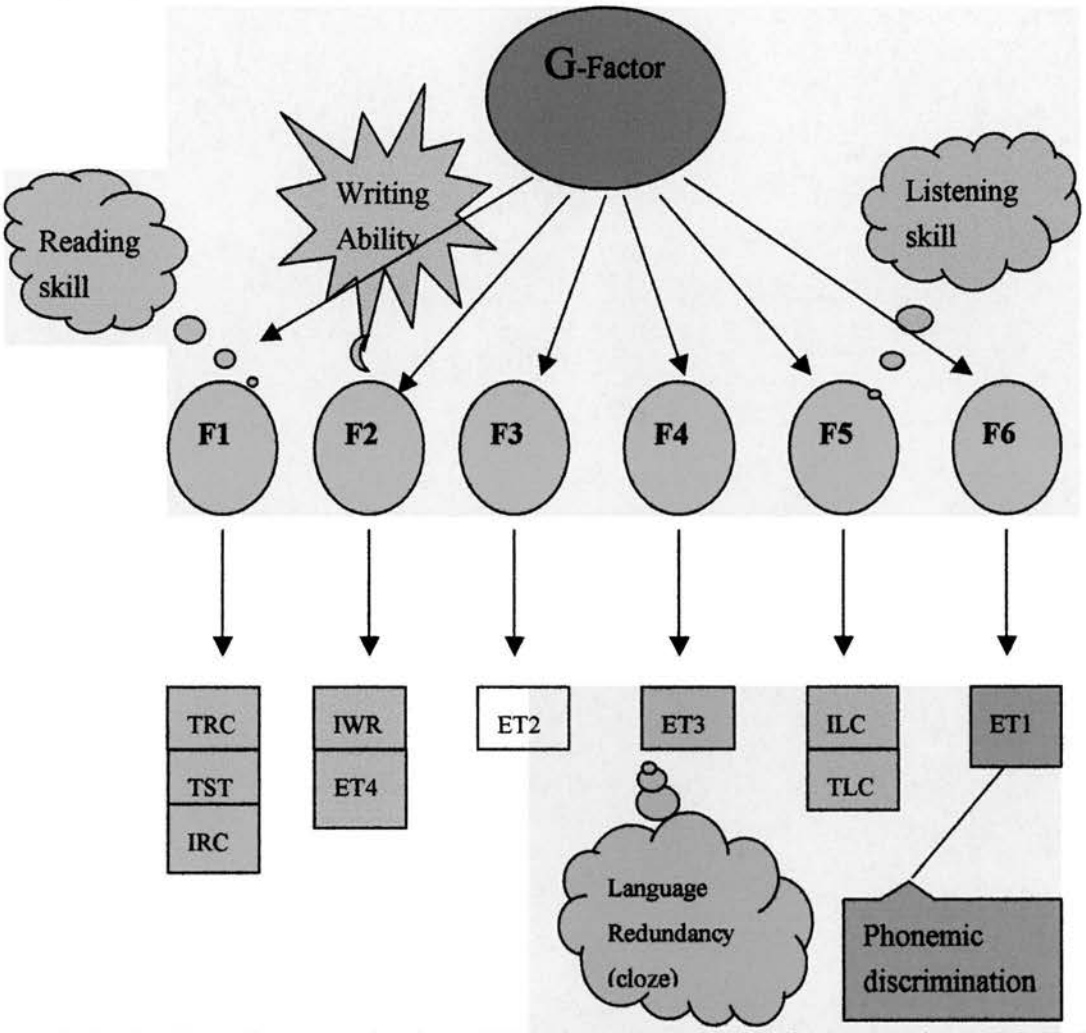
	General factor	F 1	F 2	F 3	F 4	F 5	F 6	Communality
		Factor 1: Reading comprehension and structure						
TOEFL RC	**0.706**	**0.430**	0.001	-0.184	-0.011	-0.146	-0.004	0.738
TOEFL ST	**0.655**	**0.358**	0.132	-0.331	0.036	-0.013	0.034	0.687
IELTS RC	**0.731**	**0.268**	-0.003	-0.005	0.120	0.018	0.007	0.622
		Factor 2 Writing ability						
IELTS WR	**0.566**	0.000	**0.455**	0.011	0.009	0.002	-0.043	0.529
EPTB T4	**0.706**	0.094	**0.431**	-0.007	-0.043	0.128	-0.201	0.752
		Factor 3: LC Associated with EPTB						
EPTB T2	**0.708**	0.000	0.001	**0.580**	0.017	-0.007	0.020	0.838
		Factor 4: Language redundancy						
EPTB T3	**0.594**	0.002	0.004	0.003	**0.775**	0.000	-0.004	0.954
		Factor 5: Listening comprehension						
IELTS LC	**0.751**	0.000	0.349	-0.005	0.028	**0.464**	0.044	0.905
TOEFL LC	**0.799**	0.244	0.001	0.009	-0.020	**0.404**	-0.050	0.864
		Factor 6: Phonemic discrimination						
EPTB T1	**0.337**	0.002	0.003	0.002	-0.014	0.004	**0.747**	0.672
SMSQ	4.452	0.454	0.533	0.480	0.620	0.417	0.607	7.562
% Variance	44.52	4.54	5.33	4.80	6.20	4.17	6.07	75.63

N= 105 Salient loadings on each factor are highlighted

Most of the factors correlate highly with one another, suggesting the possibility of a higher-order factor. A Schmidt-Leiman transformation to orthogonal primary factors with a second-order general factor was performed (Table 5.9). With the exception of EPTB Test 1, all the other tests loaded heavily on the second-order general factor as

expected. EPTB Test 1 is the most single-feature test in the analysis which receives a significant proportion of the total variance in the tests. Davies (personal communication, 1999) recalls that EPTB Test 1 has always behaved in this way.

Table 5.9 illustrates that the higher-order general factor accounts for a large amount of the total variance (44.52%) in the tests with the listening components loading the highest of all. Both TOEFL and IELTS tests load heavily on this general factor suggesting that they measure a common aspect of language abilities of the test takers in this study. The pattern of first-order factors follows the one explained in the DAPPER solution and can be characterised as follows: Factor 1 (4.54% of variance) – reading comprehension and structure (TOEFL RC, TOEFL ST, IELTS RC); Factor 2 (5.33 % of variance) – Writing ability (IELTS WR, EPTB T4); Factor 3 (4.80 % of variance) – listening associated with EPTB (EPTB T2); Factor 4 (6.20 % of variance) – language redundancy associated with cloze (EPTB T3); Factor 5 (4.17 %) – listening comprehension associated with TOEFL and IELTS (TOEFL LC, IELTS LC); and Factor 6 (6.07 % of variance) – phonemic discrimination (EPTB T1). Figure 5.4 visualises the interpretation.



TRC=TOEFL reading comprehension TST=TOEFL structure & written expression IRC=IELTS reading comprehension IWR=IELTS writing ET4=EPTB test 4 ET2=EPTB test 2 ET3=EPTB test 3 ILC=IELTS listening comprehension TLC=TOEFL listening comprehension ET1=EPTB test 1 F (1-6)=Hypothetical factors

Figure 5.4: Latent Traits Across Batteries

5.4 Discussion of Factor Analysis

Exploratory factor analyses within test batteries have shown that IELTS minus Speaking and TOEFL are highly unifactorial in terms of their trait structures. The test items in both tests seem to tap a common aspect of language proficiency shared by the test takers as measured by these tests. The g-factor, in the case of TOEFL, accounts for a considerable portion (63.6%) of the total variance in the subtests, while the first-order factors only account for 1.5 % of the total variance. This suggests that TOEFL

is a test of general language proficiency, which provides mainly information about the general language ability of the subjects taking the test. In addition to that, it also provides information, though to a lesser degree, about the learners' proficiency in specific skills or components, i.e., reading and listening comprehension.

The highly unifactorial structure of the TOEFL sample studied here is very similar to the results of other studies carried out on actual TOEFL tests in the past. For example, Swinton & Powers (1980, p. 15) found that TOEFL acted unifactorially for the less proficient groups with a single listening comprehension factor and a global factor underlying performance on reading and structure.

Moreover, as can be observed from the Multitrait-Multimethod Matrix (Table 5.10) the high correlation between the TOEFL reading comprehension and structure and written expression subtests (0.72) suggests that much of the information provided by these two sections overlap, questioning the validity of the independence of these sections in the battery. This may indicate the abandonment of the structure and written expression sub-test from the battery as an independent section; a similar conclusion was reported much earlier by Alderson (1993) with respect to IELTS, which led to the exclusion of the Grammar section in the final version of IELTS.

IELTS exploratory factor analysis has produced similar results. The g-factor accounts for much of the total variance (52.5%) in the sections of the tests, while the first-order factors account for 3% of the total variance. The significance of the accountability of the g-factor for the total variance is not as high as that of the TOEFL. Nevertheless, it still accounts for much of the total variance. This means that the IELTS sample, like the TOEFL sample, is very much unifactorial. This finding is consistent with the previous findings (Alderson 1993, and Clapham 1996) in the IELTS literature about its latent traits.

Based on the above results, it appears that the inclusion of both the receptive as well as productive skills in IELTS, despite providing some additional information about those skills, still ends up providing more information about the general language ability of the subjects than about the individual skills. In this respect, IELTS seems to function like TOEFL: providing information about the test takers' general language ability to cope with academic English.

Table 5.10: Multitrait-Multimethod Matrix across the Batteries

		Method 1 (IELTS)			Method 2 (TOEFL)			Method 3 (EPTB)			
	Traits	LC	RC	WR	LC	RC	ST	LC (T2)	RC (T3)	ST (T4)	T1
Method 1 IELTS	LC	(.85)									
	RC	.58	(.81)								
	WR	.53	.43	*							
Method 2 TOEFL	LC	**.76**	.65	.41	(.87)						
	RC	.44	**.63**	.45	.60	(.83)					
	ST	.52	.58	**.48**	.56	.72	(.85)				
Method 3 EPTB	LC (T2)	**.58**	.51	.37	**.64**	.38	.25	(.67)			
	RC (T3)	.51	**.50**	.45	.40	**.39**	.43	.41	(.76)		
	ST (T4)	.66	.58	**.58**	.61	.57	**.58**	.48	.47	(.78)	
	T1	.32	.25	.18	.24	.23	.26	.23	.17	.11	(.81)

LC=*Listening comprehension* RC=*Reading comprehension* WR=*Writing* ST=*Structure & Written Expression* LC(T2)=*EPTB Test2(Listening)* RC(T3)=*EPTB Test 3(Cloze)* ST(T4)=*EPTB Test 4(Grammar)* T1=*EPTB Test 1(phonemic discrimination)* *= No reliability reported for the Writing

Validity coefficients (monotrait-heteromethod) are the three diagonal sets of boldface numbers; reliability coefficients (monotrait-monomethod) are the numbers in parentheses along principal diagonal. Solid triangles enclose heterotrait-monomethod correlations; broken triangles enclose heterotrait-heteromethod correlations.

Across-battery analyses reveal more similarities between TOEFL and IELTS structures. They show that, firstly, a higher-order g-factor accounts for much of the variance in each of the batteries. Both batteries seem to measure, to a great extent, a single language ability, which may be called general language proficiency. This general factor accounts for almost half of the total variance (44.5%) in the batteries. Secondly, in addition to this general factor, each battery seems to provide some information about specific language abilities. The IELTS first-order factors, however, seem to provide relatively more information (about 47%) than the TOEFL ones (36%) about each specific language ability.

Across-battery analysis appears to suggest that the receptive skills in the two batteries provide similar information about the abilities of the test takers. The fact that similar sections of each test load on the same factor, i.e., factor 1 and factor 5 in Figure 5.4, is an indication that there is some inherent quality in both batteries, which brings them, closer to each other. TOEFL reading comprehension, TOEFL structure and written expression, and IELTS reading comprehension load heavily on factor 1, while TOEFL listening comprehension and IELTS listening comprehension load heavily on factor 5. Interestingly, IELTS listening and reading sections do not load on a single factor, as was the case in within test factor analysis.

In Chapter Two we argued that the addition of more than two different batteries in factor analysis allows us to see whether differences/similarities across tests are due to method or trait factors. The presence of a different proficiency test (EPTB) with similar sections in the final exploratory factor analysis has caused the receptive skills of IELTS to load on different factors and has helped us see how close IELTS and TOEFL internal structures are in terms of assessing the receptive skills of the test takers. The EPTB has a number of listening items but none of them loaded significantly on factor 5. This shows that factor 5 is not a general listening factor; it is a listening factor associated with TOEFL and IELTS. By referring to the Multitrait-Multimethod Matrix (Table 5.10) we can see that these two sections have the highest correlation (0.76) across the batteries. The same is true about factor 1, which is a reading factor associated with TOEFL and IELTS. The Multitrait-Multimethod Matrix (Table 5.10) shows that the correlation between the reading sections of IELTS and TOEFL is also high (0.63). We have already argued in Chapter Two, section 2.5.2 that the correlations between the monotrait-heteromethod sections of two tests represent validity coefficients. Since the validity coefficients of the listening and reading sections of IELTS and TOEFL (0.76 and 0.63) are relatively high, we could

conclude that the two test batteries measure the traits of listening comprehension and reading comprehension in much the same way. Put it in different words, the two batteries provide similar information about the listening and reading abilities of our test takers.

Additionally, it is evident that EPTB Test 4, a grammar test, does not load on factor 1, where TOEFL structure and written expression has loaded. Instead, it loads on factor 2 with the IELTS writing section. In a factor analysis one would normally expect that similar sections trait-wise load on the same factor, which are then adduced as evidence of convergent validity. The fact that EPTB Test 4 and TOEFL structure sections have not loaded on the same factor may have two implications. The first implication is that these sections of the test batteries do not provide similar information about the grammatical knowledge of our test takers; each section is associated with a different factor. The second is that the knowledge of grammar is not assessed as an independent trait by these batteries. On the one hand, the high correlation of TOEFL structure subtest with TOEFL reading (0.72) and their loadings on factor 1 along with the IELTS reading section, suggest that TOEFL structure and written expression section provides similar information to that of the reading comprehension tests. On the other hand, the loading of EPTB Test 4 (grammar test) on factor 2 along with the IELTS writing section indicates that EPTB section 4 was somehow testing some of the features associated with the writing ability as assessed by the IELTS. The validity coefficient in Table 5.10 for the IELTS writing and the EPTB Test 4 is moderately high (0.58). It is in fact the highest correlation between IELTS writing and any other subtests in TOEFL or EPTB.

This is not surprising as two thirds of the writing band for question one (coherence and sentence structure) and two fourths of the writing band for question two (word choice, form, and sentence structure) of the IELTS writing section were assessed with respect to the knowledge of grammar[8]. What is surprising is the fact that TOEFL structure and written expression section, which also appears to be associated with grammatical knowledge, does not load on factor 2. This could be due to the artefact of factor analysis techniques, which extract the factors on the basis of the intercorrelations among various items. The TOEFL structure and TOEFL reading sections have always been reported to have high correlation. This research was no

[8] See Chapter 3 section 3.6 for IELTS writing marking criteria. See also Appendix 5 for the marking instructions.

exception and as Table 5.10 illustrates there is a high correlation (0.72) for these two sections. The TOEFL Structure and Written Expression section had a strong association with the TOEFL Reading Comprehension, which could have linked them together to a single factor in the process of factor loadings.

It is also possible that the high correlation was due to the influence of the method factor. In Chapter Two, section 2.5.2 we examined the Campbell & Fiske (1959) hypothetical examples of the interactions between method and trait factors. It appears that one possible explanation for the large correlation between TOEFL structure and TOEFL reading sections (0.72) and a lower correlation between TOEFL structure and EPTB Test 4 (0.58) might be due to the function of test method variance common to the former traits and not to the latter, since the measures of the TOEFL structure and TOEFL reading were obtained by one method and that of EPTB Test 4 by another. We cannot examine whether the method was indeed a factor influencing the results here, as IELTS does not have an independent grammar component.

The Multitrait-Multimethod Matrix (Table 5.10), furthermore, reveals interesting facts about EPTB subsections. Firstly, EPTB Test 2 is measuring listening comprehension differently from the other two batteries. Despite loading on different listening factors, there are high correlations between the EPTB Test 2 and the listening sections of IELTS (0.58) and TOEFL (0.64). Secondly, the EPTB Test 3 is measuring some aspects of language proficiency, which are difficult to be associated with any particular trait in either IELTS or TOEFL. We associated this test with reading comprehension so that the Multitrait-Multimethod analysis was possible[9]. The correlations between the EPTB Test 3 and the reading sections of IELTS and TOEFL, however, do not support such association. In the case of IELTS, the highest correlation was between EPTB Test 3 and IELTS listening (0.51), whereas in the case of TOEFL, the highest correlation was with TOEFL structure and written expression (0.43). Since EPTB Test 3 is a cloze test, we may conclude that it is measuring language redundancy. Thirdly, EPTB Test 4 is measuring aspects of grammar common to all IELTS and TOEFL sections. Its highest correlations are with the listening sections of IELTS (0.66) and TOEFL (0.61). The fact that EPTB Test 4, which is a grammar test, does not correlate as high with the TOEFL structure (0.58) may have the implication that grammar is not an independent language trait. Finally, the low correlation between the EPTB Test 1 and all other subtests in the Multitrait-

[9] See the monotrait-heteromethod coefficients for the EPTB Test 3 in Table 5.10.

Multimethod Matrix is evidence of discriminant validity for this test. EPTB Test 1 is measuring a unique aspect of language proficiency that cannot be measured by any other subtests in our study. EPTB Test 1 high loading (0.75) on factor 6 (Table 5.9) further supports the idea that this test is measuring a single-featured of language proficiency, i.e., phonemic discrimination.

The factor analysis across the batteries supports the hypothesis that TOEFL and IELTS – excluding Speaking - despite their differences in test methods and language skills, measure a single general language ability of the subjects. Additionally, they measure specific skills of reading and listening comprehension in much the same way, with IELTS providing more information about these skills. IELTS also provides information about the writing ability of the test takers[10].

[10] Since Speaking was not included in the analysis, we cannot predict the likely outcome of its impact on the analysis but one could predict that the actual IELTS would also provide some information about the speaking ability of the test takers, however imperfectly that might have been. As we have already mentioned in Chapter 2, one of the reasons that the speaking section of IELTS was not included in the analysis was due to the serious problems raised in the literature about its reliability.

5.5 Results of the Analysis Of Item Difficulty

This section reports the results of the item analysis of TOEFL and IELTS based on the performance of our test takers on these tests. The results of the analysis should address question 8 of the research.

Q8: *Are the test items in TOEFL and IELTS significantly different in terms of their item difficulty?*

Having administered TOEFL and IELTS samples, we then entered the scores of the test takers into a database and analysed the reading and listening comprehension items using ITEMAN version 3.50. See Appendix 9 for the details of the analysis.

In the discussion of content analysis in Chapter 4, section 4.1.1 we examined the text difficulty of the reading passages in both batteries. The result of the readability analysis (Flesch Reading Ease)[11] showed that the IELTS passages were overall more difficult than their TOEFL counterpart, though the difference was not significant. The subjective ratings of the text difficulty, despite the judges' disagreement, also indicated that IELTS passages were judged to be more difficult. It was then necessary to examine whether the pattern of performance supported such interpretation. The item difficulties for all the reading items across the two batteries were calculated.

Table 5.11: Item Difficulty of the Reading Items

	Reading Items							
Batteries	IELTS			TOEFL				
Passages	IR1	IR2	IR3	TR1	TR2	TR3	TR4	TR5
# of items	13	15	7	7	6	6	5	6
Mean estimate: item difficulty	.38	.38	.29	.74	.48	.63	.51	.41

[11] See Chapter Four, section 4.1.1, Table 1.

Table 5.11 presents the total number of test items related to each passage and the average item difficulty estimates for those items across the batteries. The lower the item difficulty estimate, the more difficult the item is. As one can observe, the reading items associated with IELTS passages are on average more difficult than the TOEFL reading items. The Stem-and-Leaf Plots in Figure 5.5 illustrate how the reading comprehension items varied with respect to their item difficulty.

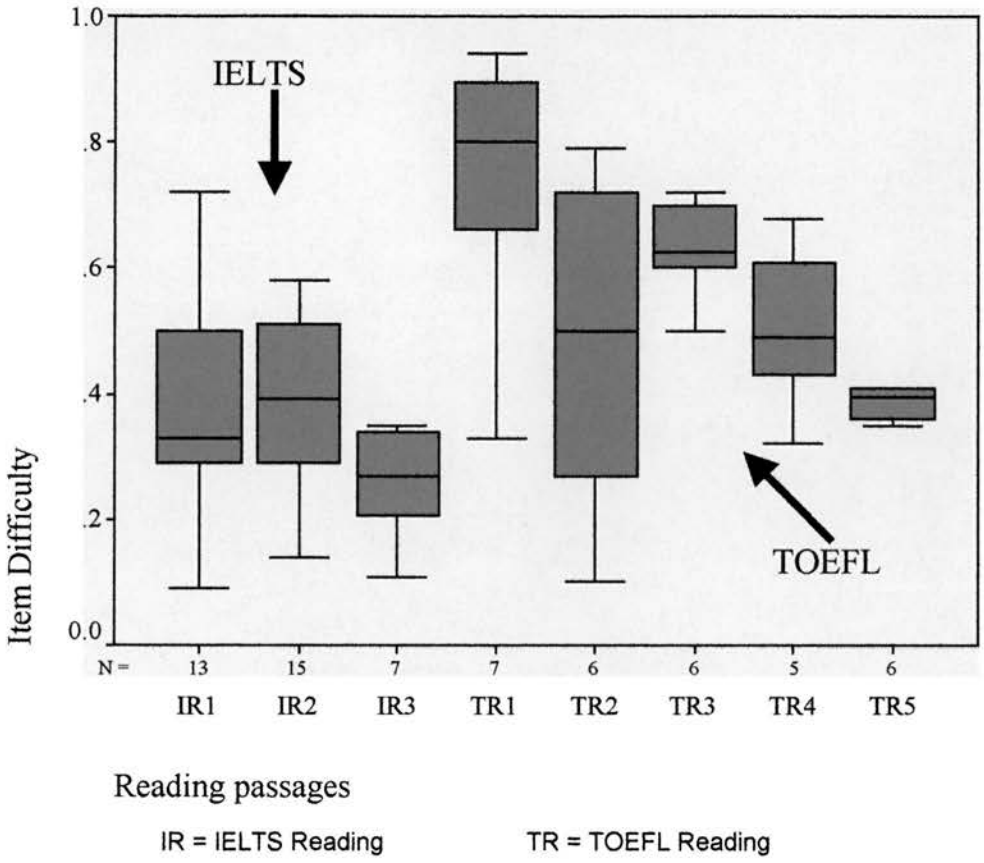


Figure 5.5: Stem-and-Leaf Plots of Item Difficulty: Reading Items

IELTS reading comprehension items were more difficult for our test takers. Reading comprehension questions related to IELTS reading passage three (IR3) were the most difficult of all for our subjects, followed by IELTS reading one and two (IR1, IR2). While the mean item difficulty estimates of the IELTS reading items seem to be of the same difficulty level across the passages, TOEFL reading items seem to vary quite a lot from one passage to another with TOEFL reading one (TR1) having the easiest items and TOEFL reading five (TR5) the most difficult. It is not appropriate to compare the estimates of item difficulty with the subjective ratings of the text

difficulty by the judges reported in Chapter Four. This is because of the unreliability of the subjective ratings of the texts. However, it seems legitimate to compare the item difficulty estimates with the readability figures of the reading texts using Flesch Reading Ease discussed in 4.1.1.

Figure 5.6 presents a comparison between item difficulty and the readability of the reading passages across the two batteries.

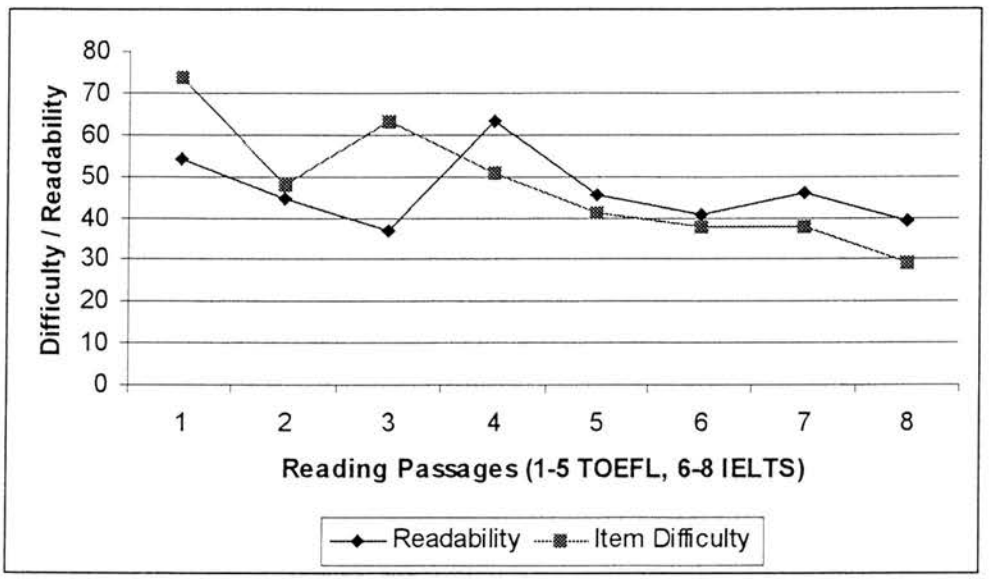


Figure 5.6: Comparing Item Difficulty With Readability Of The Reading Passages

There seems to be a close relationship between the readability of texts and the difficulty of the items associated with them. Nonetheless the correlation between the two (0.348) is not significant. It appears that with the exception of TOEFL readings three and four, the lower the readability of a text is judged to be, the more difficult the items associated with that text tend to be. The IELTS texts as well as the reading items associated with them seem to be the most difficult of all for our test takers. To test whether the difference in the item difficulty of the reading comprehension items is significant across the batteries, a *t test* was conducted on the item difficulty estimates.

Table 5.12: Comparison of Item Difficulty Means: Reading Comprehension Items

Batteries	N	Mean	Standard Deviation	*t value*	*df*	*p*(Sig.2-tailed)
TOEFL	30	0.56	0.21	4.284	63	0.000
IELTS	35	0.36	0.16			

Table 5.12 presents the number of items, means, standard deviations, *t value*, and the significance of the *t* for the item difficulty of the reading passages. The *t* obtained (4.28) is significant at $p<0.000$ suggesting that the IELTS reading items were significantly more difficult for our test takers.

The item difficulties of the listening comprehension items were also examined to observe if there was a significant difference between the two batteries in this regard. A *t test* was conducted on the item difficulties of the listening comprehension scores for the testing of the significance of difference across the batteries.

Table 5.13: Comparison of Item Difficulty Means: Listening Comprehension Items

Batteries	N	Mean	Standard Deviation	*t value*	*df*	*P* (Sig.2-tailed)
TOEFL	50	0.54	0.14	3.316	87	0.001
IELTS	39	0.40	0.24			

Table 5.13 presents the number of items, means, standard deviations, *t value*, and the significance of the *t* for the item difficulty of the listening items. The *t* observed (3.31) is significant at p <0.001 level suggesting that the IELTS listening comprehension items were significantly more difficult than their TOEFL counterpart for our test takers. IELTS reading and listening items tend to be significantly more difficult for our subjects; hence the null hypothesis $\mathbf{H_0}$ **8** cannot be sustained. This lends support to the alternative hypothesis with respect to question eight.

H_1 8: The test items in the two batteries are significantly different in terms of their item difficulty

Discussion of Item Difficulty

In this section we were trying to investigate whether the test items in the two batteries were significantly different in terms of their item difficulty. The results of the analysis of item difficulty across the batteries suggested that IELTS and TOEFL items were significantly different with respect to the item difficulty of their listening and reading comprehension items. What remains unanswered is whether the item difficulty of the reading items and indeed all the other items reflected the text difficulty of the passages.

One may argue that the item difficulty estimates are based on the performance on the tests and are the best measures of the difficulty inherent in test items. This, however, may not be the case in all situations. Item difficulty is the proportion of correct responses to the total possible responses. In most cases when a test taker does not attempt an item, we cannot be sure whether it is truly due to the difficulty of the item, or it is because they have not had the time to reach the item. In either case the response to that particular item would be considered wrong by the item analysis programme. The ITEMAN item analysis programme used here allowed us to take care of the items that were not attempted by the test takers. But this on its own could not change the fact that the item difficulty would still be calculated on the basis of the total correct responses. If a test taker does not attempt an item due to the lack of time or some other extraneous factors, one possible correct response may not be accounted in the calculation. This could inflate the difficulty estimate for that specific item.

It was then warranted to find out how much item difficulty estimates were affected by the subjects' failure to attempt all the items. It was decided to re-examine the item difficulty data sheets and observe the pattern of unattempted items. The analyses of TOEFL items show that in the case of listening comprehension items, which was the first section of TOEFL, only 1.6 % of the items were left unanswered. This figure rises as we reach the last section – the reading comprehension section- to 4.4 %. The examination of the last 6 questions of the reading comprehension section shows an increase in the number of unattempted items to 10.6 %. Incidentally, the last TOEFL reading passage associated with these questions has a mean item difficulty estimate of 0.41, making it the most difficult of the TOEFL passages. Clearly there is a link here between item difficulty and the percentage of the subjects that miss an item.

This pattern is more vivid in examining IELTS items. On average 17 % of the IELTS reading items were left unanswered, which is a high rate. This rate almost doubles to 31% when we examine the last six reading items associated with the IELTS reading passage three. These items had the highest item difficulty. The pattern of not attempting the items is similar in IELTS listening comprehension section. The average percentage of missing a listening item is 20 % in IELTS and doubles to 42 % for the last seven items. It seems that a significant proportion of the subjects did not attempt most of the final items in each section of the IELTS. The same pattern is repeated in TOEFL items though to a much lesser degree.

It is difficult to conjecture why the rate of missing items in IELTS is so high, in particular towards the end of each section when the rate doubles. It could be due to the length of the items, i.e., much longer passages. Perhaps the length had its effect on the subjects' memory and the test takers were exhausted before they reached the final items. It could also be due to the lack of subjects' familiarity with the IELTS format; very few of the subjects had previous experience of IELTS.

Another possibility is to do with the lack of clarity of the instructions, which is somehow related to the format of IELTS. For example, the last reading items were to do with a task of finding information in the texts and relating them to some graphs. This task was completely new to the test takers and confused most of them, as they did not know[12] what the question was about. The subjects were familiar with *'normal'* TOEFL-like reading comprehension questions and were not sure if the IELTS reading was really a reading test. The examination of IELTS listening items shows a similar pattern. The majority of the items that were left unanswered by the test takers were those, which required the examinees to fill in the blanks while listening to a conversation. There was no direct question to which the test takers could relate the items; they were given a task to complete.

It should also be borne in mind that by August 1994, when the IELTS test was administered, very few people knew about IELTS in Iran. Almost 60 % of the sample population (the K centre subjects) had not even heard the name prior to the administration of the test. Of those who knew about the test only 4.9 % claimed to have participated in the IELTS preparation course, which was carried out on a private

[12] The researcher acted as the chief test officer in the administration of the tests and recalls this from his own observation on one of the sites.

tuition basis. This brings up the importance of preparation courses in familiarising the test takers with the format of the proficiency tests, an issue we will shortly address.

From what has been explained above one may conclude that some of the difficulties perceived by the test takers could relate to the unfamiliarity of the subjects with the format of the IELTS, which could have been eased by IELTS preparation courses. The subjects had to spend time reading the instructions for IELTS items in order to understand what was required of them. This was not the case with the TOEFL items where they knew exactly what was required of them. Having examined the rate of unattempted items in IELTS, one cannot be sure if the difficulty of the items would have been the same had a different group of subjects, who were familiar with the IELTS, sat for the tests. This may cast doubt on the validity of the item difficulty exercise for the IELTS items. However, what cannot be refuted is the fact that in spite of the similarities of the contents of the two tests and the internal structures of the batteries, the test takers in this study found IELTS items significantly more difficult than the TOEFL ones.

5.6 Test Preparation Impact

This section reports the results of the investigation into whether test preparation courses had a significant effect on the performance of test takers on either the TOEFL or the IELTS tests. The research question to be investigated is question 9.

Q 9: Does test preparation have a significant effect on the performance of subjects on TOEFL and IELTS?

Information on test takers' participation in language proficiency courses was obtained from the Background Questionnaire[13] that was given to all the subjects at the end of the test administrations. Test takers were asked if they had participated in any of the following four preparation courses: TOEFL, FCE (First Certificate in English), CPE (Cambridge Proficiency in English), and IELTS. Figure 5.7 illustrates the number of subjects who participated in proficiency preparation courses.

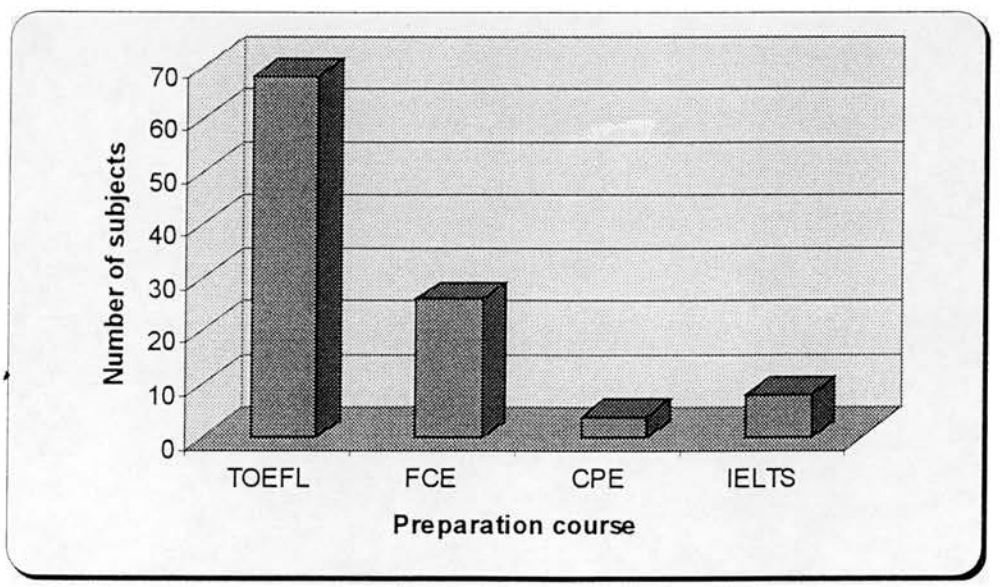


Figure 5.7: Comparison of Subjects in Preparation Courses

The number of subjects who claimed to have participated in CPE and IELTS preparation courses was too small (4 and 8 for CPE and IELTS, respectively) to be included in any meaningful analysis. Therefore, they were excluded from the analysis.

[13] See Appendix.1 for the original English version of the Background Questionnaire.

More than forty two percent of the subjects (68) stated that they had participated in TOEFL preparation courses or were currently participating in one and twenty percent of them (30) stated that they had participated in FCE preparation courses.

To examine the possible effects of TOEFL or FCE courses on test performance, hierarchical multiple linear regression analysis was applied. In each analysis, the preparation course (TOEFL or FCE) was used as a dummy variable (0 = No, 1 = Yes) and a covariate to control for differences in ability. For example, to examine the effect of a TOEFL preparation course on IELTS LC section, TOEFL LC served as the ability covariate (dependent variable), with the dummy variable representing TOEFL test preparation. Similarly, to examine TOEFL preparation effects on TOEFL LC, the IELTS LC section served as the ability covariate, with the dummy variable representing TOEFL test preparation.

There were 14 regressions in all: seven for the effects of TOEFL preparation course on IELTS LC, IELTS RC, IELTS WR, TOEFL LC, TOEFL RC, TOEFL ST, and TOEFL Total score and seven for the effects of FCE preparation course on the same tests. The ability covariate in the case of TOEFL Total score was EPTB Total score as there was no similar counterpart in the IELTS sections examined. The results of the regression analyses are given in Table 5.14 and Table 5.15. We will first discuss the effects of the TOEFL preparation course on the subjects' test performance.

5.6.1 TOEFL Course Preparation Effects On Test Performance

The tables of regression analysis (Table 5.14 & Table 5.15) contain 6 columns. Column 1 shows the variables included in each hierarchical analysis and the covariates corresponding to the ability being measured. Column 2, β refers to the standardised coefficient Beta, which is the beta weight showing the change in the dependent variable (expressed in standard deviation units) that would be produced by a positive increment of one standard deviation in the dependent variable. Column 3, **t** is a t-test for testing the regression coefficient for significance. Column 4, $\mathbf{R^2}$ is a positively biased estimate of the proportion of the variance of the dependent variable accounted for by regression. Column 5, $\mathbf{R^2}$ **change** shows the percentage of the variation in dependent variable accounted for by the addition of individual variables. Finally, the last column, **F-ratio,** is the ratio of the mean square for regression to the residual mean square to test the linearity of the relationship between the variables.

Table 5.14: Results of Multiple Linear Regression Analysis With TOEFL Preparation as a Dummy Variable (Method: Hierarchical)

Variables included	β	t	R^2	R^2 change	F-ratio
Dependent: TOEFL LC					
IELTS LC	.820	8.656*	.586	.586	144.58*
TOEFL Preparation	.142	.786	.586	.000	71.63*
PREP*IELTS LC	-.127	-.772	.589	.002	47.76*
Dependent: TOEFL RC					
IELTS RC	.671	6.807*	.403	.403	68.72*
TOEFL Preparation	.118	.621	.412	.009	35.32*
PREP* IELTS RC	-.104	-.104	.412	.000	23.32*
Dependent: TOEFL ST					
IELTS WR	.389	3.589*	.207	.207	26.38*
TOEFL Preparation	-.279	-1.031	207	.000	13.08*
PREP*	.306	1.153	.218	.010	9.19*
Dependent: IELTS LC					
TOEFL LC	-.764	8.732*	.586	.586	144.58*
TOEFL Preparation	-.004	-.020	.621	.034	82.56*
PREP*TOEFL LC	-.190	-.946	.624	.003	55.29*
Dependent: IELTS RC					
TOEFL RC	.689	7.272*	.403	.403	68.72*
TOEFL Preparation	.237	.643	.401	.058	43.18*
PREP* TOEFL RC	-.488	-1.329	.470	.009	29.59*
Dependent: IELTS WR					
TOEFL ST	.435	3.835*	.207	.207	26.38*
TOEFL Preparation	-.203	-.518	.234	.027	15.29*
PREP* TOEFL ST	-.203	.099	.234	.000	10.10*
Dependent: EPTB Total					
TOEFL Total	.688	7.805*	.517	.517	115.53*
TOEFL Preparation	-.124	-.196	.556	.039	66.86*
PREP* TOEFL Total	-.076	-.122	.556	.000	44.17*

*(p< .000) β=*Standardised coefficient Beta* **t**=*t-test for regression coefficient significance* R^2=*Variance (dependent variable)* **F-ratio**=*testing linearity of the relationships between the variables* RC=*reading comprehension* WR=*writing* LC=*listening comprehension* PREP=*preparation course effect*

As can be seen in Table 5.14, the F-ratios are highly significant for all the regression analyses, which confirms that the relationships between the variables were linear. It should be noted, however, that only an examination of the scatterplot of the variables can confirm that the relationship between the variables is genuinely linear. The scatterplots (see Appendix 10) confirm the linearity of such relationships for all the above regressions. For example, , illustrates the scatterplot for the regression analysis of TOEFL preparation effect on IELTS LC with TOEFL LC acting as covariate.

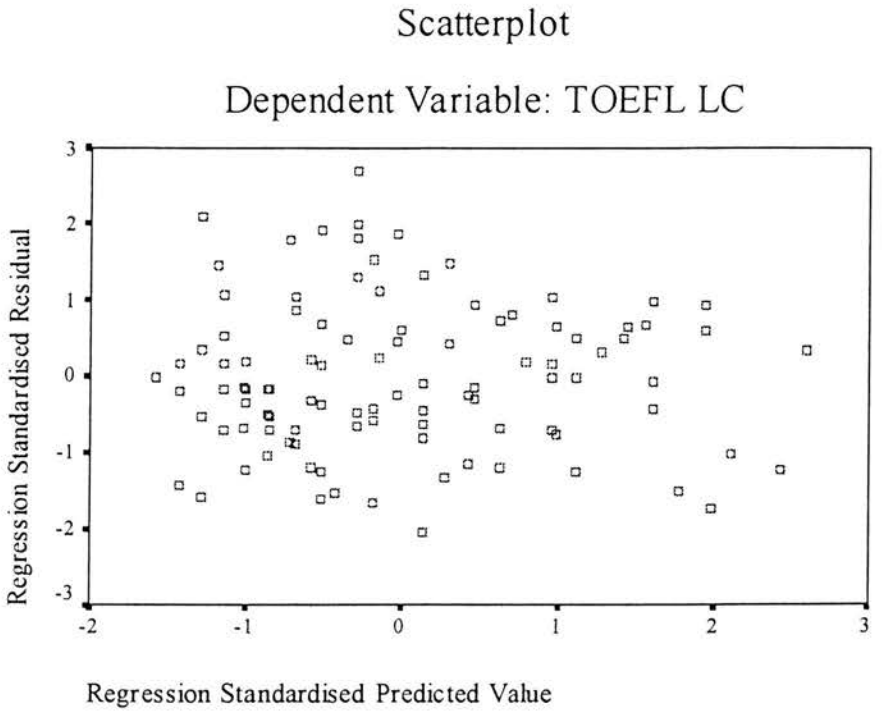


Figure 5.8: Scatterplot of Residuals Against Predicted Values

From the scatterplot of residuals against predicted values, one can see that there is no clear relationship between the residuals and the predicted values, thereby confirming that the assumptions of linearity and homogeneity of variance have been met.

The normal P-P plot of regression standardised residuals for the dependent variable (Figure 5.9) also indicates a relatively normal distribution.

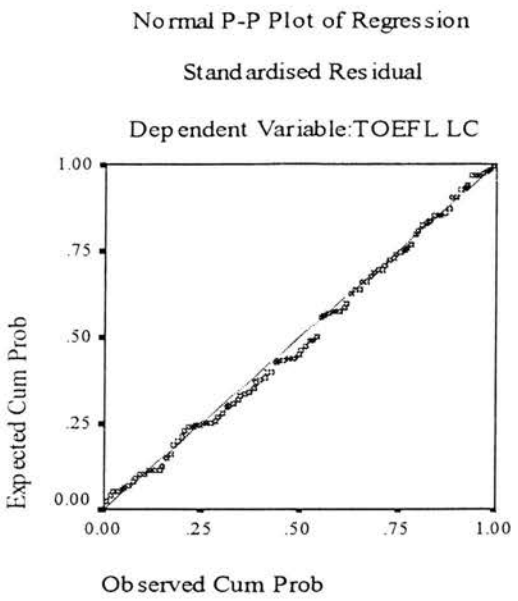


Figure 5.9: The Normal P-P Plot of Regression Standardised Residuals for The Dependent Variable

The same pattern has been observed for the rest of the scatterplots (see Appendix 10 for the details).

There seem to be two ways in which one can examine the effect of test preparation on test performance. One is to examine the *t* values of the regression coefficients for significance and the other is to look at the combined effects of test preparation variable and the interaction term accounted for the percentage of the variation in dependent variable. The *t* values in Table 5.15 indicate that none of the regression coefficients related to TOEFL test preparation are statistically significant, rejecting the effect of TOEFL test preparation on the performance of the subjects in any of the tests concerned here.

The combined effects of TOEFL test preparation and the interaction term (PREP*ABILITY) for 6 of the analyses (IELTS LC, IELTS RC, IELTS WR, TOEFL LC, TOEFL ST, and TOEFL Total) accounted for less than 4% of the variation in the dependent variable[14]. If we follow what Bachman et al. (1995, p. 75) seem to have

[14] See **R^2 change.**

adopted, any effect accounting for less than 5% of the variation in the dependent variable should be considered as having very little impact on test performance. That allows us to suggest that a TOEFL preparation course had insignificant influence on the performance of the subjects on IELTS LC, IELTS RC, IELTS WR, TOEFL LC, TOEFL ST, and TOEFL Total. For TOEFL RC, however, the combined effects of test preparation and the interaction between TOEFL preparation and the ability covariate accounted for 6.7% of the variation in test scores on TOEFL RC. Since this is beyond our 5% limit of confidence, it can allow us to say that participating in a TOEFL preparation course increased test takers' scores on TOEFL RC in comparison to test takers who did not participate in a TOEFL preparation course. Nevertheless, one should bear in mind that although the TOEFL preparation course had some positive impact on the scores of the subjects on the TOEFL RC section, it did not affect the total score on the TOEFL test. Nor did it affect the scores on any other section under investigation here.

5.6.2 FCE Course Preparation Effects On Test Performance

To examine the effect of a FCE preparation course on the performance of the subjects on TOEFL and IELTS, multiple regression analysis was used with FCE preparation as the dummy variable (0 = No, 1 = Yes) and the ability as the covariate. The method of analysis is similar to that of the TOEFL preparation effect discussed above. Table 5.15 summarises the results of such analysis using the hierarchical multiple linear regression analysis method.

Table 5.15: Results of Multiple Regression Analysis With FCE Preparation as a Dummy Variable (Method: Hierarchical)

Variables included	β	t	R^2	R^2 change	F-ratio
Dependent: TOEFL LC					
IELTS LC	.812	9.583*	.586	.586	144.58*
FCE Preparation	.213	.874	.586	.000	71.62*
PREP*IELTS LC	-.256	-.975	.590	.004	48.04*
Dependent: TOEFL RC					
IELTS RC	.616	6.408*	.403	.403	68.72*
FCE Preparation	-.095	-.461	.403	.000	34.04*
PREP* IELTS RC	.098	.436	.404	.001	22.58*
Dependent: TOEFL ST					
IELTS WR	.525	4.691*	.207	.207	26.38*
FCE Preparation	.432	1.658	.213	.006	13.54*
PREP* IELTS WR	-.407	-1.440	.229	.016	9.82*
Dependent: IELTS LC					
TOEFL LC	.637	9.146*	.586	.586	144.58*
FCE Preparation	-.065	-.225	.643	.056	90.76*
PREP*TOEFL LC	.341	1.139	.647	.005	61.12*
Dependent: IELTS RC					
TOEFL RC	.551	6.297*	.403	.403	68.72*
FCE Preparation	-.183	-.481	.442	.039	40.00*
PREP* TOEFL RC	.403	1.035	.448	.006	27.04*
Dependent: IELTS WR					
TOEFL ST	.403	4.146*	.207	.207	26.38*
FCE Preparation	.019	.039	.242	.035	15.98*
PREP* TOEFL ST	.179	.367	.243	.001	10.61*
Dependent: EPTB Total					
TOEFL Total	.635	8.222*	.517	.517	115.53*
FCE Preparation	.050	.075	.562	.045	68.53*
PREP* TOEFL Total	.178	.262	.562	.000	45.31*

*(p< .000) β=*Standardised coefficient Beta* **t**=*t-test for regression coefficient significance* **R^2**=*Variance (dependent variable)* **F-ratio**=*testing linearity of the relationships between the variables* RC=*reading comprehension* WR=*writing* LC=*listening comprehension* PREP=*preparation course effect*

As can be seen from Table 5.15 above, again the F-ratios are highly significant for all the regressions indicating that the relationships between the variables were linear. Scatterplots also confirm linear relationship (see Appendix 10). Moreover, the normal plots of regression standardised residuals for the dependent variables indicate relatively normal distributions.

The *t* values in Table 5.15 indicate that none of the regression coefficients related to FCE test preparation is significant, rejecting the effect of FCE test preparation on the performance of the subjects in any of the tests concerned here.

The combined effects of FCE test preparation and the interaction term (PREP*ABILITY) for 6 of the analyses (IELTS LC, IELTS RC, IELTS WR, TOEFL

RC, TOEFL ST, and TOEFL Total) accounted for less than 5% of the variation in the dependent variable, which allows us to suggest that FCE preparation course had insignificant influence on the performance of the subjects on IELTS LC, IELTS RC, IELTS WR, TOEFL RC, TOEFL ST, and TOEFL Total. For TOEFL LC, however, the combined effects of test preparation and the interaction between FCE preparation and the ability covariate accounted for 6.1% of the variation in test scores on TOEFL LC. Since this is beyond our 5% limit of confidence, it may lead one to say that participating in an FCE preparation course increased test takers' scores on TOEFL LC in comparison to test takers who did not participate in a FCE preparation course. This is difficult to explain, as one would expect a positive effect from TOEFL preparation courses on TOEFL subsection tests, not from participating in a different course. It could be due to the suggestion that FCE tests, as Bachman et al. (1995) conclude, are very similar to TOEFL tests. If we do accept this possibility, there still remains the question as why TOEFL preparation courses did not have that effect.

We were concerned here with the question of test preparation effect on test performance. The results of the multiple regression analyses demonstrate that TOEFL preparation courses did not have significant effect on any of the sub-tests examined here. Although the combined effects of test preparation and the interaction between TOEFL preparation and the ability covariate had a significant effect on the scores of TOEFL RC, they did not affect TOEFL total score. Moreover, the FCE preparation did not have any significant effect on the sub-tests concerned here. It can be suggested that the preparation courses, in general, did not have a significant effect on the performance of the subjects on these tests. This lends support to null-hypothesis ***H_0 9.***

H_0 9: *Test preparation has no significant effect on the performance of subjects on TOEFL and IELTS.*

A word of caution. As we have already explained in 5.6 very few subjects knew about IELTS prior to the administration of the test. IELTS test preparation courses could potentially influence the scores of the IELTS test takers. It was not possible to investigate the impact of such influence on the scores of the subjects because the subjects did not participate in such courses. Therefore the above conclusion should be limited to the impact of TOEFL and FCE preparation courses.

Chapter Six

Conclusions

6. Conclusions

We started our discussion of language proficiency testing with an attempt to define the concept of language proficiency in the field of language testing. It was argued that although various definitions of the term proficiency have been presented in the past three decades, they all seem to point to three main uses of the term:

a) a general type of knowledge of or competence in the use of language, regardless of how, where or under what conditions it has been acquired;
b) ability to do something specific in the language;
c) and performance as measured by a particular testing procedure.

We also noted that the use of two notions of language proficiency have been prevalent among applied linguists; a single faceted notion associated with general language proficiency- the dominant use during 1960s and 70s; and a multifaceted notion associated with communicative competence ever since early 1980s. These various definitions have, in turn, led to different hypotheses about language proficiency. What it is that we are trying to measure?

We referred to the theoretical question that Spolsky (Spolsky et al., 1968) raised about the nature of language proficiency thirty years ago: *What does it mean to know a language?* It was mentioned that it is not possible to develop valid language tests without a method of defining what it means to know a language. Based on different interpretations of language proficiency, we argued that the question of what it means to know a language can have many possible answers. It was suggested that there are basically three responses. One is to assume that knowledge is broken down into the individual structures- the rules and the lexical items- that make up the grammar and the lexicon of a language, so that *knowledge of these items and rules* is what needs to be measured; this is often termed in testing as *the divisible hypothesis.* A second is to assume that there is an underlying factor in all linguistic behaviour, which has a role of principal factor in understanding as well as production. This knowledge, which underlies all language skills, is often referred to as *overall proficiency.* Finally, we argued that there are good reasons that skills are as important as overall proficiency, both of which represent important aspects of language proficiency. We referred to this notion as the *eclectic hypothesis*, which accommodates both overall proficiency as well as its divisibility into skills.

The review of the literature in language proficiency testing led us to the conclusion that there is no consensus among applied linguists as to what language proficiency refers to. However, the applied linguists seem to be content that proficiency is divisible by skill and thus we have four macro-skills: reading, writing, listening, and speaking (see Alderson & Clapham 1992a).

We, furthermore, argued that even if it is possible to devise a pure theoretical model of language proficiency, practitioners are always in need of an assessment tool to measure the language behaviour of their foreign/second language learners. In real life, there are tests of language proficiency, which operationalise those aspects of language proficiency models that are relevant to the specific task and situation of the language testers. In this regard we reviewed three language proficiency tests of TOEFL, IELTS, and EPTB, exploring their underlying theoretical structures. We then discussed the effects of test methods on the performance of test takers and emphasised the importance of the use of Multitrait Multimethod designs in language testing research for separating the trait from the method factors. We also mentioned that a complete Multitrait Multimethod approach was not feasible in this research, as the three tests under examination (IELTS, TOEFL, and EPTB) do not have identical sections. Instead, we limited the scope of the Multitrait Multimethod approach to the examination of the listening and reading sections. Reviewing the two dominant American and British traditions in language test development, we suggested that despite the differences in methods, scope and populations, all language proficiency tests seem to tap the same underlying construct of proficiency, however imperfectly they may do so.

Moreover, we argued that the results of proficiency tests are almost always compared by various stakeholders for different purposes. Therefore there is a need for comparability studies. Based on Bachman's (1990) theoretical model of communicative language ability and the facets of test methods, an empirical framework was proposed for comparing language proficiency tests. We selected two popular tests of language proficiency, TOEFL and IELTS, for a comparability study so that the examination of various factors that affect test performance would be possible. In 2.3 we demonstrated that these two tests represent two radically different approaches to language testing. We expected that comparing these tests would enable us to have a better understanding of the concept of language proficiency.

The main objectives of the study were to investigate the extent to which TOEFL and IELTS are comparable in terms of the operational definitions of language proficiency on which the two tests are based and the degree to which the two batteries provide similar / different information concerning the abilities of the test takers.

To achieve the above goals three sets of empirical questions were raised for the comparison of the batteries with respect to:

a) facets of test methods,
b) components of communicative language ability,
c) and test performance.

6.1 Main Findings and Implications

We shall first list each of the research questions, which were introduced in Chapter 3, briefly commenting on the results. Then, we shall discuss the implications of the findings and relate them to the investigation into the construct of language proficiency.

6.1.1 Questions Related to Test Method Facets

Research Question 1

Are the reading passages in the two batteries significantly different in terms of their length?

The reading passages in the two batteries were compared with respect to the total number of their words. The analysis showed that IELTS passages were significantly longer than their TOEFL counterpart. The batteries, then, were not comparable in this regard. We argued that length is potentially an important factor, which might influence the performance of test takers on language tests as it might increase memory load. An instance of this can be seen in the analysis of item difficulty in Chapter 5, where the most difficult items were those associated with the longer passages. Although longer texts might put an extra burden on the memory span of the test takers, they might also ease the comprehension of the passages through recycling redundant information. Length, however, as we have seen in the analysis of the

grammar facet, does not seem to have affected the grammatical complexity of the texts.

Research Question 2

Are the reading passages in the two batteries significantly different in terms of their propositional content?

There were problems with the rating instrument used in the comparison of the propositional content of the passages. Firstly, the judges found some of the descriptions of the facets unclear and confusing. For example, the distinction of the distribution of new information through compactness or diffuseness of the load of information. They did not find the compact/diffuse continuum helpful for the distribution of new information. The judges also had problems with the descriptions of contextualisation with respect to cultural content and topic specificity. Each judge interpreted these facets differently. The follow-up interviews with the judges indicated that they had serious reservations as to the use of this instrument for rating the facets. Secondly, there were problems with the reliability of the exercise. Two caveats were in order here. The 3-point scale (0-1-2) used for the ratings was too narrow and the number of items (8 passages) was too small to produce an acceptable reliability figure. A slight discrepancy in the judgement of the raters would have skewed the results. The above problems eventually lead to the abandonment of the propositional content instrument in the course of the research.

Although the above-mentioned problems did not allow us to examine Question 2 of the research reliably, it shed light on the complexity of the propositional content of texts and how the subjective ratings of such a nature could be pursued. The propositional content analysis exercise reported above served as a pilot study for the subjective ratings, giving us insights for modifying the ratings of the rest of the facets of test methods and the components of communicative language ability. The two important outcomes of this exercise were: the need for the training of the judges, and the increase of the number of items to be rated. These two shortcomings were dealt with later in the subjective ratings of the communicative language ability components.

Research Question 3

Are the reading passages in the two batteries significantly different in terms of their organisational characteristics?

In order to answer Question 3, we first needed to answer four other questions related to the various aspects of the organisational characteristics. The questions to be answered were as follows:

Research Question 3.1

Are the reading passages in the two batteries significantly different in terms of their syntactic complexity?

Research Question 3.2

Are the reading passages in the two batteries significantly different in terms of their lexical density?

Research Question 3.3

Are the reading passages in the two batteries significantly different in terms of their text difficulty?

Research Question 3.4

Are the reading passages in the two batteries significantly different in terms of their cohesive markers?

The instrument used for the analysis of organisational characteristics was based on pure linguistic descriptions of the facets. In most cases linguistic features were counted using various techniques and their means were compared for statistical significance. Syntactic complexity was basically defined in terms of sentence complexity and voice; sentence complexity was assessed with regard to the ratios of Word per Sentence, and Clause per Sentence; while voice was measured in terms of the proportion of Passives in a text. Lexical density was rated in terms of the ratios of Character per Word and Type / Token, as well as the traditional Lexical Density measurement (ratio of Content Words to Total Words). Text difficulty of the passages was measured using Flesch readability formula. The passages were additionally rated by expert judges for text difficulty but as the judges could not agree on the ease/difficulty of the texts, their judgements were not used in the final analysis. Cohesion referred to those surface-structure features of a text, which link different parts of sentences or larger units of discourse. This facet was subdivided into six sub-facets: *Reference, Substitution, Additive, Adversatives, Causals,* and *Temporals.* Explicit cohesive markers of the above types were counted for each passage as the evidence of their cohesive comparability.

Tests of significance were applied to all the above measures, the results of which showed that the two batteries were not significantly different in terms of the above facets. This supported the null-hypothesis that the reading passages in the two batteries were not significantly different in terms of their organisational characteristics. Nevertheless, it needs to be mentioned that the scorings used in this exercise did not cover all the aspects of text difficulty and cohesive relations in a piece of discourse. It only covered a limited range of features of discourse that could be measured objectively within the limitations of the study. Examining other aspects of discourse required subjective rating instruments that are prone to reliability problems.

Research Question 4

Are the relationships of the test items to the passages significantly different in the two batteries?

All the reading comprehension as well as the listening comprehension items in the two batteries were analysed for the kind of relationship they had with the passages to which the items referred. Based on the failure of the propositional content instrument to produce reliable subjective ratings, two judges who had worked together on the instrument for a long time rated the relationship of the items to passage using a 1-5 point scale. The interrater reliability of the exercise was satisfactory, thus the results of the ratings could be used for further analysis. The mean ratings for the reading and listening items across the batteries were obtained but it was difficult to decide how to interpret the meaningfulness of a difference in mean ratings as the facet was rated on a five-point scale and the amount of variation in ratings differed across the tests. In order to set a criterion for interpreting a difference as meaningful, it was decided to interpret any difference between the mean ratings for the TOEFL and the IELTS items on a given facet as *meaningful* if that difference was greater than the standard deviations of the ratings for that facet on either of the two tests. On the basis of this criterion, there was no meaningful difference between the mean ratings for this facet on reading and listening items in the two batteries.

An important outcome of the rating of the relationship of items to passage was the involvement of the judges in the instrument design. We demonstrated in 4.2.1 that the reliability of the rating exercise improved significantly when the judges worked together for several sessions and discussed their differences about the application of the rating criteria in the exercise; this involved the modification of some of the

components of the rating instrument. We argued that the judges should also agree on the usefulness of the rating instrument in measuring the facets under investigation. Once the initial agreement on the usefulness of the instrument is achieved, it is then possible to observe how reliable the actual rating exercise is.

6.1.2 Questions Related to Communicative Language Ability

Research Question 5

Do the test items in the two batteries involve the same degree of organisational competence for successful completion of a given task?

Organisational competence was related to the abilities that control the formal organisation of language and involved two other components of grammatical and textual competence, operative at two different levels of sentence and text. Grammatical competence was subdivided into the knowledge of vocabulary, morphology, syntax, and phonology / graphology and textual competence was subdivided into the ability to structure independent sentences and utterances to form a text according to the conventions of cohesion and rhetorical organisation. To investigate research question 5, it was necessary to first examine the following questions.

Research Question 5.1

Do the test items in the two batteries require the same degree of lexical knowledge for successful completion of a given task?

Research Question 5.2

Do the test items in the two batteries require the same degree of morphological knowledge for successful completion of a given task?

Research Question 5.3

Do the test items in the two batteries require the same degree of syntactic knowledge for successful completion of a given task?

Research Question 5.4

Do the test items in the two batteries require the same degree of phonological / graphological knowledge for successful completion of a given task?

Research Question 5.5

Do the test items in the two batteries require the same degree of knowledge of cohesive relations for successful completion of a given task?

Research Question 5.6

Do the test items in the two batteries require the same degree of knowledge of rhetorical organisation features for successful completion of a given task?

Since the same rating instrument that was employed in the rating of the above facets was also used for the ratings of pragmatic competence, we list the questions related to assessing the pragmatic competence here too.

Research Question 6

Do the test items in the two batteries involve the same degree of pragmatic competence for successful completion of a given task?

Pragmatic competence was defined as the ability to produce and understand sentences or utterances, which are appropriate to the context in which they occur. Pragmatic competence comprises two abilities: illocutionary competence, which is the ability to understand the pragmatic conventions for performing acceptable language functions and the sociolinguistic competence, which is the ability to understand the sociolinguistic conventions for performing language functions appropriate in a given context. To investigate these abilities the following two questions were formed.

Research Question 6.1

Do the test items in the two batteries involve the same degree of illocutionary competence for successful completion of a given task?

Research Question 6.2

Do the test items in the two batteries involve the same degree of sociolinguistic competence for successful completion of a given task?

Finally, the tests were compared for the degree of involvement *of Strategic competence* in a successful completion of a given task. Strategic competence was associated here with test wiseness. The question to be investigated was,

Research Question 7

Do the test items in the two batteries involve the same degree of strategic competence for successful completion of a given task?

Ratings of communicative language ability facets were made on the basis of:

a) the extent to which the judges felt the ability was required for the successful completion of the task, and
b) the general level of the ability required.

With the exception of strategic competence, all the other facets were rated on five-point scales (zero to 4). Strategic competence was rated on three-point scales (zero to 2).

All the 66 reading comprehension items and the 89 listening comprehension items in the two batteries were rated by two judges to determine the degree of the involvement of the communicative language ability facets and strategic competence in successfully completing a task. Using the same criterion as was used for the investigation of Research Question 4, the mean item ratings for the above facets were compared for meaningfulness of their difference. The findings show that despite some obvious differences in the two tests, such as the length of the passages and the heavy reliance on lexical knowledge in TOEFL, there were more similarities than perceived differences between the two batteries on the facets of communicative language ability. Out of twenty six possible pair comparisons of communicative language ability facets of the listening and reading comprehension items, only two proved to have meaningful differences: lexical and phonological/graphological knowledge in listening items. In terms of the combined scores, only lexical knowledge proved to have meaningful difference across the batteries. This suggested that TOEFL and IELTS listening and reading comprehension items were not significantly different in terms of their communicative language ability and strategic competence facets and were thus comparable with respect to these facets.

Having established a basis for the content similarity of the two batteries, subjects' scores on these tests were analysed to see if patterns of performance supported such interpretation. The exploratory factor analysis showed that, firstly, the internal structures of TOEFL and IELTS (excluding the Speaking section) were very much unifactorial. Much of the variance in IELTS was accounted for by a general factor, which accounted for more than half (52.46%) of the total variance. IELTS listening

and reading comprehension loaded, to a lesser degree, on the first first-order factor and the writing on the second, indicating the specific skills they were measuring. Although the loadings on the first-order factors were salient, they comprised such a small portion of the total variance (1.7 % and 1.3 % for F1 and F2, respectively) that it was hard to say they provided any extra information other than the general ability to cope with the language. This pattern was repeated in TOEFL where its latent trait was yet more unifactorial with all the sections loading heavily on a general factor, which accounted for a large proportion (63.6%) of the total variance. TOEFL structure and reading loaded, to a lesser degree, on the first first-order factor and the listening on the second, indicating the specific skills / component they were measuring. The loadings on the first-order factors comprised such an insignificant portion of the total variance (1.16 %and 0.33 % for F1 and F2, respectively), as was the case with the IELTS, that it was hard to justify the view that they provided any additional information other than the general language ability. The dominance of the general factor in both batteries' factor matrices seems to support the view that TOEFL and IELTS are tests of general language proficiency

The factor analysis across the batteries illustrated that a higher-order general factor accounted for a large amount of the total variance (44.52%) in the tests with the listening components loading the highest of all. Both TOEFL and IELTS (minus Speaking) sections loaded heavily on this general factor suggesting that they measured a common aspect of language abilities of the test takers in this study. The pattern of first-order factors showed that TOEFL RC, TOEFL ST, and IELTS RC loaded on Factor 1 (Reading comprehension and structure) and comprised 4.54% of the total variance. IELTS WR loaded on Factor 2 (Writing ability) and accounted for the 5.33 % of the total variance. TOEFL LC and IELTS LC loaded on Factor 5 (Listening comprehension associated with TOEFL and IELTS), which accounted for 4.17 % of the total variance.

Within test battery factor analyses revealed more similarities between TOEFL and IELTS structures. They showed that, firstly, a higher-order g-factor accounted for much of the variance in each of the batteries. Both batteries seemed to measure, to a great extent, a single language ability, which may be called *general language proficiency*. This general factor accounted for almost half of the total variance (53% for IELTS and 64% for TOEFL) in each battery. One should caution here, however, that this general language proficiency might not be necessarily the same general aspect of language proficiency found in other language tests using different groups of

subjects. It is a common aspect of language proficiency shared by the subjects in this research as measured by TOEFL and IELTS subtests.

Secondly, in addition to this general factor, each battery seemed to provide some extra information about specific language abilities. The IELTS first-order factors, nevertheless, seemed to provide relatively more information (47%) about each specific language ability than the TOEFL ones (36%).

Across battery analysis appeared to suggest that the receptive skills in the two batteries provided similar information about the abilities of the test takers. The fact that similar sections of each test loaded on the same factor - TOEFL reading comprehension and structure sections and IELTS reading comprehension loaded heavily on factor 1, while TOEFL listening and IELTS listening loaded heavily on factor 5- was an indication that there was some inherent quality in both batteries, which brought them closer to each other.

6.1.3 Questions Related to Test Performance

Subjects' scores on the two batteries were analysed to see if test performance supported the interpretations of the content analysis findings. Two specific research questions were addressed in test performance analysis. The first question was related to text difficulty. It was important to investigate whether the controversial grading of text difficulty in content analysis was supported by the subjects' scores on the tests; in other words, whether the subjects found differences in the difficulty level of test items in the two batteries.

Research Question 8

Are the test items in TOEFL and IELTS significantly different in terms of their item difficulty?

The comparison between the readability of texts, estimated in content analysis through Flesch Reading Ease, and the difficulty of items associated with the texts showed that although both analyses supported the view that IELTS texts and items were more difficult than the TOEFL ones, they differed in the significance of the difference between the two. In the case of the readability, the difference between the two batteries with respect to reading passages was not significant, whereas in the case of

item difficulty, the difference between the two with respect to reading comprehension items was highly significant. Moreover, the analysis of listening comprehension items followed the same pattern indicating that IELTS listening items were significantly more difficult than the TOEFL ones.

Further analysis revealed that there was a link between the item difficulty and the number of unattempted items in each battery. It showed that the rate of unattempted items rose significantly towards the end of each section where the item difficulty rose significantly as well. This suggested that the more difficult items were those that the test takers did not reach. The rate of missing items in IELTS was much higher (17% for reading and 20 % for listening items) compared to (4.4 % for reading and 1.6% for listening items) in TOEFL.

Two possible reasons could have contributed to the high missing rate in IELTS. Firstly, the majority of test takers were not familiar with the format of IELTS, as it had just been introduced in Iran. The subjects were confused most of the time about what was required from them. Secondly, the length of the reading passages could be a factor affecting test takers' performance. It was possible that the length had its effect on the subjects' memory and they were exhausted before reaching the final items.

The second question was related to test preparation impact.

Research Question 9

Does test preparation have a significant effect on the performance of subjects on TOEFL and IELTS?

It was necessary to investigate whether test preparation had any impact on the performance of the subjects on these tests. Various researchers (Messick 1980, Alderson and Wall 1993, and Bachman et al. 1995) have discussed the effect of test preparation on test performance. Since some of the subjects in this research were participating in a TOEFL preparation course, the impact of such a course on the performance of test takers needed to be investigated.

A number of hierarchical multiple linear regression analyses were conducted on the data extracted about test takers' participation in TOEFL and FCE preparation courses. The result of the analyses showed that the preparation courses, in general,

did not have a significant effect on the performance of the subjects on the two batteries. Since the number of subjects who participated in IELTS preparation courses was not large enough to be included in the analysis, it was not possible to investigate the impact of such influence on the scores of the subjects. Therefore, the above conclusion should be limited to the impact of TOEFL and FCE preparation courses. IELTS test preparation courses could potentially influence the scores of the IELTS test takers.

Summary of the Main Findings

The main aims of this research were to find out if TOEFL and IELTS were comparable with respect to their operational definitions of language proficiency and the degree to which they provided similar / different information concerning the abilities of the test takers. Analysis of test content suggested that both batteries are based on the notion that proficiency is divisible by skill and elements of language, thus we have tests of reading, writing, listening, and speaking, as well as tests of grammar and vocabulary. However, they differ in the scope of the skills that represent the proficiency. TOEFL is based on the notion that proficiency may be represented by the receptive skills (reading and listening) and language elements (vocabulary and grammar); they are sufficient requirements for the assessment of the general language proficiency. In contrast, IELTS is based on the notion that the receptive skills as well as the productive skills (writing and Speaking) are both essential requirements of general language proficiency. This led us to the conclusion that the two test batteries differ in their scope of measuring proficiency.

The analysis also showed that the TOEFL differs from the IELTS in its method of testing. While TOEFL items all follow a pure multiple-choice discrete-point testing approach, IELTS items follow a variety of test methods from multiple-choice to integrative methods and involve various communicative tasks. Despite this, it appears that both tests measure, to a great extent, a common aspect of the subjects' language ability; therefore their internal structures are very much unifactorial. A *g-factor* (general language proficiency) comprises much of the total variance in both tests. Additional information that each test provides about competence in reading, listening, writing, and knowledge of vocabulary and grammar does not seem to contribute much to the total variance. The factor analysis of the results indicates that the knowledge of grammar, as assessed by TOEFL Structure and Written Expression, does not appear as an independent trait in factor loadings and is highly correlated with the reading

section. In other words, it does not provide much additional information other than that already provided by the TOEFL reading comprehension section. A similar finding was reported much earlier by Alderson (1993) with respect to IELTS, which led to the abandonment of the grammar section in the final version of IELTS.

The two tests, by and large, seem to provide similar information about the subjects' general language proficiency. The content analysis of the two tests indicates that there are more similarities between the reading and listening comprehension sections of the two batteries than differences. This is supported by the factor analysis of the test scores of a group of subjects on these tests.

Contribution of the Thesis

This thesis offers two contributions to the interpretation of the concept of language proficiency as measured by TOEFL and IELTS. Firstly, that there is an underlying factor in all linguistic behaviour, which has a role of principal factor in understanding and production. This factor that underlies all language skills may be called *overall proficiency*. Secondly, in addition to overall proficiency, there are language skills, which instantiate that proficiency in different forms and provide information about the specific competence in each skill. These findings support Carroll's (1983) definition of language proficiency:

> *"I have assumed that there is 'general language ability' but at the same time, that language skills have some tendency to be developed and specialised to different degrees, or at different rates, so that different language skills can be separately recognised and measured." (Carroll 1983:82)*

The thesis has illustrated that it is possible to compare language proficiency batteries that have different approaches to language test development. The communicative framework used for the comparison of the batteries offered a basis for a meaningful comparison of different facets across the batteries, though it had its shortcomings. It seems that most of the components of content analysis instrument employed allowed us to compare different aspects of the batteries with relative confidence.

Finally, the thesis demonstrated that while content analysis is an important basis for the comparability of proficiency tests, it cannot give us a complete picture of the abilities that the tests measure. Equally important aspect in test comparability is the

analysis of test performance. Thus, although the majority of the facets examined in the content analysis of the two batteries pointed towards the similarities between the two tests, it was not possible to investigate how important their contributions were to the overall similarities of the abilities that the tests tended to measure.

The factor analysis of test performance supported the similarities across the two batteries found in the content analysis. However, the analysis of item difficulty of the test scores indicated the differences in the level of the abilities that the two batteries measured. This was related to the unfamiliarity of the subjects with the IELTS methods of testing, suggesting that test methods could make an impact on the performance of test takers if the test methods were completely new to the subjects. In the case of IELTS, the unfamiliarity with the test methods could have confused the subjects and resulted in a number of unattempted items, making them look more difficult.

Although the unfamiliarity with a particular test method could have a negative impact on test scores, there is no evidence that familiarity with a test method necessarily improves the test scores of a group of subjects. In this study, TOEFL preparation courses had no positive impact on test scores.

6.2 Suggestions for Further Research

We have discussed a number of limitations in the course of this study and have suggested some ways to overcome those problems in further studies. In this section we propose further suggestions for future research into the comparability of language proficiency tests and into the quest to define language proficiency.

Prior to any suggestion, it is important to warn the researchers of the main obstacle in comparability studies. The research into the comparability of standardised language proficiency tests is politically a sensitive issue for the stakeholders as it may cast doubt on the reputation of these tests as a reliable and valid measure of language proficiency. The designers of such tests will not be happy with independent research conducted on their tests. Against this background one needs to be aware that they should not expect much cooperation from the designers of such tests. This will seriously restrict the scope of any comparability study on standardised proficiency tests.

One of the main problems in the content analysis of the tests was the use of subjective ratings for the rating of the test contents. The rating instruments used in the analysis were the modified versions of those developed in the course of CTCS (Bachman et al., 1995) and required the judgements of the expert judges. The results of the judges' ratings on the propositional content of the passages showed that the judges' ratings were not reliable and the instructions for the ratings were not transparent. The major problem with the unreliability was the fact that the judges were not trained to use this particular instrument, whereas the problem of transparency related to the validity of the instrument for the ratings of the propositional content facets. Bachman et al. (1995) found the propositional content instrument transparent for their judges, while Clapham (1996) found it to be otherwise. The judges in this study did not find the propositional content instrument transparent. We concluded that a valid rating instrument for one group of raters might not necessarily be valid for a different group of judges. Since assessing the propositional content is an important aspect of test comparability, it is suggested that further studies should examine the usefulness of the propositional content instrument for comparing tests, and in doing so make sure that their raters evaluate the subjective rating instrument and undergo extensive training prior to use.

The results of the content analysis of the tests led us to the conclusion that the test items in the two batteries do not differ significantly in terms of the degree to which they assess communicative language ability and facets of test methods. However, the analysis of item difficulty of the test scores contradicted this finding and pointed to significant differences between the two tests in terms of the difficulty of their reading and listening items. Further research should explore the text and item difficulty of the two tests at the content level and examine the factors that might influence the difficulty of the items. Such research may show us what aspects of difficulty have been ignored in our communicative framework for comparing the test contents. Further research can investigate the importance of the familiarity of the subjects with the instructions in IELTS; the effect of IELTS preparation courses on the test scores can be investigated in the same study.

Because this research was limited to examining subjects from one particular language background, i.e., Farsi speakers, it would be useful in further research to investigate subjects from different language backgrounds and language abilities.

Due to the limitation of resources, we were not able to examine all aspects of test method facets; facets of the test rubric were omitted from our study. It would be worthwhile if a similar study could revisit the test methods with greater emphasis on the examination of the facets of the test rubric. Facets such as test organisation and instructions seem to be of great importance in IELTS. If IELTS is to be compared sensibly with any other proficiency tests, one needs to examine its organisation, and instructions in detail. Salience of parts, sequence of parts, and relative importance of parts play an important part in differentiating IELTS from other tests. Equally important are the instructions in the form of specification of procedures and tasks and explicitness of criteria for correctness. These are the aspects that are suspected in our research to have caused IELTS items to look more difficult.

Finally, this research did not investigate score comparability across the tests. Since this is an aspect of the tests that concerns job opportunities and perhaps life chances of individuals taking the tests, further research should examine score comparability across the tests. In doing so, one needs to examine the fairness of the inferences we make on the basis of test scores. In other words, further research may study the consequential validity of these tests, an issue that, as Bachman (2000, p. 25) predicts, is going to dominate the language testing paradigms in the next 20 years.

References

Alderman, D.L. (1981). *Language proficiency as a moderator variable in testing academic aptitude: TOEFL Research Reports 10.* Princeton, NJ: Educational Testing Service.

Alderson, C. J. (1978). *A study of the cloze procedure with native and non-native speakers of English.* Unpublished PhD thesis. University of Edinburgh.

Alderson, C. J. (1989). Reaction paper to Bachman et al.: An investigation into the comparability of two tests of EFL: The CTCS final report. In *Cambridge TOEFL Comparability Study: Responses to the final report.* Cambridge: University of Cambridge Local Examinations Syndicate.

Alderson, C. J. (1990a). Testing reading comprehension skills: Part one. *Journal of Reading in a Foreign Language, 6(2),* 425-38.

Alderson, C. J. (1990b). Testing reading comprehension skills: Part two. *Journal of Reading in a Foreign Language, 7(1), 465-503.*

Alderson, C. J. (1993). The relationship between grammar and reading in an English for academic purposes test battery. In D. Douglas & C. Chapelle (Eds.), *A new decade of language testing research* (pp. 203-219). Virginia: TESOL.

Alderson, C.J. (1997). Bands and scores. In C. Clapham & C.J. Alderson (Eds.), *IELTS research report 3: Constructing and trialling the IELTS test,* (pp. 87-109). Cambridge: The British Council, The University of Cambridge Local Examinations Syndicate, and the International Development Program of Australian Universities and Colleges.

Alderson, C. J., & Clapham, C. (1992a). Applied linguistics and language testing: A case study of ELTS test. *Applied Linguistics,* 13(2), 149-167.

Alderson, C. J., & Clapham, C. (Eds.). (1992b). *IELTS research report 2: Examining the ELTS test; an account of the first stage of the ELTS revision project.* Cambridge: The British Council, The University of Cambridge Local Examinations Syndicate, and the International Development Program of Australian Universities and Colleges.

Alderson, C. J., & Clapham, C. 1997. The general modules: Grammar. In C. Clapham & C.J. Alderson (Eds.), *IELTS research report 3: Constructing and trialling the IELTS test,* (pp. 30-48). Cambridge: The British Council, The University of Cambridge Local Examinations Syndicate, and the International Development Program of Australian Universities and Colleges.

Alderson, C. J. & Hughes, A. (Eds.). (1981). *Issues in language testing.* London: The British Council.

Alderson, C. J., Krahnke, K.J. & Stansfield, C.W. (Eds.). (1987). *Reviews of English language proficiency tests.* Washington D.C.: TESOL.

Alderson, C. J., & Lukmani, Y. (1989). Cognition and reading: Cognitive levels as embodied in test questions. *Journal of Reading in a Foreign Language, 5(2),* 253-270.

Alderson, C. J., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14(2),* 115-129.

Allen, P., Cummins, J., Mougeon, R., & Swain, M. (1983). *Development of bilingual proficiency: Second year report.* Toronto, Ont.: The Ontario Institute for Studies in Education.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for education and psychological testing.* Washington, DC: American Educational Research Association.

Anastasi, A. (1990). *Psychological testing*. (6th Ed.). New York: Macmillan Publishing Company.

APA. (1954). *Technical recommendations for psychological tests and diagnostic techniques*. Washington, DC: American Psychological Association.

APA. (1966). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

APA. (1974). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Assessment Systems Corporation. (1993). *User's manual for the ITEMAN conventional item analysis programme*. St. Paul, Minnesota: Assessment Systems Corporation.

Bachman, L.F. (1982). The trait structure of cloze scores. *TESOL Quarterly*, 16(1), 61-70.

Bachman, L.F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L.F. (1991). What does language testing have to offer? *TESOL Quarterly*, 25(4), 671-704.

Bachman, L.F. (2000). Modern language testing at the turn of the century: assuring that what we count counts. *Language Testing*, 17(1), pp. 1-42.

Bachman, L.F., Davidson, F., & Foulkes, J. (1993). A comparison of the abilities measured by the Cambridge and Educational Testing Service EFL test batteries. In D. Douglas, & C. Chapelle, *A new decade of language testing research (pp.* 25-45). Alexandria, Virginia: TESOL.

Bachman, L.F., Davidson, F., & Milanovic, M. (1996). The use of test method characteristics in the content analysis and design of EFL proficiency tests. *Language Testing, 13(2),* 125-150.

Bachman, L.F., Davidson, F., Ryan, K. & Choi, I. (1995). *An investigation into the comparability of two tests of English as a foreign language.* Cambridge: Cambridge University Press.

Bachman, L.F., & Palmer, A.S. (1979). Convergent and discriminant validation of oral language proficiency tests. In R. Silverstein. (Ed.), *Proceedings of the Third International Conference on Frontiers in Language Proficiency and Dominance Testing (pp.* 53-62). Carbondale, Ill.: Department of Linguistics, Southern Illinois University.

Bachman, L.F., & Palmer, A.S. (1981). The construct validation of the FSI oral interview. *Language Learning,* 31(1), 67-86.

Bachman, L.F., & Palmer, A.S. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly,* 16(4), 449-465.

Bachman, L.F., & Palmer, A.S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Bachman, L.F., Vanniarajan, A.K.S., &. Lynch, B. (1988). Task and ability analysis as a basis for examining content and construct comparability in two EFL proficiency test batteries. *Language Testing,* 5(2), 128-159.

Bartlett, M.S. (1950). Tests of significance in factor analysis. *British Journal of Psychology,* 3(2), 77-85.

Bormuth, J.R. (1970). *On the theory of achievement test items.* Chicago: The University of Chicago Press.

Buel, J.G. (1993). TOEFL and IELTS as measures of academic reading ability. *Southern Illinois Working Papers in Linguistics and Language Teaching,* 2, 1-17.

Campbell, D. T., & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin,* 56, 81-105.

Canale, M. (1983). On some dimensions of language proficiency. In J. W. Jr Oller (Ed.), *Issues in language testing research* (pp. 333-42). Rowley, Mass.: Newbury House Publishers.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1,*1-47.

Carrell, P. (1987). Readability in ESL. *Reading in a Foreign Language,* 4(1), 21-40.

Carroll, B.J. (1968). The psychology of language testing. In A. Davies (Ed.), *Language Testing Symposium. A psycholinguistic perspective* (pp. 46-69). London: Oxford University Press.

Carroll, B.J. [1961(1972)]. Fundamental considerations in testing for English language proficiency in foreign students. In H.B. Allen & R.N. Campbell (Eds.), *Testing English as a second language: a book of readings* (2nd ed., pp. 313-321). N.Y.: McGraw-Hill.

Carroll, B.J. (1978). *English Language Testing Service: Specifications.* London: The British Council.

Carroll, B.J. (1980). *Testing communicative competence: An interim study.* Oxford: Pergamon.

Carroll, B.J. (1983). Psychometric theory and language testing. In J. W. Jr Oller (Ed.), *Issues in language testing research* (pp. 80-107). Rowley, Mass.: Newbury House Publishers.

Carroll, J.B. (1985). Exploratory Factor Analysis: A Tutorial. In D.K. Detterman (Ed.), *Current topics in human intelligence, 1,* (pp. 25-58). Norwood, N.J.: Ablex.

Carroll, B.J. (1986). LT + 25, and beyond? Comments. *Language Testing*, 3(2), 123-129.

Carroll, J.B. (1993). *Human cognitive abilities: a survey of factor-analytic studies.* New York: Cambridge University Press.

Carroll, J.B. (1999). *Exploratory factor analysis programmes for the IBM-PC.* Chapel Hill: Author.

Chapelle, C.A., & Abraham, R.G. (1990). Cloze method: What difference does it make? *Language Testing,* 7(2), 121-146.

Chyn S., DeVincenzi, F., Ross, J., &. Webster, R. (1992). TOEFL-2000: Update. Paper presented at the 26th Annual TESOL Convention, Vancouver, Canada.

Clapham C. (1993). Is ESP testing justified? In D. Douglas & C. Chapelle (Eds.), *A new decade of language testing research* (pp. 257-271). Virginia: TESOL.

Clapham, C. (1996). *The Development of IELTS: A study of the effect of Background knowledge on reading comprehension.* Cambridge: Cambridge University Press.

Clapham, C, & Alderson, C. J., (Eds.). (1997). *IELTS research report 3: Constructing and trialling the IELTS test.* Cambridge: The British Council, The University of Cambridge Local Examinations Syndicate, and the International Development Program of Australian Universities and Colleges.

Clark J.L.D. (1972). *Foreign language testing: Theory and Practice*. Philadelphia, Pa.: Centre for Curriculum Development, Inc.

Clark J.L.D. (1979). *An exploration of speaking proficiency measures in the TOEFL context: TOEFL Research Reports 4*. Princeton, NJ.: Educational Testing Service.

Clark, J.L.D. & Swinton, S. (1980). *The test of Spoken English as a measure of communicative ability in English-medium instructional settings.* Princeton, NJ: Educational Testing Service.

Cotton, F. & Conrow, F. (1998). An investigation of the predictive validity of IELTS amongst a group of international students studying at the University of Tasmania. In S. Wood (Ed.), *IELTS research reports 1998, V1 (*pp. 72-115). Canberra, Australia: IELTS Australia Pty Ltd.

Criper, C. & Davies, A. (1988). *ELTS Validation project report 1*. Hertford: The British Council and UCLES.

Cronbach, L.J. (1971). Test Validation. In R.L. Thorndike (Ed.), *Educational measurement.* Washington D.C.: American Council of Education.

Cronbach, L.J. (1988). Five perspectives on validation argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 34-35). Hillsdale, NJ: Erlbaum.

Cronbach, L.J. (1989). Construct validation after thirty years. In R.L. Linn (Ed.), *Intelligence: Measurement, theory, and public policy* (pp. 147-171). Chicago: University of Illinois Press.

Cronbach, L.J. (1990). *Essentials of Psychological Testing* (5th ed.). NY.: Harper and Row.

Cronbach, L.J., Gleser, G.C., Nanda, H. & Rajaratnam, N. (1972). *The Dependability of behavioural measurements: Theory of generalisability for scores and profiles.* New York: John Wiley & Sons, Inc.

Cummins, J.P. (1983). Language proficiency and academic achievement. In J.W.Jr. Oller (Ed.), *Issues in language testing research* (pp. 108-126). Rowley, Mass.: Newbury House Publishers.

Dale, E., & Chall, J.S. (1948). A formula for predicting readability. *Educational Research Bulletin*, 27, 11-20, 37-54.

Davidson, F., & Bachman, L.F. (1990). The Cambridge-TOEFL Comparability Study: An example of the cross-national comparison of language tests. *AILA Review* 7, 24-46.

Davidson, F., Turner, C.E., & Huhta, A. (1997). Language testing standards. In C. Clapham & D. Corson, (Eds.), *Encyclopaedia of language and education.* Volume 7: Language testing and assessment (pp. 303-311). Dordrecht: Kluwer Academic.

Davies, A. (1965). *Proficiency in English as a second language: the construction and use of a test of English proficiency among overseas students in Britain.* Unpublished PhD thesis. University of Birmingham.

Davies, A. (1967). The English proficiency of overseas students. *The British Journal of Educational Psychology*, 37(2), 167-175.

Davies, A. (1968). Introduction. In A. Davies (Ed.), *Language testing symposium. A psycholinguistic perspective* (pp. 1-18). London: Oxford University Press.

Davies, A. (1977). The construction of language tests. In Allen, J.P.B., & A. Davies (Eds.), *Testing and experimental methods: The Edinburgh course in applied linguistics* (pp. 38-104), *Vol. 4*. Oxford: Oxford University Press.

Davies, A. (1981a). Reaction to the Palmer and Bachman and the Vollmer papers (2). In Alderson, C. J. & A. Hughes (Eds.), *Issues in language testing* (pp. 182-187), London: The British Council.

Davies, A. (1981b). Review of J. Munby: Communicative syllabus design. *TESOL Quarterly*, 15(3), 332-336.

Davies, A. (1984). Validating three tests of English language proficiency. *Language testing*, 1(1), 50-69.

Davies, A. (1989). Comments on the CTCS final report. In *Cambridge TOEFL Comparability Study: Responses to the final report.* Cambridge: University of Cambridge Local Examinations Syndicate.

Davies A. (1990). *Principles of language testing.* Oxford: Basil Blackwell.

Davies, A. (1991). Is proficiency always achievement? Paper presented at Fourth Elicos Association Educational Conference, Monash University. Australia.

Davies, A. (1997a). Demands of being professional in language testing. *Language Testing*, 14, pp. 328-339.

Davies, A., (Ed). (1997b). Special issue: Ethics in language testing. *Language Testing*, 14(3).

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing.* Cambridge: Cambridge University Press.

Douglas, D. & Chapelle, C. (Eds.). (1993). *A new decade of language testing research*. Alexandria, Virginia: TESOL.

Dunteman, G.H. (1989). *Principal component analysis*. Beverly Hills: SAGE Publications Ltd.

Duran R.P., Canale, M., Penfield, J., Stansfield, C.W., & Liskin-Gasparro, J.E. (1985). *TOEFL from a communicative viewpoint on language proficiency: A working paper, TOEFL Research Reports 17*. Princeton, NJ.: Educational Testing Service.

Educational Testing Service (ETS). (1987). *TOEFL test and score manual supplement*. Princeton, NJ: Author.

Educational Testing Service (ETS). (1991). *TOEFL-2000: Planning for change*. Princeton, NJ.: Author.

Educational Testing Service (ETS). (1992a). *Newsline: International testing and training programs*. Princeton, NJ: Author.

Educational Testing Service (ETS). (1992b). *Bulletin of information for TOEFL/TWE and TSE, 1992-93*. Princeton, NJ.: Author.

Educational Testing Service (ETS). (1994). *TOEFL test and score manual supplement*. Princeton, NJ: Author.

Educational Testing Service (ETS). (1999). *TOEFL test and score data summary, 1999-00 edition*. Princeton, NJ.: Author.

Elder, C. (1993). Language proficiency as a predictor of performance in teacher education. *Melbourne Papers in Language Testing*, 2(1), 68-85.

Farhady, H. (1979). The disjunctive fallacy between discrete point and integrative testing. *TESOL Quarterly*, 13(3), 347-357.

Farhady, H. (1980). Justification, development, and validation of functional language testing. Unpublished PhD dissertation. University of California Los Angeles.

Farhady, H. (1983). On the plausibility of the unitary language proficiency factor. In J. W. Jr. Oller (Ed.), *Issues in language testing research* (pp. 11-28). Rowley, Mass.: Newbury House Publishers.

Ferguson, G. & White, E. (1998). A small-scale study of predictive validity. *Melbourne Papers in Language Testing,* 7(2), 15-63.

Fiocco, M. (1992). English proficiency levels of students from a non-English speaking background: a study of IELTS as an indicator of tertiary success. Unpublished research report. Perth: Curtin University of Technology.

Flesch, R. (1974). *The Art of Readable Writing.* New York: Harper & Row Publishers Inc.

Fouly, K.A., Bachman, L.F., & Cziko, G.A.. (1990). The divisibility of language competence: a confirmatory approach. *Language Learning*, 40(1), 1-21.

Fries, C.C. (1945). *Teaching and learning English as a foreign language*. Ann Arbor, MI: University of Michigan Press.

Geranpayeh, A. (1994). Are score comparisons across LP batteries justified? An IELTS-TOEFL comparability study. *Edinburgh Working Papers in Applied Linguistics,* 5, 50-60.

Gibson, C. & Rusek, W. (1992). The validity of an overall band score of 6.0 on the IELTS test as a predictor of adequate English level appropriate for successful academic study. Unpublished M A thesis. NCELTR

Gillespie, L. (1990). A comparison of two English language tests: FCE and CCSE. Unpublished MSc dissertation. University of Edinburgh.

Gorsuch, R.L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ.: Lawrence Erlbaum Associates, INC. Publishers.

Hale, G.A., Stansfield, C.W., Rock, D.A., Hicks, M.M., Butter, F.A., & Oller, J.W.Jr. (1989). The relation of multiple-choice cloze items to the Test of English as a Foreign Language. *Language Testing*, 6(1), 47-76.

Halliday, M.A.K. (1966). Lexis as a linguistic level. In C.E. Bazell, J.C. Catford, M.A.K. Halliday, & R.H. Robins, (Eds.), *In memory of J.R. Firth (pp. 148-162)*. London: Longmans, Green, & Co. Ltd.

Halliday, M.A.K. (1973). Relevant models of language. In M.A.K. Halliday, *Explorations in the functions of language*. New York: Elsevier North-Holland.

Halliday, M.A.K. (1985). *Spoken and written language.* Victoria: Deakin University.

Halliday, M.A.K., & Hassan, R. (1976). *Cohesion in English*. London: Longman.

Hamp-Lyons, L. (1997). Ethics in language testing. In C. Clapham & D. Corson, (Eds.), *Encyclopaedia of language and education.* Volume 7: Language testing and assessment (pp. 323-333). Dordrecht: Kluwer Academic.

Henning, G. (1987). *A guide to language testing*. Cambridge, Mass.: MIT Press.

Henning, G. (1988). Comments on the comparability of TOEFL and Cambridge CPE. *Language Testing*, 5(2), 217-222.

Hill, K., Storch, N., & Lynch, B. (1999). A comparison of IELTS & TOEFL as predictors of academic success. In R. Tullah (Ed.), *IELTS research reports 1999, V2 (pp. 52-63).* Canberra, Australia: IELTS Australia Pty Ltd.

Hymes, D. H. (1982). *Toward linguistic competence*. Philadelphia, Pa.: Graduate School of Education, University of Pennsylvania. (Mimeo).

Ingram, D., & Wylie, E. (1997). The general modules: Speaking. In C. Clapham & Alderson, C. J. (Eds.), *IELTS research report 3: Constructing and trialling the IELTS test,* (pp. 14-30). Cambridge: The British Council, The University of Cambridge Local Examinations Syndicate, and the International Development Program of Australian Universities and Colleges.

Ioannidou, A. (1990). A comparison evaluation of ELTS and TOEFL from a communicative point of view. Unpublished MSc dissertation. University of Edinburgh.

Irvine, A. (1991). A critical evaluation of and comparison between the Cambridge Certificate of Proficiency in English and the Certificates in Communicative Skills in English. Unpublished MSc dissertation. University of Edinburgh.

Jackendoff, R. (1983). *Semantics and cognition*. Cambridge, Mass.: MIT Press.

Klare, G.R. (1984). Readability . In P.D. Pearson (Ed.), *Handbook of Reading Research* (pp. 681-744). New York: Longman.

Lado, R. (1961). *Language testing*. NY.: McGraw-Hill.

Lewcowicz, J. (1983). Method effect in testing reading comprehension: a comparison of three methods. Unpublished MA. dissertation. University of Lancaster.

Long, J.S. (1983). *Confirmatory factor analysis: A preface to LISREL.* Beverly Hills: SAGE Publications Ltd.

McDowell, C., & Merrylees, B. (1998). Survey of receiving institutions' use and attitude to IELTS. In S. Wood (Ed.), *IELTS research reports 1998, V1 (*pp. 116-139). Canberra, Australia: IELTS Australia Pty Ltd.

Mead, R. (1982). Review of J. Munby: Communicative syllabus design. *Applied Linguistics*, 3(1), 70-78.

Messick, S.A. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist, 30,* 955-966.

Messick, S.A. (1980). Test validity and the ethics of assessment. *American Psychologist, 35,* 1012-1027.

Messick, S.A. (1989a). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). NY.: Macmillan.

Messick, S.A. (1989b). Meaning and values in test validation: The science of ethics of assessment. *Educational Researcher, 18(2),* 5-11.

Messick, S.A. (1995). Validity of psychological assessment: Validation of inferences from personal responses and performances as scientific inquiry into score meaning. *American Psychologist, 50(9),* 741-749.

Mok, M., Parr, N., Lee, T., & Wylie, E. (1998). A comparative study of IELTS & ACCESS test results. In S. Wood (Ed.), *IELTS research reports 1998, V1 (*pp. 140-208). Canberra, Australia: IELTS Australia Pty Ltd.

Montanelli, R.G., & Humphreys, L.G. (1976). Latent roots of random data correlation matrices with squared multiple correlations on the diagonal. *Psychometrika, 41*, 341-348.

Moss, P. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62(3),* 229-258.

Moss, P. (1994). Can there be validity without reliability? *Educational Researcher, 23(2),* 5-12.

Munby, J. (1978). *Communicative syllabus design.* Cambridge: Cambridge University Press.

Norton, B. (1997). Accountability in language assessment. In C. Clapham & D. Corson, (Eds.), *Encyclopaedia of language and education.* Volume 7: Language testing and assessment (pp. 313-322). Dordrecht: Kluwer Academic.

Oller, J.W.Jr. (1979). *Language tests at school: A pragmatic approach.* London: Longman.

Oller, J.W.Jr (Ed.). (1983). *Issues in language testing research.* Rowley, Mass.: Newbury House Publishers.

O'Neill, K., Steffen, M., & Broch, E. (1994). Does undergraduate major affect DIF on reading comprehension items? Paper presented at the National Convention of the American Educational Research Association. New Orleans, LA.

On Kim, J. & Mueller, C.W. (1978). *Introduction to factor analysis: What it is and how to do it.* Beverly Hills: SAGE Publications Ltd.

Peirce, B.N. (1992). Demystifying the TOEFL reading test. *TESOL Quarterly*, 26(4), 665-691.

Pike, L.W. (1979). *An evaluation of alternative item formats for testing English as a foreign language: TOEFL Research Reports 2*. Princeton, NJ.: Educational Testing Service.

Raimes, A. (1990). The TOEFL Test of Written English: Causes for concern. *TESOL Quarterly*, 24(3), 427-442.

Richards, J.C. (1985). Planning for proficiency. *Prospect*, 1(2), 1-17.

Schmidt, J. & Leiman, J. (1957). The development of hierarchical factor solutions. *Psychometrika,* 22(1), 53-61.

Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. *Language Testing,* 1(2), 147-70.

Skehan, P. (1984). Issues in the testing of English for specific purposes. *Language Testing*, 1(2), 202-220.

Skehan, P. (1989). State of the art: Language testing. *Language Teaching Abstracts* 22(1),1-13.

Spolsky, B. (1973). What does it mean to know a language? or how do you get someone to perform his competence? In J.W.Jr. Oller, & J.C. Richards (Eds.), *Focus on the learner: pragmatic perspectives for the language teacher* (pp. 164-176). Rowley, Mass.: Newbury House Publishers.

Spolsky, B. (1989). Communicative competence, language proficiency and beyond. *Applied Linguistics*, 10(2), 138-156.

Spolsky, B. (1990). The prehistory of TOEFL. *Language Testing*, 7(1), 98-118.

Spolsky, B. & Jones, R.L. (Eds.). (1975). *Testing language proficiency*. Arlington, Virginia: Centre for Applied Linguistics.

Spolsky, B., Sigurd, B., Sato, M., Walker, E., & Aterburn, C. (1968). Preliminary studies in the developments of techniques for testing overall second language proficiency. *Language Learning*, Special Issue, 3, 79-101.

SPSS Incorporated. (1998). *SPSS Base 8.0 for Windows User's Guide*. Chicago: SPSS, Inc.

Standal, T. (1987). Computer-Measured Readability. *Computers in the School,* 4(1), 123-132.

Stanley, J.C. (1971). Reliability. In R.L. Thorndike (Ed.), *Educational measurement.* Washington D.C.: American Council of Education.

Stansfield, C.W. (1993). Ethics, standards and professionalism in language testing. *Issues in Applied Linguistics,* 4(2), 15-30.

Stevenson, D.K. (1987). Test of English as a Foreign Language. In C. J Alderson,. K.J. Krahnke, & C.W. Stansfield, (Eds.), *Reviews of English language proficiency tests*. Washington D.C.: TESOL.

Swinton, S.S. & Powers, D.E. (1980). *Factor analysis of the TOEFL for several language groups*. Princeton, N.J.: Educational Testing Service.

Taylor, D.S. (1988). The meaning and use of the term `competence' in Linguistics and Applied Linguistics. *Applied Linguistics*, 9(2), 148-168.

Thorndike, R.L. (1971). *Educational Measurement.* Washington D.C.: American Council of Education.

Tucker, L.R. & Finkbeiner, C.T. (1981). *Transformation of factors by artificial personal probability functions. Research Reports 81-58.* Princeton: Educational Testing Service.

Tullah, R. (Ed.). (1999). *IELTS research reports 1999, V2.* Canberra, Australia: IELTS Australia Pty Limited.

University of Cambridge Local Examinations Syndicate (UCLES). (1999). *IELTS annual review (1998/9): Tenth anniversary edition.* Cambridge: Author.

Ure, J. (1971). Lexical density and register differentiation. In G.E. Perren, & J.L.M. Trim (Eds.), *Applications of linguistics: Selected papers of the second international congress of Applied Linguistics, Cambridge 1969* (pp. 443-452). Cambridge: Cambridge University Press.

Vollmer, H.J. (1981). Why are we interested in general language proficiency? In C. J. Alderson, & A. Hughes. (Eds.), *Issues in language testing,* (pp. 152-176). London: The British Council.

Vollmer, H.J. & Sang, F. (1983). Competing hypotheses about second language ability: a plea for caution. In Oller, J.W.Jr (Ed.), *Issues in language testing research* (pp. 29-79). Rowley, Mass.: Newbury House Publishers.

Wall, D. (1997). Impact and washback in language testing. In C. Clapham & D. Corson, (Eds.), *Encyclopaedia of language and education.* Volume 7: Language testing and assessment (pp. 291-302). Dordrecht: Kluwer Academic.

Weber, R.P. (1990). *Basic Content Analysis* (2nd ed.). Newbury Park, California: SAGE Publishers, Inc.

Weir, C.J. (1983). The Associated Examining Board's Test of English for Academic Purposes: an exercise in content validation events. In A. Hughes, & D. Porter, *Current development in language testing* (pp. 147-153). London: Academic Press.

Weir, C.J. (1987). English Language Testing Service. In C. J. Alderson, K.J. Krahnke, & C.W. Stansfield (Eds.), *Reviews of English language proficiency tests* (pp. 28-30). Washington D.C.: TESOL.

Widdowson, H.G. (1978). *Teaching language as communication.* Oxford: Oxford University Press.

Widdowson, H.G. (1983). *Learning purpose and language use.* London: Oxford University Press.

Wilson, K.M. (1982). *A comparative analysis of TOEFL examinee characteristics, 1977-1979. TOEFL Research Reports 11.* Princeton, NJ: Educational Testing Service.

Wood, S. (Ed.). (1998). *IELTS research reports 1998, V1.* Canberra, Australia: IELTS Australia Pty Limited.

Appendices

Appendix 1: Background questionnaire

Adapted from the questionnaire provided by Bachman (1995)

(Original English version)

This questionnaire is designed to provide us with information of interest for research purposes. Your answers to these questions will be kept strictly confidential. Please answer each question as accurately as you can. Thank you for your cooperation.

Surname:
Forename:
Candidate Number: Sex: Male Female Age
Highest qualification: Course of study:
Native language:

.....

1 What is your current educational status?
(A) enrolled in a secondary school
(B) enrolled part-time in a college, university or other institution of higher education
(C) enrolled full-time in a college, university or other institution of higher education
(D) enrolled in a language institute or English course given where I work
(E) not currently enrolled as a student

2 Have you ever or are you currently taking a course to prepare for the ...
(A) TOEFL
(B) FCE
(C) CPE
(D) IELTS

3 At what age did you begin to learn or use English?
(A) 1-5 years
(B) 6-9 years
(C) 10-13 years
(D) 14-17 years
(E) 18 or more years

4 How many years have you studied English in school or in a language institute?
(A) less than 1 year
(B) 1-3 years
(C) 4-6 years
(D) 7-9 years
(E) 10 or more years

5 How old were you when you first began to study English in school or in a language institute?
(A) 1-5 years
(B) 6-9 years
(C) 10-13 years
(D) 14-17 years
(E) 18 or more years

How many hours per week did you spend in English class....
6 ... in elementary school?
(A) none
(B) 1-3 hours
(C) 4-6 hours
(D) 7-9 hours
(E) 10 or more hours

7 in secondary school?
(A) none
(B) 1-3 hours
(C) 4-6 hours
(D) 7-9 hours
(E) 10 or more hours

8 ... in college and/or language institute?
(A) none
(B) 1-3 hours
(C) 4-6 hours
(D) 7-9 hours
(E) 10 or more hours

9 How many hours are you currently spending in English class?
(A) none
(B) 1-3 hours
(C) 4-6 hours
(D) 7-9 hours
(E) 10 or more hours

10 Have you used English at home with your family or friends outside the classroom?
(A) Yes
(B) No

IF YES, ANSWER QUESTIONS 11-14. IF NO, GO TO QUESTION 15

11 How many years have you used English at home?
(A) none
(B) less than 1 year
(C) 1-3 years
(D) 4-6 years
(E) 7 or more years

12 How old were you when you first started to use English at home?
(A) 1-5 years
(B) 6-9 years
(C) 10-13 years
(D) 14-17 years
(E) 18 or more years

13 How much did you use English at home?
(A) not at all
(B) a little
(C) about half the time
(D) most of the time
(E) all the time

14 How much do you currently use English at home?
(A) not at all
(B) a little
(C) about half the time
(D) most of the time
(E) all the time

15 Have you ever learned or used English while visiting or living in an English-speaking country?
(A) yes
(B) no

IF YES, ANSWER QUESTIONS 16-23. IF NO, GO TO QUESTION 27

16 How old were you when you first went to an English-speaking country?
(A) 1-5 years
(B) 6-9 years
(C) 10-13 years
(D) 14-17 years
(E) 18 or more years

17 How many years in total did you spend there?
(A) 1 year or less
(B) 2-4 years
(C) 5-7 years
(D) 8-10 years
(E) 11 or more years

18 Did you study English in school or in a language institute in the English-speaking country?
(A) yes (B) no

19 How many years did you study English in school or in a language institute in an English-speaking country?
(A) less than 1 year
(B) 1-3 years
(C) 4-6 years
(D) 7-9 years
(E) 10 or more years

20 How old were you when you first began to study English in school or in a language institute in an English-speaking country?
(A) 1-5 years
(B) 6-9 years
(C) 10-13 years
(D) 14-17 years
(E) 18 or more years

How many hours per week did you spend in English class ...

21 ... in elementary school?
(A) none
(B) 1-3 hours
(C) 4-6 hours
(D) 7-9 hours
(E) 10 or more hours

22 ... in secondary school?
(A) none
(B) 1-3 hours
(C) 4-6 hours
(D) 7-9 hours
(E) 10 or more hours

23 ... in college and/or language institute?
(A) none
(B) 1-3 hours
(C) 4-6 hours
(D) 7-9 hours
(E) 10 or more hours

24 Did you use English at home with family or friends in the English-speaking country?
(A) yes
(B) no

IF YES, ANSWER QUESTIONS 25-26. IF NO, GO TO QUESTION 27
25 How many years did you use English at home with family or friends in the English-speaking country?
(A) none
(B) less than 1 year
(C) 1-3 years
(D) 4-6 years
(E) 18 or more years

26 How old were you when you first started to use English at home in the English-speaking country?
(A) 1-5 years
(B) 6-9 years
(C) 10-13 years
(D) 14-17 years
(E) 18 or more years

The following statements describe some possible reasons why people learn English. For each statement, darken the circle on your answer sheet by the response which is most appropriate for you i.e., the response which best describes why you want to learn English.

A = strongly agree (SA)
B = agree (A)
C = disagree (D)
D = strongly disagree (SD)

	SAA D SD
27 I want to be able to read English books, articles, etc. in my field of specialisation	A B C D
28 I want to think and behave as people from America or Great Britain do	A B C D
29 I want to be able to write professional reports in English	A B C D
30 I want to fit into an English-speaking community	A B C D
31 I enjoy learning English as a second or foreign language	A B C D
32 As an international language, English is useful for communicating with other people whose native language I do not know	A B C D
33 I want to understand American or British people and culture	A B C D
34 English is important for career purposes	A B C D

Appendix 2:

IELTS Specimen Materials, Module C

Keys, & Answer Sheets

Prepared by:

University of Cambridge Local Examinations Syndicate

British Council

International Development Programme

(1992)

INTERNATIONAL DEVELOPMENT PROGRAM

Test Centre: .....

Name: .....

Number: .....

Date: .....

INTERNATIONAL ENGLISH LANGUAGE TESTING SYSTEM

LISTENING

SPECIMEN VERSION

TIME ALLOWED: 30 MINUTES
NUMBER OF QUESTIONS: 39

Instructions

You will hear a number of different recordings and you will have to answer questions on what you hear.

There will be time for you to read the instructions and questions, and you will have a chance to check your work.

All the recordings will be played once only.

The test is in four sections.

Now turn to Section 1.

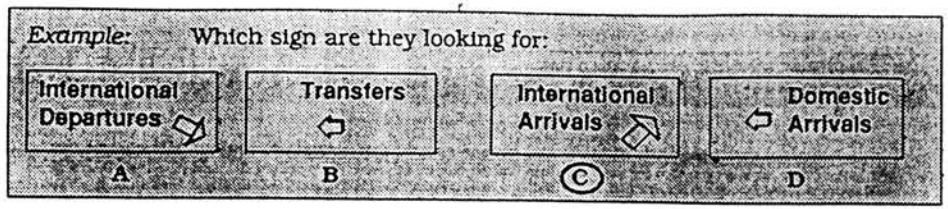
Total Marks:

Band:

University of Cambridge Local Examinations Syndicate
British Council
International Development Program

Circle the appropriate letter.

Example: Which sign are they looking for:



1. Who are they meeting?

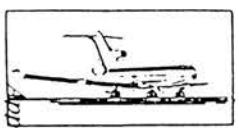

A

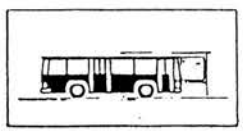

B


C


D

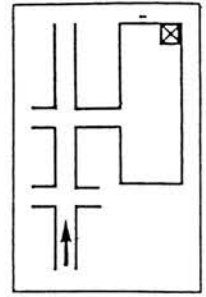
2. How are they going to travel next?

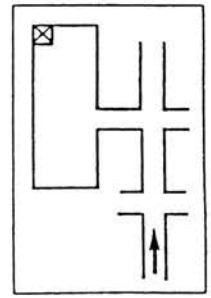

A

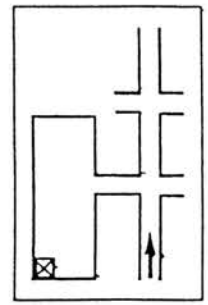

B

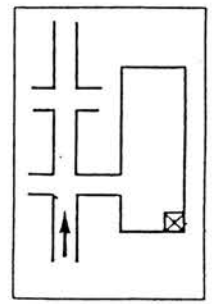
C


D

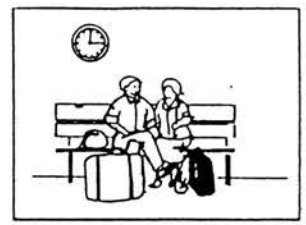

A

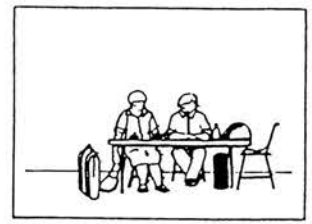

B


C

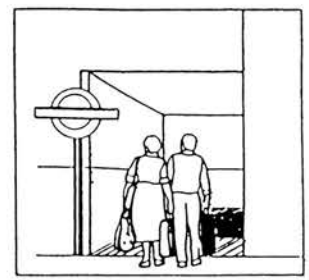

D

4. Where will they wait?


A


B

C


D

Fill in the gaps numbered 5 - 10.

Central Hotel

Registration Form

Surname Parkinson

First Name (5) .....

Nationality (6) .....

Home Address 30 Wentworth Avenue

(7) .....

(8) .....

New South Wales

Date of Arrival 1 September

Date of Departure (9) .....

Signature .....

Date 1 September

Room Number (10) .....

Fill in the gaps numbered 11 - 13.

Train leaves at (11) .....

Catch the (12) ..... (bus).

Get off at the (13) .....

Fill in the gaps numbered 14 - 21.

Flight ...(14)... bound for ...(15)... has crashed at Manchester Airport. There were ...(16)... passengers and ...(17)... crew on board. The plane took off just before 9.30 this morning. Although the pilot reported nothing wrong, the plane crashed just after ...(18)... . It appears that it crashed into ...(19)... near the airport, and there was an explosion. The ...(20)... of the plane caught fire. So far some fifteen survivors have been taken to hospital but some passengers are known to have died. People wanting information should telephone Manchester ...(21)... .

Your answers

14. .....
15. .....
16. .....
17. .....
18. .....
19. .....
20. .....
21. .....

*Show whether, **according to what you hear**, the statements are correct or not by writing **T** if the statement is true, **F** if it is false or **NI** if no information is given. Write your answers in the box next to each statement.*

Example — *Answer*

The library is at the top of the stairs on the right. — T

22. In term time the library is open until 8 p.m. on Fridays. ☐
23. If you want to study in the library you must be a member of the library. ☐
24. You have to pay a fee to join the library. ☐
25. You can sometimes keep a book for six weeks. ☐
26. Science books are on the upper floor. ☐
27. The subject index contains cards arranged alphabetically according to the title of the book. ☐
28. Exam papers are on the upper floor. ☐
29. Exam papers can be taken home if you show your identity card. ☐
30. You can borrow some foreign language newspapers. ☐

SECTION 4 *Questions 31 - 39*

Write a word or a short phrase in the space provided.

Your answers

Example
Where did she do her degree? *Open University*

31. Where was she working when she decided to do the Open University course? .....

32. Which two subjects did she study? .....

33. What surprised her about the whole course? .....

34. What surprised her about the first few months of the course? .....

35. Why did she cope well with the first few months? .....

36. Which event renewed her enthusiasm for the course? .....

37. At what times are Open University programmes broadcast? .....

38. What did she buy to make studying more convenient? .....

39. Who paid her fees? .....

UNIVERSITY OF CAMBRIDGE LOCAL EXAMINATIONS SYNDICATE
BRITISH COUNCIL
INTERNATIONAL DEVELOPMENT PROGRAM

INTERNATIONAL ENGLISH LANGUAGE TESTING SYSTEM

MODULE C

READING

SPECIMEN VERSION

TIME ALLOWED: 55 MINUTES
NUMBER OF QUESTIONS: 36

Instructions

Please note that you are **NOT** allowed to write in this booklet.
All answers must be written on the answer sheet.

In this booklet you will find 3 reading passages. Each reading passage is accompanied by some questions to test your ability to read English. Some of the questions come before the relevant reading passage, and some come after.

Start at the beginning of the booklet and work through it. There is a suggested time for each set of questions. If you cannot do a particular part of the test in the suggested time, leave that part and go on to the next. You may return to it later if you wish.

Please read the instructions for each part of the test carefully.

Passage 1 below.

QUALITY CIRCLES

In an increasingly competitive market, companies are paying more attention to the quality of the products or services they provide. Some organisations are seeking ways of encouraging employee involvement in the company, increasing job satisfaction and motivating staff.

Quality circle programmes are being increasingly adopted by British management as one method of meeting these objectives. Quality circles are small groups of employees from the same work area who meet regularly in order to analyse their work-related problems, and to find solutions to them. They make recommendations to management and, if these are approved, are usually involved in implementing their proposals.

Once dismissed as a 'flavour of the month' management technique or as a Japanese concept that would never catch on in Britain, quality circles now exist in over 400 companies throughout the UK. And they are no longer confined to the manufacturing sector - service organisations have also begun to introduce them.

In this Study we examine how a quality circle programme is initiated and organised. Practical aspects (such as whether circle meetings are held in company time, communications with non-circle members) are considered, and we describe the quality circle problem-solving process. We also look at trade union views. The Record Section (Appendix 1) contains five case studies of companies which have introduced quality circles: three from the manufacturing sector - Black & Decker, Jaguar and Tioxide; and two from the service sector - Honeywell Control Systems and the London Life Association.

The National Society of Quality Circles describes a quality circle as a 'group of four to twelve people coming from the same work area, performing similar work, who voluntarily meet on a regular basis in order to analyse and solve their own work-related problems. The circle presents solutions to management and is usually involved in implementing them'. Unlike other problem-solving groups such as task forces, quality circles have a permanent existence (i.e. they have regular meetings) and they have considerable autonomy as to what problems or ideas they consider.

Where Did They Come From?

Quality circles have gained a reputation as an important ingredient of the modern Japanese economic miracle. But the quality control

It was in Japan, however, in the early 1960s that the ideas of Deeming and Juran were first put into practice, with Professor Ishikawa, from Tokyo, becoming known as the 'father' of quality circles. Then, in the mid-1970s, the quality circle concept recrossed the Pacific when such American companies as Lockheed, General Motors and Honeywell led the first experiments in the Western world.

By the late 1970s some British companies were starting to take notice of the quality circle phenomenon. Rolls-Royce Aero Division lays claim to being the British pioneer in 1978. By 1981 it was estimated that around 100 companies in the UK had introduced quality circles, and the number today is thought to be at least 400.

THE REASONS WHY

Quality circles are only likely to succeed in companies where the programme has the strongest backing of senior management. So what do companies see as the benefits of a quality circle programme? The aims would certainly include some or perhaps all of the following:

1 to produce a higher quality product or service (increasingly important in highly competitive markets);

2 to create a greater degree of quality consciousness among all employees;

3 to reduce costs;

4 to involve employees more in the organisation;

5 to increase employees' job satisfaction;

6 to develop and motivate staff;

7 to achieve better two-way communications;

8 to encourage a more open, participative management style;

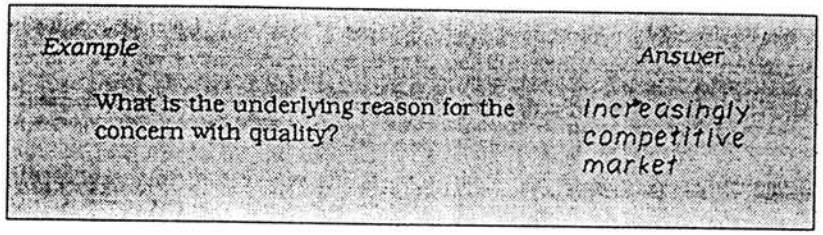
9 to encourage team-working;

10 to help create a better quality of working life.

[Turn over

Use Reading Passage 1, *Quality Circles*, on pages 30 and 31, to answer the following questions with a word or phrase **from the reading passage**. Write your answers in boxes 1 -13 on your answer sheet.

The first one has been done as an example.



Questions 1 - 4

What are the **FOUR** stages in quality circle programmes in which employees are involved?

Write your answers in boxes 1-4 on your answer sheet.

Questions 5 - 6

5. If you wanted to examine examples of quality circles, where would you look for information?

6. Was the idea of quality circles immediately popular when it was first introduced to Britain?

Questions 7 - 8

According to the text, quality circles are different from task forces in **TWO** ways. What are they?

Questions 9 - 11

9. Where did the idea behind quality circles first come from?

10. When were quality circles first introduced in the United Kingdom?

11. What, according to the text, is an important ingredient for the success of quality circles?

Questions 12 - 13

Of the aims of quality circles which are mentioned in the section 'The Reasons Why ', **TWO** seem to form a separate category from the others. Which ones are they?

Write the appropriate numbers in boxes 12 and 13 on your answer sheet.

[Turn over

THE PURPOSES OF CONTINUING EDUCATION

What are the purposes of continuing career education? Some people think about the matter very simplistically, but most people realize that several goals must be sought simultaneously and that the process of doing this is difficult. We may identify at least eight purposes of continuing professional education.

The first of these is to keep up with the new knowledge required to perform responsibly in the chosen career. Just think, for example, how much has happened in the various professions in ten years. At the start of the decade, we had just learned how to keep a man in orbit around the earth; at its end, we had sent many men to the moon. At the same time, physicians and surgeons learned to save the lives of people who in previous times would have died. Patterns of resource allocation applied by public administrators are changing as a result of new doctrines. And pharmacists - perhaps more than the members of any other profession - have had to keep up with a changing scene, as new products and new therapies succeeded one another or enlarged the range of treatments available.

Practical careers rest upon theoretical bodies of knowledge. The health professions, including pharmacy, depend upon anatomy, physiology, pathology and biochemistry; agriculture is based upon soil chemistry, meteorology and plant pathology; and social work finds its roots in sociology, psychiatry and political science. The professional does not need to become expert in these underlying bodies of knowledge, but he or she does need to learn about developments in these fields. Keeping up with changes in the relevant basic disciplines, therefore, constitutes the next aim.

The third aim is to master new ideas about the nature of the profession itself. A recent study of dentistry noted that "25 or 30 years ago, dental practice was limited to relieving pain. Today it is concerned with the overall management of oral, facial and speech defects, and with the structures and tissues of the mouth as they relate to the total health of the individual." In other occupations, an equally profound change has occured and the modern practitioner who does not understand that fact is obsolete.

The fourth purpose of career-oriented continuing education is to prepare the individual for changes in his or her career. A person may move in one of many directions, such as from generalist to specialist, from one speciality to another, upward in a hierarchy from a lesser to a more responsible job of the same sort, or into a completely new career.

In every kind of career, people go stale after a while. Consequently, maintaining freshness of outlook on one's work, so that detail is not neglected, is also important. This forms the next aim. Perhaps education is not the only way to achieve this, but it can at least be aided by educational means: for example, by supervisory training, by self-appraisal and by peer review. It can also be achieved by putting oneself in a new work or study situation which demands attention to detail.

Continuing to grow as a well-rounded person is essential as well. The mind should never be fully engaged in the practice of work, to the exclusion of all other matters. One needs to withdraw from that practice occasionally so that one can be stimulated by contemplating theory or by seeking understanding and skill in different aspects of life. Otherwise, one becomes dogmatic, not only about a single speciality but about all aspects of life. This, then, is another aim of continuing professional education.

The seventh objective is to retain the power to learn. This objective is an important adjunct to all the others. The skills of mastering knowledge are like other skills: they are learned by practice, they are lost if they are not used regularly, and they can later be regained only with difficulty.

Finally, it is essential to ensure that the individual performs effectively the role imposed by membership in a profession. The professional must learn how to take collective responsibility, to make right choices on issues, to improve and extend the delivery of service, to collaborate with other professions, and to help monitor the actions of fellow professionals.

These eight purposes suggest that a full fledged program of continuing professional education cannot be developed either easily or rapidly.

[Turn over

You are advised to spend about 20 minutes on ***questions 14-28*** *which are based on Reading Passage 2, The Purposes of Continuing Education, on the opposite page.*

Questions 14 - 20

Reading Passage 2, *The Purposes of Continuing Education*, discusses eight different aims of continuing career education. From the list of fourteen titles **(A-N)** below, choose the most suitable title for each of these aims and write the appropriate letter in boxes 14-20 on your answer sheet.

The first one has been done as an example.

NB There are more titles than aims so you will not use all of them.

Aims of continuing education

- **A** Staying fresh
- **B** Being prepared to carry out one's professional responsibilities
- **C** Enlarging the range of treatments available
- **D** Collaborating with other professions
- **E** Staying in touch with theoretical knowledge
- **F** Becoming expert in theoretical knowledge
- **G** Preparing to earn more money
- **H** Staying in touch with developments in the nature of the profession
- **I** Developing as an individual
- **J** Using knowledge regularly
- **K** Maintaining the ability to acquire new knowledge
- **L** Contemplating theory
- **M** Preparing for career change
- **N** Staying in touch with developments in professional knowledge

Example	*Answer*
Aim 1	**N**
Aim 2	**(14)**
Aim 3	**(15)**
Aim 4	**(16)**
Aim 5	**(17)**
Aim 6	**(18)**
Aim 7	**(19)**
Aim 8	**(20)**

The passage opposite is a summary of *The Purposes of Continuing Education* on page 35.

From the list of phrases below **(A-L)**, choose the most suitable phrase to complete the summary and write the appropriate letter in boxes 21-28 on your answer sheet.

The first one has been done as an example.

NB There are more phrases than gaps so you will not use all of them.

List of Phrases

A an example

B theoretical knowledge

C a thorough program

D educational in nature

E modern technology

F dismissed

G out of date

H continuing education

I career

J personal growth

K awareness

L professional knowledge

Continuing professional education has at least eight different purposes. Achieving a balance between these is not easy.

Example	*Answer*
It is important for a member of a profession to keep his or her	....L....

up to date. The need for this can be seen from space technology to medicine, from public administration to pharmacy. In all of these fields, rapid changes are taking place. Keeping one's ...**(21)**... up to date is also important. Although professionals do not need to understand in detail the developments which are taking place in the theories which underlie their profession, it is important that they should have some ...**(22)**... of what is happening. Keeping in touch with developments in the fundamental character of a profession constitutes the third objective of continuing education. In some occupations, considerable changes have occurred and a professional who is not able to adapt to such changes will soon be ...**(23)**... .

Professionals frequently need to make moves in their careers. Taking on a position with greater responsibility is ...**(24)**... . Continuing education can prepare people to make such moves.

The next aim is concerned with enabling people to stay professionally fresh. There are several ways of achieving this aim, some of which are ...**(25)**... and some of which are not. A related aim is to encourage ...**(26)**... . There are dangers in becoming too engrossed in one's work and, from time to time, the professional needs to stand back from routine work. Continuing professional education can facilitate this process.

The seventh objective is to help the individual to continue to be able to learn. This is important because, if the skill of acquiring new knowledge is not practised regularly, it may be lost. Clearly, this particular aim of ...**(27)**... is related to all the others which we have been discussing. The final aim concerns the role which a professional person has to play in society at large. The individual has to learn to fulfil all those social functions which are expected of members of the profession.

It is not a simple task to design ...**(28)**... of continuing professional education. Furthermore, such a program cannot be carried out in a short space of time.

|Turn over

You are advised to spend about 20 minutes on this question. Write your answer based on Reading Passage 3, below.

ACCESS TO HIGHER EDUCATION

1. The Record

1.1 Since 1979 the number of full-time home students in higher education in Great Britain has risen by more than 85,000 - almost three times the increase achieved during the 1970s. The size of the 18-19 year-old age group peaked in 1982; the continuing increase in student numbers reflects higher rates of participation in higher education both by 18-19 year-olds - for whom the Age Participation Index increased from 12.4 in 1979 to an estimated 14.2 in 1986 - and by mature entrants (aged 21 and over) whose numbers have grown by a quarter since 1979. The increases in participation have been particularly marked for women, who now account for about 44% of full-time students in higher education compared with less than 40% seven years ago.

1.2 Virtually all of this major increase in full-time student numbers has taken place in the polytechnics and colleges sector of higher education, more than making good the reduction in this sector in the late 1970s. The polytechnics and colleges have also accommodated over three-quarters of the 72,000 (27%) increase since 1979 in part-time student numbers. In both sectors the proportion of students on full-time science-related courses has increased - from 50% to 53% for universities and from 36% to 41% for the polytechnics and colleges.

1.3 Comparisons with higher education participation rates in other countries are not straightforward. There are differences in structures, course lengths, wastage rates and ways of counting part-time students. The best like-for-like comparison is probably of the proportion of the relevant age groups gaining degrees and higher diplomas. On that basis Britain is on a par with attainments in France and ahead of the rest of the European Community, though achievements here are not as good as those in Japan and the USA. Moreover, such comparisons take no account of quality - for which the reputation of higher education in this country is high.

2. The Future

2.1 In November 1986 the Government published *Projections of Demand for Higher Education in Great Britain 1986-2000*, which displays two projections of future student numbers. Projection P, the lower of the two, is based on the assumptions that the numbers of young people entering full-time higher education will remain a constant proportion of those gaining the traditional qualifications for entry (two or more GCE A levels of three or more Scottish Highers), and that the entry rates for mature students will also remain constant. For the higher Projection Q, it is assumed that those proportions will increase - particularly amongst young women - in part to reflect the success of the Government's policies for schools and non-advanced further education.

of Education and Science have confirmed that young people whose parents hold higher-level qualifications are proportionately more likely to apply for and obtain places in higher education and that this factor is additional to the differential demand for higher education by social class which is already allowed for in the projections. This may have a significant influence on demand for higher education in the 1990s when many of the children of those who benefited from the big increase in higher education opportunities in the 1960s will reach age 18. If so, the API for young entrants is likely to be nearer to that underlying Projection Q.

2.3 Both Projections P and Q assume that higher proportions of young people will gain the traditional qualifications for entry to higher education as a result of girls achieving parity with boys' A level attainments, and changes in the social/occupational mix of the population. Additionally, Projection Q assumes a further increase in the potential number of traditionally qualified entrants to higher education, dependent primarily on the success of the policies presented in the Government's *Better Schools* document which imply that a significantly larger proportion of pupils at age 16 will reach the standards necessary to continue on to and to succeed in A level courses. The Government believes that by the end of the century improvements in the schools will be such that the proportion of young people in Great Britain as a whole gaining two or more A levels, or three or more Scottish Highers, will have reached 20% - the proportion already attained in Scotland.

2.4 Of potentially greater impact, however, is the assumption underlying Projection Q that there will be a signficant increase in the proportion of qualified young people who enter higher education. This will depend very largely on continuing growth in the demand for higher education from young women, alongside increases in the proportion of higher education entrants with vocational and technical qualifications, for example those validated by the Business and Technician Education Council (BTEC) and the Scottish Vocational Education Council (SCOTVEC). The Government is committed to ensuring that girls have equal opportunities, throughout the education system, to develop their talents to the full. The development of the Technical and Vocational Education Initiative (TVEI) and the two-year Youth Training Scheme (YTS), and the streamlining associated with the National Council for Vocational Qualifications (NCVQ) and the Scottish Action Plan, should increase the proportion of young people gaining vocational qualifications and is likely to motivate more of them to see entry to higher education.

2.5 The Government remains committed to the modified form of the Robbins Principle. Places should be available for all who have the necessary intellectual competence, motivation and maturity to benefit from higher education and who wish to do so. Planning of higher education will need to take account, inter alia, of regular monitoring of actual demand for places and of the effects of the Government's policies to improve performance in schools and non-advanced further education on the numbers of potential entrants to higher education.

[Turn over

Choose which of the alternatives **A, B, C or D** is the correct answer and write the appropriate letter in boxes 29-32 on your answer sheet.

The first one has been done as an example.

Example *Answer*

Since 1979 the number of full-time home students in higher education in Great Britain ... .B....

A has nearly trebled.
B has increased more than in the 1970s.
C has reached 85,000.
D has grown by 25%.

29. What is mean by 'both sectors' (paragraph 1.2)?

A 'Polytechnics and Universities' and 'Colleges'
B 'Polytechnics' and 'Colleges'
C 'Polytechnics' and 'Universities'
D 'Polytechnics and Colleges' and 'Universities'

30. It is not easy to compare participation rates among different countries, because ...

A there are too many differences in the systems.

B pupils enter higher education at different ages in different countries.

C comparisons cannot be based on reputation.

D Britain and France are ahead of the rest of the European Community.

A there will be no increase in numbers of students.

B the number of women students will continue to rise.

C its educational policies to increase student numbers will succeed.

D progress in Scotland may be more rapid.

32. According to Projection Q, an increasing number of young people ...

A will definitely be more qualified.

B will have the intellectual competence to benefit from higher education.

C will have only technical or vocational qualifications.

D are probably going to be more qualified.

Questions 33 - 36

The **Figures A-D** which follow on pages 42 and 43 can be matched with the appropriate paragraphs of the reading passage. Note that the figures may refer to more than one paragraph, and that one paragraph may be referred to in more than one figure.

Choose the correct figure(s) **(A-D)** for each of the paragraphs listed below and write your answers in boxes 33-36 on your answer sheet.

The first one has been done as an example.

Example *Answer*
Paragraph 1.1 **Figures A and D**

33. Paragraph 1.2

34. Paragraph 1.3

35. Paragraph 2.1

36. Paragraph 2.5

FIGURE A

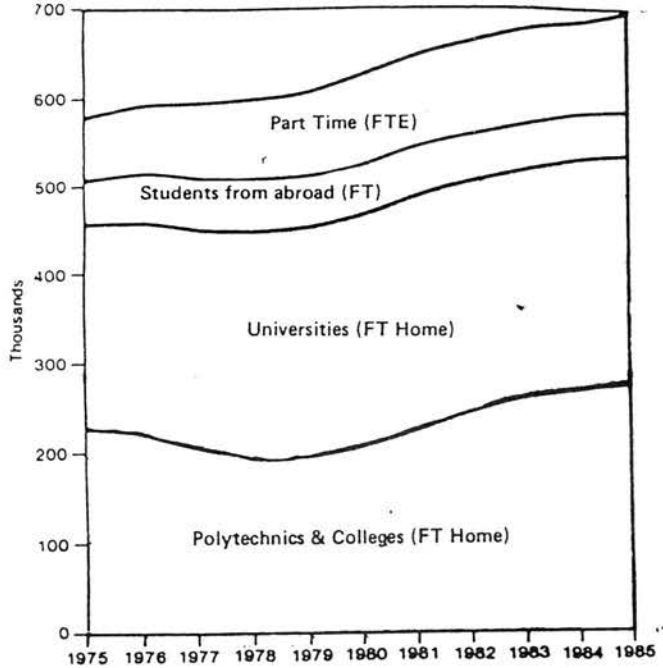
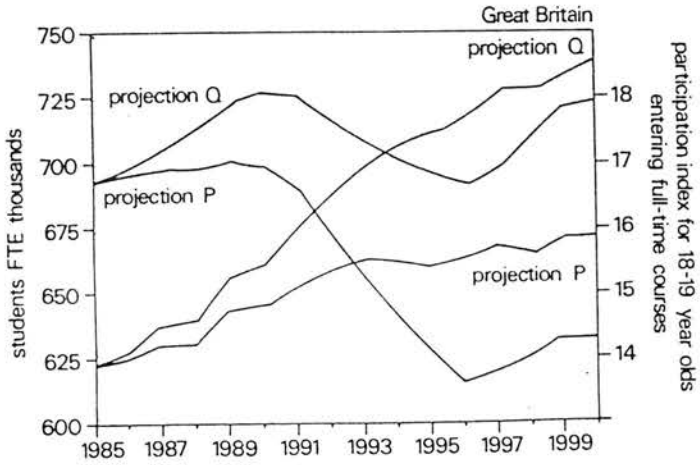


FIGURE B



Students from abroad (FT)

Qualifications awarded per 100 of relevant age group

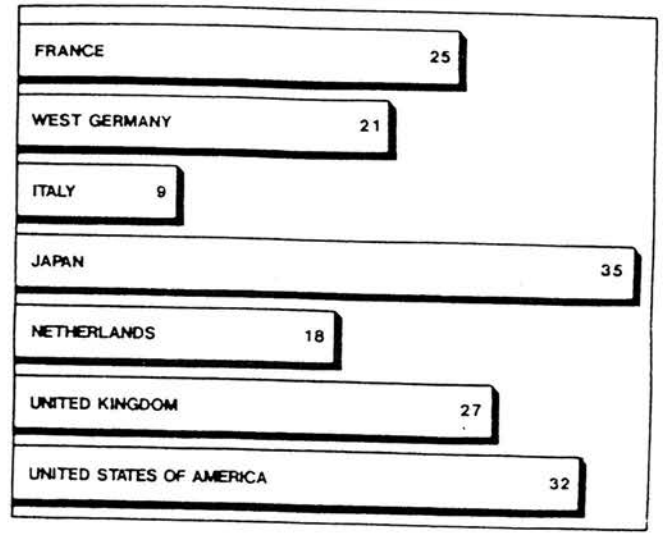
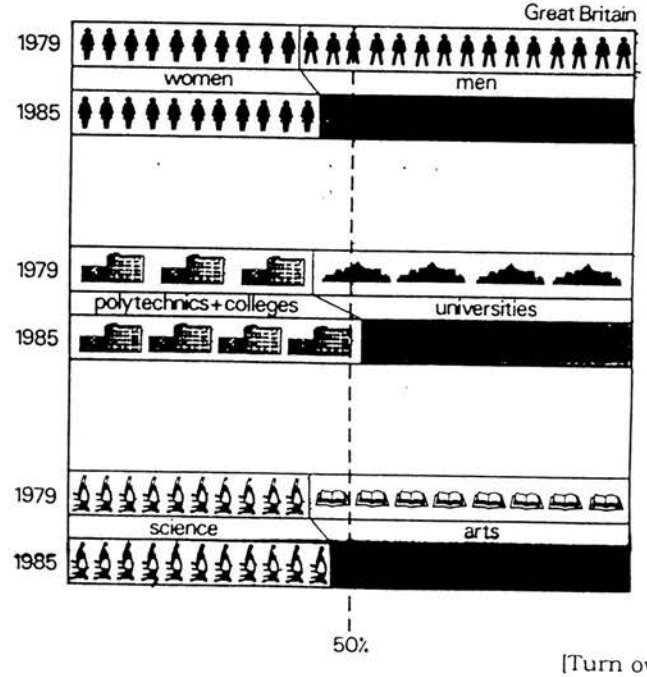


FIGURE D



[Turn over

You should spend no more than 30 minutes on this task.

TASK:
Write an essay for a university teacher on the following topic:

What are the potential benefits, to both the individual and the community, of continuing education?

You should write at least 150 words.

You may use ideas from Reading Passage 2 (page 35). You should also use your own ideas, knowledge and experience and support your arguments by examples and by relevant evidence.

NB You should **NOT** copy word for word from the Reading Passage.

[Turn over

WRITING TASK 1

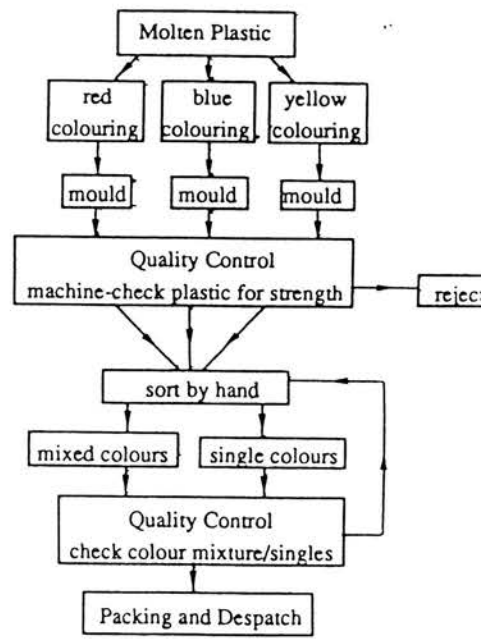
You should spend no more than 15 minutes on this task.

The diagram below illustrates some of the processes in the production of coloured paper-clips at a small factory.

TASK:
Describe the process of producing coloured plastic paper-clips, based on the information in the diagram.

You should write at least 100 words.

Glossary
molten plastic: liquid plastic
mould: hollow form or shape



Module C

Question		
	1.	analyse
	2.	solve
	3.	make recommendations
	4.	implementing (For questions 1-4 the answers can be opposite any of the 4 question numbers. As long as "analyse", "solve", "make recommendations" and "implementing" are all listed you get 4 marks. Even if you put "analyse" and "solve" on the same line, rather than on two lines, you get two marks for them.)
	5.	record section
	6.	no
	7.	permanent existence / regular meetings
	8.	autonomy
	9.	the United States / U.S.
	10.	1978 / the late 1970s
	11.	the strongest backing of senior management
	12.	to produce a higher quality product
	13.	to reduce costs
	14.	E
	15.	H
	16.	M
	17.	A
	18.	I
	19.	K
	20.	B
	21.	B
	22.	K
	23.	G
	24.	A
	25.	D
	26.	J
	27.	H
	28.	C
	29.	D
	30.	A
	31.	C
	32.	D
	33.	Figs A and D
	34.	Fig C
	35.	Fig B
	36.	No Figs

Question		
	1.	C
	2.	D
	3.	B
	4.	A
	5.	Steve
	6.	Australian
	7.	Pymble (the spelling must be correct)
	8.	Sydney 2173 (Note: for Questions 7 and 8, as long as the three pieces of information are correct - Pymble, Sydney, and 2173 - it does not matter which lines you write them on.)
	9.	2 September
	10.	249
	11.	10.45
	12.	108
	13.	(new) shopping centre
	14.	CA261
	15.	Berlin
	16.	315
	17.	12
	18.	Take-off
	19.	Trees ('Tree' is not acceptable)
	20.	Front
	21.	28723
	22.	F
	23.	F
	24.	NI
	25.	T
	26.	T
	27.	F
	28.	T
	29.	F
	30.	F
	31.	School
	32.	History, Education (both must be given)
	33.	Practical/relevant to the classroom
	34.	Not difficult/easy/coped well
	35.	Worked hard/energy/enthusiastic/keen/commitment
	36.	First summer school/meeting other students
	37.	12.00 at night (p.m.) and 6.00 in the morning (a.m.) (both times must be given)
	38.	video (recorder)
	39.	she did herself

University of Cambridge Local Examinations Syndicate | The British Council | Development Program

WRITING ANSWER SHEET

Candidate Name: ..... Candidate Number: .....

Centre Name: ..... Date: .....

Module: ..... Version: ..... Form: .....

For Marker's Use Only

Task 1 BAND ☐

Task 2 BAND ☐ FINAL BAND ☐

TE&S 91

University of Cambridge Local Examinations Syndicate | The British Council | Development Program

Candidate Name: _____ Candidate Number: _____

Centre Name: _____ Date: _____

Module: _____ Version: _____

ANSWER SHEET

1		26	
2		27	
3		28	
4		29	
5		30	
6		31	
7		32	
8		33	
9		34	
10		35	
11		36	
12		37	
13		38	
14		39	
15		40	
16		41	
17		42	
18		43	
19		44	
20		45	
21			
22			
23			
24			
25			

For centre use

Subtest: _____

Raw Score: _____

Band: _____

For UCLES use only Reading _____ Writing _____ Listening _____ Speaking _____

TE&S 91

Appendix 3:

TOEFL SAMPLE,

Answer sheet, Keys,

Score Conversion Table

This is a test of your ability to use the English language. It contains three sections. Each section has more than one part. Each section or part of the test begins with a set of specific directions and sample questions. Be sure you understand the directions before you begin to work on each section.

The supervisor will tell you when to start and stop each section. You should work quickly but carefully. Do not spend too much time on any one question. If you finish a section early, you may review your answers on that section only. You may not go on to the next section, and you may not go back to a section you have already worked on.

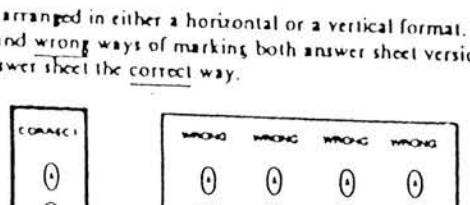
Some of the questions are more difficult than others, but try to answer every one. Your score will be based on the number of questions you answer correctly. If you are not sure of the answer to a question, make the best guess you can. It is to your advantage to answer every question, even if you have to guess.

You must mark all of your answers on the separate answer sheet. Do not mark your answer in the test book. When you mark your answers on your answer sheet, you must:

- Use a medium-soft (#2 or HB) black lead pencil.
- Mark the space that corresponds to the answer you choose for each question. Also, make sure you mark your answer in the row with the same number as the number of the question you are answering. You may not make any corrections after time is called.
- Mark only one answer to each question.
- Carefully and completely fill each intended oval with a dark mark so you cannot see the letter inside the oval; light or partial marks may not be read properly by the scoring machine.
- Erase all stray marks completely. If you decide to change an answer, completely erase your old answer and then mark your new answer.

The ovals on the answer sheet are arranged in either a horizontal or a vertical format. The examples below show the correct and wrong ways of marking both answer sheet versions. Be sure to fill in the ovals on your answer sheet the correct way.

Vertical Answer Spaces



Horizontal Answer Spaces

MCHE TOEFL Sample Test

Version A

Test Centre:

Date:

DO NOT WRITE ON THIS BOOKLET

SECTION I

LISTENING COMPREHENSION

In this section of the test, you will have an opportunity to demonstrate your ability to understand spoken English. There are three parts to this section, with special directions for each part. Do not read ahead or turn the pages while the directions are being read. Do not take notes or write in your test book at any time.

Part A

Directions: For each question in Part A, you will hear a short sentence. Each sentence will be spoken just once. The sentences you hear will not be written out for you.

After you hear each sentence, read the four choices in your test book, marked (A), (B), (C), and (D), and decide which one is closest in meaning to the sentence you heard. Then, on your answer sheet, find the number of the question and fill in the space that corresponds to the letter of the answer you have chosen. Fill in the space completely so that the letter inside the oval cannot be seen.

Listen to an example.

On the recording, you hear:

Sample Answer

Ⓐ Ⓑ ● Ⓓ

In your test book, you read:

(A) Please lock your room when you leave.
(B) Turn the key to the left to enter your room.
(C) Please return your room key before leaving.
(D) You must leave your room by four o'clock.

The woman said, "Please turn in the key to your room before you leave." Sentence (C), "Please return your room key before leaving," is closest in meaning to what the woman said. Therefore, the correct choice is (C).

Now listen to another example.

On the recording, you hear:

Sample Answer

Ⓐ ● Ⓒ Ⓓ

In your test book, you read:

(A) Will Mary be traveling tomorrow?
(B) What are Mary's plans for tomorrow?
(C) Who will be with Mary tomorrow?
(D) Does Mary have to do it tomorrow?

The man said, "What's Mary going to do tomorrow?" Sentence (B), "What are Mary's plans for tomorrow?" is closest in meaning to what the man said. Therefore, the correct choice is (B).

WAIT

1. (A) What is the topic of your term paper?
   (B) Isn't your term paper topic similar to mine?
   (C) Do you actually enjoy doing the research?
   (D) You haven't chosen a likely topic to research.

2. (A) I feel hot, but I don't have a fever.
   (B) I think I should see a doctor about my fever.
   (C) I don't feel hot, but my temperature is above normal.
   (D) I feel like turning up the temperature in this room.

3. (A) I can't get the new furnace to work.
   (B) I don't think the furniture looks good.
   (C) That new book about the future is very believable.
   (D) The furniture is more attractive than I expected.

4. (A) Michael is exceptional when it comes to solving puzzles.
   (B) Michael was the only one who had trouble solving the puzzle.
   (C) Michael was the only one who could solve the puzzle easily.
   (D) Michael easily solved all the puzzles.

5. (A) Didn't you drive here today?
   (B) I don't think this is a good driveway.
   (C) Do you want to drive or take the subway?
   (D) We should go driving today.

6. (A) He stopped eating, just for spite.
   (B) He didn't stop eating, although he wanted to.
   (C) He didn't want to stop eating.
   (D) He stopped eating because he was forced to.

7. (A) George prefers to run with a friend.
   (B) George hikes when he has time.
   (C) George tries to run farther each day.
   (D) George runs more than his partner.

8. (A) Students can request a special course.
   (B) The course will be described on the first day.
   (C) Students can sign up at the first class.
   (D) The class will have a limited number of students.

GO ON TO THE NEXT PAGE

18

22

26

9. (A) Did you see Lynn at the seminar?
   (B) Won't you apologize to Lynn?
   (C) Lynn shouldn't have said that.
   (D) Lynn will regret missing the seminar.

10. (A) He wasn't a responsible class president, was he?
    (B) Don't you believe he would be a responsible class president?
    (C) Our next class president will not be serious about his job.
    (D) He wouldn't want the responsibility of being class president.

11. (A) Peggy liked to study by herself
    (B) Peggy couldn't find the students by herself.
    (C) Peggy couldn't prove she was a student.
    (D) Peggy demonstrated her ability as a student.

12. (A) Tell me where you want me to stand.
    (B) I don't think you know what I mean.
    (C) This question is unbelievable.
    (D) I don't believe what you said.

13. (A) My aunt's book is about publishing.
    (B) It has taken my aunt a long time to finish her book.
    (C) My aunt's book will be published soon.
    (D) I've read all of my aunt's books.

14. (A) The restaurant will remain closed for remodeling.
    (B) They want the restaurant to open soon.
    (C) They have completely renovated the restaurant.
    (D) The restaurant is open until midnight.

15. (A) I asked him if he needed a ride.
    (B) I'm going to share the driving with him.
    (C) I didn't ask him for a ride since I didn't know he was driving.
    (D) I asked him for a ride as soon as I found out he was going.

16. (A) Can another person fit at this table?
    (B) Could we put another table in this room?
    (C) Is this table too big for this room?
    (D) Could one of you help me move the table?

17. (A) We should try to get there before eight o'clock.
    (B) Eight o'clock is a good time to start out.
    (C) It takes about eight hours to get there.
    (D) There isn't time to go home first.

GO ON TO THE NEXT PAGE

(A) This is what you need to find out.
(B) Tell me what I should say about the proposal.
(C) Are you recommending my proposal?
(D) Would you really like to have my opinion?

19. (A) I'd like to get out of the car.
(B) I hope you'll write to me if you can.
(C) Feel free to stop by any time.
(D) It's always best to tell the truth.

20. (A) All medicines are stocked in the infirmary storeroom.
(B) There's no doctor working in the infirmary now.
(C) Prescriptions can't be filled at the infirmary.
(D) The infirmary has ordered a supply of medicines from our stock.

GO ON TO THE NEXT PAGE

Part B

Directions: In Part B you will hear short conversations between two people. After each conversation, a third person will ask a question about what was said. You will hear each conversation and question about it only one time. After you hear a conversation and the question about it, read the four possible answers in your test book and decide which one is the best answer to the question you heard. Then, on your answer sheet, find the number of the question and fill in the space that corresponds to the letter of the answer you have chosen.

Listen to an example.

On the recording, you hear:

Sample Answer

● (B) (C) (D)

In your test book, you read:

(A) He doesn't like the painting either.
(B) He doesn't know how to paint.
(C) He doesn't have any paintings.
(D) He doesn't know what to do.

You learn from the conversation that neither the man nor the woman likes the painting. The best answer to the question "What does the man mean?" is (A), "He doesn't like the painting either." Therefore, the correct choice is (A).

WAIT

21. (A) Whether they had been at the house before.
(B) Which four friends went with her.
(C) What time they went to the house.
(D) Why they went to the house.

22. (A) The dean wants the office report.
(B) He doesn't know where the dean's office is.
(C) Perhaps the dean's office can furnish the report.
(D) Maybe the dean is in his office.

23. (A) It's excellent.
(B) The other place is far superior.
(C) It's overrated.
(D) The menu isn't very large.

24. (A) She's worried that she will make mistakes.
(B) She'd like to get started as soon as possible.
(C) It doesn't matter to her when they start.
(D) It's so far in the future that they can start anytime.

25. (A) Everyone told him to cheer up.
(B) Spending money puts him in a good mood.
(C) He had to pay a high price for his new stereo.
(D) He's very pleased with his purchase.

GO ON TO THE NEXT PAGE

26. (A) The man will work with someone else.
(B) The man must complete some paperwork.
(C) The man's application is lost for the moment.
(D) The man is not qualified for the job.

27. (A) There are too many centers already.
(B) They aren't really going to build one.
(C) He knew about the planned construction.
(D) He hasn't been to the other centers.

28. (A) She can't decide which class to take.
(B) She's having trouble getting to school.
(C) She hasn't chosen a subject for an assignment.
(D) She can't find the kind of paper she needs.

29. (A) The man hurried through the exam.
(B) The room is too warm for a sweater.
(C) The man will be late if he doesn't hurry.
(D) The man put his sweater on the wrong way.

30. (A) A retirement party.
(B) A faculty reception.
(C) A class reunion.
(D) A birthday party.

31. (A) It will probably rain tomorrow.
(B) It will rain much later in the week.
(C) He needs to buy another umbrella.
(D) The weather forecasters almost never agree.

GO ON TO THE NEXT PAGE

32. (A) He prefers shorter plays to this one.
(B) He doesn't have to go to that play.
(C) He wouldn't see the play as often as the woman had.
(D) He liked the play better the first time he saw it.

33. (A) He's unable to appear in court.
(B) He wishes he could be a better student.
(C) He plays tennis better than she does.
(D) He's not so enthusiastic about academics.

34. (A) The woman doesn't like cold weather.
(B) The snow would get dirty quickly.
(C) It wouldn't snow.
(D) All the snow would melt.

35. (A) He's already spoken to the technician.
(B) The woman should make the repairs herself.
(C) The woman should explain what needs to be repaired.
(D) The technician has already arrived.

GO ON TO THE NEXT PAGE

Part C

Directions: In this part of the test, you will hear longer conversations and talks. After each conversation or talk, you will be asked some questions. You will hear the conversations and talks and the questions about them only one time. They will not be written out for you.

After you hear a question, read the four possible answers in your test book and decide which one is the best answer to the question you heard. Then, on your answer sheet, find the number of the question and fill in the space that corresponds to the letter of the answer you have chosen. Answer the questions on the basis of what is stated or implied by the speakers in the talk or conversation.

Here is an example.

On the recording, you hear:

Now listen to a sample question.

Sample Answer

Ⓐ Ⓑ ● Ⓓ

In your test book, you read:

(A) To demonstrate the latest use of computer graphics.
(B) To discuss the possibility of an economic depression.
(C) To explain the workings of the brain.
(D) To dramatize a famous mystery story.

The best answer to the question "What is the main purpose of the program?" is (C), "To explain the workings of the brain." Therefore, the correct choice is (C).

Now listen to another sample question.

Sample Answer

Ⓐ Ⓑ Ⓒ ●

In your test book, you read:

(A) It is required of all science majors.
(B) It will never be shown again.
(C) It can help viewers improve their memory skills.
(D) It will help with course work.

The best answer to the question "Why does the speaker recommend watching the program?" is (D), "It will help with course work." Therefore, the correct choice is (D).

Remember, you are not allowed to take notes or write in your test book.

WAIT

36. (A) Summer vacation.
(B) The housing office.
(C) Resident advisers.
(D) Check-out procedures.

37. (A) At the beginning of the school year.
(B) On June 3.
(C) Near the end of the school year.
(D) After final exams.

38. (A) Register for summer school.
(B) Repair holes in room walls.
(C) Remove personal property.
(D) Call the housing office.

39. (A) Their summer addresses.
(B) Any damage to their rooms.
(C) When they plan to leave.
(D) Questions for the housing office.

40. (A) Travel expenses.
(B) A need for more printed music.
(C) An unpleasant choir member.
(D) A delay in their trip.

41. (A) Act as director.
(B) Leave for vacation.
(C) Raise money.
(D) Join the Association of Choral Directors.

42. (A) To ask for a loan.
(B) To ask for their assistance.
(C) To tell them the travel itinerary.
(D) To tell them the concert schedule.

43. (A) Fire fighting.
(B) Pest control.
(C) House construction.
(D) Plastic watches.

44. (A) It is cheaper.
(B) It is safer.
(C) It is quicker.
(D) It is available everywhere.

45. (A) To keep the heat inside.
(B) To prevent insects from escaping.
(C) To reduce the risk of fire.
(D) To keep the wood dry.

46. (A) To show that there is no danger from the treatment.
(B) To show one of the dangers of the old method.
(C) To explain one step in the new technique.
(D) To explain a compromise between old and new systems.

47. (A) He can't find his office key.
(B) He has misplaced some exams.
(C) He is unable to talk.
(D) He doesn't like his classroom.

48. (A) Mark the latest homework assignment.
(B) Put a cancellation notice on the classroom door.
(C) Make an appointment with the doctor.
(D) Return some exams to his students.

GO ON TO THE NEXT PAGE

49. (A) Teach Don's class while he's absent.
(B) Give Professor Webster the key to Don's office.
(C) Leave a message on the board in Don's classroom.
(D) Bring Don the homework that was due today.

50. (A) To put the homework on Don's desk.
(B) To leave the master key for Don.
(C) To give Don's students the next assignment.
(D) To call Don at the end of the afternoon.

THIS IS THE END OF SECTION 1, LISTENING COMPREHENSION.
STOP WORK ON SECTION 1.

DO NOT READ OR WORK ON ANY OTHER SECTION OF THE TEST.
THE SUPERVISOR WILL TELL YOU WHEN TO BEGIN WORK ON SECTION 2.

SECTION 2

STRUCTURE AND WRITTEN EXPRESSION

Time 25 minutes

This section is designed to measure your ability to recognize language that is appropriate for standard written English. There are two types of questions in this section, with special directions for each type.

Directions: Questions 1-15 are incomplete sentences. Beneath each sentence you will see four words or phrases, marked (A), (B), (C), and (D). Choose the one word or phrase that best completes the sentence. Then, on your answer sheet, find the number of the question and fill in the space that corresponds to the letter of the answer you have chosen. Fill in the space so that the letter inside the oval cannot be seen.

Example I

Sample Answer

------ are found in virtually every country in the world

Ⓐ Ⓑ ● Ⓓ

(A) Swamps and marshes which
(B) When swamps and marshes
(C) Swamps and marshes
(D) Now that swamps and marshes

The sentence should read, "Swamps and marshes are found in virtually every country in the world." Therefore, you should choose answer (C).

Example II

Sample Answer

Milk is pasteurized by heating it for thirty minutes at about 63° Centigrade, rapidly cooling it, and then ------ it at a temperature below 10° Centigrade.

Ⓐ Ⓑ Ⓒ ●

(A) to store
(B) store
(C) be stored
(D) storing

The sentence should read, "Milk is pasteurized by heating it for thirty minutes at about 63° Centigrade, rapidly cooling it, and then storing it at a temperature below 10° Centigrade." Therefore, you should choose answer (D).

Now begin work on the questions.

1. Pansies can be cultivated easily in home gardens, but ------ plenty of water and not too much sun.
   (A) to require
   (B) they require
   (C) required
   (D) requiring

2. Before 8000 B.C. wheat did not grow as prolifically ------ it does today.
   (A) like
   (B) as
   (C) for
   (D) than

3. Both nickel and iron are whitish metals ------.
   (A) that are attracted by magnets
   (B) that magnets are attracted by them
   (C) are attracted by magnets
   (D) magnets that attract them

4. The bark of some species of oak trees yields a substance used in ------ leather.
   (A) treating
   (B) to treat
   (C) its treatment
   (D) it treats

5. Although phosphorus is an essential constituent of all living creatures, ------ is among the least abundant of the mineral nutrients.
   (A) what
   (B) it
   (C) still
   (D) however

6. ------ angles of any triangle always add up to 180 degrees.
   (A) If three
   (B) The three
   (C) Three of
   (D) Three are

7. The gibbon ranges over ------ other apes do.
   (A) than an area wider
   (B) wider than the area
   (C) a wider area than
   (D) an area wider than are

8. Sarah Frances Whiting opened the ------ of physics in the United States in 1878.
   (A) undergraduate teaching was in a second laboratory
   (B) second teaching laboratory of undergraduate
   (C) undergraduate teaching laboratory was second
   (D) second undergraduate teaching laboratory

9. ------, some of the Earth's interior heat escapes to the surface.
   (A) A volcano erupts
   (B) A volcano whether erupts
   (C) A volcano erupts it
   (D) If a volcano erupts

10. Sandra Day O'Connor, the first woman member of the United States Supreme Court, believed that the courts should interpret the laws ------ legislate.
    (A) than attempt to rather
    (B) rather than attempt to
    (C) to attempt rather than
    (D) attempt rather than to

11. ------ of minerals, which are chemical elements or compounds of varying purity.
    (A) The consistency of rocks
    (B) Rocks, consisting
    (C) Rocks consist
    (D) Whereas rocks consist

12. Booker T. Washington, acclaimed as a leading educator at the turn of the century, ------ of a school that later became the Tuskegee Institute.
    (A) taking charge
    (B) took charge
    (C) charge was taken
    (D) taken charge

13. ------ white ginger, one scrapes and washes the roots before drying them.
    (A) If makes
    (B) When making
    (C) Made
    (D) The making of

14. By the time ------, Norman Rockwell had decided that he wanted to be an artist

(A) in his early teens
(B) his early teens were
(C) was his early teens
(D) he was in his early teens

15. During the eighteenth century, Little Turtle was chief of the Miami tribe whose territory became ------- is now Indiana and Ohio.

(A) there
(B) where
(C) that
(D) what

Directions In questions 16-40 each sentence has four underlined words or phrases. The four underlined parts of the sentence are marked (A), (B), (C), and (D). Identify the one underlined word or phrase that must be changed in order for the sentence to be grammatically correct. Then, on your answer sheet, find the number of the question and fill in the space that corresponds to the letter of the answer you have chosen.

Example I

Sample Answer

Meadowlarks are about the same size than robins, but they have heavier bodies, shorter tails, and longer bills.

(A) ● (C) (D)

The sentence should read, "Meadowlarks are about the same size as robins, but they have heavier bodies, shorter tails, and longer bills." Therefore, you should choose answer (B).

Example II

Sample Answer

When overall exports exceed imports, a country said to have a trade surplus

(A) (B) ● (D)

The sentence should read, "When overall exports exceed imports, a country is said to have a trade surplus." Therefore, you should choose answer (C).

Now begin work on the questions.

GO ON TO THE NEXT PAGE

16. For make adobe bricks, workers mix sand and clay or mud with water and small quantities of straw, grass, or a similar material.

17. A dictionary allows quick access to the meaning of a word only if one knows how spell the word.

18. To simulate natural sounds in music, composers often use the orchestral instrument that they feel most near approximates the sound in question.

19. Sodium is of one the few metals that will burn when heated in air.

20. Alike traditional harmony, jazz progressions are based on triads, but the special jazz sound is created by the piling up of thirds above a basic triad.

21. Maine's abundant forests and rivers has made it a haven for many kinds of wildlife.

22. In feudal times, the rank of knighthood carried no social distinction, neither any man could be a knight.

23. Ethel Harvey's career illustrates some of the challenges encountered by women scientists of her generation as they sought support for they work.

24. Before the plains were settled, prairie dog towns in many places stretch as far as the eye could see.

25. Direct mail advertising serves to acquaint customers with products, alert them to new opportunities, and paving the way for other sales activities.

GO ON TO THE NEXT PAGE

26. Animal life on Prince Edward Island is confined large to ducks, pheasants, and rabbits.
A B C D

27. Andrew Wyeth is famous for his realistic and thoughtful paintings of person and places in rural Pennsylvania and Maine.
A B C D

28. It is common knowledge that a flash of lightning is seen before a clap of thunder heard.
A B C D

29. Wild elephants are almost continuously waving their trunks, both up in the air and down aside the ground.
A B C D

30. Oriental rugs are considered valuable and because their designs are intricate and the weaving process is time-consuming.
A B C D

31. The Montreal International Exposition, "Expo '67," was applauded for displaying an degree of taste superior to that of similar expositions.
A B C D

32. A motion picture director for over twenty years, Lois Weber stamped her films with herself style and personal conviction.
A B C D

33. According to astronomers, the type cloud found most frequently in outer space consist of diffuse particles of dust and gas.
A B C D

34. Among almost seven hundred species of bamboo, some are fully grown at less than a foot high, while other can grow three feet in twenty-four hours.
A B C D

GO ON TO THE NEXT PAGE

2 • 2 • 2 • 2 • 2 • 2 • 2

35. A foreign exchange rate is a price that reflects the relative supply and demand of difference currencies.
A B C D

36. Recent studies have shown that air into a house often has higher concentrations of contaminants than heavily polluted air outside.
A B C D

37. Rock decay or weathering is the results of reactions between elements in the atmosphere and the rock's constituents.
A B C D

38. The phases of the Moon have served as primary divisions of time for thousands of years ago.
A B C D

39. The introduction of the power loom enabled weavers to produce yard goods faster, more efficiently, and less expensive.
A B C D

40. In the 1880's, when George Eastman first offered the Kodak camera and film, photography becoming a popular and individualized art.
A B C D

THIS IS THE END OF SECTION 2

IF YOU FINISH BEFORE TIME IS CALLED, CHECK YOUR WORK ON SECTION 2 ONLY.
DO NOT READ OR WORK ON ANY OTHER SECTION OF THE TEST.
THE SUPERVISOR WILL TELL YOU WHEN TO BEGIN WORK ON SECTION 3.

SECTION 3

VOCABULARY AND READING COMPREHENSION

Time - 45 minutes

This section is designed to measure your comprehension of standard written English. There are two types of questions in this section, with special directions for each type.

Directions In questions 1-30 each sentence has an underlined word or phrase. Below each sentence are four other words or phrases, marked (A), (B), (C), and (D). You are to choose the one word or phrase that best keeps the meaning of the original sentence if it is substituted for the underlined word or phrase. Then, on your answer sheet, find the number of the question and fill in the space that corresponds to the letter you have chosen. Fill in the space so that the letter inside the oval cannot be seen.

Example

Having served previously as counselor to President Richard M. Nixon, Anne Armstrong was appointed ambassador to Great Britain in 1976.

(A) loyally
(B) ably
(C) often
(D) earlier

Sample Answer

Ⓐ Ⓑ Ⓒ ●

The best answer is (D) because "Having served earlier as counselor to President Richard M. Nixon, Anne Armstrong was appointed ambassador to Great Britain in 1976" is closest in meaning to the original sentence. Therefore, you should choose answer (D).

Now begin work on the questions.

1. Writer Anne Morrow Lindbergh traveled with her husband, the famous aviator Charles Lindbergh, as a navigator on several of his pioneering flights.

(A) accompanied
(B) stopped
(C) directed
(D) interviewed

2. Although air travel has some risk, statistically it is much safer than any other means of mass transportation.

(A) downturns
(B) opponents
(C) danger
(D) inconvenience

3. George Gershwin was the first American musician whose jazz compositions were seriously appreciated by concert audiences.

(A) heard
(B) sought
(C) admired
(D) reviewed

4. Plants raised in greenhouses are tended methodically in an attempt to create the best possible conditions for their growth.

(A) systematically
(B) naturally
(C) personally
(D) lovingly

GO ON TO THE NEXT PAGE

5. The tomato is actually a fruit, although it is commonly thought of as a vegetable.

(A) really
(B) partly
(C) organically
(D) apparently

6. Fertile soil deposited by prehistoric glaciers is found in all parts of Ohio except the southeast.

(A) rich
(B) disappearing
(C) slow-moving
(D) ancient

7. The only songs officially approved by the Puritans were very simple psalms.

(A) originally
(B) reluctantly
(C) unanimously
(D) formally

8. The use of wild animals in circuses was an innovation first introduced in the United States.

(A) a number
(B) a program
(C) a musical spectacle
(D) a new idea

9. Parole is usually granted to prisoners as a reward for good conduct.

(A) paid to
(B) ignored by
(C) given to
(D) requested by

10. Some botanists fear that the worldwide transfer of plant species is threatening the Earth's biological diversity.

(A) accidental
(B) rapid
(C) illegal
(D) global

11. The Pop Art of the 1960's used imagery drawn from the everyday world.

(A) understood by
(B) approved by
(C) censored in
(D) taken from

12. Benjamin Franklin was not the first to suggest the relationship between lightning and electricity, but his experiment with a kite was original.

(A) define
(B) confirm
(C) examine
(D) propose

13. One can improve the quality of sandy soil by thoroughly mixing in a small amount of clay or compost.

(A) occasionally
(B) manually
(C) completely
(D) instantly

14. The United States Congress and the state legislatures enact thousands of laws each year.

(A) draft
(B) pass
(C) debate
(D) amend

GO ON TO THE NEXT PAGE

15. The golden pheasant, which lives in the Himalayan Mountains, has gorgeous scarlet, gold, green, and black plumage.

(A) magnificent
(B) variegated
(C) very colorful
(D) very unusual

16. In his novella *The Old Man and The Sea*, Ernest Hemingway celebrates the indomitable courage of an elderly fisherman.

(A) discusses
(B) investigates
(C) praises
(D) analyzes

17. Tornadoes are rapidly swirling columns of air that are common in the midwestern United States.

(A) rotating
(B) growing
(C) sinking
(D) appearing

18. New England town meetings, in their most highly developed form, are assemblies of the voters.

(A) protests
(B) gatherings
(C) responsibilities
(D) liabilities

19. Because of Delaware's lenient laws regarding business incorporation, many companies have their headquarters in the state's largest city, Wilmington.

(A) production plants
(B) home offices
(C) sales representatives
(D) chemical laboratories

20. The type and degree of molecular motion of a substance depend on the amount of thermal energy present.

(A) are determined by
(B) limit
(C) radiate
(D) are supported by

21. Many of Edith Wharton's best stories were completed under great personal strain.

(A) poverty
(B) privacy
(C) resentment
(D) tension

22. The malco, an exotic fowl native to Indonesia, habitually deposits its eggs in the vicinity of hot springs.

(A) saves
(B) lays
(C) positions
(D) warms

23. Less successful artists are often obliged to turn to engraving, stonecutting, sign painting, and other artisans' tasks to earn a living.

(A) inspired
(B) asked
(C) forced
(D) tempted

24. The blue whale can live for half a year without eating; it is maintained by its blubber.

(A) restrained
(B) weighed
(C) caught
(D) sustained

GO ON TO THE NEXT PAGE

25. In astronomy the Little Dipper is the figure formed by the seven most radiant stars in the constellation Ursa Minor.

(A) fascinating
(B) brilliant
(C) visible
(D) well-known

26. Old Oraibi, dating from 1100, is said to be the oldest continuously occupied settlement in the United States.

(A) community
(B) building
(C) graveyard
(D) orchard

27. Abolitionist Frederick Douglass' eloquent speeches helped him achieve great success.

(A) rigorous
(B) persuasive
(C) familiar
(D) intelligent

28. A daring experimentalist in language, Gertrude Stein wrote in a style so eccentric that early critics were uncertain whether to take her seriously.

(A) circular
(B) conservative
(C) humorous
(D) strange

29. The province of British Columbia offers visitors breathtaking views of the Canadian Rocky Mountains.

(A) distant
(B) intimate
(C) stunning
(D) high altitude

30. Alice Hamilton (1869-1970) helped bring about legislation aimed at rectifying factory conditions detrimental to the health of workers.

(A) outlawing
(B) identifying
(C) correcting
(D) studying

GO ON TO THE NEXT PAGE

Directions: In the rest of this section you will read several passages. Each one is followed by several questions about it. For questions 31-60, you are to choose the one best answer, (A), (B), (C), or (D), to each question. Then, on your answer sheet, find the number of the question and fill in the space that corresponds to the letter of the answer you have chosen.

Answer all questions following a passage on the basis of what is stated or implied in that passage.

Read the following passage:

The railroad was not the first institution to impose regularity on society, or to draw attention to the importance of precise timekeeping. For as long as merchants have set out their wares at daybreak and communal festivities have been celebrated, people have been in rough agreement with their neighbors as to the time of day. The value of this tradition is today more apparent than ever. Were it not for public acceptance of a single yardstick of time, social life would be unbearably chaotic: the massive daily transfers of goods, services, and information would proceed in fits and starts; the very fabric of modern society would begin to unravel.

Line
(5)

Example I

Sample Answer

What is the main idea of the passage? Ⓐ Ⓑ ● Ⓓ

(A) In modern society we must make more time for our neighbors.
(B) The traditions of society are timeless.
(C) An accepted way of measuring time is essential for the smooth functioning of society.
(D) Society judges people by the times at which they conduct certain activities.

The main idea of the passage is that societies need to agree about how time is to be measured in order to function smoothly. Therefore, you should choose answer (C).

Example II

Sample Answer

In line 5, the phrase "this tradition" refers to Ⓐ Ⓑ Ⓒ ●

(A) the practice of starting the business day at dawn
(B) friendly relations between neighbors
(C) the railroad's reliance on time schedules
(D) people's agreement on the measurement of time

The phrase "this tradition" refers to the preceding clause, "people have been in rough agreement with their neighbors as to the time of day." Therefore, you should choose answer (D).

Now begin work on the questions.

GO ON TO THE NEXT PAGE

Questions 31-37

With its radiant color and plantlike shape, the sea anemone looks more like a flower than an animal. More specifically, the sea anemone is formed quite like the flower for which it is named, with a body like a stem and tentacles like petals in brilliant shades of blue, green, pink, and red. Its diameter varies from about six millimeters in some species to more than ninety centimeters in the giant varieties of Australia. Like corals, hydras, and jellyfish, sea anemones are coelenterates. They can move slowly, but more often they attach the lower part of their cylindrical bodies to rocks, shells, or wharf pilings. The upper end of the sea anemone has a mouth surrounded by tentacles that the animal uses to capture its food. Stinging cells in the tentacles throw out tiny poison threads that paralyze other small sea animals. The tentacles then drag this prey into the sea anemone's mouth. The food is digested in the large inner body cavity. When disturbed, a sea anemone retracts its tentacles and shortens its body so that it resembles a lump on a rock. Anemones may reproduce by forming eggs, dividing in half, or developing buds that grow and break off as independent animals.

31. The word "shape" in line 1 is closest in meaning to which of the following?

(A) Length
(B) Grace
(C) Form
(D) Nature

32. According to the passage, which of the following statements is NOT true of sea anemones?

(A) They are usually tiny.
(B) They have flexible bodies.
(C) They are related to jellyfish.
(D) They are usually brightly colored.

33. It can be inferred from the passage that sea anemones are usually found

(A) attached to stationary surfaces
(B) hidden inside cylindrical objects
(C) floating among underwater flowers
(D) chasing prey around wharf pilings

34. The word "capture" in line 8 is closest in meaning to which of the following?

(A) Catch
(B) Control
(C) Cover
(D) Clean

35. The word "disturbed" in line 11 is closest in meaning to which of the following?

(A) Bothered
(B) Hungry
(C) Tired
(D) Sick

36. The sea anemone reproduces by

(A) budding only
(B) forming eggs only
(C) budding or dividing only
(D) budding, forming eggs, or dividing

37. Where does the author mention the sea anemone's food-gathering technique?

(A) Lines 1-2
(B) Lines 4-6
(C) Lines 7-10
(D) Lines 11-13

GO ON TO THE NEXT PAGE

Questions 38–43

Steamships were first introduced into the United States in 1807, and John Molson built the first steamship in Canada (then called British North America) in 1809. By the 1830's dozens of steam vessels were in use in Canada. They offered the traveler reliable transportation in comfortable facilities – a welcome alternative to stagecoach travel, which at the best of times could only be described as wretched. This commitment to dependable river transport became entrenched with the investment of millions of dollars for the improvement of waterways, which included the construction of canals and lock systems. The Lachine and Welland canals, two of the most important systems, were opened in 1825 and 1829, respectively. By the time that Upper and Lower Canada were united into the Province of Canada in 1841, the public debt for canals was more than one hundred dollars per capita, an enormous sum for the time. But it may not seem such a great amount if we consider that improvements allowed steamboats to remain practical for most commercial transport in Canada until the mid-nineteenth century.

38. What is the main purpose of the passage?

(A) To contrast travel by steamship and stagecoach
(B) To criticize the level of public debt in nineteenth-century Canada
(C) To describe the introduction of steamships in Canada
(D) To show how Canada surpassed the United States in transportation improvements

39. The word "reliable" in line 3 is closest in meaning to which of the following?

(A) Quick
(B) Safe
(C) Dependable
(D) Luxurious

40. Which of the following can be inferred from the passage about stagecoach travel in Canada in the 1830's?

(A) It was reasonably comfortable.
(B) It was extremely efficient.
(C) It was not popular.
(D) It was very practical.

GO ON TO THE NEXT PAGE

3 • 3 • 3 • 3 • 3 • 3 • 3

41. According to the passage, when was the Welland Canal opened?

(A) 1807
(B) 1809
(C) 1825
(D) 1829

42. The word "sum" in line 10 is closest in meaning to which of the following?

(A) Size
(B) Cost
(C) Payment
(D) Amount

43. According to the passage, steamships became practical means of transportation in Canada because of

(A) improvements in the waterways
(B) large subsidies from John Molson
(C) a relatively small population
(D) the lack of alternate means

GO ON TO THE NEXT PAGE

Archaeology is a source of history, not just a humble auxiliary discipline. Archaeological data are historical documents in their own right, not mere illustrations to written texts. Just as much as any other historian, an archaeologist studies and tries to reconstitute the process that has created the human world in which we live—and us ourselves in so far as we are each creatures of our age and social environment. Archaeological data are all changes in the material world resulting from human action or, more succinctly, the fossilized results of human behavior. The sum total of these constitute what may be called the archaeological record. This record exhibits certain peculiarities and deficiencies the consequences of which produce a rather superficial contrast between archaeological history and the more familiar kind based upon written records.

Not all human behavior fossilizes. The words I utter and you hear as vibrations in the air are certainly human changes in the material world and may be of great historical significance. Yet they leave no sort of trace in the archaeological records unless they are captured by a dictaphone or written down by a clerk. The movement of troops on the battlefield may "change the course of history," but this is equally ephemeral from the archaeologist's standpoint. What is perhaps worse, most organic materials are perishable. Everything made of wood, hide, wool, linen, grass, hair, and similar materials will decay and vanish in dust in a few years or centuries, save under very exceptional conditions. In a relatively brief period the archaeological record is reduced to mere scraps of stone, bone, glass, metal, and earthenware. Still modern archaeology, by applying appropriate techniques and comparative methods, aided by a few lucky finds from peat bogs, deserts, and frozen soils, is able to fill up a good deal of the gap.

44. What is the author's main purpose in the passage?
    (A) To point out the importance of recent advances in archaeology
    (B) To describe an archaeologist's education
    (C) To explain how archaeology is a source of history
    (D) To encourage more people to become archaeologists

45. According to the passage, the archaeological record consists of
    (A) spoken words of great historical significance
    (B) the fossilized results of human activity
    (C) organic materials
    (D) ephemeral ideas

46. The word "they" in line 13 refers to
    (A) scraps
    (B) words
    (C) troops
    (D) humans

47. Which of the following is NOT mentioned as an example of an organic material?
    (A) Stone
    (B) Wool
    (C) Grass
    (D) Hair

GO ON TO THE NEXT PAGE

48. The author mentions all of the following archaeological discovery sites EXCEPT
    (A) urban areas
    (B) peat bogs
    (C) very hot and dry lands
    (D) earth that has been frozen

49. The paragraph following the passage most probably discusses
    (A) techniques for recording oral histories
    (B) certain battlefield excavation methods
    (C) some specific archaeological discoveries
    (D) building materials of the nineteenth and twentieth centuries

GO ON TO THE NEXT PAGE

Questions 50-54

Many artists late in the last century were in search of a means to express their individuality. Modern dance was one of the ways some of these people sought to free their creative spirit. At the beginning there was no exacting technique, no foundation from which to build. In later years trial, error, and genius founded the techniques and the principles of the movement. Eventually, innovators even drew from what they considered the dread ballet, but first they had to discard all that was academic so that the new could be discovered. The beginnings of modern dance were happening before Isadora Duncan, but she was the first person to bring the new dance to general audiences and see it accepted and acclaimed.

Her search for a natural movement form sent her to nature. She believed movement should be as natural as the swaying of the trees and the rolling waves of the sea, and should be in harmony with the movements of the Earth. Her great contributions are in three areas.

First, she began the expansion of the kinds of movements that could be used in dance. Before Duncan danced, ballet was the only type of dance performed in concert. In the ballet the feet and legs were emphasized, with virtuosity shown by complicated, codified positions and movements. Duncan performed dance by using all her body in the freest possible way. Her dance stemmed from her soul and spirit. She was one of the pioneers who broke tradition so others might be able to develop the art.

Her second contribution lies in dance costume. She discarded corset, ballet shoes, and stiff costumes. These were replaced with flowing Grecian tunics, bare feet, and unbound hair. She believed in the natural body being allowed to move freely, and her dress displayed this ideal.

Her third contribution was in the use of music. In her performances she used the symphonies of great masters, including Beethoven and Wagner, which was not the usual custom.

She was as exciting and eccentric in her personal life as in her dance.

50. Which of the following would be the best title for the passage?
    (A) The Evolution of Dance in the Twentieth Century
    (B) Artists of the Last Century
    (C) Natural Movement in Dance
    (D) A Pioneer in Modern Dance

51. According to the passage, what did nature represent to Isadora Duncan?
    (A) Something to conquer
    (B) A model for movement
    (C) A place to find peace
    (D) A symbol of disorder

52. Which of the following is NOT mentioned in the passage as an area of dance that Isadora Duncan worked to change?
    (A) The music
    (B) The stage sets
    (C) Costumes
    (D) Movements

GO ON TO THE NEXT PAGE

53. Compared to those of the ballet, Isadora Duncan's costumes were less
    (A) costly
    (B) colorful
    (C) graceful
    (D) restrictive

54. What does the paragraph following the passage most probably discuss?
    (A) Isadora Duncan's further contribution to modern dance
    (B) The music customarily used in ballet
    (C) Other aspects of Isadora Duncan's life
    (D) Audience acceptance of the new form of dance

GO ON TO THE NEXT PAGE

Questions 55-60

The theory of plate tectonics describes the motions of the lithosphere, the comparatively rigid outer layer of the Earth that includes all the crust and part of the underlying mantle. The lithosphere is divided into a few dozen plates of various sizes and shapes; in general the plates are in motion with respect to one another. A mid-ocean ridge is a boundary between plates where new lithospheric material is injected from below. As the plates diverge from a mid-ocean ridge they slide on a more yielding layer at the base of the lithosphere.

Since the size of the Earth is essentially constant, new lithosphere can be created at the mid-ocean ridges only if an equal amount of lithospheric material is consumed elsewhere. The site of this destruction is another kind of plate boundary: a subduction zone. There one plate dives under the edge of another and is reincorporated into the mantle. Both kinds of plate boundary are associated with fault systems, earthquakes and volcanism, but the kinds of geologic activity observed at the two boundaries are quite different.

The idea of sea-floor spreading actually preceded the theory of plate tectonics. In its original version, in the early 1960's, it described the creation and destruction of the ocean floor, but it did not specify rigid lithospheric plates. The hypothesis was substantiated soon afterward by the discovery that periodic reversals of the Earth's magnetic field are recorded in the oceanic crust. As magma rises under the mid-ocean ridge, ferromagnetic minerals in the magma become magnetized in the direction of the geomagnetic field. When the magma cools and solidifies, the direction and the polarity of the field are preserved in the magnetized volcanic rock. Reversals of the field give rise to a series of magnetic stripes running parallel to the axis of the rift. The oceanic crust thus serves as a magnetic tape recording of the history of the geomagnetic field that can be dated independently; the width of the stripes indicates the rate of the sea-floor spreading.

55. What is the main topic of the passage?
    (A) Magnetic field reversal
    (B) The formation of magma
    (C) The location of mid-ocean ridges
    (D) Plate tectonic theory

56. According to the passage, there are approximately how many lithospheric plates?
    (A) Six
    (B) Twelve
    (C) Twenty-four or more
    (D) One thousand nine hundred

57. Which of the following is true about tectonic plates?
    (A) They are moving in relationship to one other.
    (B) They have unchanging borders.
    (C) They are located far beneath the lithosphere.
    (D) They have the same shape.

58. According to the passage, which of the following statements about the lithosphere is LEAST likely to be true?
    (A) It is a relatively inflexible layer of the Earth.
    (B) It is made up entirely of volcanic ash.
    (C) It includes the crust and some of the mantle of the Earth.
    (D) It is divided into plates of various shapes and sizes.

GO ON TO THE NEXT PAGE

3 • 3 • 3 • 3 • 3 • 3 • 3

59. What does the author imply about the periodic reversal of the Earth's magnetic field?
    (A) It is inexplicable.
    (B) It supports the hypothesis of sea-floor spreading.
    (C) It was discovered before the 1960's.
    (D) It indicates the amount of magma present.

60. The author states that the width of the stripes preserved in magnetized volcanic rock give information about the
    (A) date of a volcanic eruption
    (B) speed of sea-floor spreading
    (C) width of oceanic crust
    (D) future behavior of the geomagnetic field

THIS IS THE END OF SECTION 3

IF YOU FINISH BEFORE TIME IS CALLED, CHECK YOUR WORK ON SECTION 3 ONLY.
DO NOT READ OR WORK ON ANY OTHER SECTION OF THE TEST.

STOP STOP STOP STOP STOP STOP STOP

NO TEST MATERIAL ON THIS PAGE

NO TEST MATERIAL ON THIS PAGE

Test of English as a Foreign Language

Be sure to fill in completely the oval that corresponds to your answer choice.
Completely erase errors or stray marks.

CORRECT ● | WRONG ⊘ | WRONG ⊗ | WRONG ○ | WRONG ◎

NAME (Print).

Appendix 3

Section 1-Listening Comprehension

SECTION 1

1	11	21	31	41
2	12	22	32	42
3	13	23	33	43
4	14	24	34	44
5	15	25	35	45
6	16	26	36	46
7	17	27	37	47
8	18	28	38	48
9	19	29	39	49
10	20	30	40	50

Section 2-Structure and Written Expression

SECTION 2

	9	17	25	33
	10	18	26	34
	11	19	27	35
	12	20	28	36
	13	21	29	37
	14	22	30	38
	15	23	31	39
	16	24	32	40

Section 3-Vocabulary and Reading Comprehension

SECTION 3

	13	25	37	49
	14	26	38	50
	15	27	39	51
	16	28	40	52
	17	29	41	53
	18	30	42	54
	19	31	43	55
	20	32	44	56
	21	33	45	57
	22	34	46	58
	23	35	47	59
	24	36	48	60

AFTER YOU TAKE THE TEST

Use the answer key on this page to determine which questions you answered correctly. Count the number of correct answers in each section of the test and write the total number of correct answers for each section in the appropriate box below.

Section 1	Section 2	Section 3

Answer Key

Question Number	Correct Answer	Question Number	Correct Answer	Question Number	Correct Answer	Question Number	Correct Answer	Question Number	Correct Answer	Question Number	Correct Answer
1	B	25	D	1	C	21	A	1	C	30	B
2	B	26	B	2	A	22	C	2	A	31	A
3	B	27	D	3	D	23	B	3	A	32	B
4	A	28	A	4	A	24	A	4	A	33	D
5	C	29	C	5	D	25	C	5	D	34	A
6	D	30	A	6	B	26	D	6	B	35	C
7	D	31	D	7	D	27	B	7	B	36	C
8	B	32	A	8	B	28	C	8	C	37	C
9	C	33	C	9	B	29	D	9	B	38	B
10	B	34	C	10	B	30	B	10	A	39	D
11	C	35	A	11	A	31	D	11	D	40	A
12	A	36	C	12	D	32	B	12	B	41	D
13	B	37	A	13	B	33	B	13	D	42	C
14	D	38	D	14	D	34	D	14	A	43	A
15	D	39	A	15	C	35	C	15	C	44	C
16	C	40	B	16	A	36	A	16	B	45	D
17	A	41	D	17	A	37	A	17	A	46	D
18	A	42	A	18	D	38	D	18	C	47	B
19	B	43	C	19	C	39	C	19	C	48	C
20	C	44	A	20	A	40	A	20	D	49	B
21	A	45	A					21	C	50	A
22	C	46	D					22	C	51	A
23	B	47	D					23	D	52	C
24	B	48	D					24	C	53	B
		49	B					25	D	54	B
		50	C					26	A	55	A
								27	B	56	D
								28	D	57	D
								29	C	58	B
										59	D
										60	A

The number of correct answers for each section is your "number-right" score for that section. Your number-right score is not the same as your TOEFL score for that section. TOEFL section scores are reported on a uniform scale of 20 to 80, and total scores are reported on a scale of 200 to 800. Statistical procedures are used to change the number-right score for each section to the "scaled" or "converted" score. In order to determine the equivalent scores for persons of equal ability regardless of the level of difficulty of the particular test they happen to take, TOEFL converted scores are equated by statistical methods based on "item response theory." As a result, a total score of 550, for example, on one edition of the test represents the same level of English ability as a score of 550 on another edition of the test.

When you have written your number-right scores for each section of the test in the boxes, look at the "converted scores" chart below. The first column gives ranges of number-right scores. The second, third, and fourth columns give ranges of converted scores.

60			67
57-59			64-66
54-56			61-63
51-53			59-61
48-50	64-68		57-58
45-47	61-63		55-56
42-44	58-60		53-54
39-41*	56-57	65-68	51-52
36-38	53-55	61-64	49-50
33-35	51-53	57-59	47-48
30-32	49-51	53-56	45-47
27-29	48-49	50-52	43-45
24-26	46-47	47-49	41-43
21-23	44-45	44-46	39-40
18-20	42-43	41-43	36-38
15-17	40-42	38-40	33-35
12-14	37-39	35-37	30-32
9-11	33-36	32-34	27-29
6-8	30-32	28-31	25-26
3-5	28-29	24-26	23-24
0-2	25-27	20-23	20-22

*For Section 2, this range is 39-40.

In the column marked "Number-Right Score Range," find the score range that includes your number-right score for Section 1. In the column marked "Section 1," find the range of converted scores for your number-right score. Write your converted score range for Section 1 in the appropriate box below. Do the same for your number-right scores for Sections 2 and 3.

Section 1	Section 2	Section 3

Appendix 4:

EPTB Short Version Form C

Keys, & Score Conversion Table

Prepared by

Alan Davies & Alan Moller

(1973)

English Proficiency Test Battery

Short Version Form C

Test	R	S
1		
2		
3		
4		
Total	///	

Candidate Name: ..... Candidate Number: .....

Centre Name: ..... Date: .....

SEX: MALE / FEMALE
(Please circle one)

You have now completed the details on the cover of this test booklet.

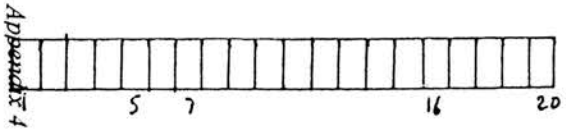
The instructions for each test will be read to you and, if necessary, repeated. If you do not understand the instructions, put up your hand. The Test Administrator will help you. You must not ask anything after starting the questions.

Try all the examples as you hear or read them. Answer each question as quickly as you can. If you do not know the answer, make a guess. Then go straight on to the next question. If you delay, you may miss the next question.

The first two tests are Listening Tests. The next two are Reading Tests. Each time you will be told when to begin and when to stop. If you finish a test before time , you must not go back to try to complete an earlier test. The test is now starting.

GO ON TO TEST 1

For official use only.



TEST 1

In this Test there are 58 questions. In each question you will hear three words, 1, 2, and 3. You must decide which words are the same. Sometimes all three words are the same, sometimes two are the same, sometimes no words are the same. You decide on your answer quickly. The figures 1, 2 and 3 stand for the three words. Show your answers by a circle around the groups of figures 12, 23, 123, or 13, indicating the words you think are the same. If you think all three words are different, put a circle round the figure 0.

Here are some examples:

Example 1

(12) 23 123 13 0

Here 1 and 2 are the same. That is why there is a circle round 12.

Example 2

12 23 123 13 0

Put in your circle. Did you circle 13? The first and third words are the same.

Example 3

12 23 123 13 0

Put in your circle. Is your circle round 123? All three words are the same.

312

Example 4

12 23 123 13 0

Put in your circle. Did you put a circle round the figure 0 ?All three words are different.

Example 5

12 23 123 13 0

Put in your circle. Is it round 23? The second and third words are the same.

(Now listen to the Instructions again).

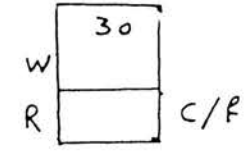
Remember you must answer each question quickly.

Now go on to Question 1 of the Test on the next page.

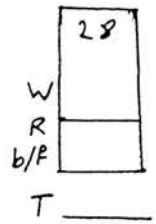
Appendix 4

1.	**12**	**23**	**123**	**13**	**0**
2.	**12**	**23**	**123**	**13**	**0**
3.	**12**	**23**	**123**	**13**	**0**
4.	**12**	**23**	**123**	**13**	**0**
5.	**12**	**23**	**123**	**13**	**0**
6.	**12**	**23**	**123**	**13**	**0**
7.	**12**	**23**	**123**	**13**	**0**
8.	**12**	**23**	**123**	**13**	**0**
9.	**12**	**23**	**123**	**13**	**0**
10.	**12**	**23**	**123**	**13**	**0**
11.	**12**	**23**	**123**	**13**	**0**
12.	**12**	**23**	**123**	**13**	**0**
13.	**12**	**23**	**123**	**13**	**0**
14.	**12**	**23**	**123**	**13**	**0**
15.	**12**	**23**	**123**	**13**	**0**
16.	**12**	**23**	**123**	**13**	**0**
17.	**12**	**23**	**123**	**13**	**0**
18.	**12**	**23**	**123**	**13**	**0**
19.	**12**	**23**	**123**	**13**	**0**
20.	**12**	**23**	**123**	**13**	**0**
21.	**12**	**23**	**123**	**13**	**0**
22.	**12**	**23**	**123**	**13**	**0**
23.	**12**	**23**	**123**	**13**	**0**
24.	**12**	**23**	**123**	**13**	**0**
25.	**12**	**23**	**123**	**13**	**0**
26.	**12**	**23**	**123**	**13**	**0**
27.	**12**	**23**	**123**	**13**	**0**
28.	**12**	**23**	**123**	**13**	**0**
29.	**12**	**23**	**123**	**13**	**0**
30.	**12**	**23**	**123**	**13**	**0**

GO ON TO PAGE 4



31.	12	23	123	13	0
32.	12	23	123	13	0
33.	12	23	123	13	0
34.	12	23	123	13	0
35.	12	23	123	13	0
36.	12	23	123	13	0
37.	12	23	123	13	0
38.	12	23	123	13	0
39.	12	23	123	13	0
40.	12	23	123	13	0
41.	12	23	123	13	0
42.	12	23	123	13	0
43.	12	23	123	13	0
44.	12	23	123	13	0
45.	12	23	123	13	0
46.	12	23	123	13	0
47.	12	23	123	13	0
48.	12	23	123	13	0
49.	12	23	123	13	0
50.	12	23	123	13	0
51.	12	23	123	13	0
52.	12	23	123	13	0
53.	12	23	123	13	0
54.	12	23	123	13	0
55.	12	23	123	13	0
56.	12	23	123	13	0
57.	12	23	123	13	0
58.	12	23	123	13	0



THAT IS THE END OF TEST 1. NOW GO ON TO TEST 2

TEST 2

In this Test you will hear 24 pieces of conversation between two students, John and Mary. First one speaks and then the other. Each piece of conversation is numbered, and after each piece you should answer the question on your answer paper. For each question there are two statements. Sometimes both are right, sometimes both are wrong; sometimes either the first or the second is right and the other wrong. When you have listened twice to the piece of conversation, read the statements. If you think a statement is right, or true, put a circle round the letter **T** after the statement; if you think a statement is wrong, or false, put a circle round the letter **F** which follows the statement.

Let us listen to this example.

Example 1:

Statements: 1. John is not sure if he's late. (T) F
2. Mary is not sure if she's early. T (F)

Here John is asking a question. He is not sure if he's late. Therefore statement (1) is right. That is why the letter **T** after the statement has a circle round it. Mary is not asking a question; she knows that she is early. Therefore statement (2) is wrong. That is why the letter **F** after the statement has a circle round it. The answer to this question, then, is a circle round the **T** for statement (1) and a circle round the **F** for statement (2), as shown above.

Now try these two examples. Listen.

Example 2:

Statements: 1. John is sure that Mary has sewn the buttons on. T F
2. Mary has sewn the buttons on. T F

Put circles round the appropriate letters.

Here John is again asking a question. He is not really sure if Mary has sewn the buttons on or not. Therefore statement (1) is wrong. Did you put a circle round the letter **F** after statement (1) ? Mary's answer indicates that she has finished sewing. Statement (2) is therefore right. Did you put a circle round the letter **T** after statement (2) ?

TURN TO PAGE 6

Example 3 :

Statements: 1. Mary knows who the person is. T F
2. John knows who the person is. T F

Put circles round the appropriate letters.

Here Mary shows by not pausing after 'look' that she knows who the person is. And the use of 'I'd never have believed it' by John shows that he too knows who the person is. Therefore both (1) and (2) are right. If Mary had been uncertain she would probably have said: .....

Did you put a circle round the letter T after both (1) and (2) ?

(Now listen to the Instructions again.)

Listen very carefully. You will hear each piece of conversation twice. Put your circles round the letters T or F as soon as you have heard the conversation repeated.

Now go on to Question 1 on page 7.

2. Mary is sure that John is going with her. T F
John is sure that he is going with her. T F

3. Mary wants a cup of tea. T F
Tea is not made around here. T F

4. John saw Helen yesterday. T F
There is a parcel for Mary. T F

5. Mary has no more compulsory lectures on politics. T F
John attended the lecture. T F

6. Mary went to see the film last night. T F
John went to see the film last night. T F

7. Mary is talking about one particular student. T F
There was probably only one student on Mary's bus. T F

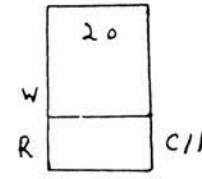
8. John's father writes amusing letters to him. T F
Mary's father writes to her regularly. T F

9. Mary would like to play tennis. T F
Mary will go shopping with Helen. T F

10. The new lecturer teaches French. T F
The new lecturer is a Frenchman. T F

11. Mary has understood the book. T F
John has understood the book. T F

GO ON TO PAGE 8



13. Mary has heard Bill's results. T F
Bill has passed his exam. T F

14. John and Mary think that Dick wants to improve his French. T F
Dick is going to France. T F

15. Helen is going to wear a new dress to the party. T F
Mary is going to wear a new dress to the party. T F

16. John is hungry. T F
Mary knows John is hungry. T F

17. John's younger sister has entered a beauty competition. T F
It is certain that John's sister will win the competition. T F

18. John saw a programme on up-to-date methods of teaching languages. T F
Mary seriously wants to know who invented television. T F

19. John enjoyed the last dance. T F
The professor had a new wife. T F

20. John does not want Mary to go. T F
John wants to know the aunt's name. T F

21. Mary is talking about the title of a film. T F
John sees two people lying on the grass. T F

22. Mary is anxious to get her essay back. T F
John is anxious to get his essay back. T F

THAT IS THE END OF TEST 2 AND OF THE LISTENING TESTS
The Test Administrator will read you the Instructions for Tests 3 and 4.
DO NOT GO ON UNTIL TOLD

	18
W	
R	
b/f	
T	

This is a Test of your understanding of written English. Here are two passages taken from fairly recent publications. In each passage some of the words are shown only by their initial letter and some dots. Complete these words to show that you understand these passages.

Now look at the example below. If you read the sentences you will see that they make some sort of sense but that five of the words are incomplete. Try to complete them.

Example

I always get u..... at seven o'clock i..... the morning. After breakfast I pack m..... lunch in my bag a..... walk round the corner t..... the bus stop.

Have you succeeded? The completed words are up, in, my, and, to.

DO NOT GO ON UNTIL TOLD

Passage 1

Furthermore, it was clear t..... i..... spite of the indispensable role played b..... scientists i..... t..... development of nuclear weapons, t..... had been labouring u..... a profound illusion in supposing t..... this w..... give them a..... affective voice i..... t..... use o..... these weapons. Scientists w..... still t..... b..... regarded as t..... 'backroom boys', t..... supply t..... ideas a..... t..... new gadgets, but to b..... kept i..... t..... place. 'Scientists s..... be on tap b..... not o..... top', a..... one British politician cynically put it.

R

Passage 2

A..... potters t..... Wedgwoods w..... well known and quite successful within t..... narrow limits o..... the Staffordshire trade. T..... invention o..... salt glaze i..... the 1690s a..... t..... proximity o..... the great Cheshire salt deposits gave t..... district a fillip i..... t..... early years o..... t..... eighteenth century. B..... the time, h....., Wedgwood w..... growing t..... manhood, the industry w..... losing ground.

R
T

For each item in this Test there is a sentence containing three choices. A native English speaker would choose only one of these. You must choose the one which you think would be used by a native English speaker. Put a circle round the number of your choice.

Example A

	1.	Do you like	
	(2)	Would you like	to come to tea tomorrow?
	3.	Could you like	

Answer 2 is what the native English speaker would write or say. That is why number 2 is circled.

Now try the following examples:

Example B

	1.	has been going on	
This banging noise	2.	went on	since Christmas.
	3.	goes on	

Example C

	1.	there are	
How many books	2.	are there	on the reading list?
	3.	is there	

Answer 1 is the right answer in Example B and Answer 2 is the right answer in Example C. Did you circle number 1 in Example B and number 2 in Example C?

The Test Administrator will tell you when to go on to Question 1 on page 12.

3. from

2. If I lived outside the city, 1. would I have / 2. will I have / 3. did I have to buy a car?

3. After six months the needle will need 1. to replace / 2. replace / 3. replacing

4. He has repaired the tap that was found 1. to leak / 2. leaking / 3. leaked

5. Would you like 1. to see / 2. see / 3. seeing the article when I've finished it?

6. I've sold a ticket to 1. every / 2. any / 3. all person in the room.

7. Because of lack of time we shall put 1. **away** / 2. on / 3. off the test until next term.

8. I should like a cigarette. Who 1. does have / 2. is having / 3. has one to give me?

9. I hear that 1. fishing / 2. fish / 3. to fish in this river will be prohibited next year.

TURN TO PAGE 13

3. have not decided

11. The accident 1. did happen / 2. has happened / 3. happened because it was very foggy that evening.

12. Jane would have passed the exam if she 1. had got / 2. got / 3. has got two more marks.

13. The teachers 1. do check / 2. are checking / 3. check attendance only about twice a week.

14. He's certain to see 1. through / 2. in / 3. out that unfounded argument very quickly.

15. I was told that the trains 1. did not might / 2. may not / 3. might not run on Thursday, but they did.

16. My sister thought it was expensive, and so 1. thought I / 2. did I / 3. I did

17. You have ironed 1. my trouser / 2. trousers / 3. my trousers very well.

18. The bus leaves at 10, 1. doesn't it? / 2. isn't it? / 3. is it?

TURN TO PAGE 14

3. ought not to

20. Give your clock to John. He'll be able to make it
1. go
2. to go
3. going

21. His ambition is to learn
1. read
2. to read
3. reading
Chinese.

22. Her brother is
1. doctor
2. a doctor
3. the doctor
in the largest hospital in the city.

23. The team
1. is coming
2. come
3. comes
here for their training session every Thursday.

24. Has your son taken
1. to
2. from
3. after
his new teacher readily?

25.
1. Did
2. Has
3. Had
he made a will before he died?

26. Does every student
1. has
2. have
3. having
his own room?

27. The weather is
1. very
2. so
3. too
cold that I have to wear two sweaters.

TURN TO PAGE 15

3. advise

29. That was
1. a first
2. the first
3. her first
time she had seen snow.

30. I'm coming back late, so don't wait
1. up
2. out
3. in
for me.

31. I am still waiting for him to reply
1. me
2. at me
3. to me

32. It's such a long name, but it's still
1. too easy
2. too easy to pronounce.
3. very easy to pronounce.

33. Before you cross the road, look
1. at the right,
2. to the right,
3. out the right,
then left, then right again.

34. There is a loud bang when a plane breaks
1. up
2. out
3. through
the sound barrier.

35. I shall need
1. to renew
2. renewing
3. renew
my car licence after four months.

36. 'No more cigarettes', the doctor said. 'You must cut
1. off
2. out
3. down
smoking'.

TURN TO PAGE 16

3. he goes

38. They must not begin until 1. instructed 2. they instruct to do so. 3. they instructed

39. 'What are you doing with that camera?' 'I'm taking 1. the 2. few pictures.' 3. some

40. Young people 1. ought to 2. must show more respect for older people. 3. have to

41. 1. Fewer 2. Less students are studying for an Arts degree this year than last year. 3. Few

42. 1. Some witnesses 2. The witnesses we were talking about have been to the police station. 3. Witnesses

43. She has 1. dark brown big 2. big dark brown eyes. 3. dark big brown

44. He had worked so well that he 1. might have 2. should have passed. But he didn't. 3. would have

45. Ten years ago there was a lot of traffic in Britain. Now there is 1. too much 2. a lot 3. much

TURN TO PAGE 17

3. itself

47. I thought 1. a 2. that man in the green hat was very rude to your wife last night. 3. some

T

That is the end of this English Proficiency Test

SHORT VERSION FORM C 1973
KEY - PART I

Test 1

1.	12	16.	13	31.	0	46.	23
2.	123	17.	0	32.	123	47.	12
3.	0	18.	23	33.	12	48.	13
4.	23	19.	13	34.	13	49.	23
5.	13	20.	12	35.	23	50.	0
6.	13	21.	13	36.	23	51.	23
7.	0	22.	12	37.	12	52.	0
8.	23	23.	23	38.	13	53.	13
9.	23	24.	0	39.	13	54.	12
10.	123	25.	12	40.	123	55.	0
11.	13	26.	0	41.	23	56.	12
12.	12	27.	0	42.	0	57.	12
13.	0	28.	123	43.	12	58.	123
14.	23	29.	12	44.	13		
15.	12	30.	12	45.	0		

Test 2

1.	– F	12.	– T
2.	F F	13.	T F
3.	– F	14.	F T
4.	T F	15.	– F
5.	T F	16.	T T
6.	F T	17.	– F
7.	T F	18.	– F
8.	F F	19.	F F
9.	T F	20.	T F
10.	T F	21.	T T
11.	T F	22.	F F

Test 3

(1)	that	in		
	by	in	the	
	they	under		
	that	would	an	
	in	the	of	were
	to	be	the	to
	the	and	the	be
	in	their	should	but
	on	as		

(28)

(2)	as	the	were	
	the	of		
	the	of	in	
	and	the	of	
	the	in	the	of
	the	by	however	
	was	to	was	

(21)

Test 4

1.	1	11.	3	21.	2	31.	3	41.	1
2.	1	12.	1	22.	2	32.	3	42.	2
3.	3	13.	3	23.	2	33.	2	43.	2
4.	2	14.	1	24.	1	34.	3	44.	2
5.	1	15.	3	25.	3	35.	1	45.	1
6.	1	16.	2	26.	2	36.	2	46.	2
7.	3	17.	3	27.	2	37.	3	47.	2
8.	3	18.	1	28.	1	38.	1		
9.	1	19.	3	29.	2	39.	3		
10.	1	20.	1	30.	1	40.	1		

ENGLISH PROFICIENCY TEST BATTERY
SHORT VERSION FORM C 1973

TABLE FOR CONVERSION OF RAW SCORES TO STANDARDISED SCORES

Raw Score	Standardised score			
	Test 1	Test 2	Test 3	Test 4
1	.1	2.4	4.6	.3
2	.1	2.8	4.8	.6
3	.1	3.1	5.0	.8
4	.1	3.5	5.2	1.1
5	.4	3.8	5.4	1.4
6	.6	4.1	5.6	1.7
7	.9	4.5	5.8	2.0
8	1.1	4.8	6.0	2.2
9	1.4	5.2	6.2	2.5
10	1.6	5.5	6.4	2.8
11	1.9	5.8	6.6	3.1
12	2.1	6.2	6.8	3.4
13	2.4	6.5	7.0	3.6
14	2.6	6.9	7.2	3.9
15	2.9	7.2	7.4	4.2
16	3.1	7.5	7.6	4.5
17	3.4	7.9	7.8	4.8
18	3.6	8.2	8.0	5.0
19	3.9	8.6	8.2	5.3
20	4.1	8.9	8.4	5.6
21	4.4	9.2	8.6	5.9
22	4.6	9.6	8.8	6.2
23	4.9	9.9	9.0	6.4
24	5.1	10.3	9.2	6.7

Raw Score	Standardised score			
	Test 1	Test 2	Test 3	Test 4
25	5.4	10.6	9.4	7.0
26	5.6	10.9	9.6	7.3
27	5.9	11.3	9.8	7.6
28	6.1	11.6	10.0	7.8
29	6.4	12.0	10.2	8.1
30	6.6	12.3	10.4	8.4
31	6.9	12.6	10.6	8.7
32	7.1	13.0	10.8	9.0
33	7.4	13.3	11.0	9.2
34	7.6	13.7	11.2	9.5
35	7.9	14.0	11.4	9.8
36	8.1	14.3	11.6	10.1
37	8.4	14.7	11.8	10.4
38	8.6	15.0	12.0	10.6
39	8.9		12.2	10.9
40	9.1		12.4	11.2
41	9.4		12.6	11.5
42	9.6		12.8	11.8
43	9.9		13.0	12.0
44	10.1		13.2	12.3
45	10.4		13.4	12.6
46	10.6		13.6	12.9
47	10.9		13.8	13.2
48	11.1		14.0	
49	11.4		14.2	
50	11.6			
51	11.9			
52	12.1			
53	12.4			
54	12.6			
55	12.9			
56	13.1			
57	13.4			
58	13.6			

Appendix 5:

IELTS Band Scores (1999)

Examiner's Writing Mark Sheet (1995)

Final Band Conversion Grid for Writing (1995)

Profile Band Descriptors for Writing Task 1 (1995)

Profile Band Descriptors for Writing Task 2 (1995)

Writing Samples (1995)

IELTS Band Scores (Source: UCLES, 1999, p. 6)

IELTS scores are reported on a nine-band scale. The nine bands and their descriptive statements are as follows:

9 **Expert User**. Has fully operational command of the language: appropriate, accurate and fluent with complete understanding.

8 **Very Good User.** Has fully operational command of the language with only occasional unsystematic inaccuracies and inappropriacies. Misunderstandings may occur in unfamiliar situations. Handles complex detailed argumentation well.

7 **Good User**. Has operational command of the language, though occasional inaccuracies, inappropriacies and misunderstandings in some situations. Generally handles complex language well and understands detailed reasoning.

6 **Competent User.** Has generally effective command of the language despite some inaccuracies, inappropriacies and misunderstandings. Can use and understand fairly complex language, particularly in familiar situations.

5 **Modest User**. Has partial command of the language, coping with overall meaning in most situations, though is likely to make many mistakes. Should be able to handle basic communication in own field.

4 **Limited User.** Basic competence is limited to familiar situations. Has frequent problems in understanding and expression. Is not able to use complex language.

3 **Extremely Limited User.** Conveys and understands only general meaning in very familiar situations. Frequent breakdowns in communication occur.

2 **Intermittent User.** No real communication is possible except for the most basic information using isolated words or short formulae in familiar situations and to meet immediate needs. Has great difficulty understanding spoken and written English.

1 **Non User**. Essentially has no ability to use the language beyond possibly a few isolated words.

0 Did not attempt the test. No assessable information.

REVERSE SIDE OF EXAMINER'S WRITING MARK SHEET
ACADEMIC MODULES A, B, C
(Provided by UCLES, 1995)

Question 1 Sub-scales	Band
Task fulfilment	.....
Coherence and cohesion	+ .....
Sentence structure	+ .....

(Round mark to nearest whole number.
Scores of .5 are rounded up.) Total ..... + 3 = ☐ Global Band

Question 2 Sub-scales	Band
Communicative quality	.....
Arguments, ideas & evidence	+ .....
Word choice, form & spelling	+ .....
Sentence structure	+ .....

(Round mark to nearest whole number.
Scores of .5 are rounded up.) Total ..... + 3 = ☐ Global Band

Final Band Conversion Grid

Question 1 band	Question 2 band 0	1	2	3	4	5	6	7	8	9
0	0	1	1	2	3	3	4	5	5	6
1	0	1	2	2	3	4	4	5	6	6
2	1	1	2	3	3	4	5	5	6	7
3	1	2	2	3	4	4	5	6	6	7
4	1	2	3	3	4	5	5	6	7	7
5	2	2	3	4	4	5	6	6	7	8
6	2	3	3	4	5	5	6	7	7	8
7	2	3	4	4	5	6	6	7	8	8
8	3	3	4	5	5	6	7	7	8	9
9	3	4	4	5	6	6	7	8	8	9

Examiner's name (capitals): .....

Final Band ☐

Final Band Conversion Grid for Writing Modules A, B and C

(Provided by UCLES, 1995)

Question 1 band	Question 2 band									
	0	1	2	3	4	5	6	7	8	9
0	0	1	1	2	3	3	4	5	5	6
1	0	1	2	2	3	4	4	5	6	6
2	1	1	2	3	3	4	5	5	6	7
3	1	2	2	3	4	4	5	6	6	7
4	1	2	3	3	4	5	5	6	7	7
5	2	2	3	4	4	5	6	6	7	8
6	2	3	3	4	5	5	6	7	7	8
7	2	3	4	4	5	6	6	7	8	8
8	3	3	4	5	5	6	7	7	8	9
9	3	4	4	5	6	6	7	8	8	9

The Conversion Grid weights the question 1/3 for task 1 and 2/3 for tank 2.

The examiner should read the score awarded for task 1 on the vertical (side) line and the score awarded for task 2 on the horizontal (top) line to arrive at the final score. For example, if Band 6 Is awarded for task 1 and Band 4 for tank 2, the Final Band for writing is 5.0. The final writing adore should be put in the box labelled "Final Band".

The examiner should write his/her name legibly and sign the mark sheet. Monitoring cannot take place if this in not done.

There are occasions when it may seem that the final band score does not represent a candidate's ability. For example, a candidate might get a Band 8 for task 1, but a Band 3 for task 2 (perhaps because the answer is too short). The Final Band Score would then be a 5.0. This might seem inappropriate for what is potentially a good student. However, the second task is more linguistically demanding than the first and if the student has not been able to display the language required for tank 2, for whatever reason, it cannot he assumed that he or she is capable of producing it. A Final Band of 5.0 is reasonable therefore, although it may be counter-intuitive.

NOTE that if a candidate has attended other parts of the tent but has not attended or attempted this part of the test (i.e. has not submitted an answer paper with his/her name on), a score of 0 should be, recorded.

PROFILE BAND DESCRIPTORS — TASK 1

Task Fulfilment	Coherence & Cohesion	Vocabulary & Sentence Structure
9. The writing fulfils the task in a way which satisfies all requirements.	The message can be followed effortlessly. Coherence and cohesion are so skilfully managed that they attract no attention.	A wide range of vocabulary and sentence structures is used accurately and appropriately.
8. The writing fulfils the task in a very satisfactory manner.	The message can be followed with ease. Coherence and cohesion are very good.	The range of vocabulary and sentence structures used is good, and well controlled for accuracy and appropriacy. There are no significant errors in word formation or spelling.
7. The writing generally addresses the task relevantly, appropriately and accurately, however it could be more fully developed.	The message can be followed throughout and usually with ease. Information is generally arranged coherently, and cohesion within and between sentences is well managed.	A satisfactory range of vocabulary and sentence structures occurs, usually used appropriately. There are only occasional minor flaws in word formation and in control of sentence structure. Spelling errors may occur, but they are not intrusive.
6. The writing mostly addresses the task. However, the reader notices some irrelevant, inappropriate or inaccurate information in areas of minor importance. Minor details may be missing.	The message can be followed throughout. Information is generally arranged coherently, but cohesion within and/or between sentences may be faulty with misuse, overuse or omission of cohesive devices.	Vocabulary and sentence structures are generally adequate and appropriate, but the reader may feel that control is achieved through the use of a restricted range. In contrast, examples of the use of a wider range of structures are not marked by the same level of accuracy. Some errors in word choice, word formation and spelling may occur, but they are only slightly intrusive.
5. The writing is generally adequate, but the inclusion of irrelevant, inappropriate or inaccurate material in key areas detracts from its fulfilment of the task. There may be some details missing.	The message can generally be followed, although sometimes with difficulty. Both coherence and cohesion may be faulty.	The range of vocabulary and the appropriacy of its use are limited. There is a limited range of sentence structures and the greatest accuracy is achieved on short, simple sentences. Inappropriate choice of words and errors in areas such as agreement of tenses or subject/verb agreement are noticeable. Word formation and spelling errors may be quite intrusive.
4. The writing attempts to fulfil the task but is prevented from doing so adequately by omission of key details, and by irrelevance, inappropriacy, or inaccuracy.	The message is difficult to follow. Information is not arranged coherently, and cohesive devices are inadequate or missing.	The range of vocabulary is often inadequate and/or inappropriate and limited control of sentence structures, even short, simple ones, is evident. Choice of words can cause significant problems for the reader. Errors in such areas as agreement of tenses or subject/verb agreement, word formation and spelling can cause severe strain for the reader.
3. The seriousness of the problems in the writing makes it difficult to judge in relation to the task.	There are only occasional glimpses of a message. Neither coherence nor cohesion are apparent.	Control of vocabulary and sentence structure is evident only occasionally and errors predominate.
2. The writing does not appear to be related to the task.	There is no recognizable message.	There is little or no evidence of control of sentence structure, vocabulary, word form or spelling.
1. The writing appears to be by a virtual non-writer, containing no assessable strings of English writing. If an answer is wholly, or almost wholly copied from the source materials it is scored as band 1. Answers of less than two lines are automatically scored as band 1.		
0. Should only be used where a candidate did not attend or attempt the question in any way.		

PROFILE BAND DESCRIPTORS — TASK 2

Communicative Quality	Arguments, Ideas & Evidence	Vocabulary & Sentence Structure
9. The reader finds the writing completely satisfactory.	A clear point of view is presented and developed. The argument proceeds logically through the text, with a very clear progression of ideas. There is plentiful material	A wide range of vocabulary and sentence structure is used accurately and appropriately
8. The reader finds the writing communicates fluently.	A clear point of view is presented and developed. The argument proceeds logically through the text with a clear progression of ideas.	The range of vocabulary and sentence structures used is good, and well controlled for accuracy and appropriacy. There are no significant errors in word formation or spelling.
7. The reader finds the writing satisfactory and that it generally communicates fluently with only occasional lapses.	A generally clear point of view is presented. The argument has a clear progression overall, and ideas and evidence are relevant and sufficient, although there may be minor isolated problems in these areas.	A satisfactory range of vocabulary and sentence structures occurs, usually used appropriately. There are only occasional minor flaws in word formation and in control of sentence structure. Spelling errors may occur, but they are not intrusive.
6. The reader finds the writing mainly satisfactory and that it communicates with some degree of fluency. Although there is occasional strain for the reader, control of organisational patterns and devices is evident.	A point of view is presented, although it may become unclear in places. The progression of the argument is generally clear. The relevance of some ideas or evidence may be dubious and more specific support may seem desirable.	Vocabulary and sentence structures are generally adequate and appropriate, but the reader may feel that control is achieved through the use of a restricted range. In contrast, examples of the use of a wider range of structures are not marked by the same level of accuracy. Some errors in word choice, word formation and spelling may occur, but they are only slightly intrusive.
5. The writing sometimes causes strain for the reader. While the reader is aware of an overall lack of fluency, there is a sense of an answer which has an underlying coherence.	The writing introduces ideas, although they may be limited in number or insufficiently developed. A point of view may be evident, but arguments my lack clarity, relevance, consistency or support.	The range of vocabulary and the appropriacy of its use are limited. There is a limited range of sentence structures and the greatest accuracy is achieved on short, simple sentences. Inappropriate choice of words and errors in areas such as agreement of tenses or subject/verb agreement are noticeable. Word formation and spelling errors may be quite intrusive.
4. The writing attempts communication but the meaning may come through only after considerable effort by the reader.	There are signs of a point of view, but main ideas are difficult to distinguish from supporting material and the amount of support is inadequate. Such evidence and ideas as are presented may not be relevant. There is no clear progression to the argument.	The range of vocabulary is often inadequate and/or inappropriate and limited control of sentence structures, even short, simple ones, is evident. Choice of words can cause significant problems for the reader. Errors in such areas as agreement of tenses or subject/verb agreement, word formation and spelling can cause severe strain for the reader.
3. The seriousness of the problems in the writing prevents meaning from coming through more than spasmodically.	The writing has few ideas and no apparent development. Such evidence and ideas as are presented are largely irrelevant. There is little comprehensible point of view.	Control of vocabulary and sentence structures is evident only occasionally and errors predominate.
2. The writing displays almost no ability to communicate.	There may be glimpses of one or two ideas without development.	There is little or no evidence of control of sentence structure, vocabulary, word form or spelling.
1. The writing appears to be by a virtual non-writer, containing no assessable strings of English writing. If an answer is wholly, or almost wholly, copied from the source materials it is scored as band 1. Answers of less than two lines are automatically scored as band 1.		
0. Should only be used where a candidate did not attend or attempt the question in any way.		

Writing Samples

Band 8: Very Good Writer

The dentition of each mammal is determined by its feeding habits and diet. The sheep differs from the dog in its dental formula, in particular, the absence of canines and in upper-jaw ~~canine~~ incisors, and in the manner of jaw movement. The premolars and molars serve a grinding function with the lower jar taking a circular path about the jaw point. In the dog, the emphasis shifts to the canine and carnassial teeth which are used to ~~attack~~ chop up meat into chunks, ~~and not good~~ Movement of the lower jaw is in the vertical plane about the jaw point in the dog.

and 7: Good Writer

Though interior heating is not necessary for a large part of the year in a Tropical country such as India, heat pumps can still be gainfully employed not only during the colder periods of the year but also in those instances where heating is necessary, irrespective of the weather. With more efficient and economically viable heat pumps, the refrigerator, hitherto regarded as a luxury item, can be brought to the lower income homes.

In places like Delhi, Srinagar, Jaipur and other such places in the North, where the weather is quite cold in the winter, heat pumps can be used to warm up the rooms or buildings

Band 6: Competent Writer

As a plant breeder, I will join the rest people to produce high yielding, well adapted and many disease resistant varaties for my country. This is indeed a great benefit which is very likely to put my country in a very good stead as one of the leading agricultural countries in the world. Without my coming to Britain, the acquisition of the technology used in producing hybrids, which has been a giant step in crop improvement, could have probably not come to me. This is not because the course is not done in other countries but language has always been the problem. In Britain I have been introduced to modern techniques and I have shared knowledge with a well informed Supervisor. All these will be of considerable help to me in future in my endeavours to improve agriculture in my country.

Band 5: Modest Writer

I ~~like~~ shall choose to read sizes 'Prisons I Have Known". Because here I may get the real information of the Prison source. The author described here own experience and it is about the female prisoners. She was the Governor of the 'Open' prison. It is also interesting thing that the prison is open. On the other hand 'Portrait in Grey' by Hignett is not well written He under-estimated the idealism of members of the prison source. I do not, perhaps get any real picture as information in it.

Band 4: Marginal Writer

First of all a am very interested to lear
ong English language. I believe that a good langu
age could be learn in own country with mixing
British people, seeing their own Caulcra and
cuctumes. I always had in diffuculty to read
and understand ~~as~~ medical books in English.
As we know a lot of medical books are
written in English. A lot of medical recuarch
are done in ~~Great~~. England and U.S.A. And
of course their ~~puplished~~ in English. In my
own country if you don't know any
~~no~~ foreign language as a doctor you have
to wait somebody tranlate a book from English
to Turbish. This may take years.

Band 3: Extremely Limited Writer

If pt with limb = The breast and had been
removed and She develp any sing of depres. first
of all to tell her the True about The lumb which
removed and give her a hope of complete quiet
life. if it benin and if it is malegnent, give her a
Proper treatment and not to loss a hope for the best

Band 2: Intermittent Writer

she must go to Doctor who is speacliced
and to detail of the oprotion is nessery to do
it it depend on Mr doctor who will going to
treat her.

Appendix 6:

Instructions & Checklist for Propositional Content Instrument

Propositional content

Degree of contextualisatoin:

In rating this facet, consider the relative proportion of "new" to "contextual" information. "New information" is that which is not known to the test taker and cannot be predicted from the context. "Contextual information" is that which is developed in the passage itself. Thus, a passage is "not at all contextualised" if there is a lot of new information in the passage that is not explained through definition, example, paraphrase, etc. The passage is "highly contextualised" if there is either not much new information, or if the new information is explained.

Input can be contextualised in terms of two types of information: Cultural and that which is topic specific. Ratings on this facet should be as follows:

	Not at all contextualised		Highly contextualised
With respect to cultural content (CTXCULT)	0	1	2
With respect to specific topical content (CTXSPEC)	0	1	2

Distribution of new information:
This facet characterises the distribution of the new information that must be processed and manipulated in order for the test taker to successfully complete a given test task. Discourse in which new information is distributed over a relatively short space or time may be called *compact*, while discourse with new information distributed over a relatively long space or time may be called *diffuse*. For example, a highly diffuse passage is the one that test takers have ample time to read the passage in its entirety and to re-read as necessary to answer the questions based on the passage.

	Highly compact		Highly diffuse
Distribution of new information (C/D)	-1	0	+1

Type of information:

These facets should be rated in terms of the relative degree of abstractness or negativeness of the passage. What is of concern here is the information contained in the text, and not how the test taker is expected to process that information. For the "negative" (NEG) rating, consider only explicitly marked negatives, recognising that negatives may be explicitly marked in English in a variety of ways.

Abstract (ABSTRACT) 0 1 2 (Concrete)

Negative (NEG) 0 1 2 (Positive)

.....

Topic:
This facet has to do with the topic of the text. All the texts are considered to be academic. Decide whether they are biased towards a specific descipline or they are just general academic .

	Discipline Specific	General Academic
Specialised topic (TOPSPEC)	1	2

****

Please enter your ratings in the following table.

	TR1	TR2	TR3	TR4	TR5	IR1	IR2	IR3
CTXCULT								
CTXSPEC								
C/D								
ABSTRACT								
NEG								
TOPSPEC								

Appendix 7:

Examples of Cohesive Markers

Summary Table of Conjunctive Relations: Halliday & Hassan, Cohesion in English, 1976: 242-3.

	External/internal	Internal (unless otherwise specified)		
Additive	Additive, simple: Additive *and, and also* Negative *nor, and ... not* Alterna- *or, or else* tive	Complex, emphatic: Additive *furthermore, in addition, besides* Alternative *alternatively* Complex, de-emphatic: After- *incidentally,* thought *by the way*	Apposition: Expository *that is, I mean, in other words* Exemplifi- *for instance,* catory *thus*	Comparison: Similar *likewise similarly, in the same way* Dissimilar *on the other hand, by contrast*
Adversative	Adversative 'proper': Simple *yet, though,* Contain- *but only* ing 'and' Emphatic *however, nevertheless, despite this*	Contrastive: Avowal *in fact, actually, as a mater of fact* Contrastive (external): Simple *but, and* Emphatic *however, on the other hand, at the same time*	Correction: Of meaning *instead, rather, on the contrary* Of wording *at least, rather, I mean*	Dismissal: Closed *in any case, in either case, which ever way it is* Open-ended *in any case, anyhow, at any rate, however it is*
Causal	Causal, general: Simple so, *then, hence,* *therefore* Emphatic *consequently, because of this* Causal, specific: Reason *for this reason, on account of this* Result *as a result, in consequence* Purpose *for this purpose, with this in mind*	Reversed causal: Simple *for, because* Causal, specific: Reason *it follows, on this basis* Result *arising out of this* Purpose *to this end*	Conditional (also external): Simple *then* Emphatic *in that case, in such an event, that being so* Generalized *under the circumstances* Reversed polarity *otherwise, under other circumstances*	Respective: Direct *in this respect, in this regard, with reference to this* Reversed polarity *otherwise, in other respects, aside from this*
Temporal	Temporal, simple (external only): Sequential *then, next, after that* Simul- *just then, at* taneous *the same time* Preceding *previously, before that* Conclusive: Simple *finally, at last* Correlative forms: Sequential *first ... then* Conclu- *at first ... in* sive *the end*	Complex (external only): Immediate *at once, thereupon* Interrupted *soon, after a time* Repetitive *next time, on another occasion* Specific *next day, an hour later* Durative *meanwhile* Terminal *until then* Punctiliar *at this moment*	Internal temporal: Sequential *then, next, secondly* Conclusive *finally, in* Correlative forms: Sequential *first ... next* Conclusive . . finally	'Here and now': Past *up to now, hitherto* Present *at this point, here* Future *from now on, hence forward* Summary: Sum marizing *to sum up, in short, briefly to* Resumptive *resume, to return to the point*

Appendix 8:

Communicative Language Ability Rating Instrument & Checklists

Appendix 8

Communicative Language Ability (CLA) Instrument

Provided by Bachman (1995), personal communication

Ratings of CLA should be made on the basis of a) the extent to which you feel the ability is required for the successful completion of the task and b) the general "level" of that ability required, according to the following scale:

Not Required	Somewhat Involved	Critical Basic	Critical Intermediate	Critical Advanced
0	1	2	3	4

If you feel the ability is not required for successful completion of the task, write "0"; if the ability may be involved, but is not critical to successful completion of the task, write "1"; if the ability is critical to successful completion of the task, and at a basic level, write "2"; if critical, but intermediate level, write "3"; and if critical and advanced level, write "4".

In the following example, no ability other than lexical competence is critical to correctly answering the item:

> **Instructions**: Choose the word from the choices for each item that means most nearly the same as the underlined word in the item. Circle the letter for your choice.
>
> "Sylvia Plath's *The Bell Jar* was written a decade after the occurrence of the events it describes."
>
> a. a short time
> b. several months
> c. ten years
> d. a century

This is because the test taker can simply match the underlined phrase "a decade" with its meaning, "ten years", without even reading the rest of the stem.

Components to be rated

Grammatical Competence

LEX: Lexis
MOR: Morphology
STX: Syntax
PG: Phonology/Graphology

NB: PG will always be "2" for written tests.

Textual Competence

COH: Cohesion

Cohesion refers to explicitly marked relationships across clauses within the same sentence or across sentences. This explicit marking may be in the form of lexical connectors or of specific grammatical patterns that provide appropriate topicalisation. Types of cohesion include reference, substitution, ellipsis, conjunction and lexical cohesion.

ORG: Rhetorical Organisation

Conventions of rhetorical organisation include common methods of development such as narration, description, comparison, classification, argumentation and process analysis.

NB: ORG will always be "0" for single-sentence items.

Illocutionary Competence

Illocutionary competence pertains to the ability to use language functionally, or to perform speech acts, or language functions. These functions can be grouped into four "macro-functions", each of which is described briefly below. Functions can be performed with varying degrees of directness, ranging, for example, from very direct forms such as "I request you to open the window," to less direct forms such as "I wonder if someone could open the window" and "It's really hot in here." In general the level of ability required to interpret the realisation (utterance, sentence, text) of a given function can be considered to be determined by two factors: 1) the amount of information in and complexity of the function and 2) the degree of directness or indirectness with which it is expressed. Thus "basic" realisations of functions would be those that express simple functions and which realise the functions quite directly, while "advanced" realisations are those that express complex functions or that state the functions very indirectly.

IDE: Ideational Functions

Ideational functions are those by which we express meaning in terms of our experience of the real world, that is, by which we communicate information – ideas or feelings.

MAN: Manipulative Functions

Manipulative functions are those in which the primary purpose is to affect the world around us. These include:

> Instrumental functions: use of language to get things done
> Regulatory functions: use of language to control the behaviour of others, or to manipulate persons in the environment.
> Interactional functions: use of language to form, maintain or change interpersonal relationships

HEU: Heuristic Functions

Heuristic functions pertain to the use of language to extend our knowledge of the world around us, such as in problem-solving, teaching and learning.

IMG: Imaginative Functions

Imaginative functions enable us to use language to create or extend our own environment for humorous or aesthetic purposes, where the value derives from the way in which the language itself is used. Examples include telling/listening to jokes, constructing and communicating fantasies, creating/interpreting metaphors or other figures of speech, as well as attending plays or films and reading literary works such as novels, short stories or poetry for enjoyment.

> NB: IMG does not refer to how imaginative the text is, but to whether the test taker must perceive an imaginative function in order to correctly answer the item. For example, the following item does not require the test taker to perceive the imaginative function in order to answer the item correctly, even though it includes a figure of speech, a simile:

"Rough diamonds look like pebbles made of glass."

a. gems
b. bubbles
c. stones
d. beads

Sociolinguistic Competence

DIA: Dialect

REG: Register

Strategic Competence (STC)

Strategic competence refers to the mental capacity that enables language users to implement the components of language competence (listed above) in contextualised communicative language use. That is, strategic competence enables language users to relate the features of the language use context to the intended illocutionary force to produce an appropriate utterance form, or to relate an utterance form to the features of the context to interpret its illocutionary force appropriately.

	Very much		Not at all
Degree to which engaged	2	1	0

NB: Include in STC the skills associated with "test-wiseness", and the processing of non-verbal visual information, such as pictures, graphs or charts.

Communicative Language Abilities checklist

Test: IELTS
Rater: _____

Part: Listening Comprehension
Date: _____

0 = not involved, 1 = involved, 2 = critical basic, 3 = critical intermediate, 4 = critical advanced

Item #	LEX	MOR	STX	PG	COH	ORG	IDE	MAN	HEU	IMG	DIA	REG	STC
1.													
2.													
3.													
4.													
5.													
6.													
7.													
8.													
9.													
10.													
11.													
12.													
13.													
14.													
15.													
16.													
17.													
18.													
19.													
20.													
21.													
22.													
23.													
24.													
25.													
26.													
27.													
28.													
29.													
30.													
31.													
32.													
33.													
34.													
35.													
36.													
37.													
38.													
39.													
40													

Communicative Language Abilities checklist

Test: IELTS Part: Reading Comprehension1
Rater: _____ Date: _____

0 = not involved, 1 = involved, 2 = critical basic, 3 = critical intermediate, 4 = critical advanced

Item #	LEX	MOR	STX	PG	COH	ORG	IDE	MAN	HEU	IMG	DIA	REG	STC
1.	___	___	___	___	___	___	___	___	___	___	___	___	___
2.	___	___	___	___	___	___	___	___	___	___	___	___	___
3.	___	___	___	___	___	___	___	___	___	___	___	___	___
4.	___	___	___	___	___	___	___	___	___	___	___	___	___
5.	___	___	___	___	___	___	___	___	___	___	___	___	___
6.	___	___	___	___	___	___	___	___	___	___	___	___	___
7.	___	___	___	___	___	___	___	___	___	___	___	___	___
8.	___	___	___	___	___	___	___	___	___	___	___	___	___
9.	___	___	___	___	___	___	___	___	___	___	___	___	___
10.	___	___	___	___	___	___	___	___	___	___	___	___	___
11.	___	___	___	___	___	___	___	___	___	___	___	___	___
12.	___	___	___	___	___	___	___	___	___	___	___	___	___
13.	___	___	___	___	___	___	___	___	___	___	___	___	___
14.	___	___	___	___	___	___	___	___	___	___	___	___	___
15.	___	___	___	___	___	___	___	___	___	___	___	___	___
16.	___	___	___	___	___	___	___	___	___	___	___	___	___
17.	___	___	___	___	___	___	___	___	___	___	___	___	___
18.	___	___	___	___	___	___	___	___	___	___	___	___	___
19.	___	___	___	___	___	___	___	___	___	___	___	___	___
20.	___	___	___	___	___	___	___	___	___	___	___	___	___

Communicative Language Abilities checklist

Test: IELTS Part: Reading Comprehension 2
Rater: _____ Date: _____

0 = not involved, 1 = involved, 2 = critical basic, 3 = critical intermediate, 4 = critical advanced

Item #	LEX	MOR	STX	PG	COH	ORG	IDE	MAN	HEU	IMG	DIA	REG	STC
21.	___	___	___	___	___	___	___	___	___	___	___	___	___
22.	___	___	___	___	___	___	___	___	___	___	___	___	___
23.	___	___	___	___	___	___	___	___	___	___	___	___	___
24.	___	___	___	___	___	___	___	___	___	___	___	___	___
25.	___	___	___	___	___	___	___	___	___	___	___	___	___
26.	___	___	___	___	___	___	___	___	___	___	___	___	___
27.	___	___	___	___	___	___	___	___	___	___	___	___	___
28.	___	___	___	___	___	___	___	___	___	___	___	___	___

Test: IELTS Part: Reading Comprehension 3
Rater: _____ Date: _____

0 = not involved, 1 = involved, 2 = critical basic, 3 = critical intermediate, 4 = critical advanced

Item #	LEX	MOR	STX	PG	COH	ORG	IDE	MAN	HEU	IMG	DIA	REG	STC
29.	___	___	___	___	___	___	___	___	___	___	___	___	___
30.	___	___	___	___	___	___	___	___	___	___	___	___	___
31.	___	___	___	___	___	___	___	___	___	___	___	___	___
32.	___	___	___	___	___	___	___	___	___	___	___	___	___
33.	___	___	___	___	___	___	___	___	___	___	___	___	___
34.	___	___	___	___	___	___	___	___	___	___	___	___	___
35.	___	___	___	___	___	___	___	___	___	___	___	___	___
36.	___	___	___	___	___	___	___	___	___	___	___	___	___

Communicative Language Abilities checklist

Test: TOEFL Part: Reading Comprehension
Rater: _____ Date: _____

0 = not involved, 1 = involved, 2 = critical basic, 3 = critical intermediate, 4 = critical advanced

Item #	LEX	MOR	STX	PG	COH	ORG	IDE	MAN	HEU	IMG	DIA	REG	STC
31.	___	___	___	___	___	___	___	___	___	___	___	___	___
32.	___	___	___	___	___	___	___	___	___	___	___	___	___
33.	___	___	___	___	___	___	___	___	___	___	___	___	___
34.	___	___	___	___	___	___	___	___	___	___	___	___	___
35.	___	___	___	___	___	___	___	___	___	___	___	___	___
36.	___	___	___	___	___	___	___	___	___	___	___	___	___
37.	___	___	___	___	___	___	___	___	___	___	___	___	___
38.	___	___	___	___	___	___	___	___	___	___	___	___	___
39.	___	___	___	___	___	___	___	___	___	___	___	___	___
40.	___	___	___	___	___	___	___	___	___	___	___	___	___
41.	___	___	___	___	___	___	___	___	___	___	___	___	___
42.	___	___	___	___	___	___	___	___	___	___	___	___	___
43.	___	___	___	___	___	___	___	___	___	___	___	___	___
44.	___	___	___	___	___	___	___	___	___	___	___	___	___
45.	___	___	___	___	___	___	___	___	___	___	___	___	___
46.	___	___	___	___	___	___	___	___	___	___	___	___	___
47.	___	___	___	___	___	___	___	___	___	___	___	___	___
48.	___	___	___	___	___	___	___	___	___	___	___	___	___
49.	___	___	___	___	___	___	___	___	___	___	___	___	___
50.	___	___	___	___	___	___	___	___	___	___	___	___	___
51.	___	___	___	___	___	___	___	___	___	___	___	___	___
52.	___	___	___	___	___	___	___	___	___	___	___	___	___
53.	___	___	___	___	___	___	___	___	___	___	___	___	___
54.	___	___	___	___	___	___	___	___	___	___	___	___	___
55.	___	___	___	___	___	___	___	___	___	___	___	___	___
56.	___	___	___	___	___	___	___	___	___	___	___	___	___
57.	___	___	___	___	___	___	___	___	___	___	___	___	___
58.	___	___	___	___	___	___	___	___	___	___	___	___	___
59.	___	___	___	___	___	___	___	___	___	___	___	___	___
60.	___	___	___	___	___	___	___	___	___	___	___	___	___

Communicative Language Abilities checklist

Test: TOEFL
Part: Listening comprehension
Rater: _____
Date: _____

0 = not involved, 1 = involved, 2 = critical basic, 3 = critical intermediate, 4 = critical advanced

Item #	LEX	MOR	STX	PG	COH	ORG	IDE	MAN	HEU	IMG	DIA	REG	STC
1.													
2.													
3.													
4.													
5.													
6.													
7.													
8.													
9.													
10.													
11.													
12.													
13.													
14.													
15.													
16.													
17.													
18.													
19.													
20.													
21.													
22.													
23.													
24.													
25.													
26.													
27.													
28.													
29.													
30.													
31.													
32.													
33.													
34.													
35.													
36.													
37.													
38.													
39.													
40.													
41.													
42.													

43. ___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___
44. ___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___
45. ___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___
46. ___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___
47. ___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___
48. ___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___
49. ___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___
50. ___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___

Appendix 9:

The Details of Item Analysis: TOEFL, IELTS, & EPTB

Item and Test Analysis Program -- ITEMAN (tm) Version 3.50
Scale Definition Codes: DICHOT = Dichotomous MPOINT = Multipoint/Survey

TOEFL Item Analysis

Scale:	0	1	2
Type of Scale	DICHOT	DICHOT	DICHOT
N of Items	50	40	60
N of Examinees	127	127	127

***** CONFIGURATION INFORMATION *****
Type of Correlations: Point-Biserial
Correction for Spuriousness: YES
Ability Grouping: YES
Subgroup Analysis: NO
Express Endorsements As: PROPORTIONS
Score Group Interval Width: 2
*** Correlations have been corrected for spuriousness **

		Item Statistics			Alternative Statistics					
Seq. No.	Scale -Item	Prop. Correct	Disc. Index	Point Biser.	Alt.	Prop. Total	Endorsing Low	Endorsing High	Point Biser.	Key
1	0-1	.70	.20	.14	1	.09	.21	.05	-.21	
					2	.05	.06	.05	-.09	
					3	.70	.59	.79	.14	*
					4	.15	.15	.11	-.10	
					Other	.01	.00	.00	-.04	
2	0-2	.69	.32	.27	1	.69	.50	.82	.27	*
					2	.06	.09	.05	-.10	
					3	.20	.32	.13	-.26	
					4	.03	.09	.00	-.22	
					Other	.02	.00	.00	-.09	
3	0-3	.53	.47	.34	1	.07	.15	.03	-.28	
					2	.30	.41	.18	-.24	
					3	.09	.18	.03	-.20	
					4	.53	.26	.74	.34	*
					Other	.01	.00	.00	.06	
4	0-4	.67	.74	.54	1	.09	.21	.03	-.24	
					2	.67	.24	.97	.54	*
					3	.12	.24	.00	-.30	
					4	.13	.32	.00	-.40	
					Other	.00	.00	.00		
5	0-5	.61	.63	.45	1	.09	.15	.00	-.23	
					2	.17	.38	.05	-.37	
					3	.10	.21	.03	-.25	
					4	.61	.26	.89	.45	*
					Other	.03	.00	.00	.01	
6	0-6	.66	.23	.18	1	.09	.18	.00	-.26	
					2	.66	.56	.79	.18	*

					3	.13	.12	.16	-.05	
					4	.09	.15	.00	-.25	
					Other	.03	.00	.00	.10	
7	0-7	.57	.74	.58	1	.57	.18	.92	.58	*
					2	.05	.09	.03	-.16	
					3	.07	.18	.00	-.23	
					4	.30	.50	.05	-.49	
					Other	.02	.00	.00	-.16	
8	0-8	.69	.49	.35	1	.14	.35	.05	-.31	
					2	.13	.24	.08	-.22	
					3	.69	.38	.87	.35	*
					4	.04	.03	.00	-.14	
					Other	.01	.00	.00	-.05	
9	0-9	.71	.37	.31	1	.09	.15	.03	-.23	
					2	.15	.24	.05	-.27	
					3	.03	.06	.00	-.12	
					4	.71	.53	.89	.31	*
					Other	.02	.00	.00	-.07	
10	0-10	.48	.47	.36	1	.17	.24	.05	-.22	
					2	.48	.26	.74	.36	*
					3	.14	.09	.11	-.07	
					4	.17	.35	.11	-.31	
					Other	.03	.00	.00	-.17	
11	0-11	.72	.42	.29	1	.08	.09	.03	-.05	
					2	.09	.21	.03	-.30	
					3	.10	.21	.03	-.25	
					4	.72	.50	.92	.29	*
					Other	.01	.00	.00	-.04	
12	0-12	.83	.39	.37	1	.03	.06	.00	-.20	
					2	.83	.59	.97	.37	*
					3	.06	.21	.03	-.29	
					4	.07	.12	.00	-.22	
					Other	.01	.00	.00	-.13	
13	0-13	.57	.18	.10	1	.22	.12	.26	.08	
					2	.17	.32	.03	-.34	
					3	.57	.50	.68	.10	*
					4	.04	.06	.03	-.11	
					Other	.00	.00	.00		
14	0-14	.47	.53	.44	1	.47	.18	.71	.44	*
					2	.08	.03	.11	-.00	
					3	.22	.38	.13	-.34	
					4	.22	.41	.05	-.34	
					Other	.01	.00	.00	-.04	
15	0-15	.50	.64	.46	1	.09	.21	.00	-.24	
					2	.11	.12	.03	-.16	
					3	.50	.21	.84	.46	*
					4	.28	.44	.13	-.36	
					Other	.01	.00	.00	-.15	
16	0-16	.34	.43	.33	1	.34	.15	.58	.33	*
					2	.39	.32	.32	-.06	
					3	.17	.38	.05	-.38	
					4	.09	.15	.03	-.23	
					Other	.02	.00	.00	.03	

17	0-17	.37	.24	.16	1	.25	.26	.26	-.05	
					2	.37	.26	.50	.16	*
					3	.06	.09	.05	-.11	
					4	.28	.29	.18	-.16	
					Other	.04	.00	.00	-.20	
18	0-18	.53	.66	.45	1	.09	.12	.00	-.18	
					2	.13	.29	.00	-.35	
					3	.24	.29	.08	-.25	
					4	.53	.26	.92	.45	*
					Other	.01	.00	.00	-.12	
19	0-19	.58	.24	.12	1	.08	.12	.05	-.06	
					2	.58	.47	.71	.12	*
					3	.21	.26	.21	-.10	
					4	.12	.12	.03	-.19	
					Other	.01	.00	.00	-.12	
20	0-20	.35	.18	.15	1	.24	.24	.13	-.21	
					2	.24	.15	.32	.09	
					3	.35	.26	.45	.15	*
					4	.16	.32	.11	-.24	
					Other	.02	.00	.00	-.12	
21	0-21	.80	.22	.19	1	.06	.06	.00	-.11	
					2	.10	.18	.03	-.23	
					3	.04	.06	.05	-.09	
					4	.80	.71	.92	.19	*
					Other	.01	.00	.00	-.04	
22	0-22	.50	.61	.39	1	.08	.18	.03	-.23	
					2	.29	.38	.16	-.24	
					3	.50	.18	.79	.39	*
					4	.10	.24	.00	-.30	
					Other	.02	.00	.00	-.01	
23	0-23	.62	.24	.19	1	.62	.47	.71	.19	*
					2	.14	.09	.11	-.10	
					3	.19	.35	.16	-.20	
					4	.04	.06	.03	-.15	
					Other	.01	.00	.00	-.15	
24	0-24	.69	.47	.35	1	.09	.21	.03	-.23	
					2	.69	.50	.97	.35	*
					3	.14	.18	.00	-.28	
					4	.06	.09	.00	-.14	
					Other	.02	.00	.00	-.13	
25	0-25	.50	.36	.23	1	.13	.15	.08	-.12	
					2	.18	.29	.03	-.32	
					3	.18	.21	.18	-.07	
					4	.50	.35	.71	.23	*
					Other	.00	.00	.00		
26	0-26	.47	.67	.51	1	.18	.15	.16	-.09	
					2	.47	.15	.82	.51	*
					3	.18	.44	.00	-.47	
					4	.15	.26	.03	-.28	
					Other	.02	.00	.00	-.06	
27	0-27	.50	.55	.43	1	.50	.29	.84	.43	*
					2	.16	.26	.08	-.26	
					3	.15	.15	.05	-.19	
					4	.16	.26	.03	-.29	

					Other	.03	.00	.00	-.10	
28	0-28	.73	.54	.42	1	.06	.12	.03	-.18	
					2	.06	.12	.00	-.21	
					3	.73	.41	.95	.42	*
					4	.16	.35	.03	-.38	
					Other	.00	.00	.00		
29	0-29	.46	.55	.42	1	.15	.24	.00	-.25	
					2	.12	.21	.03	-.26	
					3	.25	.26	.16	-.22	
					4	.46	.26	.82	.42	*
					Other	.02	.00	.00	-.08	
30	0-30	.39	.50	.34	1	.39	.35	.21	-.19	
					2	.09	.15	.03	-.16	
					3	.13	.29	.05	-.30	
					4	.39	.21	.71	.34	*
					Other	.00	.00	.00		
31	0-31	.60	.35	.27	1	.60	.41	.76	.27	*
					2	.06	.15	.00	-.24	
					3	.05	.06	.03	-.11	
					4	.28	.32	.21	-.20	
					Other	.02	.00	.00	-.19	
32	0-32	.39	.20	.14	1	.23	.26	.11	-.14	
					2	.20	.24	.16	-.17	
					3	.39	.32	.53	.14	*
					4	.18	.18	.21	-.05	
					Other	.01	.00	.00	-.04	
33	0-33	.29	.33	.20	1	.06	.12	.00	-.21	
					2	.53	.47	.53	-.01	
					3	.13	.29	.03	-.32	
					4	.29	.12	.45	.20	*
					Other	.00	.00	.00		
34	0-34	.28	.30	.24	1	.09	.12	.11	-.05	
					2	.21	.21	.21	-.04	
					3	.28	.15	.45	.24	*
					4	.39	.53	.21	-.32	
					Other	.02	.00	.00	.03	
35	0-35	.30	.16	.11	1	.45	.44	.47	-.07	
					2	.09	.18	.03	-.21	
					3	.30	.24	.39	.11	*
					4	.16	.15	.11	-.08	
					Other	.01	.00	.00	-.05	
36	0-36	.44	.42	.32	1	.23	.29	.11	-.20	
					2	.13	.15	.08	-.16	
					3	.20	.32	.16	-.24	
					4	.44	.24	.66	.32	*
					Other	.00	.00	.00		
37	0-37	.50	.45	.25	1	.24	.24	.24	-.02	
					2	.09	.24	.00	-.31	
					3	.50	.24	.68	.25	*
					4	.17	.29	.08	-.26	
					Other	.00	.00	.00		

38	0-38	.64	.62	.47	1	.09	.18	.00	-.32	
					2	.13	.21	.00	-.29	
					3	.64	.35	.97	.47	*
					4	.14	.26	.03	-.26	
					Other	.00	.00	.00		
39	0-39	.59	.46	.30	1	.06	.12	.00	-.20	
					2	.59	.41	.87	.30	*
					3	.15	.29	.05	-.23	
					4	.20	.15	.08	-.17	
					Other	.01	.00	.00	-.13	
40	0-40	.51	.46	.30	1	.51	.35	.82	.30	*
					2	.06	.09	.03	-.03	
					3	.20	.24	.11	-.17	
					4	.22	.32	.05	-.34	
					Other	.01	.00	.00	-.04	
41	0-41	.31	.21	.18	1	.17	.12	.18	.01	
					2	.09	.21	.00	-.33	
					3	.31	.29	.50	.18	*
					4	.39	.35	.32	-.12	
					Other	.02	.00	.00	-.07	
42	0-42	.36	.04	.01	1	.25	.21	.39	.06	?
					2	.36	.32	.37	.01	*
	CHECK THE KEY				3	.22	.24	.18	-.08	
	2 was specified, 1 works better				4	.15	.24	.05	-.23	
					Other	.02	.00	.00	-.00	
43	0-43	.58	.80	.58	1	.16	.41	.03	-.42	
					2	.58	.15	.95	.58	*
					3	.24	.35	.03	-.35	
					4	.02	.09	.00	-.24	
					Other	.00	.00	.00		
44	0-44	.65	.35	.33	1	.09	.15	.05	-.23	
					2	.65	.47	.82	.33	*
					3	.06	.09	.03	-.14	
					4	.20	.29	.11	-.28	
					Other	.01	.00	.00	-.05	
45	0-45	.39	.29	.11	1	.39	.24	.53	.11	*
					2	.28	.24	.32	.07	
					3	.23	.35	.11	-.28	
					4	.09	.15	.05	-.12	
					Other	.01	.00	.00	-.12	
46	0-46	.40	.65	.52	1	.40	.09	.74	.52	*
					2	.18	.38	.05	-.40	
					3	.13	.15	.05	-.19	
					4	.25	.35	.16	-.19	
					Other	.03	.00	.00	-.14	
47	0-47	.65	.57	.42	1	.08	.15	.00	-.22	
					2	.20	.32	.08	-.29	
					3	.65	.35	.92	.42	*
					4	.05	.12	.00	-.22	
					Other	.02	.00	.00	-.20	
48	0-48	.59	.69	.55	1	.13	.15	.11	-.12	
					2	.17	.35	.00	-.43	
					3	.10	.29	.00	-.40	
					4	.59	.21	.89	.55	*

					Other	.01	.00	.00	-.05	
49	0-49	.54	.38	.26	1	.13	.09	.03	-.18	
					2	.54	.35	.74	.26	*
					3	.20	.47	.11	-.30	
					4	.13	.09	.13	-.05	
					Other	.02	.00	.00	-.06	
50	0-50	.56	.41	.31	1	.05	.03	.05	.01	
					2	.21	.35	.05	-.35	
					3	.56	.32	.74	.31	*
					4	.17	.29	.16	-.18	
					Other	.01	.00	.00	-.05	
51	1-1	.76	.35	.25	1	.02	.03	.00	-.14	
					2	.76	.62	.97	.25	*
					3	.11	.24	.00	-.34	
					4	.12	.11	.03	-.13	
					Other	.00	.00	.00		
52	1-2	.91	.22	.30	1	.06	.11	.00	-.27	
					2	.91	.78	1.00	.30	*
					3	.01	.03	.00	-.18	
					4	.02	.08	.00	-.24	
					Other	.01	.00	.00	-.01	
53	1-3	.84	.35	.32	1	.84	.65	1.00	.32	*
					2	.06	.16	.00	-.35	
					3	.08	.14	.00	-.17	
					4	.01	.03	.00	-.11	
					Other	.02	.00	.00	-.15	
54	1-4	.94	.14	.28	1	.94	.86	1.00	.28	*
					2	.02	.05	.00	-.25	
					3	.04	.08	.00	-.24	
					4	.00	.00	.00		
					Other	.00	.00	.00		
55	1-5	.80	.30	.24	1	.02	.05	.00	-.15	
					2	.80	.62	.92	.24	*
					3	.11	.14	.08	-.09	
					4	.06	.16	.00	-.32	
					Other	.01	.00	.00	-.19	
56	1-6	.83	.32	.26	1	.02	.05	.00	-.12	
					2	.83	.68	1.00	.26	*
					3	.11	.22	.00	-.33	
					4	.03	.05	.00	-.13	
					Other	.01	.00	.00	-.01	
57	1-7	.77	.46	.44	1	.09	.22	.03	-.35	
					2	.08	.19	.00	-.36	
					3	.77	.51	.97	.44	*
					4	.05	.08	.00	-.17	
					Other	.01	.00	.00	-.01	
58	1-8	.85	.24	.20	1	.03	.05	.00	-.14	
					2	.10	.22	.08	-.18	
					3	.01	.03	.00	-.15	
					4	.85	.68	.92	.20	*
					Other	.01	.00	.00	-.21	

59	1-9	.80	.51	.48	1	.16	.35	.00	-.41	
					2	.04	.14	.00	-.36	
					3	.01	.03	.00	-.17	
					4	.80	.49	1.00	.48	*
					Other	.00	.00	.00		
60	1-10	.70	.38	.27	1	.01	.03	.00	-.11	
					2	.70	.54	.92	.27	*
					3	.13	.16	.03	-.13	
					4	.14	.24	.05	-.33	
					Other	.02	.00	.00	-.16	
61	1-11	.68	.41	.28	1	.07	.08	.05	-.10	
					2	.20	.35	.05	-.36	
					3	.68	.46	.86	.28	*
					4	.04	.11	.00	-.18	
					Other	.02	.00	.00	.02	
62	1-12	.42	.41	.26	1	.41	.41	.22	-.21	
					2	.42	.27	.68	.26	*
					3	.06	.22	.00	-.43	
					4	.10	.11	.08	-.06	
					Other	.01	.00	.00	.05	
63	1-13	.60	.65	.52	1	.06	.08	.05	-.14	
					2	.60	.24	.89	.52	*
					3	.14	.27	.00	-.32	
					4	.19	.41	.05	-.46	
					Other	.01	.00	.00	-.02	
64	1-14	.58	.46	.38	1	.38	.62	.16	-.49	
					2	.00	.00	.00		
					3	.03	.03	.03	-.07	
					4	.58	.35	.81	.38	*
					Other	.01	.00	.00	-.02	
65	1-15	.18	.19	.19	1	.04	.03	.00	-.07	
					2	.67	.68	.68	-.10	
					3	.10	.16	.03	-.26	
					4	.18	.11	.30	.19	*
					Other	.01	.00	.00	-.11	
66	1-16	.85	.43	.46	1	.85	.57	1.00	.46	*
					2	.03	.11	.00	-.23	
					3	.06	.19	.00	-.39	
					4	.03	.08	.00	-.22	
					Other	.02	.00	.00	-.18	
67	1-17	.72	.51	.43	1	.14	.32	.00	-.44	
					2	.01	.03	.00	-.07	
					3	.13	.22	.05	-.26	
					4	.72	.43	.95	.43	*
					Other	.00	.00	.00		
68	1-18	.79	.41	.38	1	.05	.05	.00	-.13	
					2	.08	.14	.05	-.18	
					3	.79	.54	.95	.38	*
					4	.09	.27	.00	-.46	
					Other	.00	.00	.00		
69	1-19	.81	.00	-.05	1	.81	.84	.84	-.05	*
					2	.03	.03	.03	-.04	
					3	.03	.03	.03	-.04	

					4	.13	.11	.11	-.04	
					Other	.00	.00	.00		
70	1-20	.78	.27	.28	1	.78	.62	.89	.28	*
					2	.02	.05	.00	-.17	
					3	.13	.14	.11	-.14	
					4	.07	.19	.00	-.37	
					Other	.00	.00	.00		
71	1-21	.80	.27	.19	1	.09	.11	.05	-.10	
					2	.80	.65	.92	.19	*
					3	.06	.16	.00	-.29	
					4	.06	.08	.03	-.14	
					Other	.00	.00	.00		
72	1-22	.55	.57	.31	1	.06	.08	.03	-.12	
					2	.27	.51	.11	-.37	
					3	.55	.24	.81	.31	*
					4	.12	.16	.05	-.14	
					Other	.00	.00	.00		
73	1-23	.80	.19	.19	1	.02	.05	.00	-.17	
					2	.09	.19	.05	-.28	
					3	.08	.05	.08	-.06	
					4	.80	.68	.86	.19	*
					Other	.01	.00	.00	-.07	
74	1-24	.75	.32	.21	1	.07	.08	.05	-.12	
					2	.75	.57	.89	.21	*
					3	.09	.14	.03	-.18	
					4	.09	.22	.03	-.23	
					Other	.00	.00	.00		
75	1-25	.63	.22	.14	1	.20	.19	.16	-.08	
					2	.08	.22	.00	-.41	
					3	.63	.51	.73	.14	*
					4	.08	.05	.11	.04	
					Other	.01	.00	.00	-.19	
76	1-26	.69	.54	.40	1	.18	.30	.08	-.23	
					2	.09	.24	.00	-.46	
					3	.69	.38	.92	.40	*
					4	.03	.03	.00	-.05	
					Other	.02	.00	.00	-.22	
77	1-27	.69	.49	.34	1	.02	.08	.00	-.27	
					2	.12	.19	.03	-.24	
					3	.69	.43	.92	.34	*
					4	.14	.22	.05	-.18	
					Other	.02	.00	.00	-.30	
78	1-28	.63	.51	.38	1	.07	.08	.05	-.09	
					2	.18	.32	.05	-.33	
					3	.10	.19	.03	-.27	
					4	.63	.35	.86	.38	*
					Other	.02	.00	.00	-.22	
79	1-29	.69	.43	.40	1	.13	.14	.11	-.09	
					2	.06	.11	.03	-.23	
					3	.10	.27	.03	-.43	
					4	.69	.41	.84	.40	*
					Other	.02	.00	.00	-.26	
80	1-30	.57	.65	.46	1	.57	.32	.97	.46	*

					2	.10	.16	.03	-.31	
					3	.25	.35	.00	-.31	
					4	.05	.05	.00	-.15	
					Other	.03	.00	.00	-.28	
81	1-31	.76	.57	.50	1	.02	.05	.00	-.25	
					2	.76	.41	.97	.50	*
					3	.02	.08	.00	-.24	
					4	.14	.30	.03	-.35	
					Other	.05	.00	.00	-.31	
82	1-32	.53	.54	.37	1	.09	.11	.05	-.17	
					2	.31	.41	.14	-.26	
					3	.53	.27	.81	.37	*
					4	.02	.08	.00	-.30	
					Other	.05	.00	.00	-.26	
83	1-33	.70	.38	.33	1	.70	.46	.84	.33	*
					2	.16	.14	.14	-.11	
					3	.06	.22	.00	-.42	
					4	.04	.05	.03	-.03	
					Other	.04	.00	.00	-.34	
84	1-34	.61	.59	.38	1	.13	.27	.00	-.36	
					2	.14	.22	.05	-.21	
					3	.61	.27	.86	.38	*
					4	.09	.14	.08	-.08	
					Other	.03	.00	.00	-.28	
85	1-35	.64	.51	.42	1	.03	.11	.00	-.24	
					2	.07	.22	.00	-.45	
					3	.24	.30	.16	-.21	
					4	.64	.32	.84	.42	*
					Other	.02	.00	.00	-.18	
86	1-36	.58	.49	.38	1	.58	.30	.78	.38	*
					2	.09	.22	.00	-.31	
					3	.20	.24	.14	-.25	
					4	.08	.11	.08	-.07	
					Other	.05	.00	.00	-.28	
87	1-37	.65	.46	.34	1	.19	.30	.08	-.24	
					2	.65	.35	.81	.34	*
					3	.09	.19	.08	-.25	
					4	.02	.03	.03	-.01	
					Other	.06	.00	.00	-.30	
88	1-38	.35	.46	.35	1	.09	.11	.05	-.11	
					2	.28	.35	.27	-.16	
					3	.21	.24	.03	-.28	
					4	.35	.16	.62	.35	*
					Other	.06	.00	.00	-.22	
89	1-39	.37	.27	.18	1	.17	.08	.27	.15	
					2	.24	.27	.08	-.23	
					3	.18	.27	.08	-.29	
					4	.37	.27	.54	.18	*
					Other	.05	.00	.00	-.17	
90	1-40	.80	.35	.32	1	.01	.00	.03	.09	
					2	.05	.08	.03	-.08	
					3	.10	.24	.00	-.41	
					4	.80	.57	.92	.32	*

					Other	.05	.00	.00	-.21	
91	2-1	.86	.26	.34	1	.86	.71	.97	.34	*
					2	.00	.00	.00		
					3	.08	.17	.03	-.29	
					4	.06	.12	.00	-.29	
					Other	.00	.00	.00		
92	2-2	.89	.20	.13	1	.01	.02	.00	-.06	
					2	.08	.12	.00	-.13	
					3	.89	.80	1.00	.13	*
					4	.02	.05	.00	-.17	
					Other	.00	.00	.00		
93	2-3	.95	.05	.10	1	.01	.02	.00	-.17	
					2	.02	.02	.00	-.09	
					3	.95	.93	.97	.10	*
					4	.01	.02	.00	-.18	
					Other	.01	.00	.00	.14	
94	2-4	.91	.09	.06	1	.91	.85	.94	.06	*
					2	.08	.12	.06	-.06	
					3	.00	.00	.00		
					4	.00	.00	.00		
					Other	.02	.00	.00	-.18	
95	2-5	.84	.21	.23	1	.84	.73	.94	.23	*
					2	.00	.00	.00		
					3	.10	.17	.03	-.19	
					4	.05	.10	.03	-.26	
					Other	.01	.00	.00	-.05	
96	2-6	.87	.17	.27	1	.06	.02	.03	-.07	
					2	.02	.05	.00	-.23	
					3	.05	.10	.00	-.25	
					4	.87	.80	.97	.27	*
					Other	.02	.00	.00	-.18	
97	2-7	.79	.15	.15	1	.15	.20	.14	-.12	
					2	.04	.10	.00	-.25	
					3	.01	.00	.03	.13	
					4	.79	.68	.83	.15	*
					Other	.02	.00	.00	-.18	
98	2-8	.90	.22	.36	1	.02	.02	.00	-.11	
					2	.06	.15	.00	-.33	
					3	.02	.02	.00	-.15	
					4	.90	.78	1.00	.36	*
					Other	.02	.00	.00	-.18	
99	2-9	.69	.43	.38	1	.15	.17	.06	-.15	
					2	.05	.12	.03	-.28	
					3	.69	.49	.92	.38	*
					4	.09	.20	.00	-.31	
					Other	.02	.00	.00	-.18	
100	2-10	.63	.42	.24	1	.08	.10	.06	-.10	
					2	.20	.24	.08	-.15	
					3	.08	.20	.00	-.29	
					4	.63	.44	.86	.24	*
					Other	.01	.00	.00	-.21	
101	2-11	.61	.37	.28	1	.11	.17	.00	-.19	

					2	.14	.17	.08	-.14	
					3	.11	.12	.06	-.21	
					4	.61	.49	.86	.28	*
					Other	.02	.00	.00	-.19	
102	2-12	.73	.30	.26	1	.09	.12	.08	-.14	
					2	.08	.15	.03	-.22	
					3	.08	.12	.00	-.20	
					4	.73	.59	.89	.26	*
					Other	.02	.00	.00	-.12	
103	2-13	.96	.12	.32	1	.02	.05	.00	-.27	
					2	.02	.05	.00	-.17	
					3	.96	.88	1.00	.32	*
					4	.01	.02	.00	-.20	
					Other	.00	.00	.00		
104	2-14	.53	.25	.22	1	.09	.15	.06	-.18	
					2	.53	.39	.64	.22	*
					3	.24	.24	.22	-.15	
					4	.13	.22	.08	-.21	
					Other	.01	.00	.00	-.02	
105	2-15	.38	.14	.09	1	.38	.24	.39	.09	*
					2	.11	.15	.06	-.19	
					3	.38	.44	.47	-.06	
					4	.13	.17	.08	-.12	
					Other	.00	.00	.00		
106	2-16	.65	.27	.16	1	.13	.12	.06	-.08	
					2	.07	.17	.00	-.28	
					3	.65	.54	.81	.16	*
					4	.15	.17	.14	-.12	
					Other	.00	.00	.00		
107	2-17	.78	.36	.35	1	.78	.56	.92	.35	*
					2	.06	.12	.03	-.21	
					3	.06	.10	.03	-.21	
					4	.09	.20	.03	-.26	
					Other	.01	.00	.00	-.15	
108	2-18	.73	.28	.22	1	.08	.20	.03	-.30	
					2	.73	.59	.86	.22	*
					3	.11	.17	.03	-.23	
					4	.07	.02	.08	.06	
					Other	.01	.00	.00	-.11	
109	2-19	.45	.13	.11	1	.03	.05	.00	-.11	
					2	.45	.32	.44	.11	*
					3	.48	.51	.56	-.10	
					4	.04	.12	.00	-.29	
					Other	.00	.00	.00		
110	2-20	.50	.44	.29	1	.50	.37	.81	.29	*
					2	.08	.12	.06	-.17	
					3	.06	.10	.00	-.20	
					4	.35	.41	.14	-.26	
					Other	.00	.00	.00		
111	2-21	.65	.34	.25	1	.13	.20	.06	-.13	
					2	.11	.22	.08	-.28	
					3	.10	.12	.08	-.14	
					4	.65	.44	.78	.25	*
					Other	.01	.00	.00	-.10	

112	2-22	.55	.14	.15	1	.30	.39	.28	-.14	
					2	.55	.44	.58	.15	*
					3	.11	.10	.11	-.14	
					4	.03	.05	.03	-.15	
					Other	.01	.00	.00	-.17	
113	2-23	.79	.31	.28	1	.09	.17	.00	-.33	
					2	.06	.10	.00	-.12	
					3	.79	.63	.94	.28	*
					4	.06	.10	.06	-.13	
					Other	.00	.00	.00		
114	2-24	.57	.41	.25	1	.28	.29	.19	-.13	
					2	.02	.05	.00	-.12	
					3	.13	.29	.06	-.30	
					4	.57	.34	.75	.25	*
					Other	.01	.00	.00	-.21	
115	2-25	.62	.45	.35	1	.05	.07	.00	-.22	
					2	.62	.41	.86	.35	*
					3	.22	.32	.11	-.29	
					4	.08	.12	.03	-.12	
					Other	.03	.00	.00	-.21	
116	2-26	.44	.10	.03	1	.44	.34	.44	.03	*
					2	.35	.37	.31	-.02	
					3	.11	.12	.14	-.11	
					4	.08	.10	.11	-.02	
					Other	.02	.00	.00	-.31	
117	2-27	.34	.41	.29	1	.17	.07	.19	.11	
					2	.34	.15	.56	.29	*
					3	.14	.27	.08	-.26	
					4	.31	.41	.17	-.25	
					Other	.04	.00	.00	-.28	
118	2-28	.46	.16	.11	1	.02	.00	.03	.09	
					2	.16	.22	.14	-.13	
					3	.35	.41	.33	-.18	
					4	.46	.34	.50	.11	*
					Other	.01	.00	.00	-.15	
119	2-29	.31	.36	.27	1	.13	.12	.08	-.11	
					2	.21	.22	.28	-.03	
					3	.31	.17	.53	.27	*
					4	.31	.39	.11	-.25	
					Other	.03	.00	.00	-.28	
120	2-30	.47	.48	.31	1	.24	.32	.19	-.15	
					2	.20	.22	.06	-.19	
					3	.47	.27	.75	.31	*
					4	.06	.10	.00	-.20	
					Other	.03	.00	.00	-.28	
121	2-31	.80	.36	.32	1	.03	.05	.00	-.12	
					2	.04	.10	.00	-.24	
					3	.80	.61	.97	.32	*
					4	.13	.22	.03	-.28	
					Other	.01	.00	.00	-.11	
122	2-32	.33	.20	.15	1	.33	.22	.42	.15	*
					2	.14	.24	.06	-.21	
					3	.41	.32	.50	.06	
					4	.09	.15	.03	-.26	

					Other	.02	.00	.00	-.22	
123	2-33	.57	.62	.46	1	.57	.24	.86	.46	*
					2	.14	.20	.06	-.25	
					3	.13	.22	.06	-.18	
					4	.11	.22	.03	-.28	
					Other	.04	.00	.00	-.31	
124	2-34	.89	.19	.28	1	.89	.76	.94	.28	*
					2	.03	.05	.00	-.20	
					3	.06	.12	.06	-.19	
					4	.01	.02	.00	-.15	
					Other	.02	.00	.00	-.17	
125	2-35	.75	.25	.21	1	.75	.61	.86	.21	*
					2	.10	.22	.06	-.23	
					3	.09	.07	.06	-.08	
					4	.05	.07	.03	-.17	
					Other	.01	.00	.00	-.11	
126	2-36	.94	.15	.27	1	.00	.00	.00		
					2	.02	.02	.00	-.12	
					3	.03	.10	.00	-.30	
					4	.94	.85	1.00	.27	*
					Other	.01	.00	.00	-.11	
127	2-37	.90	.22	.37	1	.00	.00	.00		
					2	.02	.05	.00	-.21	
					3	.90	.76	.97	.37	*
					4	.06	.12	.03	-.27	
					Other	.02	.00	.00	-.27	
128	2-38	.61	.37	.28	1	.14	.20	.11	-.11	
					2	.06	.05	.03	-.07	
					3	.61	.44	.81	.28	*
					4	.17	.27	.06	-.31	
					Other	.02	.00	.00	-.22	
129	2-39	.10	.01	-.02	1	.07	.10	.08	-.10	
					2	.76	.71	.75	-.00	
					3	.10	.10	.11	-.02	*
					4	.06	.10	.06	-.07	
					Other	.00	.00	.00		
130	2-40	.27	.04	.01	1	.40	.34	.33	-.06	
					2	.11	.10	.11	-.10	
					3	.27	.32	.36	.01	*
					4	.22	.24	.19	-.07	
					Other	.00	.00	.00		
131	2-41	.72	.41	.40	1	.02	.05	.00	-.23	
					2	.06	.12	.00	-.22	
					3	.19	.24	.06	-.29	
					4	.72	.54	.94	.40	*
					Other	.02	.00	.00	-.24	
132	2-42	.39	.10	.01	1	.01	.02	.00	-.06	
					2	.32	.39	.36	-.08	
					3	.28	.27	.22	-.09	
					4	.39	.32	.42	.01	*
					Other	.00	.00	.00		
133	2-43	.79	.14	.09	1	.79	.78	.92	.09	*
					2	.04	.07	.03	-.10	

					3	.07	.12	.00	-.26	
					4	.09	.00	.06	.05	
					Other	.01	.00	.00	-.11	
134	2-44	.70	.30	.19	1	.16	.15	.08	-.12	
					2	.11	.22	.00	-.30	
					3	.70	.59	.89	.19	*
					4	.01	.02	.00	-.06	
					Other	.02	.00	.00	-.03	
135	2-45	.64	.39	.28	1	.17	.24	.06	-.26	
					2	.64	.44	.83	.28	*
					3	.10	.17	.00	-.22	
					4	.06	.10	.08	-.08	
					Other	.02	.00	.00	-.10	
136	2-46	.61	.65	.48	1	.02	.07	.00	-.22	
					2	.61	.29	.94	.48	*
					3	.09	.17	.03	-.28	
					4	.26	.41	.03	-.36	
					Other	.02	.00	.00	-.14	
137	2-47	.72	.48	.42	1	.72	.44	.92	.42	*
					2	.09	.12	.03	-.24	
					3	.06	.17	.00	-.33	
					4	.11	.20	.06	-.21	
					Other	.02	.00	.00	-.17	
138	2-48	.60	.63	.45	1	.60	.32	.94	.45	*
					2	.17	.27	.00	-.29	
					3	.08	.15	.03	-.24	
					4	.09	.15	.00	-.26	
					Other	.06	.00	.00	-.16	
139	2-49	.50	.21	.19	1	.24	.20	.31	-.01	
					2	.17	.24	.11	-.22	
					3	.50	.34	.56	.19	*
					4	.04	.10	.00	-.24	
					Other	.05	.00	.00	-.20	
140	2-50	.43	.56	.44	1	.24	.29	.14	-.20	
					2	.16	.15	.06	-.19	
					3	.14	.24	.03	-.28	
					4	.43	.22	.78	.44	*
					Other	.04	.00	.00	-.23	
141	2-51	.68	.42	.36	1	.09	.15	.06	-.19	
					2	.68	.44	.86	.36	*
					3	.13	.22	.06	-.28	
					4	.06	.10	.03	-.15	
					Other	.04	.00	.00	-.21	
142	2-52	.61	.52	.39	1	.08	.07	.03	-.14	
					2	.61	.37	.89	.39	*
					3	.20	.29	.08	-.26	
					4	.03	.07	.00	-.16	
					Other	.08	.00	.00	-.31	
143	2-53	.49	.69	.49	1	.14	.17	.03	-.17	
					2	.09	.12	.00	-.26	
					3	.20	.29	.08	-.25	
					4	.49	.20	.89	.49	*
					Other	.08	.00	.00	-.33	

144	2-54	.32	.39	.29	1	.24	.22	.17	-.15	
					2	.17	.22	.08	-.19	
					3	.32	.24	.64	.29	*
					4	.20	.12	.11	-.02	
					Other	.08	.00	.00	-.31	
145	2-55	.53	.53	.38	1	.13	.22	.06	-.28	
					2	.14	.12	.11	-.07	
					3	.11	.22	.06	-.21	
					4	.53	.24	.78	.38	*
					Other	.09	.00	.00	-.28	
146	2-56	.36	.18	.07	1	.09	.17	.03	-.23	
					2	.43	.29	.53	.12	?
	CHECK THE KEY				3	.36	.27	.44	.07	*
	3 was specified, 2 works better				4	.03	.07	.00	-.11	
					Other	.09	.00	.00	-.31	
147	2-57	.39	.45	.27	1	.39	.27	.72	.27	*
					2	.20	.17	.17	-.08	
					3	.27	.27	.11	-.11	
					4	.03	.07	.00	-.19	
					Other	.11	.00	.00	-.32	
148	2-58	.35	.35	.21	1	.20	.20	.25	-.05	
					2	.35	.12	.47	.21	*
					3	.26	.37	.22	-.11	
					4	.08	.10	.03	-.12	
					Other	.12	.00	.00	-.27	
149	2-59	.41	.34	.20	1	.10	.20	.06	-.23	
					2	.41	.22	.56	.20	*
					3	.24	.24	.22	-.08	
					4	.13	.12	.14	.01	
					Other	.12	.00	.00	-.27	
150	2-60	.40	.44	.26	1	.14	.12	.14	-.01	
					2	.40	.20	.64	.26	*
					3	.21	.32	.11	-.19	
					4	.13	.15	.08	-.15	
					Other	.11	.00	.00	-.27	

There were 127 examinees in the data file.

Scale Statistics

Scale:	0 = LC 0	1 = ST 1	2 = RC 2
N of Items	50	40	60
N of Examinees	127	127	127
Mean	26.819	27.386	36.638
Variance	76.337	43.686	64.074
Std. Dev.	8.737	6.610	8.005
Skew	0.212	-0.409	-0.159
Kurtosis	-0.944	-0.638	-0.451
Minimum	9.000	13.000	18.000
Maximum	47.000	40.000	57.000
Median	27.000	29.000	37.000
Alpha	0.867	0.846	0.829
SEM	3.192	2.595	3.308
Mean P	0.536	0.685	0.611
Mean Item-Tot.	0.315	0.319	0.251
Mean Biserial	0.404	0.437	0.346
Max Score (Low)	19	23	32
N (Low Group)	34	37	41
Min Score (High)	33	32	42
N (High Group)	38	37	36

Scale Intercorrelations

	0	1	2
0	1.000	0.595	0.592
1	0.595	1.000	0.758
2	0.592	0.758	1.000

SCALE # 0 Score Distribution Table

Score Interval	Frequency	Cum Freq	PR	PCT	
. . . No examinees below this score . . .					
7 - 8	0	0	1	0	|
9 - 10	1	1	1	1	+#
11 - 12	1	2	2	1	|#
13 - 14	5	7	6	4	|####
15 - 16	9	16	13	7	|#######
17 - 18	12	28	22	9	|#########
19 - 20	11	39	31	9	+#########
21 - 22	8	47	37	6	|######
23 - 24	11	58	46	9	|#########
25 - 26	5	63	50	4	|####
27 - 28	12	75	59	9	|#########
29 - 30	7	82	65	6	+######
31 - 32	7	89	70	6	|######
33 - 34	6	95	75	5	|#####
35 - 36	11	106	83	9	|#########
37 - 38	10	116	91	8	|########
39 - 40	3	119	94	2	+##
41 - 42	4	123	97	3	|###
43 - 44	2	125	98	2	|##
45 - 46	1	126	99	1	|#
47 - 48	1	127	99	1	|#
49 - 50	0	127	99	0	+

```
|----+----+----+----+--
  5    10   15   20  25
 Percentage of Examinees
```

SCALE # 1 Score Distribution Table

Score Interval	Freq-uency	Cum Freq	PR	PCT	
. . . No examinees below this score . . .					
11 - 12	0	0	1	0	|
13 - 14	4	4	3	3	|###
15 - 16	8	12	9	6	|######
17 - 18	3	15	12	2	|##
19 - 20	9	24	19	7	+#######
21 - 22	5	29	23	4	|####
23 - 24	11	40	31	9	|#########
25 - 26	10	50	39	8	|########
27 - 28	11	61	48	9	|#########
29 - 30	21	82	65	17	+##################
31 - 32	16	98	77	13	|##############
33 - 34	12	110	87	9	|#########
35 - 36	10	120	94	8	|########
37 - 38	4	124	98	3	|###
39 - 40	3	127	99	2	+##
					|----+----+----+----+---

-+

5 10 15 20 25
Percentage of Examinees

SCALE # 2 Score Distribution Table

Score Interval	Freq-uency	Cum Freq	PR	PCT	
. . . No examinees below this score . . .					
15 - 16	0	0	1	0	|
17 - 18	1	1	1	1	|#
19 - 20	3	4	3	2	+##
21 - 22	3	7	6	2	|##
23 - 24	4	11	9	3	|###
25 - 26	4	15	12	3	|###
27 - 28	4	19	15	3	|###
29 - 30	10	29	23	8	+########
31 - 32	12	41	32	9	|#########
33 - 34	8	49	39	6	|######
35 - 36	14	63	50	11	|###########
37 - 38	7	70	55	6	|######
39 - 40	13	83	65	10	+##########
41 - 42	13	96	76	10	|##########
43 - 44	7	103	81	6	|######
45 - 46	10	113	89	8	|########
47 - 48	7	120	94	6	|######
49 - 50	4	124	98	3	+###
51 - 52	2	126	99	2	|##
53 - 54	0	126	99	0	|
55 - 56	0	126	99	0	|
57 - 58	1	127	99	1	|#
59 - 60	0	127	99	0	+
					|----+----+----+----+---

-+

5 10 15 20 25
Percentage of Examinees

IELTS Listening comprehension Item Analysis

********** ANALYSIS SUMMARY INFORMATION **********

Scale Definition Codes: DICHOT = Dichotomous MPOINT = Multipoint/Survey

	IELTS LC
Scale:	1
Type of Scale	DICHOT
N of Items	39
N of Examinees	131

***** CONFIGURATION INFORMATION *****

Type of Correlations:	Point-Biserial
Correction for Spuriousness:	YES
Ability Grouping:	YES
Subgroup Analysis:	NO
Express Endorsements As:	PROPORTIONS
Score Group Interval Width:	1

*** Correlations have been corrected for spuriousness **

		Item Statistics			Alternative Statistics					
Seq. No.	Scale -Item	Prop. Correct	Disc. Index	Point Biser.	Alt.	Prop. Total	Endorsing Low	Endorsing High	Point Biser.	Key
1	1-1	.72	.58	.45	1	.12	.30	.02	-.37	
					2	.05	.13	.00	-.25	
					3	.72	.38	.95	.45	*
					4	.08	.15	.02	-.23	
					Other	.03	.00	.00	-.18	
2	1-2	.36	.39	.32	1	.12	.20	.02	-.24	
					2	.12	.15	.07	-.18	
					3	.37	.48	.34	-.23	
					4	.36	.15	.54	.32	*
					Other	.02	.00	.00	-.06	
3	1-3	.79	.33	.25	1	.04	.03	.00	-.10	
					2	.79	.60	.93	.25	*
					3	.12	.23	.07	-.21	
					4	.03	.08	.00	-.23	
					Other	.02	.00	.00	-.20	
4	1-4	.51	.38	.30	1	.51	.38	.76	.30	*
					2	.31	.30	.22	-.20	
					3	.13	.23	.00	-.32	
					4	.05	.10	.02	-.16	
					Other	.01	.00	.00	-.06	
5	1-5	.92	.18	.21	1	.08	.18	.00	-.29	
					2	.92	.83	1.00	.21	*
					Other	.00	.00	.00		
6	1-6	.80	.38	.28	1	.15	.30	.00	-.36	
					2	.80	.63	1.00	.28	*
					Other	.05	.00	.00	-.14	
7	1-7	.07	.12	.19	1	.77	.73	.80	.01	
					2	.07	.03	.15	.19	*

					Other	.16	.00	.00	-.30	
8	1-8	.17	.37	.37	1	.41	.45	.32	-.14	
					2	.17	.00	.37	.37	*
					Other	.42	.00	.00	-.32	
9	1-9	.21	.54	.54	1	.18	.30	.10	-.23	
					2	.21	.03	.56	.54	*
					Other	.60	.00	.00	-.42	
10	1-10	.37	.46	.39	1	.27	.33	.17	-.17	
					2	.37	.15	.61	.39	*
					Other	.37	.00	.00	-.42	
11	1-11	.78	.45	.32	1	.16	.30	.05	-.30	
					2	.78	.50	.95	.32	*
					Other	.06	.00	.00	-.30	
12	1-12	.47	.66	.46	1	.34	.58	.15	-.38	
					2	.47	.10	.76	.46	*
					Other	.19	.00	.00	-.34	
13	1-13	.47	.73	.58	1	.18	.30	.05	-.31	
					2	.47	.10	.83	.58	*
					Other	.34	.00	.00	-.51	
14	1-14	.23	.34	.26	1	.57	.63	.46	-.20	
					2	.23	.08	.41	.26	*
					Other	.20	.00	.00	-.24	
15	1-15	.24	.41	.37	1	.27	.30	.20	-.18	
					2	.24	.05	.46	.37	*
					Other	.49	.00	.00	-.33	
16	1-16	.16	.39	.35	1	.58	.58	.44	-.17	
					2	.16	.03	.41	.35	*
					Other	.26	.00	.00	-.30	
17	1-17	.24	.44	.43	1	.19	.23	.10	-.19	
					2	.24	.05	.49	.43	*
					Other	.57	.00	.00	-.38	
18	1-18	.38	.44	.28	1	.29	.45	.29	-.18	
					2	.38	.13	.56	.28	*
					Other	.33	.00	.00	-.32	
19	1-19	.56	.65	.47	1	.21	.33	.07	-.31	
					2	.56	.25	.90	.47	*
					Other	.23	.00	.00	-.43	
20	1-20	.11	.24	.39	1	.54	.45	.54	-.05	
					2	.11	.03	.27	.39	*
					Other	.35	.00	.00	-.37	
21	1-21	.69	.35	.28	1	.24	.35	.15	-.29	
					2	.69	.50	.85	.28	*
					Other	.07	.00	.00	-.28	
22	1-22	.89	.18	.16	1	.89	.80	.98	.16	*
					2	.11	.20	.02	-.26	
					3	.00	.00	.00		
					Other	.00	.00	.00		
23	1-23	.39	.31	.18	1	.39	.25	.56	.18	*

					2	.49	.65	.29	-.38	
					3	.11	.10	.15	.09	
					Other	.02	.00	.00	-.08	
24	1-24	.24	.17	.04	1	.45	.43	.49	.03	
					2	.31	.43	.20	-.28	
					3	.24	.15	.32	.04	*
					Other	.00	.00	.00		
25	1-25	.47	.51	.34	1	.51	.75	.22	-.47	
					2	.47	.25	.76	.34	*
					3	.02	.00	.02	-.02	
					Other	.00	.00	.00		
26	1-26	.75	.20	.20	1	.18	.20	.10	-.22	
					2	.75	.68	.88	.20	*
					3	.06	.10	.00	-.24	
					Other	.02	.00	.00	-.05	
27	1-27	.19	.14	.11	1	.19	.13	.27	.11	*
					2	.79	.85	.73	-.20	
					3	.02	.03	.00	-.09	
					Other	.00	.00	.00		
28	1-28	.56	.43	.31	1	.23	.33	.12	-.27	
					2	.56	.35	.78	.31	*
					3	.19	.28	.10	-.26	
					Other	.02	.00	.00	-.18	
29	1-29	.44	.48	.36	1	.44	.25	.73	.36	*
					2	.36	.53	.17	-.40	
					3	.20	.23	.10	-.19	
					Other	.00	.00	.00		
30	1-30	.44	.29	.26	1	.44	.30	.59	.26	*
					2	.43	.60	.32	-.36	
					3	.11	.08	.10	-.10	
					Other	.02	.00	.00	-.04	
31	1-31	.37	.51	.42	1	.47	.48	.29	-.26	
					2	.37	.20	.71	.42	*
					Other	.15	.00	.00	-.43	
32	1-32	.19	.24	.16	1	.61	.75	.66	-.07	
					2	.19	.05	.29	.16	*
					Other	.20	.00	.00	-.29	
33	1-33	.09	.15	.30	1	.41	.40	.27	-.20	
					2	.09	.05	.20	.30	*
					Other	.50	.00	.00	-.15	
34	1-34	.24	.61	.54	1	.37	.48	.22	-.29	
					2	.24	.00	.61	.54	*
					Other	.39	.00	.00	-.36	
35	1-35	.05	.10	.22	1	.53	.43	.59	.05	
					2	.05	.00	.10	.22	*
					Other	.42	.00	.00	-.30	
36	1-36	.13	.32	.38	1	.24	.25	.15	-.14	
					2	.13	.00	.32	.38	*
					Other	.63	.00	.00	-.30	
37	1-37	.20	.51	.54	1	.48	.50	.32	-.23	

					2	.20	.03	.54	.54	*
					Other	.32	.00	.00	-.39	
38	1-38	.27	.36	.32	1	.24	.28	.29	-.02	
					2	.27	.10	.46	.32	*
					Other	.48	.00	.00	-.44	
39	1-39	.44	.51	.39	1	.34	.35	.22	-.21	
					2	.44	.25	.76	.39	*
					Other	.22	.00	.00	-.44	

There were 131 examinees in the data file.

Scale Statistics

Scale:	IELTS LC 1
N of Items	39
N of Examinees	131
Mean	15.611
Variance	40.818
Std. Dev.	6.389
Skew	0.525
Kurtosis	-0.579
Minimum	5.000
Maximum	32.000
Median	14.000
Alpha	0.848
SEM	2.489
Mean P	0.400
Mean Item-Tot.	0.326
Mean Biserial	0.456
Max Score (Low)	11
N (Low Group)	40
Min Score (High)	19
N (High Group)	41

IELTS LC
SCALE # 1 Score Distribution Table

Number Correct	Freq-uency	Cum Freq	PR	PCT	
. . . No examinees below this score . . .					
4	0	0	1	0	|
5	1	1	1	1	+#
6	5	6	5	4	|####
7	3	9	7	2	|##
8	9	18	14	7	|#######
9	6	24	18	5	|#####
10	7	31	24	5	+#####
11	9	40	31	7	|#######
12	8	48	37	6	|######
13	11	59	45	8	|########
14	8	67	51	6	|######
15	6	73	56	5	+#####
16	7	80	61	5	|#####
17	6	86	66	5	|#####
18	4	90	69	3	|###
19	6	96	73	5	|#####
20	4	100	76	3	+###
21	3	103	79	2	|##
22	6	109	83	5	|#####
23	5	114	87	4	|####
24	1	115	88	1	|#
25	2	117	89	2	+##
26	4	121	92	3	|###
27	3	124	95	2	|##
28	4	128	98	3	|###
29	1	129	98	1	|#
30	0	129	98	0	+
31	1	130	99	1	|#
32	1	131	99	1	|#
33	0	131	99	0	|
34	0	131	99	0	|
. . . No examinees above this score . . .					|

|----+----+----+----+---
-+

5 10 15 20 25
Percentage of Examinees

IELTS Reading comprehension Item Analysis

********** ANALYSIS SUMMARY INFORMATION **********

Scale Definition Codes: DICHOT = Dichotomous MPOINT = Multipoint/Survey

Scale:	IELTS RC 2
Type of Scale	DICHOT
N of Items	35
N of Examinees	127

***** CONFIGURATION INFORMATION *****

Type of Correlations: Point-Biserial
Correction for Spuriousness: YES
Ability Grouping: YES
Subgroup Analysis: NO
Express Endorsements As: PROPORTIONS

Score Group Interval Width: 1

*** Correlations have been corrected for spuriousness **

		Item Statistics			Alternative Statistics					
Seq. No.	Scale -Item	Prop. Correct	Disc. Index	Point Biser.	Alt.	Prop. Total	Endorsing Low	Endorsing High	Point Biser.	Key
1	2-1	.29	.54	.43	1	.54	.63	.30	-.38	
					2	.29	.08	.62	.43	*
					Other	.17	.00	.00	-.25	
2	2-2	.31	.60	.48	1	.54	.66	.27	-.43	
					2	.31	.05	.65	.48	*
					Other	.16	.00	.00	-.25	
3	2-3	.33	.44	.31	1	.50	.53	.32	-.28	
					2	.33	.16	.59	.31	*
					Other	.17	.00	.00	-.27	
4	2-4	.33	.44	.36	1	.43	.50	.27	-.32	
					2	.33	.16	.59	.36	*
					Other	.24	.00	.00	-.26	
5	2-5	.09	.22	.33	1	.68	.71	.54	-.24	
					2	.09	.03	.24	.33	*
					Other	.24	.00	.00	-.13	
6	2-6	.69	.10	.02	1	.21	.26	.19	-.13	
					2	.69	.63	.73	.02	*
					Other	.09	.00	.00	-.13	
7	2-7	.54	.44	.35	1	.22	.32	.08	-.32	
					2	.54	.29	.73	.35	*
					Other	.24	.00	.00	-.31	
8	2-8	.50	.49	.38	1	.22	.32	.11	-.31	
					2	.50	.21	.70	.38	*
					Other	.28	.00	.00	-.34	

9	2-9	.36	.36	.23	1	.54	.68	.43	-.26	
					2	.36	.16	.51	.23	*
					Other	.10	.00	.00	-.25	
10	2-10	.72	.29	.24	1	.19	.29	.11	-.27	
					2	.72	.55	.84	.24	*
					Other	.09	.00	.00	-.25	
11	2-11	.16	.14	.15	1	.50	.42	.59	.01	
					2	.16	.11	.24	.15	*
					Other	.35	.00	.00	-.34	
12	2-12	.46	.25	.17	1	.35	.32	.22	-.16	
					2	.46	.39	.65	.17	*
					Other	.20	.00	.00	-.28	
13	2-13	.17	.14	.21	1	.62	.61	.62	-.14	
					2	.17	.11	.24	.21	*
					Other	.20	.00	.00	-.25	
14	2-14	.39	.44	.26	1	.56	.74	.38	-.35	
					2	.39	.16	.59	.26	*
					Other	.06	.00	.00	-.17	
15	2-15	.51	.47	.39	1	.44	.63	.24	-.47	
					2	.51	.26	.73	.39	*
					Other	.05	.00	.00	-.20	
16	2-16	.50	.52	.36	1	.46	.66	.19	-.45	
					2	.50	.26	.78	.36	*
					Other	.04	.00	.00	-.18	
17	2-17	.54	.55	.40	1	.39	.68	.24	-.46	
					2	.54	.18	.73	.40	*
					Other	.07	.00	.00	-.22	
18	2-18	.15	.27	.29	1	.77	.76	.62	-.23	
					2	.15	.08	.35	.29	*
					Other	.08	.00	.00	-.26	
19	2-19	.30	.51	.39	1	.65	.84	.41	-.43	
					2	.30	.05	.57	.39	*
					Other	.06	.00	.00	-.21	
20	2-20	.30	.46	.32	1	.63	.79	.46	-.33	
					2	.30	.05	.51	.32	*
					Other	.07	.00	.00	-.26	
21	2-21	.51	.57	.35	1	.44	.66	.19	-.41	
					2	.51	.21	.78	.35	*
					Other	.05	.00	.00	-.25	
22	2-22	.58	.65	.44	1	.35	.63	.14	-.48	
					2	.58	.18	.84	.44	*
					Other	.06	.00	.00	-.27	
23	2-23	.30	.30	.18	1	.63	.66	.49	-.24	
					2	.30	.16	.46	.18	*
					Other	.07	.00	.00	-.19	
24	2-24	.28	.51	.47	1	.63	.76	.41	-.45	
					2	.28	.03	.54	.47	*
					Other	.09	.00	.00	-.24	

25	2-25	.14	.35	.41	1	.72	.66	.57	-.21	
					2	.14	.03	.38	.41	*
					Other	.13	.00	.00	-.35	
26	2-26	.28	.43	.39	1	.55	.55	.41	-.26	
					2	.28	.08	.51	.39	*
					Other	.17	.00	.00	-.35	
27	2-27	.39	.62	.41	1	.47	.61	.19	-.37	
					2	.39	.11	.73	.41	*
					Other	.14	.00	.00	-.30	
28	2-28	.53	.71	.50	1	.34	.55	.08	-.48	
					2	.53	.16	.86	.50	*
					Other	.13	.00	.00	-.31	
29	2-29	.11	.08	.12	1	.04	.00	.05	.07	
					2	.65	.55	.76	.06	
					3	.02	.03	.00	-.11	
					4	.11	.03	.11	.12	*
					Other	.19	.00	.00	-.39	
30	2-30	.57	.44	.30	1	.57	.34	.78	.30	*
					2	.15	.18	.08	-.18	
					3	.06	.05	.03	-.11	
					4	.02	.00	.03	.05	
					Other	.20	.00	.00	-.41	
31	2-31	.27	.17	.08	1	.13	.16	.16	-.05	
					2	.34	.26	.43	.05	
					3	.27	.11	.27	.08	*
					4	.04	.03	.05	.05	
					Other	.23	.00	.00	-.41	
32	2-32	.27	.03	-.07	1	.18	.16	.22	-.04	
					2	.18	.05	.30	.26	?
CHECK THE KEY					3	.07	.05	.08	.04	
4 was specified, 2 works better					4	.27	.21	.24	-.07	*
					Other	.30	.00	.00	-.42	
33	2-33	.14	.11	.05	1	.46	.34	.54	.13	?
CHECK THE KEY					2	.14	.05	.16	.05	*
2 was specified, 1 works better					Other	.40	.00	.00	-.37	
34	2-34	.35	.36	.23	1	.24	.24	.19	-.09	
					2	.35	.16	.51	.23	*
					Other	.41	.00	.00	-.36	
35	2-35	.33	.38	.24	1	.22	.18	.11	-.10	
					2	.33	.18	.57	.24	*
					Other	.45	.00	.00	-.35	

There were 127 examinees in the data file.

Scale Statistics

	IELTS RC
Scale:	2
N of Items	35
N of Examinees	127
Mean	12.661
Variance	32.917
Std. Dev.	5.737
Skew	0.441
Kurtosis	0.216
Minimum	1.000
Maximum	30.000
Median	13.000
Alpha	0.806
SEM	2.530
Mean P	0.362
Mean Item-Tot.	0.290
Mean Biserial	0.387
Max Score (Low)	9
N (Low Group)	38
Min Score (High)	16
N (High Group)	37

SCALE # 2 Score Distribution Table

Number Correct	Freq-uency	Cum Freq	PR	PCT	
1	1	1	1	1	|#
2	2	3	2	2	|##
3	1	4	3	1	|#
4	1	5	4	1	|#
5	9	14	11	7	+#######
6	10	24	19	8	|########
7	6	30	24	5	|#####
8	3	33	26	2	|##
9	5	38	30	4	|####
10	7	45	35	6	+######
11	6	51	40	5	|#####
12	10	61	48	8	|########
13	9	70	55	7	|#######
14	10	80	63	8	|########
15	10	90	71	8	+########
16	8	98	77	6	|######
17	10	108	85	8	|########
18	2	110	87	2	|##
19	5	115	91	4	|####
20	2	117	92	2	+##
21	2	119	94	2	|##
22	1	120	94	1	|#
23	1	121	95	1	|#
24	1	122	96	1	|#
25	1	123	97	1	+#
26	1	124	98	1	|#
27	1	125	98	1	|#
28	0	125	98	0	|
29	1	126	99	1	|#
30	1	127	99	1	+#
31	0	127	99	0	|
32	0	127	99	0	|
. . . No examinees above this score . . .					|

\|----+----+----+----+---

-+

5 10 15 20 25

Percentage of Examinees

EPTB Item Analysis

********** ANALYSIS SUMMARY INFORMATION **********
Scale Definition Codes: DICHOT = Dichotomous MPOINT = Multipoint/Survey

Scale:	0	1	2
Type of Scale	DICHOT	DICHOT	DICHOT
N of Items	58	44	47
N of Examinees	134	134	134

***** CONFIGURATION INFORMATION *****

Type of Correlations: Point-Biserial
Correction for Spuriousness: YES
Ability Grouping: YES
Subgroup Analysis: NO
Express Endorsements As: PROPORTIONS
Score Group Interval Width: 1

*** Correlations have been corrected for spuriousness **

		Item Statistics			Alternative Statistics					
Seq. No.	Scale -Item	Prop. Correct	Disc. Index	Point Biser.	Alt.	Prop. Total	Endorsing Low	Endorsing High	Point Biser.	Key
1	0-1	.75	.09	.10	1	.75	.69	.78	.10	*
					2	.04	.07	.03	-.07	
					3	.03	.07	.00	-.19	
					4	.00	.00	.00		
					5	.16	.14	.16	-.03	
					Other	.01	.00	.00	-.37	
2	0-2	.85	.09	.16	1	.04	.02	.03	-.04	
					2	.07	.10	.03	-.10	
					3	.85	.83	.92	.16	*
					4	.01	.02	.00	-.18	
					5	.00	.00	.00		
					Other	.02	.00	.00	-.31	
3	0-3	.67	.19	.07	1	.00	.00	.00		
					2	.04	.05	.00	-.07	
					3	.01	.02	.00	-.07	
					4	.26	.26	.19	-.04	
					5	.67	.62	.81	.07	*
					Other	.01	.00	.00	-.51	
4	0-4	.86	.26	.36	1	.01	.02	.00	-.20	
					2	.86	.71	.97	.36	*
					3	.03	.02	.00	-.06	
					4	.07	.19	.00	-.30	
					5	.03	.02	.03	-.05	
					Other	.01	.00	.00	-.56	
5	0-5	.78	.23	.22	1	.01	.02	.03	.02	
					2	.01	.02	.00	-.16	
					3	.01	.00	.00	-.00	
					4	.78	.67	.89	.22	*
					5	.16	.24	.08	-.19	
					Other	.02	.00	.00	-.37	

6	0-6	.58	.45	.35	1	.06	.10	.08	-.03	
					2	.09	.17	.03	-.17	
					3	.06	.10	.03	-.17	
					4	.58	.31	.76	.35	*
					5	.20	.31	.11	-.28	
					Other	.01	.00	.00	-.56	
7	0-7	.77	.40	.28	1	.01	.00	.00	-.00	
					2	.02	.05	.00	-.15	
					3	.01	.02	.00	-.02	
					4	.17	.31	.03	-.22	
					5	.77	.57	.97	.28	*
					Other	.01	.00	.00	-.54	
8	0-8	.90	.14	.29	1	.00	.00	.00		
					2	.90	.83	.97	.29	*
					3	.03	.05	.03	-.10	
					4	.01	.02	.00	-.10	
					5	.05	.07	.00	-.17	
					Other	.01	.00	.00	-.56	
9	0-9	.76	.47	.39	1	.01	.02	.00	-.07	
					2	.76	.48	.95	.39	*
					3	.08	.17	.00	-.25	
					4	.04	.05	.05	-.06	
					5	.10	.26	.00	-.28	
					Other	.01	.00	.00	-.56	
10	0-10	.78	.30	.27	1	.07	.10	.08	-.05	
					2	.09	.19	.03	-.23	
					3	.78	.60	.89	.27	*
					4	.01	.00	.00	-.01	
					5	.01	.05	.00	-.09	
					Other	.02	.00	.00	-.49	
11	0-11	.51	.24	.12	1	.02	.05	.00	-.12	
					2	.07	.07	.08	.00	
					3	.13	.26	.05	-.23	
					4	.51	.40	.65	.12	*
					5	.23	.17	.22	-.02	
					Other	.02	.00	.00	-.37	
12	0-12	.77	.32	.26	1	.77	.60	.92	.26	*
					2	.07	.12	.00	-.20	
					3	.01	.02	.00	-.04	
					4	.03	.05	.03	-.06	
					5	.12	.19	.05	-.16	
					Other	.01	.00	.00	-.56	
13	0-13	.54	.40	.26	1	.10	.17	.03	-.20	
					2	.03	.05	.03	-.03	
					3	.01	.05	.00	-.20	
					4	.29	.38	.22	-.16	
					5	.54	.33	.73	.26	*
					Other	.02	.00	.00	-.35	
14	0-14	.60	.21	.09	1	.01	.02	.00	-.12	
					2	.60	.52	.73	.09	*
					3	.00	.00	.00		
					4	.02	.02	.03	-.10	
					5	.36	.38	.24	-.08	
					Other	.01	.00	.00	-.43	
15	0-15	.96	.12	.34	1	.96	.88	1.00	.34	*

					2	.01	.02	.00	-.05	
					3	.00	.00	.00		
					4	.00	.00	.00		
					5	.01	.05	.00	-.14	
					Other	.01	.00	.00	-.48	
16	0-16	.59	.40	.29	1	.10	.07	.14	.04	
					2	.07	.14	.03	-.19	
					3	.08	.12	.00	-.20	
					4	.59	.38	.78	.29	*
					5	.14	.24	.05	-.20	
					Other	.01	.00	.00	-.51	
17	0-17	.79	.40	.30	1	.00	.00	.00		
					2	.02	.05	.00	-.11	
					3	.01	.05	.00	-.20	
					4	.16	.26	.00	-.18	
					5	.79	.60	1.00	.30	*
					Other	.01	.00	.00	-.51	
18	0-18	.81	.15	.19	1	.01	.00	.05	.09	
					2	.81	.71	.86	.19	*
					3	.01	.02	.00	-.20	
					4	.07	.12	.05	-.13	
					5	.09	.10	.03	-.11	
					Other	.01	.00	.00	-.45	
19	0-19	.70	.26	.22	1	.01	.02	.00	-.08	
					2	.04	.10	.03	-.14	
					3	.01	.02	.00	-.09	
					4	.70	.55	.81	.22	*
					5	.22	.26	.16	-.14	
					Other	.01	.00	.00	-.51	
20	0-20	.90	.21	.34	1	.90	.79	1.00	.34	*
					2	.04	.05	.00	-.11	
					3	.01	.02	.00	-.23	
					4	.00	.00	.00		
					5	.03	.07	.00	-.10	
					Other	.02	.00	.00	-.46	
21	0-21	.77	.42	.38	1	.01	.02	.03	-.01	
					2	.04	.10	.00	-.19	
					3	.00	.00	.00		
					4	.77	.50	.92	.38	*
					5	.16	.33	.05	-.28	
					Other	.01	.00	.00	-.54	
22	0-22	.87	.11	.27	1	.87	.81	.92	.27	*
					2	.01	.02	.00	-.07	
					3	.01	.02	.00	-.22	
					4	.01	.02	.00	-.20	
					5	.10	.10	.08	-.11	
					Other	.01	.00	.00	-.56	
23	0-23	.85	.18	.29	1	.01	.00	.00	.01	
					2	.85	.76	.95	.29	*
					3	.01	.02	.00	-.22	
					4	.04	.05	.03	-.11	
					5	.09	.14	.03	-.19	
					Other	.01	.00	.00	-.56	
24	0-24	.67	.19	.15	1	.27	.31	.19	-.14	
					2	.02	.05	.00	-.13	

					3	.00	.00	.00		
					4	.03	.02	.03	-.07	
					5	.67	.60	.78	.15	*
					Other	.01	.00	.00	-.56	
25	0-25	.93	.06	.23	1	.93	.88	.95	.23	*
					2	.01	.00	.00	-.00	
					3	.01	.00	.03	.05	
					4	.00	.00	.00		
					5	.05	.10	.03	-.16	
					Other	.01	.00	.00	-.56	
26	0-26	.78	.23	.23	1	.01	.02	.00	-.06	
					2	.01	.02	.00	-.11	
					3	.01	.05	.00	-.19	
					4	.17	.19	.11	-.15	
					5	.78	.67	.89	.23	*
					Other	.01	.00	.00	-.46	
27	0-27	.57	.36	.15	1	.04	.05	.00	-.12	
					2	.27	.24	.16	-.04	
					3	.07	.14	.00	-.25	
					4	.05	.07	.00	-.05	
					5	.57	.48	.84	.15	*
					Other	.01	.00	.00	-.56	
28	0-28	.89	.16	.30	1	.04	.10	.03	-.15	
					2	.04	.05	.03	-.06	
					3	.89	.79	.95	.30	*
					4	.01	.02	.00	-.18	
					5	.01	.02	.00	-.17	
					Other	.01	.00	.00	-.56	
29	0-29	.92	.21	.35	1	.92	.79	1.00	.35	*
					2	.01	.02	.00	-.04	
					3	.04	.10	.00	-.25	
					4	.00	.00	.00		
					5	.02	.05	.00	-.07	
					Other	.01	.00	.00	-.45	
30	0-30	.93	.14	.35	1	.93	.86	1.00	.35	*
					2	.01	.02	.00	-.07	
					3	.01	.05	.00	-.21	
					4	.00	.00	.00		
					5	.04	.05	.00	-.14	
					Other	.01	.00	.00	-.56	
31	0-31	.27	.08	.05	1	.01	.00	.05	.10	?
					2	.01	.02	.00	-.22	
CHECK THE KEY					3	.01	.05	.00	-.15	
5 was specified, 1 works better					4	.69	.69	.65	-.03	
					5	.27	.21	.30	.05	*
					Other	.01	.00	.00	-.56	
32	0-32	.82	.25	.30	1	.08	.17	.03	-.16	
					2	.03	.02	.05	-.05	
					3	.82	.64	.89	.30	*
					4	.03	.05	.03	-.10	
					5	.02	.07	.00	-.25	
					Other	.01	.00	.00	-.45	
33	0-33	.90	.09	.17	1	.90	.83	.92	.17	*
					2	.01	.02	.00	-.07	
					3	.01	.02	.00	-.06	

					4	.01	.02	.00	-.22	
					5	.07	.07	.08	.00	
					Other	.01	.00	.00	-.56	
34	0-34	.92	.07	.14	1	.01	.00	.00	-.01	
					2	.04	.02	.03	.00	
					3	.01	.02	.00	-.09	
					4	.92	.90	.97	.14	*
					5	.01	.02	.00	-.04	
					Other	.01	.00	.00	-.56	
35	0-35	.72	.37	.31	1	.00	.00	.00		
					2	.72	.57	.95	.31	*
					3	.18	.26	.03	-.28	
					4	.07	.12	.03	-.09	
					5	.02	.02	.00	-.15	
					Other	.01	.00	.00	-.56	
36	0-36	.92	.19	.41	1	.01	.00	.00	-.01	
					2	.92	.81	1.00	.41	*
					3	.00	.00	.00		
					4	.02	.07	.00	-.28	
					5	.04	.10	.00	-.22	
					Other	.01	.00	.00	-.56	
37	0-37	.77	.30	.33	1	.77	.62	.92	.33	*
					2	.01	.02	.00	-.11	
					3	.11	.17	.05	-.19	
					4	.04	.05	.03	-.05	
					5	.04	.10	.00	-.20	
					Other	.01	.00	.00	-.53	
38	0-38	.65	.41	.27	1	.02	.05	.00	-.20	
					2	.01	.00	.00	-.00	
					3	.01	.02	.00	-.11	
					4	.65	.43	.84	.27	*
					5	.31	.48	.16	-.24	
					Other	.01	.00	.00	-.56	
39	0-39	.73	.17	.19	1	.01	.00	.00	-.00	
					2	.01	.02	.00	-.22	
					3	.02	.07	.00	-.26	
					4	.73	.64	.81	.19	*
					5	.22	.21	.19	-.07	
					Other	.01	.00	.00	-.48	
40	0-40	.87	.28	.33	1	.04	.07	.03	-.13	
					2	.02	.07	.00	-.17	
					3	.87	.69	.97	.33	*
					4	.04	.12	.00	-.18	
					5	.01	.02	.00	-.05	
					Other	.01	.00	.00	-.56	
41	0-41	.43	.29	.10	1	.02	.02	.03	-.01	
					2	.43	.31	.59	.10	*
					3	.15	.24	.03	-.24	
					4	.16	.17	.22	.03	
					5	.23	.21	.14	-.03	
					Other	.01	.00	.00	-.52	
42	0-42	.81	.26	.29	1	.06	.14	.03	-.25	
					2	.01	.02	.00	-.19	
					3	.01	.05	.00	-.11	
					4	.09	.07	.03	-.05	

					5	.81	.69	.95	.29	*
					Other	.01	.00	.00	-.56	
43	0-43	.81	.40	.41	1	.81	.57	.97	.41	*
					2	.00	.00	.00		
					3	.09	.24	.00	-.28	
					4	.06	.12	.03	-.26	
					5	.04	.05	.00	-.08	
					Other	.01	.00	.00	-.56	
44	0-44	.74	.22	.25	1	.01	.00	.03	.06	
					2	.04	.02	.00	-.04	
					3	.11	.17	.08	-.20	
					4	.74	.64	.86	.25	*
					5	.08	.14	.03	-.21	
					Other	.01	.00	.00	-.41	
45	0-45	.96	.02	.19	1	.00	.00	.00		
					2	.00	.00	.00		
					3	.00	.00	.00		
					4	.03	.02	.03	.02	
					5	.96	.95	.97	.19	*
					Other	.01	.00	.00	-.56	
46	0-46	.65	.36	.23	1	.01	.02	.00	-.17	
					2	.65	.45	.81	.23	*
					3	.28	.40	.16	-.21	
					4	.02	.05	.00	-.08	
					5	.02	.00	.03	.01	
					Other	.02	.00	.00	-.41	
47	0-47	.62	.18	.14	1	.62	.50	.68	.14	*
					2	.01	.02	.00	-.09	
					3	.01	.00	.00	-.00	
					4	.03	.07	.00	-.16	
					5	.31	.36	.32	-.09	
					Other	.02	.00	.00	-.43	
48	0-48	.32	.04	.01	1	.03	.02	.03	-.06	
					2	.21	.21	.32	-.04	
					3	.13	.14	.05	-.08	
					4	.32	.31	.35	.01	*
					5	.31	.29	.24	-.03	
					Other	.01	.00	.00	-.56	
49	0-49	.82	.23	.26	1	.01	.02	.00	-.07	
					2	.82	.69	.92	.26	*
					3	.05	.07	.03	-.20	
					4	.02	.05	.00	-.15	
					5	.08	.12	.05	-.07	
					Other	.01	.00	.00	-.45	
50	0-50	.94	.17	.42	1	.01	.02	.00	-.22	
					2	.01	.05	.00	-.15	
					3	.02	.05	.00	-.15	
					4	.01	.02	.00	-.06	
					5	.94	.83	1.00	.42	*
					Other	.01	.00	.00	-.56	
51	0-51	.63	.48	.32	1	.07	.07	.03	-.15	
					2	.63	.38	.86	.32	*
					3	.16	.31	.05	-.24	
					4	.04	.07	.03	-.07	
					5	.07	.12	.03	-.12	

					Other	.01	.00	.00	-.48	
52	0-52	.95	.14	.48	1	.01	.02	.00	-.22	
					2	.00	.00	.00		
					3	.01	.02	.00	-.13	
					4	.01	.05	.00	-.17	
					5	.95	.86	1.00	.48	*
					Other	.01	.00	.00	-.53	
53	0-53	.65	.39	.29	1	.01	.02	.00	-.14	
					2	.01	.02	.00	-.04	
					3	.01	.05	.00	-.29	
					4	.65	.48	.86	.29	*
					5	.30	.38	.14	-.21	
					Other	.01	.00	.00	-.45	
54	0-54	.47	.29	.15	1	.47	.36	.65	.15	*
					2	.03	.00	.08	.10	
					3	.42	.52	.19	-.22	
					4	.02	.02	.03	-.01	
					5	.04	.02	.05	-.05	
					Other	.02	.00	.00	-.39	
55	0-55	.37	.30	.17	1	.59	.62	.46	-.09	
					2	.02	.07	.00	-.26	
					3	.00	.00	.00		
					4	.01	.05	.00	-.20	
					5	.37	.24	.54	.17	*
					Other	.01	.00	.00	-.56	
56	0-56	.84	.13	.21	1	.84	.76	.89	.21	*
					2	.04	.05	.05	-.14	
					3	.02	.02	.03	.00	
					4	.01	.02	.00	-.11	
					5	.08	.12	.03	-.12	
					Other	.01	.00	.00	-.56	
57	0-57	.75	.12	.13	1	.75	.67	.78	.13	*
					2	.01	.02	.00	-.22	
					3	.01	.02	.00	-.18	
					4	.00	.00	.00		
					5	.22	.24	.22	-.05	
					Other	.01	.00	.00	-.45	
58	0-58	.93	.12	.34	1	.01	.02	.03	.01	
					2	.03	.05	.00	-.15	
					3	.93	.86	.97	.34	*
					4	.01	.02	.00	-.18	
					5	.01	.02	.00	-.17	
					Other	.01	.00	.00	-.56	
59	1-1	.97	.05	.26	1	.02	.02	.00	-.10	
					2	.97	.95	1.00	.26	*
					Other	.01	.00	.00	-.47	
60	1-2	.90	.15	.22	1	.90	.83	.98	.22	*
					2	.09	.14	.02	-.20	
					Other	.01	.00	.00	-.47	
61	1-3	.72	-.00	.04	1	.72	.69	.69	.04	*
					2	.27	.29	.31	-.13	
					Other	.01	.00	.00	-.47	
62	1-4	.72	.39	.29	1	.72	.55	.94	.29	*

					2	.26	.43	.06	-.37	
					Other	.01	.00	.00	-.34	
63	1-5	.99	.05	.37	1	.01	.02	.00	-.12	
					2	.99	.95	1.00	.37	*
					Other	.01	.00	.00	-.47	
64	1-6	.74	.30	.26	1	.74	.57	.88	.26	*
					2	.25	.40	.13	-.33	
					Other	.01	.00	.00	-.47	
65	1-7	.54	.15	.10	1	.45	.52	.40	-.22	
					2	.54	.45	.60	.10	*
					Other	.01	.00	.00	-.47	
66	1-8	.66	.31	.21	1	.66	.52	.83	.21	*
					2	.33	.45	.17	-.30	
					Other	.01	.00	.00	-.47	
67	1-9	.59	.25	.13	1	.37	.45	.27	-.22	
					2	.59	.48	.73	.13	*
					Other	.04	.00	.00	-.31	
68	1-10	.43	.18	.16	1	.43	.38	.56	.16	*
					2	.54	.57	.44	-.26	
					Other	.02	.00	.00	-.34	
69	1-11	.50	.45	.28	1	.50	.26	.71	.28	*
					2	.48	.69	.29	-.36	
					Other	.02	.00	.00	-.36	
70	1-12	.65	.45	.32	1	.33	.55	.15	-.39	
					2	.65	.40	.85	.32	*
					Other	.02	.00	.00	-.36	
71	1-13	.43	.25	.14	1	.54	.62	.44	-.22	
					2	.43	.31	.56	.14	*
					Other	.02	.00	.00	-.41	
72	1-14	.47	.15	.12	1	.47	.31	.46	.12	*
					2	.51	.62	.54	-.20	
					Other	.02	.00	.00	-.41	
73	1-15	.39	.06	-.04	1	.39	.36	.42	-.04	*
					2	.60	.62	.58	-.08	
					Other	.01	.00	.00	-.47	
74	1-16	.76	.34	.28	1	.76	.60	.94	.28	*
					2	.23	.38	.06	-.34	
					Other	.01	.00	.00	-.47	
75	1-17	.53	.33	.23	1	.46	.60	.29	-.33	
					2	.53	.38	.71	.23	*
					Other	.01	.00	.00	-.47	
76	1-18	.15	.20	.23	1	.15	.07	.27	.23	*
					2	.84	.90	.73	-.25	
					Other	.01	.00	.00	-.47	
77	1-19	.68	.22	.11	1	.30	.40	.21	-.20	
					2	.68	.57	.79	.11	*
					Other	.02	.00	.00	-.31	
78	1-20	.54	.48	.32	1	.54	.31	.79	.32	*

					2	.45	.67	.21	-.41	
					Other	.01	.00	.00	-.35	
79	1-21	.99	.05	.37	1	.01	.02	.00	-.12	
					2	.99	.95	1.00	.37	*
					Other	.01	.00	.00	-.47	
80	1-22	.24	.10	.02	1	.24	.19	.29	.02	*
					2	.75	.79	.71	-.10	
					Other	.01	.00	.00	-.47	
81	1-23	.85	.22	.25	1	.14	.26	.06	-.28	
					2	.85	.71	.94	.25	*
					Other	.01	.00	.00	-.47	
82	1-24	.50	-.00	-.11	1	.49	.45	.48	-.01	
					2	.50	.52	.52	-.11	*
					Other	.01	.00	.00	-.47	
83	1-25	.47	.36	.19	1	.52	.71	.38	-.30	
					2	.47	.26	.63	.19	*
					Other	.01	.00	.00	-.47	
84	1-26	.34	.59	.47	1	.34	.10	.69	.47	*
					2	.65	.88	.31	-.53	
					Other	.01	.00	.00	-.47	
85	1-27	.48	.60	.38	1	.48	.19	.79	.38	*
					2	.51	.79	.21	-.47	
					Other	.01	.00	.00	-.47	
86	1-28	.75	.32	.28	1	.23	.36	.08	-.34	
					2	.75	.60	.92	.28	*
					Other	.01	.00	.00	-.38	
87	1-29	.87	.02	.05	1	.12	.12	.13	-.06	
					2	.87	.86	.88	.05	*
					Other	.01	.00	.00	-.37	
88	1-30	.53	.26	.11	1	.53	.40	.67	.11	*
					2	.46	.57	.33	-.22	
					Other	.01	.00	.00	-.37	
89	1-31	.81	.09	.11	1	.18	.21	.15	-.16	
					2	.81	.76	.85	.11	*
					Other	.01	.00	.00	-.47	
90	1-32	.27	.24	.23	1	.71	.83	.65	-.29	
					2	.27	.12	.35	.23	*
					Other	.02	.00	.00	-.34	
91	1-33	.87	.24	.28	1	.11	.21	.02	-.27	
					2	.87	.74	.98	.28	*
					Other	.01	.00	.00	-.41	
92	1-34	.12	.12	.14	1	.12	.05	.17	.14	*
					2	.87	.90	.83	-.12	
					Other	.01	.00	.00	-.41	
93	1-35	.91	.08	.16	1	.08	.10	.04	-.14	
					2	.91	.88	.96	.16	*
					Other	.01	.00	.00	-.47	
94	1-36	.51	.31	.15	1	.51	.36	.67	.15	*

					2	.49	.62	.33	-.27	
					Other	.01	.00	.00	-.47	
95	1-37	.75	.10	.13	1	.75	.71	.81	.13	*
					2	.23	.26	.19	-.20	
					Other	.02	.00	.00	-.32	
96	1-38	.39	.15	.08	1	.39	.33	.48	.08	*
					2	.60	.64	.52	-.18	
					Other	.01	.00	.00	-.37	
97	1-39	.54	.42	.24	1	.43	.64	.29	-.31	
					2	.54	.29	.71	.24	*
					Other	.02	.00	.00	-.40	
98	1-40	.49	.46	.22	1	.49	.19	.65	.22	*
					2	.49	.74	.35	-.29	
					Other	.02	.00	.00	-.40	
99	1-41	.46	.34	.27	1	.52	.67	.33	-.38	
					2	.46	.31	.65	.27	*
					Other	.01	.00	.00	-.30	
100	1-42	.74	.21	.17	1	.25	.33	.13	-.26	
					2	.74	.64	.85	.17	*
					Other	.01	.00	.00	-.30	
101	1-43	.20	.04	-.01	1	.20	.17	.21	-.01	*
					2	.78	.79	.79	-.04	
					Other	.02	.00	.00	-.33	
102	1-44	.46	.07	.03	1	.46	.43	.50	.03	*
					2	.52	.52	.50	-.14	
					Other	.02	.00	.00	-.33	
103	2-1	.51	.35	.11	1	.51	.38	.73	.11	*
					2	.46	.56	.27	-.17	
					3	.01	.04	.00	-.12	
					Other	.01	.00	.00	-.42	
104	2-2	.89	-.02	.01	1	.89	.89	.86	.01	*
					2	.02	.04	.00	-.11	
	CHECK THE KEY				3	.08	.04	.14	.05	?
	1 was specified, 3 works better				Other	.01	.00	.00	-.42	
105	2-3	.43	.66	.36	1	.47	.60	.14	-.32	
					2	.09	.20	.03	-.24	
					3	.43	.18	.84	.36	*
					Other	.01	.00	.00	-.42	
106	2-4	.63	.34	.24	1	.16	.22	.05	-.22	
					2	.63	.47	.81	.24	*
					3	.19	.27	.11	-.21	
					Other	.02	.00	.00	-.24	
107	2-5	.81	.19	.13	1	.81	.76	.95	.13	*
					2	.08	.13	.00	-.19	
					3	.10	.07	.05	-.04	
					Other	.01	.00	.00	-.34	
108	2-6	.53	.49	.38	1	.53	.27	.76	.38	*
					2	.29	.40	.24	-.25	
					3	.17	.31	.00	-.34	
					Other	.01	.00	.00	-.42	

109	2-7	.67	.53	.31	1	.30	.47	.00	-.34	
					2	.02	.04	.00	-.13	
					3	.67	.47	1.00	.31	*
					Other	.01	.00	.00	-.42	
110	2-8	.69	.36	.29	1	.24	.40	.08	-.35	
					2	.06	.04	.03	-.06	
					3	.69	.53	.89	.29	*
					Other	.01	.00	.00	-.42	
111	2-9	.88	.11	.21	1	.88	.78	.89	.21	*
					2	.09	.16	.11	-.18	
					3	.02	.04	.00	-.10	
					Other	.01	.00	.00	-.42	
112	2-10	.63	.51	.41	1	.63	.38	.89	.41	*
					2	.07	.11	.03	-.25	
					3	.29	.49	.08	-.37	
					Other	.01	.00	.00	-.42	
113	2-11	.78	.20	.18	1	.09	.07	.14	.10	
					2	.12	.24	.00	-.37	
					3	.78	.67	.86	.18	*
					Other	.01	.00	.00	-.42	
114	2-12	.89	.15	.25	1	.89	.80	.95	.25	*
					2	.06	.09	.05	-.08	
					3	.04	.09	.00	-.26	
					Other	.01	.00	.00	-.42	
115	2-13	.76	.37	.34	1	.10	.13	.05	-.15	
					2	.13	.27	.00	-.36	
					3	.76	.58	.95	.34	*
					Other	.01	.00	.00	-.42	
116	2-14	.57	.25	.14	1	.57	.51	.76	.14	*
					2	.16	.27	.08	-.22	
					3	.25	.20	.16	-.09	
					Other	.01	.00	.00	-.42	
117	2-15	.75	.23	.20	1	.04	.07	.03	-.10	
					2	.19	.29	.16	-.21	
					3	.75	.58	.81	.20	*
					Other	.02	.00	.00	-.31	
118	2-16	.89	.20	.37	1	.00	.00	.00		
					2	.89	.78	.97	.37	*
					3	.10	.20	.03	-.36	
					Other	.01	.00	.00	-.42	
119	2-17	.79	.42	.42	1	.14	.29	.00	-.33	
					2	.04	.07	.00	-.25	
					3	.79	.58	1.00	.42	*
					Other	.02	.00	.00	-.35	
120	2-18	.93	.03	.21	1	.93	.91	.95	.21	*
					2	.06	.04	.05	-.13	
					3	.01	.02	.00	-.12	
					Other	.01	.00	.00	-.42	
121	2-19	.29	.11	.08	1	.05	.07	.00	-.24	
					2	.64	.64	.65	-.04	

					3	.29	.24	.35	.08	*
					Other	.01	.00	.00	-.35	
122	2-20	.66	.50	.33	1	.66	.44	.95	.33	*
					2	.18	.31	.03	-.30	
					3	.16	.22	.03	-.23	
					Other	.01	.00	.00	-.42	
123	2-21	.29	.11	.02	1	.01	.04	.00	-.18	
					2	.29	.24	.35	.02	*
					3	.69	.69	.65	-.05	
					Other	.01	.00	.00	-.42	
124	2-22	.87	.19	.24	1	.05	.11	.00	-.22	
					2	.87	.73	.92	.24	*
					3	.07	.13	.08	-.13	
					Other	.01	.00	.00	-.42	
125	2-23	.21	.10	.03	1	.10	.20	.00	-.28	
					2	.21	.20	.30	.03	*
	CHECK THE KEY				3	.69	.58	.70	.07	?
	2 was specified, 3 works better				Other	.01	.00	.00	-.42	
126	2-24	.41	.48	.33	1	.41	.22	.70	.33	*
					2	.34	.51	.14	-.37	
					3	.24	.22	.16	-.10	
					Other	.01	.00	.00	-.33	
127	2-25	.79	.37	.39	1	.08	.24	.00	-.40	
					2	.11	.13	.05	-.17	
					3	.79	.58	.95	.39	*
					Other	.01	.00	.00	-.38	
128	2-26	.73	.21	.20	1	.23	.31	.22	-.14	
					2	.73	.58	.78	.20	*
					3	.03	.09	.00	-.33	
					Other	.01	.00	.00	-.42	
129	2-27	.69	.45	.40	1	.10	.16	.05	-.18	
					2	.69	.44	.89	.40	*
					3	.20	.38	.05	-.40	
					Other	.01	.00	.00	-.42	
130	2-28	.71	.45	.39	1	.71	.47	.92	.39	*
					2	.07	.07	.05	-.03	
					3	.20	.42	.03	-.44	
					Other	.01	.00	.00	-.46	
131	2-29	.92	.05	.13	1	.04	.07	.00	-.09	
					2	.92	.87	.92	.13	*
					3	.03	.02	.08	.08	
					Other	.01	.00	.00	-.46	
132	2-30	.41	.52	.36	1	.41	.16	.68	.36	*
					2	.21	.27	.16	-.18	
					3	.36	.51	.16	-.29	
					Other	.02	.00	.00	-.41	
133	2-31	.28	.18	.12	1	.66	.69	.59	-.07	
					2	.04	.07	.03	-.23	
					3	.28	.20	.38	.12	*
					Other	.01	.00	.00	-.46	
134	2-32	.52	.26	.11	1	.07	.11	.08	-.06	

					2	.40	.51	.32	-.15	
					3	.52	.33	.59	.11	*
					Other	.01	.00	.00	-.46	
135	2-33	.18	.29	.24	1	.74	.82	.51	-.24	
					2	.18	.11	.41	.24	*
					3	.07	.02	.08	.02	
					Other	.01	.00	.00	-.46	
136	2-34	.31	.32	.18	1	.31	.33	.24	-.07	
					2	.37	.38	.19	-.20	
					3	.31	.24	.57	.18	*
					Other	.01	.00	.00	-.46	
137	2-35	.55	.20	.14	1	.55	.44	.65	.14	*
					2	.27	.29	.22	-.16	
					3	.16	.22	.14	-.10	
					Other	.01	.00	.00	-.46	
138	2-36	.07	-.01	.04	1	.69	.69	.59	-.16	
					2	.07	.09	.08	.04	*
	CHECK THE KEY				3	.23	.18	.32	.11	?
	2 was specified, 3 works better				Other	.01	.00	.00	-.46	
139	2-37	.71	.43	.36	1	.11	.16	.03	-.24	
					2	.15	.29	.05	-.22	
					3	.71	.49	.92	.36	*
					Other	.03	.00	.00	-.45	
140	2-38	.31	.52	.39	1	.31	.16	.68	.39	*
					2	.54	.53	.30	-.23	
					3	.12	.22	.03	-.23	
					Other	.03	.00	.00	-.46	
141	2-39	.74	.30	.25	1	.04	.07	.00	-.12	
					2	.20	.24	.08	-.18	
					3	.74	.62	.92	.25	*
					Other	.02	.00	.00	-.50	
142	2-40	.38	-.08	-.01	1	.38	.38	.30	-.01	*
					2	.37	.24	.49	.10	?
	CHECK THE KEY				3	.22	.31	.22	-.18	
	1 was specified, 2 works better				Other	.02	.00	.00	-.50	
143	2-41	.68	.46	.31	1	.68	.40	.86	.31	*
					2	.13	.18	.05	-.12	
					3	.16	.33	.08	-.25	
					Other	.03	.00	.00	-.51	
144	2-42	.76	.42	.38	1	.08	.13	.00	-.19	
					2	.76	.56	.97	.38	*
					3	.10	.20	.03	-.20	
					Other	.05	.00	.00	-.46	
145	2-43	.63	.42	.39	1	.17	.22	.11	-.23	
					2	.63	.42	.84	.39	*
					3	.16	.27	.05	-.25	
					Other	.03	.00	.00	-.51	
146	2-44	.59	.43	.30	1	.13	.13	.08	-.05	
					2	.59	.38	.81	.30	*
					3	.25	.38	.11	-.27	
					Other	.04	.00	.00	-.53	

147	2-45	.69	.11	.15	1	.69	.64	.76	.15	*
					2	.10	.18	.00	-.28	
					3	.17	.07	.24	.11	
					Other	.04	.00	.00	-.49	
148	2-46	.63	.09	.06	1	.14	.18	.11	-.11	
					2	.63	.53	.62	.06	*
	CHECK THE KEY				3	.16	.13	.27	.08	?
	2 was specified, 3 works better				Other	.07	.00	.00	-.43	
149	2-47	.82	.21	.26	1	.10	.11	.05	-.12	
					2	.82	.71	.92	.26	*
					3	.01	.02	.00	-.07	
					Other	.07	.00	.00	-.41	

There were 134 examinees in the data file.

Scale Statistics

Scale:	0	1	2
N of Items	58	44	47
N of Examinees	134	134	134
Mean	43.299	25.903	28.866
Variance	45.687	24.655	38.221
Std. Dev.	6.759	4.965	6.182
Skew	-2.266	-0.705	-0.911
Kurtosis	11.220	4.282	2.680
Minimum	0.000	0.000	0.000
Maximum	55.000	37.000	40.000
Median	44.000	25.000	29.000
Alpha	0.808	0.673	0.783
SEM	2.963	2.839	2.879
Mean P	0.747	0.589	0.614
Mean Item-Tot.	0.250	0.186	0.236
Mean Biserial	0.384	0.278	0.324
Max Score (Low)	41	23	26
N (Low Group)	42	42	45
Min Score (High)	48	28	33
N (High Group)	37	48	37

Scale Intercorrelations

	0	1	2
0	1.000	0.467	0.412
1	0.467	1.000	0.609
2	0.412	0.609	1.000

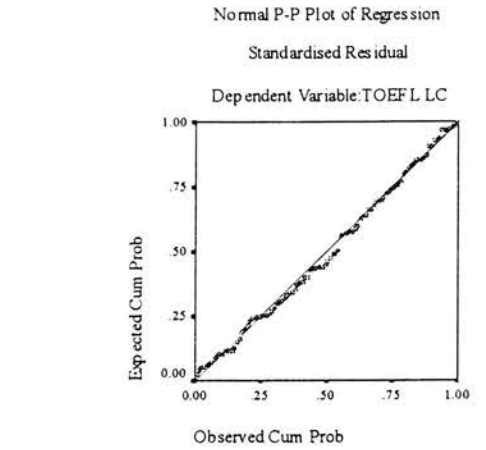
MicroCAT (tm) Testing System

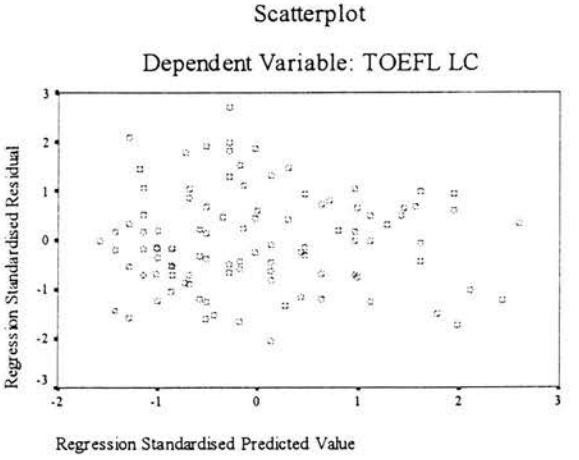
Item and Test Analysis Program -- ITEMAN (tm) Version 3.50

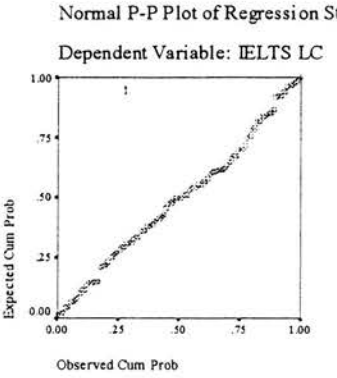
Appendix 10: Test Preparation Impact

Scatter plots and Normal P-P Plot of Regression Standardised Residual

TOEFL PREP COURSE Impact

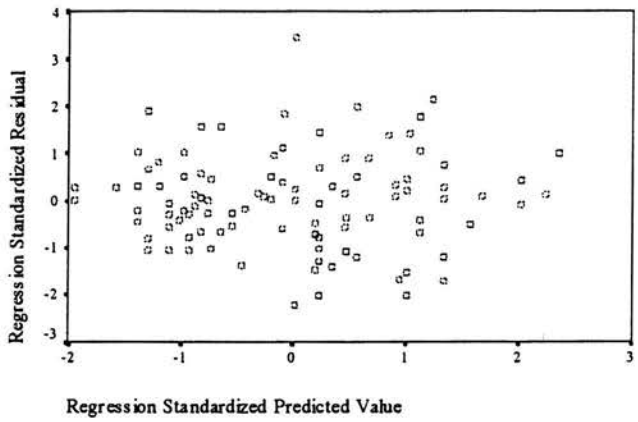






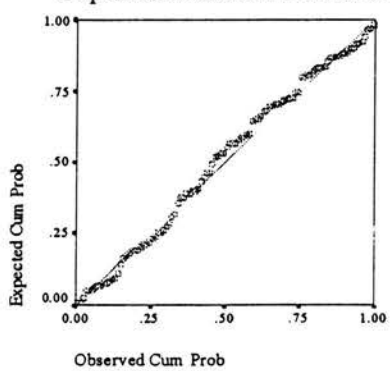
Scatterplot

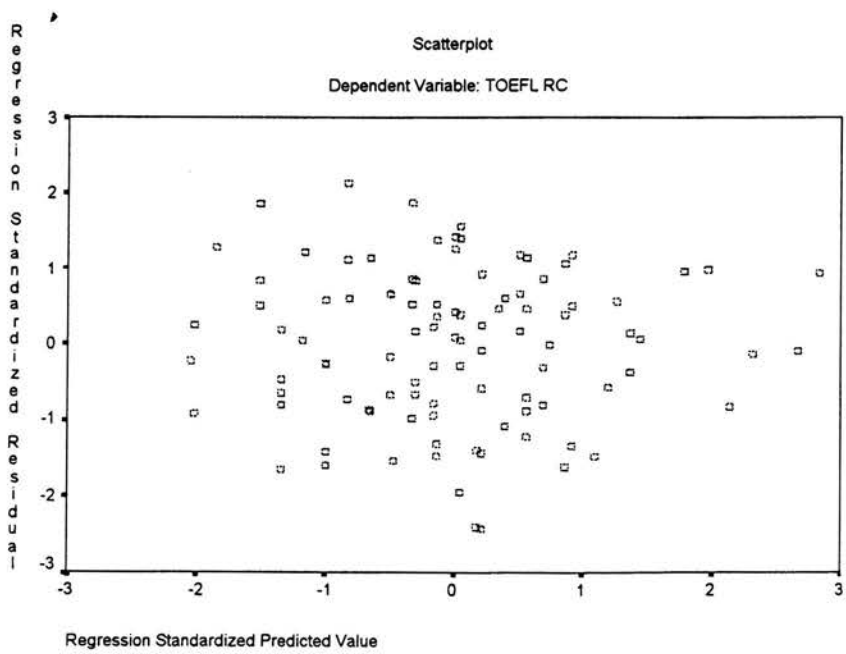
Dependent Variable: IELTS LC

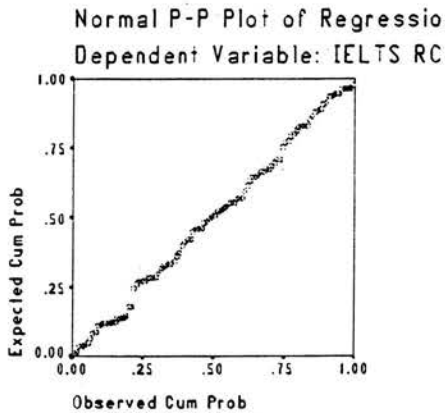


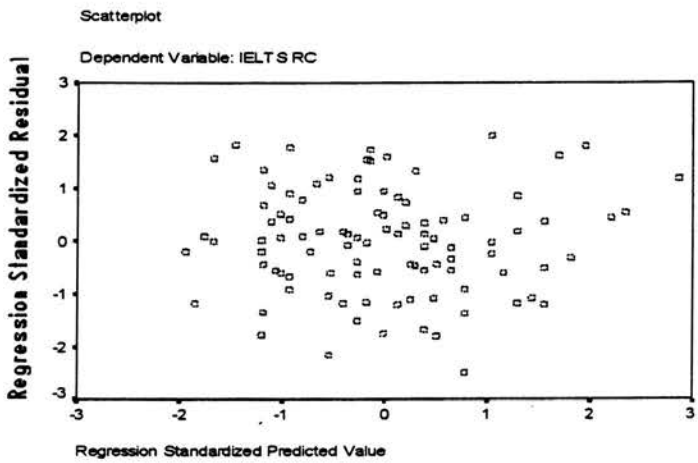
Normal P-P Plot of Regression Standardized Residua

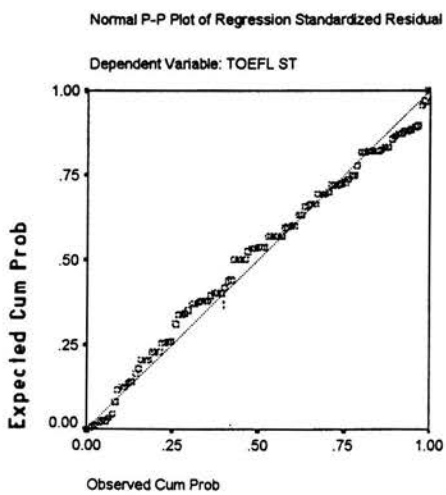
Dependent Variable: TOEFL RC

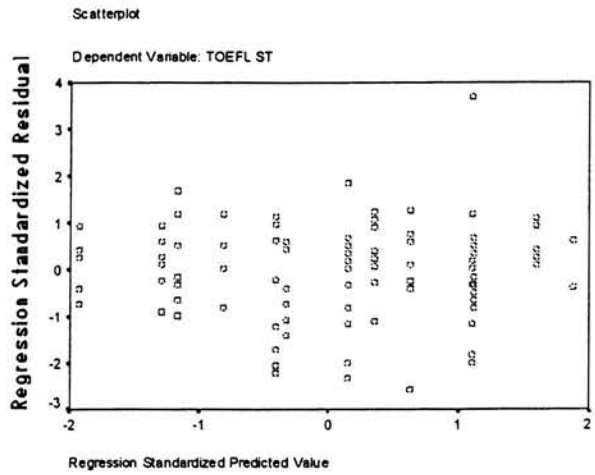


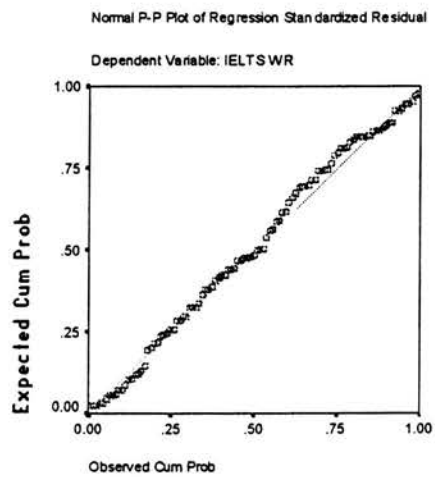


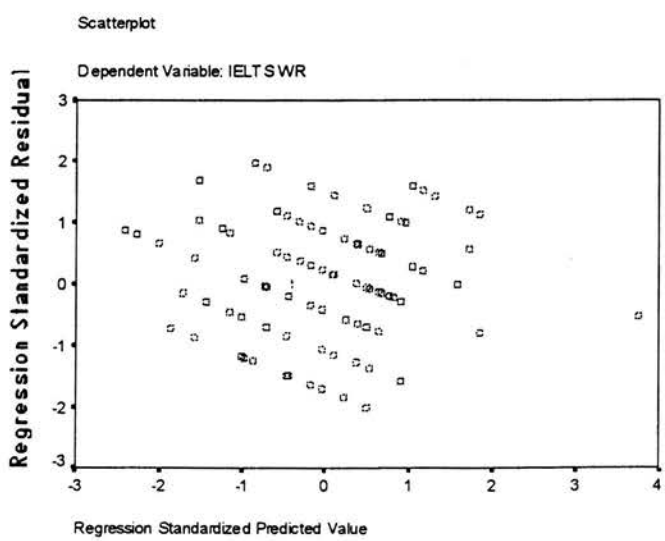

Normal P-P Plot of Regressior
Dependent Variable: IELTS RC
Expected Cum Prob
Observed Cum Prob
0.00
.25
.50
.75
1.00


Scatterplot
Dependent Variable: IELTS RC
Regression Standardized Residual
Regression Standardized Predicted Value
-3
-2
-1
0
1
2
3


Normal P-P Plot of Regression Standardized Residual
Dependent Variable: TOEFL ST
Expected Cum Prob
Observed Cum Prob
0.00
.25
.50
.75
1.00


Scatterplot
Dependent Variable: TOEFL ST
Regression Standardized Residual
Regression Standardized Predicted Value

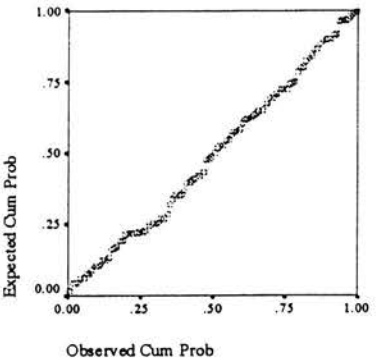

Normal P-P Plot of Regression Standardized Residual
Dependent Variable: IELTS WR
Expected Cum Prob
Observed Cum Prob


Scatterplot
Dependent Variable: IELTS WR
Regression Standardized Residual
Regression Standardized Predicted Value

FCE PREP COURSE Impact

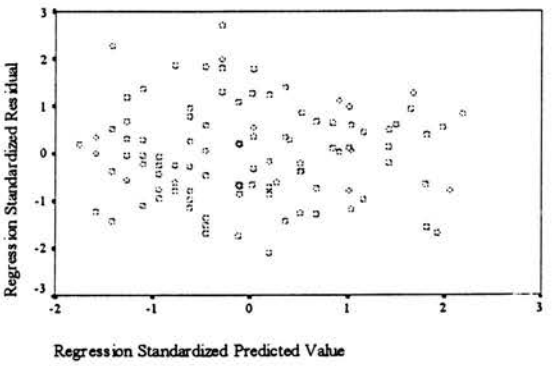
Normal P-P Plot of Regression Standard:

Dependent Variable: TOEFL LC



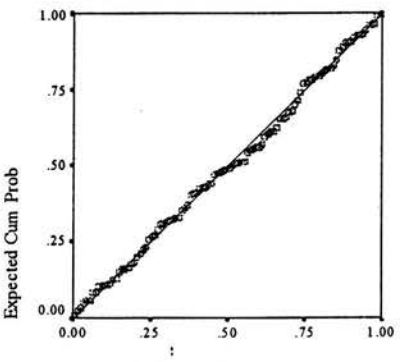
Scatterplot

Dependent Variable: TOEFL LC



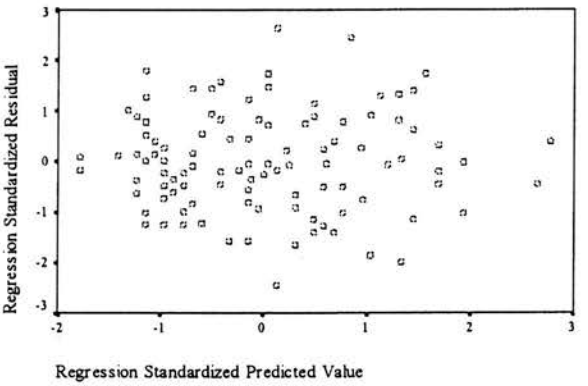
Normal P-P Plot of Regression Standard:

Dependent Variable: IELTS LC



Scatterplot

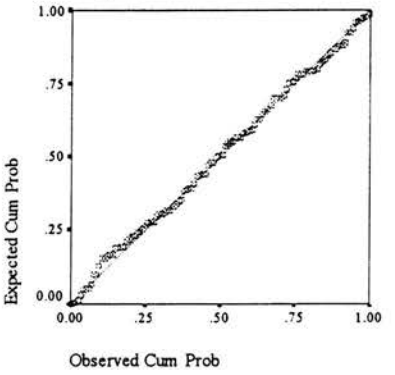
Dependent Variable: IELTS LC



TOEFL PREP Impact on TOEFL Total

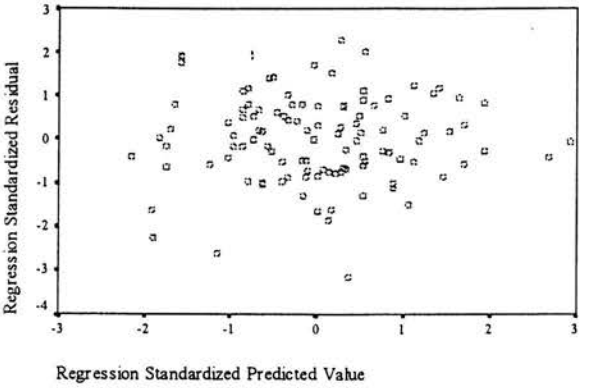
Normal P-P Plot of Regression Standard:

Dependent Variable: EPTB Total Score



Scatterplot

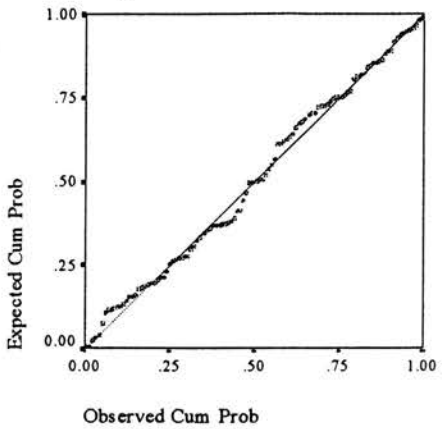
Dependent Variable: EPTB Total Score



FCE PREP Impact on TOEFL Total

Normal P-P Plot of Regression Standarc

Dependent Variable: EPTB Total Score



Scatterplot

Dependent Variable: EPTB Total Score