LANGUAGE IN INDIA
Strength for Today and Bright Hope for Tomorrow
Volume 6 : 1 January 2006
Editor: M. S. Thirumalai, Ph.D.
Associate Editors: B. Mallikarjun, Ph.D.
Sam Mohanlal, Ph.D.
B. A. Sharada, Ph.D.
A. R. Fatihi, Ph.D.
Lakhan Gusain, Ph.D.

FACULTY OF FOREIGN LANGUAGES

TEXT FAMILIARITY, READING TASKS, AND ESP TEST PERFORMANCE: A STUDY ON IRANIAN LEP AND NON-LEP UNIVERSITY STUDENTS

by
Muhammad Ali Salmani-Nodoushan

Supervisor: Professor Parviz Birjandi

Reader: Dr. Seyyed Mohammad Alavi

Dissertation submitted to the Graduate Studies Office in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Teaching English as a Foreign Language (TEFL)

(Azar, 1381)
November, 2002

In the name of God!

We hereby certify that we have read this dissertation written by Muhammad Ali Salmani-Nodoushan, entitled *Text Familiarity, Reading Tasks, and ESP Test Performance: A Study on Iranian LEP and Non-LEP University Students*, and that it is satisfactory in scope and quality as a dissertation for the degree of Doctor of Philosophy in Teaching English as a Foreign Language.

Dr. P. Birjandi (Supervisor) ……

Dr. S. M. Alavi (Reader) ……

Dr. M. Nowroozi Khiabani (External Examiner) ……

Dr. P. Maftoon (External Examiner) ……

Dr. A. A. Farahani (Internal Examiner) ……

Dr. M. Nemati (Internal Examiner) ……

May 27, 2002

ACKNOWELEDGEMENTS

I have to thank so many people for the important role they played in the development of this study that I cannot do justice to everybody's contribution. First of all, I am greatly indebted to Professor Parviz Birjandi whose help and support during this study were manifested in more ways than I can say without writing a whole other dissertation on the subject. I also acknowledge the cooperation and assistance of Dr. Seyyed Mohammad Alavi, who not only ushered me through the maze of test development and validation, but also taught me how to analyze and interpret the data, and provided me with useful comments which enabled me to convert turgid prose into readable English. Special thanks also go to the examiners who evaluated the end-product of the study both qualitatively and quantitavely.

Special thanks also go to my fellow university instructors at Shahid Bahonar University of Kerman, Shiraz University, and Azad University of Bushehr who helped me in the painstaking process of data collection. I would also like to thank the junior and senior students (majoring in electronics) at these universities who spent hours of their precious time responding to the research instruments.

Finally, I wish to thank my wife, Arezou, for suffering through the birth pangs of this study. She insisted, much to my chagrin, that variety is not the spice of life when it comes to Ph.D. dissertation. I hope she will regard the result as worth her sufferings.

To all, many thanks again for your assistance and encouragement. May we all continue to learn from each other.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

In a study of the effects of text familiarity, task type, and language proficiency on university students' LSP test and task performance, 541 senior and junior university students majoring in electronics took the Task-Based Reading Test (TBRT). The results indicated that the effect of each of these factors on subjects' test and performances was statistically significant. Moreover, the impact of the interactions between any given pair and also among all three of these factors on subjects' test performance was statistically significant.

Subjects' performance on different tasks at the same level of text familiarity afforded statistically significant results. The semi-and no-proficient subjects did not perform significantly different in the following contexts: (a) true-false, sentence-completion, and writer's-view tasks in partially familiar tests; (b) outlining, writer's-view, true-false, and sentence completion tasks in totally unfamiliar tests; and (c) sentence-completion, outlining, and writer's-view tasks in totally familiar tests. The differences found in subjects' performances on the same tasks at different levels of text familiarity were also significant. However, the difference between semi- and non-proficient subjects' performance was not statistically significant when they performed (a) the true-false task in partially familiar versus totally familiar contexts, and (b) outlining, sentence-completion, and writer's-view tasks along the text-familiarity cline. In a comparison of different tasks, subjects' performance of the sentence-completion task was found to be significantly different from their performance of the other four tasks in question along the text-familiarity cline. Moreover, subjects' performances of the writer's-view and the true-false tasks in totally unfamiliar contexts differed significantly. In addition, regression analyses revealed that the greatest influence on subjects' overall and differential test and task performance was due to language proficiency.

CHAPTER ONE
PRELIMINARIES

1. Introduction

Individuals are not only speakers, but also receivers, consumers, readers and interpreters of language. The extent to which one knows and uses a foreign language may be crucial to one's existence, education, relationships, and careers. As soon as the world entered the new millennium, the ability and the need to understand and communicate with others became increasingly important, at times even urgent. Today, an international exchange of ideas—from environmental issues such as the thinning ozone layer and the warming of the planet, to medical topics such as genetic engineering, to political crises—is essential. To meet these communication needs, more and more individuals have highly specific academic and professional reasons for seeking to improve their language skills: for these students, usually adults, courses that fall under the heading English for Special Purposes (ESP) hold particular appeal. ESP can help people to become better professionals, which may reflect on their whole lives.

The recent increase in ESP publications, conference presentations, professional gatherings, Web sites, e-mail lists, invited lectures, consulting requests, and model program study-tours adequately testifies to the fact that ESP has gained a significant place in the world. Motivated by new policies and priorities at national and local levels, many universities worldwide are now in the midst of rethinking their English language curriculums and searching for better options. Therefore, for more effective instructional content, ESP is considered an intelligent option. Beginning in the early 1960s, English for Specific Purposes (ESP) has grown to become one of the most prominent areas of

EFL teaching today. Its development is reflected in the increasing number of universities offering an MA in ESP (e.g. The University of Birmingham, and Aston University in the UK) and in the number of ESP courses offered to overseas students in English speaking countries. There is now a well-established international journal dedicated to ESP discussion (*English for Specific Purposes: An international journal*), and the ESP SIG groups of the IATEFL and TESOL are always active at their national conferences.

2. The Difference between ESP and EGP

What are the differences between English for Specific Purposes (ESP) and English for General Purposes (EGP)? Hutchinson and Waters (1987) answer this quite simply: "in theory, nothing, in practice, a great deal." On the face of it, ESP differs from EGP in the sense that the words and sentences learned, the subject matter discussed, all relate to a particular field or discipline—for example, a lawyer writing a brief, or a diplomat preparing a policy paper. ESP courses make use of vocabulary and tasks related to the field such as negotiation skills and effective techniques for oral presentations. The entire program is designed to meet the specific professional or academic needs of the learner. A balance is created between educational theory and practical considerations. ESP also increases students' skills and confidence in using English.

A closer look at EGP and ESP is, however, vital. English for General Purposes (EGP) is essentially the English language education in junior and senior high schools. Students are introduced to the sounds and symbols of English, as well as to the lexical/grammatical/rhetorical elements that compose spoken and written discourse. EGP also focuses on applications in general situations: appropriate dialogue with restaurant staff, bank tellers, postal clerks, telephone operators, English teachers, and party guests as well as lessons on how to read and/or write the English typically found in textbooks, newspaper and magazine articles, telephone books, shopping catalogues, application forms, personal letters, e-mail, and home pages. Supplementary information about appropriate gestures, cultural conventions, and cultural taboos is also normally included in EGP curriculums. EGP conducted in English-speaking countries is typically called ESL, and EGP conducted in non-English-speaking countries is normally called EFL. Pedagogically, a solid understanding of basic EGP should precede higher-level instruction in ESP if ESP programs are to yield satisfactory results.

English for Specific Purposes (ESP), however, is research and instruction that builds on EGP and is designed to prepare students or working adults for the English used in specific disciplines, vocations, or professions to accomplish specific purposes. ESP researchers have come to the conclusion that ESP is defined to meet specific needs of the learners. ESP makes use of methodology and activities of the discipline it serves, and is centered on the language appropriate to these activities. According to Hutchinson and Waters (1987:19), "ESP is an approach to language teaching in which all decisions as to content and method are based on the learner's reason for learning." In this connection, it is interesting to note that, as Dudley-Evans (1998) explains, ESP may not always focus on the language of one specific discipline or occupation, such as English for Law or English for Physics. University instruction that introduces students to common features of academic discourse in the sciences or humanities, frequently called English for Academic Purposes (EAP), is equally ESP.

3. ESP Revisited

ESP has had a relatively long time to mature and so one would expect the ESP community to have a clear idea about what ESP means. Strangely, however, this does not seem to be the case. Recently a very heated debate took place on the TESP-L e-mail discussion list about whether or not English for Academic Purposes (EAP) could be considered part of ESP in general. At the Japan Conference on ESP (1997) also, clear differences in how people interpreted the meaning of ESP could be seen. Some ESP scholars described ESP as simply being the teaching of English for any purpose that could be specified. Others, however, were more precise, describing it as the teaching of English used in academic studies or the teaching of English for vocational or professional purposes. At this conference, Dudley-Evans clarified the meaning of ESP, giving an extended definition of ESP in terms of 'absolute' and 'variable' characteristics.

Strevens (1988) defines ESP in terms of its absolute and variable characteristics. His definition makes a distinction between four absolute and two variable characteristics. In terms of absolute characteristics, ESP consists of English language teaching which is (1) designed to meet specified needs of the learner, (2) related in content (i.e. in its themes and topics) to particular disciplines, occupations and activities, (3) centered on the language appropriate to those activities in syntax, lexis, discourse, semantics, etc., and analysis of this discourse, and (4) in contrast with General English. In terms of variable

characteristics, ESP may be, but is not necessarily, (1) restricted as to the language skills to be learned (e.g. reading only), and (2) not taught according to any pre-ordained methodology.

Anthony (1997) refers to the considerable recent debate on the meaning of ESP despite the fact that it is an approach which has been widely used over the last three decades. Dudley-Evans (1997) offered a modified definition for ESP. The revised definition Dudley-Evans and St. John postulate is the extension of the definition proposed by Strevens (1988) in terms of absolute and variable characteristics. According to Dudley-Evans and St. John, in terms of absolute characteristics, ESP (1) is defined to meet specific needs of the learner, (2) makes use of the underlying methodology and activities of the discipline it serves, and (3) is centered on the language (grammar, lexis, register), skills, discourse and genres appropriate to these activities. In terms of the variable characteristics, ESP (1) may be related to or designed for specific disciplines, (2) may use, in specific teaching situations, a different methodology from that of general English, (3) is likely to be designed for adult learners, either at a tertiary level institution or in a professional work situation, and could also be for learners at secondary school level, (4) is generally designed for intermediate or advanced students, (5) assume some basic knowledge of the language system, and (6) can be used with beginners.

A comparison of this latter definition with that of Strevens reveals that Dudley-Evans and St. John have removed the absolute characteristic that 'ESP is in contrast with General English' and added more variable characteristics. They assert that ESP is not necessarily related to a specific discipline. Furthermore, ESP is likely to be used with adult learners although it could be used with young adults in a secondary school setting. The definition Dudley-Evans offered is clearly influenced by that of Strevens (1988), although he has improved it substantially by removing the absolute characteristic that ESP is "in contrast with 'General English'" (Johns and Dudley-Evans, 1991: 298), and has included more variable characteristics. The division of ESP into absolute and variable characteristics, in particular, is very helpful in resolving arguments about what is and is not ESP. From Dudley-Evans' definition, one can see that ESP can be (though not necessarily so) concerned with a specific discipline, nor does it have to be aimed at a certain age group or ability range. ESP should be seen simply as an 'approach' to teaching, or what Dudley-Evans describes as an attitude of mind. This is a similar conclusion to that made by Hutchinson and Waters (1987:19)

who state, “ESP is an approach to language teaching in which all decisions as to content and method are based on the learner’s reason for learning.”

Along the same lines, Hutchinson and Waters (1987) proposed a broader definition of ESP. They theorize ESP to be an approach to language teaching in which all decisions as to content and method are based on the learner’s reason for learning. Anthony (1997) noted that, with this broad definition, it is not clear where general English courses end and ESP courses begin. Numerous non-specialist ESL instructors use an ESP approach in that their syllabuses are based on analysis of learner needs and their own personal specialist knowledge of using English for real communication.

In general, “special language” and “specialized aim” are viewed as similar concepts although they are two entirely different notions. Perren (1974) noted that confusion arises over these two notions. Mackay and Mountford (1978) stated that the only practical way in which we can understand the notion of “special language” is as a restricted repertoire of words and expressions selected from the whole language because that restricted repertoire covers every requirement within a well-defined context, task or vocation. On the other hand, a “specialized aim” refers to the purpose for which learners learn a language, not the nature of the language they learn. Consequently, the focus of the word “special” in ESP is on the purpose for which learners learn and not on the specific jargon or registers they learn. As such, all instances of language learning might be considered ESP.

4. Background

Hutchinson and Waters (1987) succinctly identified three key reasons they believe are common to the emergence of all ESP: the demands of a Brave New World, a revolution in linguistics, and focus on the learner. They noted that two key historical periods breathed life into ESP. First, the end of the Second World War brought with it an age of enormous and unprecedented expansion in scientific, technical and economic activity on an international scale. For various reasons, most notably the economic power of the United States in the post-war world, the role of international language fell to English. Second, the Oil Crisis of the early 1970s resulted in Western money and knowledge flowing into the oil-rich countries. The language of this knowledge became English. The general effect of all this development was to exert pressure on the language teaching profession to deliver the required goods. Whereas English had

previously decided its own destiny, it now became subject to the wishes, needs, and demands of people other than language teachers.

The second key reason that had a tremendous impact on the emergence of ESP was a revolution in linguistics. Whereas traditional linguists set out to describe the features of language, revolutionary pioneers in linguistics began to focus on the ways in which language is used in real communication. Hutchinson and Waters (1987) point out that one significant discovery was in the ways that spoken and written English vary. In other words, given the particular context in which English is used, the variant of English will change. This idea was taken one step farther. If language in different situations varies, then tailoring language instruction to meet the needs of learners in specific contexts is also possible. Hence, in the late 1960s and the early 1970s there were many attempts to describe English for Science and Technology (EST).

The final reason which Hutchinson and Waters (1987) cite to have influenced the emergence of ESP has less to do with linguistics and everything to do with psychology. Rather than simply focusing on the method of language delivery, more attention was given to the ways in which learners acquire language and the differences in the ways language is acquired. Learners were seen to employ different learning strategies, use different skills, enter with different learning schemata, and be motivated by different needs and interests. Therefore, focus on the learners' needs became equally paramount as the methods employed to disseminate linguistic knowledge. As such, designing specific courses to better meet individual needs was a natural extension of "learner-centered" or "learning-centered" perspectives on ESP.

5. Types of ESP

Carver (1983) identifies three types of ESP: English as a Restricted Language, English for Academic and Occupational Purposes (EAOP), and English with Specific Topics. The language used by air traffic controllers or by waiters are examples of English as a restricted language. Mackay and Mountford (1978: 4-5) clearly illustrate the difference between restricted language and language with this statement: "… the language of international air-traffic control could be regarded as 'special', in the sense that the repertoire required by the controller is strictly limited and can be accurately determined situationally, as might be the linguistic needs of a dining-room waiter or air-hostess. However, such restricted repertoires are not languages, just as a tourist phrase book is not grammar. Knowing a restricted 'language' would not allow the speaker to

communicate effectively in novel situation, or in contexts outside the vocational environment."

The second type of ESP is English for Academic and Occupational Purposes. Carver (1983) indicates that this English should be at the heart of ESP although he refrains from developing it any further. Hutchinson and Waters (1987), on the other hand, have developed a "Tree of ELT" in which the subdivisions of ESP are clearly illustrated. ESP is broken down into three branches: English for Science and Technology (EST), English for Business and Economics (EBE), and English for Social Studies (ESS). Each of these subject areas is further divided into two branches: English for Academic Purposes (EAP) and English for Occupational Purposes (EOP). An example of EOP for the EST branch is "English for Technicians" whereas an example of EAP for the EST branch is English for Medical Studies.

Hutchinson and Waters (1987) do note that there is not a clear-cut distinction between EAP and EOP on the basis of the considerations that (1) people can work and study simultaneously, and that (2) it is also likely that in many cases the language learnt for immediate use in a study environment will be used later when the student takes up, or returns to, a job. Perhaps this explains the rationale for categorizing EAP and EOP under the same type of ESP. It appears that the end purposes of both EAP and EOP are the same: employment. This cannot be contested. However, despite the end purpose being identical, the means taken to achieve the end is very different indeed. One has no other choice than to contend that EAP and EOP are different in terms of focus on Cummins' (1979) notions of cognitive academic proficiency versus basic interpersonal skills.

The third and final type of ESP is English with specific topics. It is only here where emphasis shifts from purpose to topic. This type of ESP is uniquely concerned with anticipated future English needs of, for example, scientists requiring English for postgraduate reading studies, attending conferences or working in foreign institutions. However, it is possible to argue that this is not a separate type of ESP. Rather it is an integral component of ESP courses or programs which focus on situational language. This situational language has been determined based on the interpretation of results from needs analysis of authentic language used in target workplace settings.

In brief, there are three features common to ESP: (a) authentic materials, (b) purpose-related orientation, and (c) self-direction. These features are indeed useful in attempting to formulate one's own understanding of ESP. If we revisit Dudley-Evans'

(1997) claim that ESP should be offered at an intermediate or advanced level, use of authentic learning materials is entirely feasible. The use of authentic content materials, modified or unmodified in form, is indeed a feature of ESP, particularly in self-directed study and research tasks. Purpose-related orientation, on the other hand, refers to the simulation of communicative tasks required of the target setting, for example, student simulation of a conference, involving the preparation of papers, reading, notetaking, and writing. Finally, self-direction is characteristic of ESP courses in that the point of including self-direction is that ESP is concerned with turning learners into users. In order for self-direction to occur, the learners must have a certain degree of freedom to decide when, what, and how they will study. There must also be a systematic attempt by teachers to teach the learners how to learn by teaching them about learning strategies (Hutchinson and Waters, 1987; Dudley-Evans, 1997 and 1998; Shohamy, 1995; Douglas, 2000).

6. Five Concepts Underlying ESP

Five conceptions function as what might be called the foundations, essential features or basic principles of ESP. Swales (1990) uses the term `enduring conceptions' to refer to them. As originally formulated, these five conceptions are: authenticity, research-base, language/text, need, and learning/methodology. According to Coffey (1984), the main consideration in ESP is that of authenticity. It includes authentic texts, and authentic tasks. Swales (1990), in explaining what is meant by the research-base of ESP, reviews the ESP literature and observes a trend towards papers that rely on some kind of data-base (textual or otherwise). In addition, Strevens (1977) alludes to the importance of the "specific language" of ESP. that is, only those items of vocabulary, patterns of grammar, and functions of language which are required by the learner's purposes are included in ESP. Strevens also alludes to the importance of learner needs in discussions of ESP. Finally, ESP draws on the methodology or learning theories which are appropriate to the learning/teaching situation. In other words, specific purpose language teaching (SPLT) is not itself a methodology. According to Strevens (1977), this characteristic of ESP makes the materials both more relevant and more interesting to the student due to the varied and ingenious exploitation of the opportunities provided by ESP settings.

These five conceptions all have dual and potentially conflicting origins in both the real world (the 'target situation' of the ESP) and in ESP pedagogy. It is therefore crucial

to discuss each of them in an attempt to survey the development and directions of ESP as it has evolved. Such a survey will identify five major approaches to ESP, each of which has focused on one of the major conceptions and thus contributed to the concept of ESP itself. However, it is also evident that as each approach to ESP has itself evolved, its particular enduring conception has also evolved, bringing ESP practitioners towards their current thinking in each of the five areas.

The earliest concept to emerge from the development of ESP was that of authenticity. The first generation of ESP materials that appeared in the mid-1960s took skills as their principal means of selection, arguing that ESP teachers would need to establish the skills priorities of students in order to develop appropriate ESP teaching materials. The definition of skill was somewhat broad, establishing little more than the ranking of the four usual language skills of reading, writing, listening and speaking. Of these, it was almost always reading that was singled out in early materials. Almost all such materials consist of collections of specialist texts with accompanying comprehension and language exercises. As Close (1992) rightly argues, the conception of authenticity was central to the approach taken to the reading skill.

Authenticity has gained so much significance that even today, most ESP programs focus on developing communicative competence in specific fields, such as aviation, business, technology, etc. Some courses prepare students for various academic programs. For example, Yale University offers a seminar for Graduate Teaching Assistants (GTA) that emphasizes training in public speaking skills and uses videos to improve teaching and lecturing styles. Others prepare students for work in fields such as law, medicine, engineering, tourism or graphic design. Many courses now focus on the Internet. Still there is a gap between students' real life needs and what a common ESP course book can suggest. One inherent flaw of this short-sighted view of authenticity is that very often, instead of conducting interviews with specialists in the field, analyzing the language that is required in the profession or even conducting students' needs analysis, many ESP teachers become dependent only on the published textbooks available.

This conception of authenticity was limited in several ways: it was confined to authenticity of text, with no differentiation between different kinds of scientific/technical texts. A closer examination of the texts reveals that authenticity was being contrasted with simplification, in that the vocabulary and grammar were not simplified in any way. However, it is also apparent that authenticity did not exclude

editing by omitting long sections, especially if these involved complex language. Moreover, there was no thought of authenticity of task in the early conception

Later skills-based approaches to ESP enlarged the conception of authenticity in two principal ways. First, authenticity of text was both broadened to include texts other than written texts and narrowed to differentiate between the different types of text generated by each skill. Reading, for example, could be sub-divided into reading reports, reading technical journals, reading instruction manuals, etc. Secondly, the conception of authenticity was enlarged to embrace authenticity of task. In effect, this meant designing tasks requiring students to process texts as they would in the real world. In other words, ESP learners were required to use ESP materials which employed the same skills and strategies as would be required in the target situation (Morrow, 1980).

Halliday, McIntosh and Strevens (1964) were the first scholars who pointed to the importance of, and the need for, a research base for ESP, set out in one of the earliest discussions of ESP. This was a call for a program of research into ESP registers which was taken up by several early ESP materials writers, such as Herbert (1965) or Ewer and Latorre (1969), who analyzed large corpora of specialist texts in order to establish the statistical contours of different registers. The principal limitation of this approach was not its research base but its conception of text as register, restricting the analysis to the word and sentence levels as register was invariably defined in these terms. The procedure adopted for the analysis was twofold. The main structural words and non-structural vocabulary were identified by visual scanning. For the main sentence patterns, a small representative-sample count was made.

In the 1990s, there has been a number of ESP projects which have grown out of concerns for international safety and security. The first of these was SEASPEAK. It was a practical project in applied linguistics and language engineering. According to Strevens and Johnson (1983), SEASPEAK was the establishment for the first time of an International Maritime English. SEASPEAK was published in 1987-88 and followed by AIRSPEAK (1988) and POLICESPEAK (1994), with RAILSPEAK in preparation. Each of these projects involved a substantial research phase with linguists and technical specialists cooperating. The NEWSPEAK research shared the large-scale base of the register-analysis approach but the principal advance was that it was now applied to a more sophisticated, four-level concept of text: purposes of maritime communication, operational routines, topics of maritime communication, and discourse procedures. Although register analysis remains small-scale and restricted to native-speaker

encounters, later research demonstrated the gap between ESP materials designers' intuitions about language and the language actually used in ESP situations (Williams, 1988; Mason, 1989; Lynch and Anderson, 1991; Jones, 1990).

The reaction against register analysis in the early 1970s concentrated on the communicative values of discourse rather than the lexical and grammatical properties of register. The approach was clearly set out by two of its principal advocates, Allen and Widdowson (1974). They specifically argued that one might usefully distinguish two kinds of ability which an English course at ESP level should aim at developing. The first is the ability to recognize how sentences are used in the performance of acts of communication, or the ability to understand the rhetorical functioning of language in use. The second is the ability to recognize and manipulate the formal devices which are used to combine sentences to create continuous passages of prose. One might say that the first has to do with rhetorical coherence of discourse, the second with the grammatical cohesion of text.

In practice, however, the discourse-analysis approach tended to concentrate on 'how sentences are used in the performance of acts of communication' and to generate materials based on functions. The main shortcoming of the approach was that its treatment remained fragmentary, identifying the functional units of which discourse was composed at sentence/utterance level but offering limited guidance on how functions and sentences/utterances fit together to form text.

As an offspring of discourse analysis, the genre-analysis approach seeks to see text as a whole rather than as a collection of isolated units. According to Johnson (1995), this is achieved by seeking to identify the overall pattern of the text through a series of phases or 'moves'. The major difference between discourse analysis and genre analysis is that, while discourse analysis identifies the functional components of text, genre analysis enables the materials writer to sequence these functions into a series to capture the overall structure of such texts. The limitation of genre analysis has been a disappointing lack of application of research to pedagogy. There are few examples of teaching materials based on genre-analysis research.

ESP is driven by the specific learning needs of the language learner. The first step for ESP is research to identify the specific learning needs of students, for these will inform the decisions made about ESP programs. Before beginning a needs analysis, however, one must first answer the following crucial question: "Will the students use English at university or in their jobs after graduation?" If the answer is no, then ESP is

not a reasonable option for the university's English language program. The university will have to justify its existence and improve the program via other means. If the answer is yes, however, then ESP is probably the most intelligent option for the university curriculum. ESP needs analysis lays a solid foundation for a stable ESP program. ESP also begins with some basic questions to survey what will be needed. Will students use English at university or in their jobs after graduation? In what situations? For what purposes? What language skills will be required (reading, writing, listening, speaking)? What are the significant characteristics of the language in these situations (lexicon, grammar, spoken scripts, written texts, other characteristics)? What extralinguistic knowledge of academia, specific disciplines, specific vocations, or specific professions is required for successful English usage in these areas?

Needs analysis was firmly established in the mid-1970s as course designers came to see learners' purposes rather than specialist language as the driving force behind ESP. Early instruments, notably Munby's (1978) model, established needs by investigating the target situation for which learners were being prepared. Munby's model clearly established the place of needs as central to ESP, indeed the necessary starting point in materials or course design. However, his model has been widely criticized for two apparently conflicting reasons: (1) its over-fullness in design, and (2) what it fails to take into account (that is, socio-political considerations, logistical considerations, administrative considerations, psycho-pedagogic, and methodological considerations).

To counter the shortcomings of target-situation needs analysis, various forms of pedagogic needs have been identified to give more information about the learner and the educational environment. These forms of needs analysis should be seen as complementing target-situation needs analysis and each other, rather than being alternatives. They include deficiency analysis, strategy analysis, and means analysis. Deficiency analysis gives us information about what the learners' learning needs are (i.e., which of their target-situation needs they lack or feel they lack). This view of needs analysis gains momentum when we consider that the question of priorities is ignored by standard needs analysis. In discussing learners' perceptions of their needs, deficiency analysis takes into account lacks and wants, as well as objective needs of the learners (Allwright, 1982). Strategy analysis seeks to establish how the learners wish to learn rather than what they need to learn. By investigating learners' preferred learning styles and strategies, strategy analysis provides a picture of the learner's conception of learning. Means analysis, on the other hand, investigates precisely those considerations

that Munby excluded. These relate to the educational environment in which the ESP course is to take place. (Swales, 1989).

The attention to strategy analysis gave rise to a new generation of ESP materials which was founded as much on conceptions of learning as on conceptions of language or conceptions of need. The concern in ESP was no longer with language use—although this would help to define the course objectives. The concern was rather with language learning. It was no longer simply assumed that describing and exemplifying what people do with language would enable someone to learn it. A truly valid approach to ESP would be based on an understanding of the processes of language learning. Hutchinson and Waters (1987) called this approach the learning-centered approach and stressed the importance of a lively, interesting and relevant teaching/learning style in ESP materials. The first ESP materials to adopt a conscious model of learning were probably those of the Malaysian UMESPP project in the late 1970s, but the approach has received its widest circulation in the papers and materials of Hutchinson and Waters, and, more recently, Waters and Waters (1992).

In sum, the five conceptions underlying ESP have endured, although there have clearly been points at which they have taken new directions as they have evolved. While these conceptions have now reached a maturity which serves ESP well, there continue to be tensions arising from their application to practical materials design. In part, these tensions derive from a conflict between real-world and pedagogic conceptions, which frequently involves a 'trade-off' between the two. These tensions include some of the following areas: target needs versus pedagogic needs, target authenticity versus materials design, language as text versus pedagogic texts, method/learning style versus content-driven materials, and research-data findings versus materials design. It may be that these tensions are inevitable and irreconcilable. However, ESP teachers and materials designers should be aware of them and make informed decisions based on their knowledge of the language, the target situation, the educational environment, and the learners. The same informed decisions are also needed when LSP testers set out to develop LSP tests (Douglas, 2000; Clapham, 1996; Dudley-Evans, 1998).

7. The Need for ESP Testing

In this rapidly changing world when teaching English for Specific Purposes (ESP) has grown to become one of the most prominent areas of teaching English as a foreign language (EFL), ESP practitioners face new opportunities and challenges. As the ability

to speak more than one language becomes more important, so too does the need to assess the language abilities of second language learners. In the classroom, assessment can be seen as an ongoing process, in which the teacher uses various tools to measure the progress of the learner. Among those tools are portfolios, self-assessment, and, of course, tests. If assessment can be seen as a movie, then a test is a freeze frame: it gives a picture of the learner's language at a particular point in time. Used properly, these tools can help the teacher develop a full picture of the learner's progress. It is important to note that all types of testing and assessment are important in gathering information about students.

Testing has traditionally been the most widely used assessment tool in the classroom, and in many classrooms, it still is. Moreover, testing has applications outside of the classroom. Foreign language programs test students for placement, colleges and universities test students for credit, and employers test the abilities of prospective employees. In any testing situation, it is important to consider which of the four skills (speaking, listening, reading, writing) needs to be assessed, who will be taking the test, and for what purposes the test results will be used. Clearly, a test which is appropriate in one situation may be inappropriate in another: a test designed to measure the reading abilities of elementary school learners will not be appropriate for college placement. Thus, when choosing a test to use, it is important to define the testing situation, and then to find or develop a test that fits the situation. It is also important to know the reliability and validity of the test, especially if the test is to be used for high-stakes purposes, such as entrance into a college or university. Reliability measures the consistency of the test; validity is the extent to which the test measures what it claims to measure. Large scale standardized tests have more reliability and validity requirements than classroom tests, and many books, articles, research projects, and other materials have been devoted to this issue.

The need for tests of ESP has grown out of the conceptions that function as the pedestals upon which ESP stands. As was pointed above an analysis of the communicative needs of students is at the heart of ESP. identifying the language targets toward which students must aim is, in turn, at the heart of such needs analysis. It does not, however, comprise the whole of needs analysis. ESP practitioners must also discover (1) where ESP students currently stand and (2) how much distance lies between them and the target before they can begin to determine (3) where instruction is necessary. Many good publications exist on language testing, so the issue need not be

discussed here in detail; however, it should be noted that the goal of testing for ESP instruction is to determine (4) what portions of the target language students don't know—not to test their knowledge of EGP. TOEFL and other popular tests of English language proficiency can be, as they claim to be, useful for testing how much general English students know. However, they can't provide adequate data on student competence in the spoken scripts and written texts characteristic of a specific discipline or vocation, such as electronics, accounting, medicine or shipbuilding. The vocabulary and grammatical/rhetorical structures that surface most frequently in many work situations lie beyond the narrow range of English tested in popular standardized tests.

These concerns were the backbone of ESP tests. Specifically developed ESP tests that could be used for purposes of selection, achievement, and content-area proficiency would, needless to say, intrigue anyone. As Douglas (2000) argues, the rapid interest in performance assessment made ESP tests even more intriguing. Ethical considerations as well as the notion of authenticity in testing provided further support for the claim that ESP tests were not only needed but also vital. These as well as many practical considerations such as economy, time limits, and the rapid developments of scientific fields made ESP tests even more vital. Businesses needed to select the fittest students for their future job vacancies. University departments needed to measure exactly "how fit" their students were. ESP teachers needed to know how much their students had attained the objectives of ESP courses. Policy makers, too, needed to have an exact estimation of the amount of money they were supposed to invest in language programs, and also a precise estimation of the amount of return they hoped to gain.

8. The Need for Research into ESP Testing

The rapid expansion in ESP teaching was not accompanied by a similar increase in ESP testing. Perhaps the earliest attempts at testing ESP dates back to the time when the ELTS was launched. At that time, in 1980, there had been little or no research into the validity of giving academic students English proficiency tests that were based on different subject areas. Alderson (1981), in a discussion on ESP testing, questioned many of the principles behind this approach. He agreed that since different university departments placed different demands on their students, there were some good arguments for including ESP tests in an EAP test battery. He felt that a comparison between performance on academically specific tests and the communicative needs of the relevant area might provide useful diagnostic information. He also accepted that

ESP tests would have a really high face validity for both content-area students and university lectures. However, he questioned whether it was possible to produce a test which would be equally suitable for students in all branches of a discipline. For example, he wondered whether it would be possible to have a test for engineers that was of the same level of appropriacy for all engineers, regardless of their specialization. This highlights one of the main difficulties with English for Specific Academic Purposes (ESAP) testing.

Another difficulty with ESP tests was delineated in Alderson's (1981) question 'how specific is specific?'. Since it was at that time usually impossible to give each student a test which was tailor-made for a unique set of circumstances, any ESP test had to be a compromise and in the case of EAP, where many disciplines would be subsumed under one broad subject area, these areas would cover so wide a field that some students would not fit into any of the groupings. Alderson (1981) cited the example of the student in urban studies who would not know whether to choose a test in science or in social studies.

In relation to the reading comprehension section of ESP tests, Alderson (1988b) also asked what was meant by the term 'general text'. General to whom? Were 'general' texts so neutral that their subject matter was unfamiliar to all or were they intended to be neutral, but actually based on arts-based topics which might turn out to be more appropriate for arts than for science students?

Until there were answers to the above questions, Alderson wondered how much point there was in having specific EAP tests, since they were time-consuming and expensive to produce, and since it was so difficult to make equivalent tests in different subject areas genuinely parallel. The only way we could know, Alderson (1981: 133) said, was to carry out empirical studies. Although there has been some response to Alderson's plea for more research since 1981, there is still room, and need, for more research. The present study is another response to Alderson's (1981) plea.

9. Statement of the Problem

Over the past two decades, there have been several studies into the effect of background knowledge on EAP test performance. Three articles by Alderson and Urquhart (1983, 1985a, and 1985b) aroused considerable interest and led to several follow-up studies. These articles described three studies carried out with students attending English classes in Britain in preparation for going to British universities. In

each, Alderson and Urquhart compared students' scores on reading texts related to their own field of study with those on texts in other subject areas. The students' scores on the modules were somewhat contradictory. On the one hand, for example, science and engineering students taking the technology module of ELTS did better than the business and economics students who took the same test, and as well as the liberal arts students, although their language proficiency was lower. On the other hand, the business and economic students did no better than the science and engineering group on the social studies module. Alderson and Urquhart concluded that background knowledge had some effect on test scores, but that this was not consistent, and that future studies should take account of linguistic proficiency and other factors as well.

Along the same lines, Shoham, Peretz, and Vorhaus (1987) concluded that, while students in the biological and physical sciences did better at the scientific texts, the humanities and social science students did not do better on the test in their own subject area. In a similar study, Peretz and Shoham (1990) had similar results. Their explanation for this was that the texts were only indirectly related to the students' specialized fields of study, and suggested that this might support Lipson's suggestion (1984) that "a totally unfamiliar text is often easier to comprehend than a text with a partially familiar content." This contention of Lipson was indeed radical. If supported by further research, it would be an almost unassailable reasons for dropping ESP testing. If Lipson's idea were taken to its logical conclusion, of course, proficiency tests would have to contain materials outside any candidates experience. The Joint Matriculation Board (JMB) University Test in English for Speakers of Other Languages followed just such an approach, with passages in esoteric subjects. As a result, item writers had difficulty finding suitable texts and the ensuing materials were often excessively dull.

In the same vein, the main aims of the present study can be categorized into two classes. On the one hand, an attempt will be made to determine if university students' test and task performances are related to language proficiency, text familiarity, or task type in any meaningful way. On the other hand, subjects' test and task performances are compared to determine which of these factors accounts for a larger percentage of their score variances. As such, the present study has a number of purposes. First, it will try to manifest whether there is any meaningful difference between Iranian LSP university students on the same reading tasks due to language proficiency, text familiarity, or task type. Second, it will investigate whether these students manifest meaningful differences

in their performance on different reading tasks due to the same variables. Moreover, it will also try to manifest which type of text familiarity (i.e., partial familiarity, total familiarity, or total unfamiliarity) will result in better performance on reading tests and reading tasks across different proficiency levels. Finally it will determine the size of the impact of each of the independent variables on subjects' test and task performances.

10. Research Questions and Hypotheses

The present study is an attempt at answering a few questions that pertain to university students' performance on reading comprehension tests in LSP (Language for Specific Purposes) contexts. The objectives of the investigation can be expressed in the following research questions:

1. Is LSP students' overall and differential test and task performance a function of text familiarity, task type, language proficiency, or the interaction of these variables?
2. Which variable accounts for a greater share of LSP students' test and task score variance: task type, text familiarity, or language proficiency?
3. Does LSP students' level of language proficiency significantly affect their overall test and task performance as well as their test and task performance across different levels of the text familiarity cline?
4. Does LSP students' degree of familiarity with the propositional content of LSP tests significantly affect their overall test and task performance and their test and task performance across different levels of language proficiency?

All these questions can be expressed in terms of the following research hypotheses. For each hypothesis, its negative counterpart will represent the null hypothesis.

H1. LSP students' overall and differential test and task performance is a function of text familiarity, task type, language proficiency, and the interaction of these variables.

H2. Text familiarity accounts for a greater share of LSP students' test and task score variance than task type or language proficiency.

H3. LSP students' level of language proficiency significantly affects their overall test and task performance as well as their test and task performance across different levels of the text familiarity cline.

H4. LSP students' degrees of familiarity with the propositional content of LSP tests significantly affect their overall test and task performance and their test and task performance across different levels of language proficiency.

11. Definition of Key Terms and Concepts

In this section, definitions of the terms and concepts that have a key role in the present study are presented. The aim of this section is to clarify the scope of the present study and the limits within which the findings of the investigation should be interpreted.

TEXT FAMILIARITY: Studies into the effect of background knowledge on ESL reading comprehension can be divided into two kinds: those concerned with world (that is, content and cultural knowledge), and those relating to knowledge of the formal or linguistic structure of texts. This study will be mainly concerned with content knowledge, and in particular with subject or domain specific content knowledge (that is, the knowledge acquired from schooling, interests, and hobbies).

TEXT FAMILIARITY CLINE: The degree to which LSP students are familiar with the propositional content of the passages that appear in the different modules of the Task-Based Reading Test (TBRT) used in the present study will indicate their stance on the text familiarity continuum. This continuum is referred to as the text familiarity cline.

LIMITED ENGLISH PROFICIENT (LEP): According to Lockwood (2001), Limited English proficiency (LEP) students are defined as those who perform in the bottom one-half to one-quarter on tests that measure knowledge of the English language. Through US Government's Bilingual Education Act of 1968, LEP students are those who are often provided with instruction in the public schools that supplements or replaces regular classroom instruction, including English as a Second Language (ESL) classes which emphasize English language learning, and long- and short-term bilingual programs that facilitate the move to the regular classroom. In the United States of America, according to the Federal Register definition adopted by the State, a limited English proficient student is any student whose primary language is other than English and who is insufficiently proficient in the English language to receive instruction exclusively from regular educational programs and to function on an academic par with his/her peers. According to the Bilingual Education Act of 1968, amended in 1988, a limited English proficient student is one who:

(I) meets one or more of the following conditions:

(a) the student was born outside of the United States or whose native language is not English;

(b) the student comes from an environment where a language other than English is dominant; or

(c) the student is American Indian or Alaskan Native and comes from an environment where a language other than English has had a significant impact on his/her level of English language proficiency; and

(II) has sufficient difficulty speaking, reading, writing, or understanding the English language to deny him or her the opportunity to learn successfully in English-only classrooms.

For purposes of the present study, based on their distribution around the mean on the IELTS test, subjects were categorized into four classes: proficient, fairly proficient, semi-proficient, and non-proficient. The last two classes, semi-proficient and non-proficient, contained Limited English Proficient (LEP) subjects, and any finding that pertains to these sub-groups can also be interpreted as pertaining to LEP students.

TASK: Candlin (1987: 10) defines task as "One of a set of differentiated, sequencable, problem-posing activities involving learners and teachers in some joint selection from a range of varied cognitive and communicative procedures applied to existing and new knowledge in the collective exploration and pursuance of foreseen or emergent goals within a social milieu." Kunnan (1998: 20-21) defines task as "… the open-ended stimulus serving to elicit the examinee's performance to be evaluated. In language testing, examples of performance include a written response to an essay prompt, an oral response to an interviewer's questions or instructions to a role play, or a physical response to instructions given in the target language." Ur (1996: 145) considers "… setting questions to answer, whether before or after a text" and "… giving the learners a set of titles together with a set of extracts from different newspaper articles or stories and asking them to match the titles to the appropriate extracts" as examples of reading tasks.

12. Final Remarks

Administration of a language proficiency test in English is the most common method used to determine whether a student is limited English proficient (LEP) (Hopstock, Bucaro, Fleischman, Zehler, and Eu, 1993). The English proficiency tests most frequently used to identify LEP students are the Language Asssessment Scales (LAS), the Idea Proficiency Test (IPT), the Maculaitis Assessment Program (MAC), the Bilingual Syntax Measure (BSM), the Peabody Picture Vocabulary Test (PPVT), and

the Language Assessment Battery (LAB) (Hopstock, *et al.*, 1993). Locally developed tests are also frequently used. For purposes of the present study, LEP students were identified on the basis of their performances on the IELTS. To this end, the general training reading module of the test was used. The justification for this choice lies in the "reliability and validity" claims for the IELTS made by the University of Cambridge Local Examinations Syndicate (UCLES). For further information, please refer to chapter two, section 6.2.1.

The reading tasks employed in this study are based on the definition given by Antony John Kunnan (1998). They include the following:

1) Answering true/false/not-given items on the basis of the information subjects read in the corresponding reading passages ("true-false task" hereafter);
2) Completing open-ended statements with two possible endings on the basis of the information present in the reading passage (" sentence-completion task" hereafter);
3) Selecting appropriate summaries for paragraphs ("outlining task" hereafter);
4) Answering yes/no/not-given items on the basis of the claims made by the writer of the corresponding passage ("writer's-view task" hereafter); and
5) Completing sentences and tables with the information subjects gain through skimming the corresponding passages (" skimming task" hereafter).

In all these tasks, the subjects will use the information that is present in the reading passages even though that information might run against their schemata or prior knowledge.

It is also noteworthy that definitions for the key terms and concepts, other than those presented above, may be found in the literature. However, the investigator purposefully chose to use the above-mentioned definitions since they are operational. As such, they will make the quantification of the data more accurate, and the results and their interpretations more precise.

Finally, the validity and reliability of the instrument (i.e., Task-Based Reading Test or TBRT) developed for the collection of data for the present study are based on the pilot and trial administrations of the instrument. It is, therefore, important to refrain from overinterpreting the findings of the study.

CHAPTER TWO

REVIEW OF THE LITERATURE

1. Introduction

Perhaps ESP is one of the major activities around the world today. It is an enterprise involving education, training and practice. ESP draws upon three major realms of knowledge: language, pedagogy, and the students'/participants' specialist areas of interest. ESP teachers generally have a great variety of often-simultaneous roles, such as researchers, course designers, materials developers, testers, evaluators, as well as classroom teachers. It is, therefore, more reasonable to use the term ESP practitioner. ESP practitioners need training in ways of describing language, training in teaching language, and-training in designing language courses. In addition, and unlike those involved in EGP (English for general purposes), they need some knowledge of, or at least access to information on, whatever it is with which students are professionally involved, for example economics, physics, nursing, catering. Authentic materials (for example texts, recorded discussions, interviews, and lectures) may be needed from these work or study situations to be developed as classroom materials. Thus, ESP may be seen as dependent for its successful implementation on help and materials from specialists in many other areas of professional activity.

ESP may be seen as pluralistic, because many approaches to it are simultaneously being followed around the world today. The full form of 'ESP' is generally given as 'English for specific purposes', and this would imply that what is specific and appropriate in one part of the globe may well not be elsewhere. Thus, it is impossible to

produce a universally applicable definition of ESP. Strevens (1980: 109) suggests that 'a definition of ESP that is both simple and watertight is not easy to produce, and Hutchinson and Waters (1987: 18-19) prefer to say what ESP is 'not'. In the same vein, Dudley-Evans (1998) proposes a loose definition for ESP to include almost all instances of language learning.

ESP is protean, as it is responsive to developments in all three realms of language, pedagogy, and content studies. The changing interpretations of ESP over the years and in different parts of the world represent changing relationships between, and changing fashions in, these three realms of knowledge. Continental European studies in ESP, for example, have seemed relatively unconcerned with pedagogy but very active in aspects of language description. "The pedagogy of ESP has always been important in Britain and North America, however, with Britain taking the lead in matters of syllabus and course design, practitioners in the USA and in Canada leading the way in matters of classroom-based practice and research" (Swales, 1989: 79-83). Currently, it could be suggested, there is a greater interest in the content with which ESP must be involved, the subject matter that ESP students have to study and work with through English. According to Tickoo (1988), content-based approaches to language teaching and testing seem to be more discussed now, and not just in North America, where they have been most developed.

There are many types of ESP and many acronyms. A major distinction is often drawn between EOP (English for occupational purposes), involving work-related needs and training, and EAP (English for academic purposes), involving academic study needs. Cutting across these is EST (English for science and technology), mainly used for ESP work in the USA (especially the pioneering work of Selinker, *et al.* (1970), which can refer to both work-and study-related needs. A further important distinction must be made between those students who are newcomers to their field of work or study and those who are already expert (or on the way to becoming so), perhaps via the medium of their own language. This distinction, as Strevens (1988: 139-40) notes, "is between English which is instructional and English which is operational." Students who are newcomers to their field need some instruction in the concepts and practices of that field. Experienced students 'require operational ESP materials, where the knowledge, the concepts, the instruction and the training are taken for granted, and where it is the ability to function in English which is being imparted'. Each situation has implications

for the kind of content knowledge that the ESP teacher may need to deploy and for the degree of generality or specificity of the ESP course.

2. The Origin of ESP

As with most developments in human activity, ESP was not a planned and coherent movement, but rather a phenomenon that grew out of a number of converging trends. These trends have operated in a variety of ways around the world, but we can identify three main reasons common to the emergence of all ESP.

2.1. Worldwide Demands

The end of the Second World War in 1945 heralded an age of enormous and unprecedented expansion in scientific, technical and economic activity on an international scale. This expansion created a world unified and dominated by two forces, technology and commerce, which in their relentless progress soon generated a demand for an international language. For various reasons, most notably the economic power of the United States in the post-war world, this role fell to English. The effect was to create a whole new mass of people wanting to learn English, not for the pleasure or prestige of knowing the language, but because English was the key to the international currencies of technology and commerce. Previously the reasons for learning English (or any other language) had not been well defined. A knowledge of a foreign language had been generally regarded as a sign of a well rounded education, but few had really questioned why it was necessary. Learning a language was, so to speak, its own justification. But as English became the accepted international language of technology and commerce, it created a new generation of learners who knew specifically why they were learning a language. Businessmen and women wanted to sell their products; mechanics had to read instruction manuals; doctors needed to keep up with developments in their field; and a whole range of students whose course of study included textbooks and journals only available in English also needed ESP. All these and many others needed English and, most importantly, they knew why they needed it.

This development was accelerated by the Oil Crises of the early 1970s, which resulted in a massive flow of funds and Western expertise into the oil rich countries. English suddenly became big business, and commercial pressures began to exert an influence. Time and money constraints created a need for cost effective courses with clearly defined goals.

The general effect of all this development was to exert pressure on the language teaching profession to deliver the required goods. Whereas English had previously decided its own destiny, it now became subject to the wishes, needs and demands of people other than language teachers. English had become accountable to the scrutiny of the wider world. The traditional leisurely and purpose-free stroll through the landscape of the English language seemed no longer appropriate in the harsher realities of the market place.

2.2. A Revolution in Linguistics

At the same time as the demand was growing for English courses tailored to specific needs, influential new ideas began to emerge in the study of language. Traditionally the aim of linguistics had been to describe the rules of English usage, that is, the grammar. However, the new studies shifted attention away from defining the formal features of language usage to discovering the ways in which language is actually used in real communication (Widdowson, 1978). One finding of this research was that the language we speak and write varies considerably, and in a number of different ways, from one context to another. In English language teaching this gave rise to the view that there are important differences between, say, the English of commerce and that of engineering. These ideas married up naturally with the development of English courses for specific groups of learners. The idea was simple: if language varies from one situation of use to another, it should be possible to determine the features of specific situations and then make these features the basis of the learners' course.

Swales (1985) presents an article by C. L. Barber on the nature of Scientific English which was published as early as 1962. However, it was the late 1960s and early 1970s which saw the greatest expansion of research into the nature of particular varieties of English including descriptions of written scientific and technical English by Ewer and Latorre (1969), Swales (1971), Selinker and Trimble (1976) and others. Most of the work at this time was in the area of English for Science and Technology (EST) and for a time ESP and EST were regarded as almost synonymous. But there were studies in other fields too, such as the analysis of doctor-patient communication by Candlin, Bruton and Leather (1976).

In short, the view gained ground that the English needed by a particular group of learners could be identified by analyzing the linguistic characteristics of their specialist

area of work or study. Hutchinson and Waters (1987) point to the fact that 'Tell me what you need English for and I will tell you the English that you need' became the guiding principle of ESP.

2.3. Focus on the Learner

New developments in educational psychology also contributed to the rise of ESP, by emphasizing the central importance of the learners and their, attitudes to learning (Rodgers, 1969). Learners were seen to have different needs and interests, which would have an important influence on their motivation to learn and, therefore, on the effectiveness of their learning. This lent support to the development of courses in which 'relevance' to the learners' needs and interests was paramount. The standard way of achieving this was to take texts from the learners' specialist area (e.g., texts about biology for biology students etc.). The assumption underlying this approach was that the clear relevance of the English course to their needs would improve the learners' motivation and thereby make learning better and faster. The growth of ESP, then, was brought about by a combination of three important factors: (a) the expansion of demand for English to suit particular needs, (b) developments in the fields of linguistics, and (c) developments of educational psychology. All three factors seemed to point towards the need for increased specialization in language learning.

3. The Development of ESP

From its early beginnings in the 1960s ESP has undergone three main phases of development. It is now in a fourth phase with a fifth phase starting to emerge. It will provide a useful perspective to give a brief summary of these five phases here. It should be pointed out first of all that ESP is not a monolithic universal phenomenon. ESP has developed at different speeds in different countries, and examples of different approaches to ESP can be found operating somewhere in the world at the present time. The summary here must, therefore, be very general in its focus. It will be noticeable in the following overview that one area of activity has been particularly important in the development of ESP. This is the area usually known as EST (English for Science and Technology). Swales (1985), in fact, uses the development of EST to illustrate the development of ESP in general. With one or two exceptions, English for Science and Technology (EST) has always set and continues to set the trend in theoretical

discussion, in ways of analyzing language, and in the variety of actual teaching materials.

3.1. Register Analysis

This stage took place mainly in the 1960s and early 1970s and was associated in particular with the work of Strevens (Halliday, Mclntosh and Strevens, 1964), Ewer (Ewer and Latorre, 1969) and Swales (1971). Operating on the basic principle that the English needed in one scientific field constituted a specific register different from those of other fields of science, or General English, register analysis sought to identify the grammatical and lexical features of different scientific registers. Teaching materials then took these linguistic features as their syllabus. A good example of such a syllabus is that of *A Course in Basic Scientific English* by Ewer and Latorre (1969). In fact, as Ewer and Latorre's syllabus shows, register analysis revealed that there was very little that was distinctive in the sentence grammar of Scientific English beyond a tendency to favor particular forms such as the present simple tense, the passive voice, and nominal compounds. It did not, for example, reveal any forms that were not found in General English. But we must be wary of making unfair criticism. Although there was an academic interest in the nature of registers of English *per se,* the main motive behind register analyses such as Ewer and Latorre's was the pedagogic one of making the ESP course more relevant to learners' needs. The aim was to produce a syllabus which gave high priority to the language forms students would meet in their science studies and, in turn, would give low priority to forms they would not meet. Ewer and Hughes-Davies (1971), for example, compared the language of the texts their science students had to read with the language of some widely used school textbooks. They found that the school textbooks neglected some of the language forms commonly, found in Science texts, for example, compound nouns, passives, conditionals, anomalous finites (i.e. modal verbs). Their conclusion was that the ESP course should, therefore, give precedence to these forms. Today, register analysis is one of the most significant areas of research for LSP scholars.

3.2. Rhetorical Discourse Analysis

There were serious flaws in the register-analysis-based syllabus, but, as it happened, register analysis as a research procedure was rapidly overtaken by developments in the world of linguistics. In the first stage of its development, ESP had focused on language at the sentence level. As ESP became closely involved with the emerging field of

discourse or rhetorical analysis, the second phase of ESP development shifted attention to the level above the sentence. The leading lights in this movement were Widdowson in Britain and the so called Washington School of Selinker, Trimble, Lackstrom and Trimble in the United States. The basic hypothesis of this stage is succinctly expressed by Allen and Widdowson (1974) who took the view that the difficulties which the students encounter arise not so much from a defective knowledge of the system of English, but from an unfamiliarity with English use, and that consequently their needs cannot be met by a course which simply provides further practice in the composition of sentences, but only by one which develops a knowledge of how sentences are used in the performance of different communicative acts.

Register analysis had focused on sentence grammar, but now attention shifted to understanding how sentences were combined in discourse to produce meaning. The concern of research, therefore, was to identify the organizational patterns in texts and to specify the linguistic means by which these patterns are signaled. These patterns would then form the syllabus of the ESP course.

As in stage 1, there was a more or less tacit assumption in this approach that the rhetorical patterns of text organization differed significantly between specialist areas of use: the rhetorical structure of science texts was regarded as different from that of commercial texts, for example. However, this point was never very clearly examined (Swales, 1985) and indeed paradoxically, the results of the research into the discourse of subject-specific academic texts were also used to make observations about discourse in general (Widdowson, 1978). The typical teaching materials based on the discourse approach taught students to recognize textual patterns and discourse markers mainly by means of text diagramming exercises. The *English in Focus* series (Oxford University Press) is a good example of this approach.

3.3. TLU Situation Analysis

The stage of Target Language Use (TLU) Situation Analysis did not really add anything new to the range of knowledge about ESP. What it aimed to do was to take the existing knowledge and set it on a more scientific basis, by establishing procedures for relating language analysis more closely to learners' reasons for learning. Given that the purpose of an ESP course is to enable learners to function adequately in a target situation, that is, the situation in which the learners will use the language they are learning, then the ESP course design process should proceed by first identifying the

target situation and then carrying out a rigorous analysis of the linguistic features of that situation. The identified features will form the syllabus of the ESP course. This process is usually known as “needs analysis.” However, Chambers (1980) prefers to use the term ‘target situation analysis’, since it is a more accurate description of the process concerned.

The most thorough explanation of target situation analysis is the system set out by Munby (1978) in *Communicative Syllabus Design*. The Munby model produces a detailed profile of the learners’ needs in terms of communication purposes, communicative setting, language skills, functions, structures, the means of communication, etc. The target situation analysis stage marked a certain ‘coming of age’ for ESP. What had previously been done very much in a piecemeal way, was now systematized, and learners’ needs were apparently placed at the heart of the course design process It proved in the event to be a false dawn since the concept of needs on which it was based was far too simple.

3.4. Skills and Strategies

It should be noted that in the first two stages of the development of ESP all the analysis had been of the surface forms of the language (whether at sentence level, as in register analysis, or above, as in discourse analysis). The target situation analysis approach did not really change this, because in its analysis of learner needs it still looked mainly at the surface linguistic features of the target situation.

The fourth stage of ESP development has seen an attempt to look below the surface and to consider not the language itself but the thinking processes that underlie language use. Before the 1990s, there was no dominant figure in this movement, although one might mention the work of Grellet (1981), Nuttall (1982), and Alderson and Urquhart (1984) as having made significant contributions to work on reading skills. Most of the work in the area of skills and strategies, however, has been done close to the ground in schemes such as the National ESP Project in Brazil, the University of Malaya ESP Project. Both these projects were set up to cope with study situations where the medium of instruction is the mother tongue but students need to read a number of specialist texts which are available only in English. The projects have, therefore, concentrated their efforts on reading strategies. It is interesting to note, however, that not all such projects have such a focus. The ESP project at King Mongkut’s Institute of Technology in Bangkok, Thailand, for example, aimed at coping with a very similar study situation,

but the focus here was on the full range of skills (reading, writing, listening, and speaking).

The principal idea behind the skills-centered approach is that underlying all language use there are common reasoning and interpreting processes, which, regardless of the surface forms, enable us to extract meaning from discourse. There is, therefore, no need to focus closely on the surface forms of the language. The focus should rather be on the underlying interpretive strategies, which enable the learner to cope with the surface forms, for example, guessing the meaning of words from context, using visual layout to determine the type of text, exploiting cognates (i.e., words which are similar in the mother tongue and the target language), etc. A focus on specific subject registers is unnecessary in this approach, because the underlying processes are not specific to any subject register. It was argued that reading skills are not language-specific but universal and that there is a core of language (for example, certain structures of argument and forms of presentation) which can be identified as “academic” and which is not subject-specific (Chitravelu, 1980).

As has been noted, in terms of materials this approach generally puts the emphasis on reading or listening strategies. The characteristic exercises get the learners to reflect on and analyze how meaning is produced in and retrieved from written or spoken discourse. Taking their cue from cognitive learning theories, the language learners are treated as thinking beings who can be asked to observe and verbalize the interpretive processes they employ in language use.

3.5. A Learning-Centered Approach

In outlining the origins of ESP, it was mentioned earlier that there were three forces behind ESP, which we might characterize as (a) need, (b) new ideas about language, and (c) new ideas about learning. It should have become clear that in its subsequent development, however, scant attention has been paid to the last of these forces—learning. All of the stages outlined so far have been fundamentally flawed, in that they are all based on descriptions of language use. Whether this description is of surface forms, as in the case of register analysis, or of underlying processes, as in the skills and strategies approach, the concern in each case is with describing what people do with language. although this will help to define the course objectives, the concern of ESP has begun to be with language learning. One cannot simply assume that describing and exemplifying what people do with language will enable someone to learn it. If that were

so, one would need to do no more than read a grammar book and a dictionary in order to learn a language. Many scholars are recently noticing that a truly valid approach to ESP must be based on an understanding of the processes of language learning. This brings us to the fifth stage of ESP development or "the learning-centered approach." ESP proponents are striving to delineate the importance and the implications of the distinction that they have made between language use and language learning.

4. LSP Ability

Douglas (2000) reviews research on the nature of communicative language ability, and summarizes our current best understanding of this complex concept, including a discussion of background knowledge in specific fields of interest. His goal is to work towards a clearer understanding of the construct of "Language for Specific Purposes (LSP) ability," for that, of course, is what we are attempting to measure with LSP tests. He especially considers the relationship between language ability and SP background knowledge. This is an important issue, for if language is learned in communicative contexts, then it follows that those contexts must affect the very nature of the language that is thus acquired. Chapelle (1998: 43) points out, in an elaboration of what she calls an "interactionist view" of construct definition, that merely taking into account both the traits of the language user and the features of the context is not enough; rather, we must allow for the interaction between the two. This inevitably means that the quality of each changes: Trait components can no longer be defined in context-independent, absolute terms and contextual features cannot be defined without reference to their impact on underlying characteristics. This notion suggests that there is such a thing as LSP knowledge, and that the nature of language knowledge may be different from one domain to another. Douglas takes the position, again following Chapelle (1998: 15), that what is required is a theory of "how the context of a particular situation within a broader context of culture, constrains the linguistic choices a language user can make during a linguistic performance." Bearing in mind that external context is a major factor in the engagement of SP communicative language ability, he considers how language ability and SP background knowledge interact with each other,

4.1. Communicative Language Ability

Spolsky (1973) asked the fundamental question "What does it mean to know a language?" Researchers have been working to answer it ever since. We now know more about the nature of communicative language ability than we did then, but we are still far

from understanding with any precision what language knowledge consists of and how the various bits and pieces work together to produce communicative utterances and written text.

As has become clear in recent years through empirical studies conducted by language testers and others, language knowledge is multicomponential. However, what is extremely unclear is precisely what those components may be and how they interact in actual language use. As Alderson (1991: 12) has pointed out, the answer to the question of what it means to know a language, "depends upon why one is asking the question, how one seeks to answer it, and what level of proficiency one might be concerned with." In the case of LSP testing, Douglas (2000: 26) adds the expression, "and in what specific situational context one is interested," to the quotation from Alderson.

In an attempt to make sense of the various models of communicative competence and communicative language ability, Henning and Cascallar (1992) turn to the field of cartography for a metaphor:

> Various kinds of two-dimensional maps have been devised as aids to navigation. Some maps are useful geographical models for ocean navigation, others for automobile navigation, and still others for wilderness trekking … none of these two-dimensional maps provides a completely accurate representation of three-dimensional reality, nor does any one kind of two-dimensional map serve every navigational purpose equally well. (Henning and Cascallar, 1992: 4)

So it is with models of language ability. The framework Douglas develops is not offered in opposition to any others. He tries to design a map to help achieve a particular purpose: navigating in the realm of LSP use, and understanding the abilities that underlie it.

4.2. Communicative Competence

The term "communicative competence" has been invoked for nearly three decades now to encompass the notion that language competence involves more than Chomsky's (1965) rather narrowly-defined linguistic competence. As Hymes (1971, 1972) originally formulated the concept 'communicative competence' involves judgements about what is systemically possible (in other words, what the grammar will allow). This is, however, not the whole story. It also includes what is psycholinguistically feasible (what the mind will allow), and socioculturally appropriate (what society will allow). Moreover, communicative competence affords information about the probability of

occurrence of a linguistic event and what is entailed in the actual accomplishment of it. It is important to remember that, for Hymes, competence is more than knowledge: "Competence is dependent upon both [tacit] *knowledge* and [ability for] *use*" (Hymes 1972: 282). Hymes's formulation that communicative competence consists of language knowledge and ability for use has become something of a classic in the field of applied linguistics. However, it is important to note early in the discussion that communicative competence is not to be confused with communicative success: ability for use is not the same as use. Speakers may have sufficient knowledge to address a communicative task and yet, for reasons of their own, or perhaps owing to factors outside their control, choose not to address the task or not accomplish their communicative goal (Hornberger, 1989). This is an important point for LSP testing. However, it must remain clear that what is being tested is not the success of the performance, but rather the underlying traits that produce the performance, i.e., communicative competence, or what Douglas (2000: 27) calls "LSP ability."

The problem with many LSP tests, according to Douglas, is that we do not distinguish between a language performance and the abilities that underlie it. This will make generalizing from performance in one context or situation to any other situation problematic. In other words, we observe performances that we elicit under controlled test conditions, and make inferences about the abilities that produce these performances. It should, however, be noted that failure to succeed in a test task does not automatically indicate a lack of communicative competence; it may indicate an impossible situation set up by the test developer. By the same token, it may also be true that success in accomplishing a test task does not automatically guarantee that the test taker possesses communicative competence, since it is possible to accomplish certain tasks by bringing other types of knowledge, such as background knowledge, to bear on the problem. The key point to remember is the importance of distinguishing performance on tasks from the abilities that make the performance possible.

Others have since reformulated Hymes's notion of communicative competence, and the current, most well-known, framework is that of Bachman (1990), elaborated by Bachman and Palmer (1996). They postulate two components of communicative language ability: "language knowledge" and "strategic competence." Strategic competence serves as a mediator between the internal traits of background knowledge and language knowledge and the external context, controlling the interaction between them. The engagement of strategic competence, then, is of central concern in LSP

testing, for this cognitive aspect is responsible for assessing the characteristics of the language use situation.

The framework presented by Chapelle and Douglas (1993) for communicative language ability combines Bachman and Palmer's formulation of the components of language knowledge with a modified formulation of strategic competence. In this framework, language knowledge consists of grammatical knowledge, textual knowledge, functional knowledge, and sociolinguistic knowledge. Strategic competence comprises the processes of assessment, goal setting, planning, and control of execution. From the perspective of the LSP tester, such a framework can be employed in two ways: in defining the construct to be measured, and in interpreting test performances as evidence of SP language (LSP) ability.

4.3. The Construct of LSP Ability

Douglas (2000) takes the position that what is required in LSP testing is an understanding of how SP background knowledge interacts with language knowledge to produce a communicative performance in SP contexts. As such, Douglas's framework for LSP ability: (a) includes Specific Purpose (SP) background knowledge as a component of communicative language ability, and (b) gives a central role to the cognitive construct of discourse domain. This is where the language user interprets what is referred to as contextualization cues (also referred to in chapter 5), the features of the external communicative context that language users attend to in determining where they are and what type of communicative activity they are engaged in.

Douglas further argues that at some point in the test design process, testers will need to finally decide precisely what components of LSP ability they will attempt to measure. This is the task of construct definition. There are four aspects they will need to consider: (a) the level of detail necessary in the definition, (b) whether to include strategic competence or not, (c) the treatment of the four skills, and (d) whether to distinguish between language knowledge and SP background knowledge. It is important to note that in LSP testing, we need to make a distinction between the construct of LSP ability as it is analyzed in the TLU situation and as it is realized in an LSP test. Douglas maintains that language ability is far richer and more complex than can be effectively measured in any test.

Actual language use in SP contexts involves a complex interaction among the components of LSP ability (i.e., all the features of language knowledge, strategic

competence, and background knowledge) but it is not possible to actually score or rate all these components in a test. Therefore, while a communicative performance, whether in a target situation or in a test, may require a wide range of linguistic, strategic, and content knowledge, in assessing the performance in the test situation, we normally focus on only a small set of features. Moreover, these features are context-specific and vary from one context to another. Perhaps the most thorough treatment of the construct of LSP ability has been proposed by Douglas (2000). The following few sections describe some of the characteristics of his model.

4.3.1. Level of Detail

There are situations where a broader, less detailed definition will be sufficient. If the purpose of the test is to determine whether candidates' English language ability is sufficient for them to be admitted to a medical residency program, then a broad definition of language ability without distinguishing its components will be enough for test users to interpret the results. On the other hand, if the situation calls for a profile of component abilities as input to a remedial course of instruction, then a more detailed specification of the construct will be necessary. For example, the definition of the construct to be measured in the TEACH test (discussed below) includes grammatical knowledge, familiarity with the cultural code, rhetorical development, listening ability, and question handling and responding, in addition to other, non-language performance characteristics.

4.3.2. Strategic Competence

Strategic competence serves as a mediator and interpreter between the external situational context and the internal language and background knowledge required to respond to the communicative situation, and its engagement in test tasks is central to the LSP enterprise. As a link between context and language knowledge, strategic competence is assumed to operate in all communicative situations. However, it may not be necessary for certain testing purposes to measure strategic competence. If 'whether candidates for certification have adequate English skills to perform the job' is all test users want to know, then the definition of the construct to be measured may well include only the language ability components. It can, therefore, be assumed that strategic competence is implicitly a part of the performance. However, if the test users want to know how well candidates for certification can adjust to changing situational conditions, then strategic competence would be a relevant focus of the measurement. In

his framework of LSP ability, Douglas (2000) noted the distinction between what components of LSP ability may be described in a construct definition, and what aspects of it are actually scored separately. For example, it may be that strategic competence, while a part of the theoretical construct of language ability in a SP test, may not be given a separate score since the test users are not interested in receiving one.

4.3.3. The Four Skills

It has been traditional in language testing to categorize tests and subtests according to the four skills of reading, writing, listening, and speaking. We also find reference to tests of reading ability or listening ability, which would seem to merge the concepts of skill and ability. Douglas (2000), in an explanation of his LSP ability framework, draws on a new perspective and treatment of the four linguistic skills. What he refers to as LSP ability must be manifested in the performance of tasks involving independent skills, or some combination of them. As intriguing as it may seem, Douglas claims that the four skills will not be considered to be a part of LSP ability, but rather the means by which that ability is realized in the performance of tasks in TLU situations, including LSP tests. Therefore, he tries to avoid using terms such as speaking ability or reading ability. That is, he focuses on the interaction between LSP ability and the characteristics of the tasks in which that ability is engaged. He also develops a framework for describing tasks in TLU situations and LSP tests. The task characteristics will include the format of the input, which may be visual or auditory, and of the response to it, which may be spoken, written, or physical. Thus, while the four skills are obviously an important consideration in language use, the primary focus in LSP testing is the interaction between LSP ability and the characteristics of language use tasks in SP situations.

4.3.4. LSP Background Knowledge

A final aspect in the definition of LSP ability is that the construct contains, by definition, SP background knowledge. The very essence of LSP tests is that they require the test takers to engage themselves authentically in test tasks that are demonstrably related to the TLU situation. Therefore, relevant background knowledge will necessarily be called upon in the interpretation of the communicative situation and in the formulation of a response. A question that needs to be considered in this regard is whether it will be necessary to distinguish between the two types of knowledge: language knowledge and SP background knowledge. In some testing situations it may be, while in others it may not. For example, when test takers are known to possess a

high level of expertise in the specific field, then it will not be necessary to disambiguate language and background knowledge. On the other hand, when expertise in the field cannot be taken as given, it may be desirable to devise a test of SP knowledge as a way of determining the source of a poor performance on the language test (Bachman and Palmer, 1996).

The distinction between language knowledge and background knowledge has long been a problem for language testers, since there is the difficulty of distinguishing between them in interpreting test results. There are a few studies which suggest that, under some conditions, background knowledge does not influence language test performance to any significant degree. On the other hand, several other studies have found significant interactions between background knowledge and language test performance. Thus, it appears that, under some conditions at least, background knowledge makes a difference to language test performance.

LSP ability is most likely to be engaged when test content and tasks are sufficiently specified, and when subjects' levels of language knowledge are sufficiently high to enable them to make use of the situational information. A problem for LSP testers is to understand the conditions that influence test performance. Until such features are understood and controlled, true LSP test development, authenticity in test performance, and valid interpretation of language test results will be elusive goals. In other words, if there is to be a congruence between the elicitation of language performances and the interpretation of those performances, there needs to be a congruence between the types of knowledge and tasks the test requires and the types of knowledge and tasks demanded by TLU situation. Douglas claims that LSP test developers need to be aware of this aspect of the relationship between background knowledge and language knowledge.

5. LSP Testing

According to Douglas (2000: 1), "testing language for specific purposes (LSP) refers to that branch of language testing in which the test content and test methods are derived from an analysis of a specific language use situation, such as Spanish for Business, Japanese for Tour Guides, Italian for Language Teachers, or English for Air Traffic Control." In this sense, LSP tests are usually contrasted with general purpose language tests like TOEFL (Educational Testing Service, 1965), in which "purpose" is more broadly defined. It is, therefore, noteworthy that tests are not readily classifiable as

either “general purpose” or “specific purpose.” All tests are developed for some purpose. There is, however, a continuum of specificity from very general to very specific, and a given test may fall at any point on the continuum. As such, LSP testing may be considered a special case of communicative language testing, since both are based on a theoretical construct of contextualized communicative language ability. Moreover, LSP tests are no different in terms of the qualities of good testing practice from other types of language tests.

Over the years since its beginnings, specific purpose language testing has been criticized on a number of grounds. The criticism of LSP testing can be summarized to include: (a) specific purpose language proficiency is really just general purpose language proficiency with technical vocabulary thrown in; (b) we don’t need specific purpose tests since, if we test general language knowledge, specific uses will take care of themselves; (c) specific purpose language tests are unreliable and invalid since subject knowledge interferes with the measurement of language knowledge; (d) there is no theoretical justification for specific purpose language testing; and (e) specific purpose language testing is impossible anyway, since the logical end of specificity is a test for one person at one point in time. In spite of all these attacks, Douglas (2000) seeks to refute these and other arguments in favor of the view that specific purpose language tests are indeed necessary, reliable, valid, and theoretically well-motivated.

Typically, LSP tests have been construed as those involving language for academic purposes and for occupational or professional purposes. The following publications provide further information on the field of language for specific purposes, of which LSP testing is certainly a part: Swales (1985) affords a discussion of the development of the field. Dudley-Evans and St. John (1998) provide a discussion of current developments in ESP testing. Douglas (2000) focuses on two aspects of LSP testing that may be said to distinguish it from more general purpose language testing: (a) authenticity of task, and (b) the interaction between language knowledge and specific purpose content knowledge.

Authenticity of task means that the LSP test tasks should share critical features of tasks in the target language use situation of interest to the test takers. The intent of linking the test tasks to non-test tasks in this way is to increase the likelihood that the test taker will carry out the test task in the same way as the task would be carried out in the actual target situation. In this sense, LSP testing draws on the principles of performance assessment.

The interaction between language knowledge and content, or background knowledge is perhaps the clearest defining feature of LSP testing. In more general purpose language testing, the factor of background knowledge is usually seen as a confounding variable, contributing to measurement error and to be minimized as much as possible. In LSP testing, on the other hand, background knowledge is, as Douglas (2000) notices, a necessary, integral part of the concept of specific purpose language ability.

5.1. The History of LSP Testing

LSP testing, like LSP teaching, has a relatively short history. A case could be made for the beginning of LSP testing as early as 1913, with the establishment of the University of Cambridge Local Examinations Syndicate's (UCLES) Certificate of Proficiency in English. It was a test designed for prospective English teachers to demonstrate their proficiency in the language (University of Cambridge Local Examinations Syndicate, 1995). Another candidate for the title of first LSP test might be the College Entrance Examination Board's English Competence examination in the US. This was a test for international applicants to US colleges and universities introduced in 1930 (Spolsky, 1995). Both of these tests have clearly defined purposes related to vocational and academic English, respectively, and thus in a sense qualify as examples of LSP tests. However, as Hutchinson and Waters (1987) rightly claim, LSP testing requires first, an analysis of a target language use situation, from which characteristics of test content and method are derived. To this, Douglas (2000) adds the need for an interaction between language knowledge and specific purpose content knowledge. Clearly, not all examples of what are called LSP tests manage to meet these criteria completely. One can, however, argue that a theory of LSP testing establishes these two characteristics as fundamental goals. The UCLES and the College Board tests were not developed on the basis of analyses of language teaching or academic situations, nor did the tasks on the tests bear much relationship to the kinds of tasks required of either teachers or students (except when taking language tests).

So, when might we say that true LSP testing began? A strong candidate is the Temporary Registration Assessment Board (TRAB) examination. This test was introduced in 1975 by the British General Medical Council for the purpose of evaluating the professional and language abilities of physicians trained outside the UK applying for temporary registration to practice medicine in Britain (Rea-Dickins, 1987). The

examination consisted of an assessment of both professional competence and ability to communicate in English. The language component comprised a taped listening test, a written essay, and an oral interview in which both professional knowledge and language ability were assessed. The TRAB language component was based on an analysis of the language, both spoken and written, actually used by physicians, nurses, and patients in British hospitals. As Douglas (2000) discussed, this analytical approach is a critical feature of LSP test development. In addition, the language testing specialists who developed the language component of the TRAB test were not solely responsible for its development, but worked together with medical experts in constructing the tests. This is an important aspect of specific purpose test development. As Rea-Dickins (1987: 196) put it in discussing the TRAB development process, collaboration with practitioners in the specialist area "would seem to be a pre-requisite for the design of a "special purposes" test as the domains incorporated within the specialist area go beyond those in, which the linguist—independently—is competent to make judgements." Thirdly, the TRAB developers attempted to promote the engagement of the test takers' language ability and background knowledge in the test tasks by providing appropriate and rich contextual features in the test material. For example, in the writing tasks, the test takers were presented with authentic information about a patient's case history, and the tasks were linked systematically to the problems presented. Typical writing tasks included the following:

> Write a letter to Dr Jones summarising the case and giving your recommendations for Mr Brown's after-care.
>
> Complete the x-ray request card for this examination.
>
> When the patient is admitted to hospital, what written instructions would you leave the night nurse in charge of the ward regarding management?
>
> Rea-Dickins (1987: 195)

One can see in this early example of an LSP test the embodiment of the critical features of LSP test development: analysis of the target language use situation, authenticity of task, and interaction between language and content knowledge. The TRAB was later revised (its name changed to PLAB—Professional and Linguistic Assessment Board), and is at present no longer in use, but it stands as a worthy prototype of the art of LSP test development.

Some LSP proponents also accept another test as another early LSP test. This test is called the English Language Teaching Development Unit (ELTDU) test. It was introduced in 1976 as an assessment of vocational English (North, 1994).

5.2. The Need for LSP Testing

You might reasonably ask the question, however, as to why LSP testing is necessary, or even desirable. As in all good language testing projects, LSP test development begins with a problem to be solved.

In order to establish the importance of LSP testing, Douglas (2000) draws on a typical language teaching situation. Suppose we want to determine whether people involved in international trade know English well enough to conduct their business. In such a situation, we might reasonably decide to devise a test of English for international business purposes. We would begin our task as test developers by interviewing experienced business people, as well as company supervisors, heads of international divisions, and an assortment of middle level managers who typically deal with international colleagues. We might observe actual negotiating sessions and business meetings, and tape record participants' use of English in the various situations they find themselves in: large meetings, one-to-one discussions in offices, individual and conference telephone calls, the ubiquitous business lunch and other business-related social occasions, and so on. Our goal would be to describe the situations in which international business people conduct their work, and the characteristics of the language they use and of the tasks they must perform in English.

We would need to make some decisions about the scope and content of our test. For example, how important is it to test ability to communicate about food or travel? Should we require the test candidates to demonstrate knowledge of their field of business as well as their abilities in English? Such decisions would have to be made in consultation with the sponsors of the test, for their purposes in wishing to give the test—and their willingness to pay for a longer and more varied test—will help determine what aspects of the milieu of international business we will include in our test. Eventually, however, we would be in a position to produce test specifications, a blueprint of the test we intend to develop, including a statement of the purpose of the test, a description of what it is we intend to measure, a description of the contexts and tasks we intend to include in the test (based on our analysis of the features of the international business domain), details

of how the test will be scored, and an indication of how scores on the test should be interpreted.

On the basis of these specifications, we would then actually produce test tasks and assemble a specific purpose test of English for international business. After trying the new test out, perhaps by giving it to a group of business people, and revising it, we would offer it to our target group of prospective international traders. We would interpret their performance on our test as evidence that they could, or could not, use English well enough to succeed in the tasks required of them in the marketplace.

But why go to all the trouble of devising a new test? Why not just turn to an existing test of English language ability, one such as the Educational Testing Service's Test of English as a Foreign Language (TOEFL), or the Cambridge University Local Examinations Syndicate's Certificate of Proficiency in English (CPE). These, after all, are well-known international tests, with well-known measurement properties. The TOEFL is a multiple-choice test of listening, structure, reading comprehension, and writing, and is often taken by people who wish to demonstrate English language ability for international communication. TOEFL candidates can opt to take a speaking test as well, to further demonstrate their ability to use English. The CPE is a general test of English reading, writing, structure, listening, and speaking, and is used by many businesses to certify the English language skills of their employees, in addition to its main purpose for university admissions.

5.2.1. The Role of Context

Researchers are pretty much in agreement that language performances vary with both context and test task. Therefore, our interpretations of a test taker's language ability must vary from performance to performance. For example, if we give test takers a reading test based on a passage about square-rigged sailing ships, followed by one based on a passage about micro-chips in computers, they will probably perform somewhat differently on the two tests, particularly if they are studying computer engineering. However, it is not enough merely to give test takers topics relevant to the field they are studying or working in. Rather, the material the test is based on must engage test takers in a task in which both language ability and knowledge of the field interact with the test content in a way similar to the target language use situation. In other words, the test task must be authentic for it to represent a specific purpose field in any measurable way. That is, LSP testing requires the use of field specific content in

tasks which might plausibly be carried out in those fields. Returning to the business English example presented above, it would not be enough to provide test takers with listening texts about the work of international commerce, but rather it would be necessary to provide test tasks that share similar characteristics with the tasks that international traders actually perform in their work, both in the processing of information and in responding to it. Thus we must keep in mind that an important reason for using specific purpose measures is that if we wish to interpret a person's test performance as evidence of language ability in a specific language use situation, we must engage the test taker in tasks which are authentically representative of that situation.

As Douglas (2000) eloquently puts it, there is quite a bit of research which suggests that this interaction between the test taker's language ability and specific purpose content knowledge and the test task is a necessary condition in LSP tests. This suggests that there may be no such animal as a pure language test. Measures of language ability are always colored by such factors as background knowledge and test method. It has also been found, however, that the advantage due to specific purpose content knowledge may be quite negligible unless the passage and tasks are sufficiently specific to engage the test takers in authentic language use.

5.2.2. The Precision of LSP

A second reason for preferring LSP tests over more general ones is that technical language (that which is used in any academic, professional or vocational field, including cooking, law, physics, chemistry, air traffic control, scuba diving, religion, stamp collecting, or language teaching) has specific characteristics that people who work in the field must control. What we often refer to as jargon or even gobbledygook has a specific communicative function within that field, namely precision. There are lexical, semantic, syntactic, and even phonological characteristics of language peculiar to any field. Moreover, these characteristics allow for people in that field to speak and write more precisely about aspects of the field that outsiders sometimes find impenetrable. It is this precision that is a major focus of specific purpose language use and is a major factor arguing in favor of specific purpose language tests. A classic example of the need for precise, specific purpose language comes from the field of law. We frequently deplore what we call legalese, the arcane lexis, the convoluted syntax, the use of Latin terminology, and the interminable cross-references to previous laws and cases in legal

texts. Yet, legal language was purposefully developed and is used dynamically by members of the legal profession to communicate among themselves the precise meaning of the law. Therefore, if, for whatever reason, we wanted to measure a lawyer's control of English to conduct the business of law, it would not seem to be sufficient to use texts and tasks which were not specific to the legal profession. There may be perfectly good reasons to include language and tasks not so strictly related to the legal register in the test, but certainly if our goal is to measure a test taker's ability to use language within a specific vocation, profession, or academic field, and that is the focus of this book, then specific purpose texts and tasks will be needed.

5.3. LSP Tests and Other Test Types

Speaking of precision, it is, of course, necessary to be more precise about the nature of specific purpose language tests. For the moment, let us agree to define our object of interest as tests which attempt to measure language ability for specific vocational, professional, and academic purposes. A more precise definition will be presented later. However, before arriving at a useful definition of specific purpose language testing, we need to discuss a number of related concepts in language testing that form the background to LSP testing. According to Douglas (2000), these include (a) communicative testing, (b) general proficiency testing, (c) criterion-referenced testing, and (d) the notion of authenticity. Any of these characteristics is highly significant in its own way and, therefore, it is not possible to give priority to any of them over the rest. Moreover, these characteristics are inclusive.

5.3.1. LSP Tests and Communicative Tests

Particularly since the publication in 1978 of Widdowson's book, *Teaching language as communication,* and in 1980 of Canale and Swain's paper, "Theoretical bases of communicative approaches to second language teaching and testing," the related fields of language pedagogy and language assessment have been characterized by the communicative paradigm, the communicative approach, and communicative language teaching. But even before the publication of Hymes's (1972) classic paper, "On communicative competence," which provided much of the impetus for the communicative approach, language testers were discussing 'productive communication testing' (Upshur, 1971), and teachers and testers have been fascinated with the notion for over a quarter of a century now. Specific purpose language tests are by definition

communicative. Indeed, Sajavaara (1992: 123), in a discussion of LSP test design, assumes from the outset that "It is impossible to distinguish LSP testing theoretically from communicative language testing." Weir (1990: 7), in his book *Communicative language testing*, argues that "In testing communicative language ability we are evaluating samples of performance, in certain specific contexts of use, created under particular test constraints, for what they can tell us about a candidate's communicative capacity or language ability."

In his definition, Weir employs a number of key terms: communicative language ability, specific contexts of use, test constraints, and capacity. Since specific purpose language testing involves all these concepts, Douglas (2000) conceives of LSP testing as a special case of communicative language testing. The first of Weir's terms, communicative language ability (CLA), was introduced by Bachman as a framework for describing language knowledge and the capacity for implementing it "in appropriate, contextualized communicative language use" (Bachman, 1990: 84). This leads us to Weir's second key term, specific contexts of use, requiring us to take account of the many features of context that influence communication, features such as the physical and temporal setting, the role(s) of the test taker and the interlocutor(s)/audience, the purposes of the communication, the topic and content of the message, its tone and manner, and the channels, codes, and genres being employed (Hymes, 1974). The third key term in Weir's definition, test constraints, reminds us that the methods we employ in eliciting a language performance will influence the nature of the performance and thus the interpretations we might make on the basis of it. Tests are, after all, contrived language use events, and even the most cleverly contrived test tasks limit to some degree the generalizability of our interpretations concerning the test takers' specific purpose language abilities.

Finally, Weir refers to capacity, a term employed by Widdowson (1983: 25), as "the ability to use knowledge of language as a resource for the creation of meaning." The term 'capacity' is intended to be understood from the perspective of the language user rather than that of the language analyst (or, indeed, the language tester).

Douglas (2000) uses the terms "communicative language ability" and "specific purpose language ability" to capture the notion of 'capacity' as Weir and Widdowson use the term. The point that is crucial in the testing of language ability in specific purpose contexts is understanding that ability from the perspective of the language user.

That is, not only are LSP testers interested in measuring communicative language ability rather than language performances *per se*, but they also try to interpret test performance from the point of view of language users in the specific purpose situation of interest. Thus, specific purpose language testing, as Widdowson points out with regard to specific purpose language teaching, is essentially an exercise rooted in an understanding of human activity from the point of view of the participants in the activity. Therefore, as Douglas (2000) in the second chapter of his book, *Assessing Language for Specific Purposes*, mentions, the concepts of "grounded ethnography" and "indigenous assessment," are useful approaches for understanding the nature of LSP test performance from the point of view of the language users.

5.3.2. LSP Tests and LGP Tests

Having defined the construct of LSP language ability as a sub-component of CLA, it is now important to decide whether we can dichotomize between general purpose tests and LSP tests. Some ESP practitioners would argue against this dichotomy. Lipson (1984), for example, argues that ESP testing is not very much justifiable. Peretz and Shoham (1990) appear to assume a more or less similar position. According to Douglas (2000), however, the two types of tests are indeed different.

5.3.2.1. Defining Purpose

Widdowson (1983) points out that although all language courses (and tests) are purposeful, there is a difference in how purpose is defined. He suggests that in general purpose language courses, a distinction is made between aims, the eventual target behaviors of the learners, and objectives, pedagogical constructs which will enable the learners to achieve the behavioral targets. The goal, he says, of general purpose language courses, is to provide learners with an ability to solve, on their own, the profusion of communication problems they will encounter when they leave the language learning classroom. On the other hand, designers of specific purpose language courses, Widdowson suggests, often collapsed the distinction between aims and objectives, so that descriptions of target behaviors, usually derived from a needs analysis of a specific purpose language situation, become the course content. In other words, Widdowson argues, specific purpose language teaching suffered from a lack of theoretical motivation for course design, and became a very narrowly focused training exercise in which learners were taught specific behaviors but not strategies enabling them to adapt

to new, unspecifiable situations. Although many LSP courses are nowadays more strategically oriented, language teachers can tell anecdote after anecdote about learners who demonstrate an ability to perform the required language functions in the context of the classroom, but as soon as they walk out of the door, fail miserably to transfer the skills to the requirements of communication outside the classroom. There seems often to be a gap between what students can do in the classroom and what they can do in the real world.

In discussing the issue of a lack of theory in LSP testing, Davies (1990: 62) argues that "Tests of LSP/ESP are indeed possible, but they are distinguished from one another on non-theoretical terms. Their variation depends on practical and ad hoc distinctions that cannot be substantiated." Douglas (2000) tries to provide some theoretical justification and framework for LSP testing in an attempt to take LSP testing out of the realm of narrowly focused behavioral assessment and to bring it more in line with the theoretical underpinnings of communicative language testing. Such an approach, as Douglas eloquently puts it, will lead to the assessment of the abilities that underlie communicative performance which will be generalizable from one situation to the next and from the test situation to non-test target situations. For example, if test takers demonstrated in a test that they could successfully read a patient's case history and then could use that information to write a letter of referral to another physician, we want to be certain that they could carry out a similar task in a non-test situation. More importantly, we must ask how many different communicative situations might the test takers potentially need to deal with in their work as physicians which need to be tested in a test of medical English. However, it is practically impossible not only to sample the variety adequately, but even to list all the possible permutations of communicative events that language users must cope with. The problem of generalizing to real life is a central one in LSP testing. LSP practitioners wish to avoid the potential problem of producing a test on which performance is interpretable only in terms of that test. This situation can come about when we equate target behaviors with test content without noting the features of the target situation that are shared by the test tasks. That is, it is only by taking note of the features of the target situations and comparing them with those of the test task, that we can make that inference with any certainty. Bachman and Palmer's book *Language testing in practice* (1996), for example, adopts an approach to

test design and development that is useful in informing and carrying out any LSP testing enterprise.

5.3.2.2. Generalizing to Real Life

It has proven very difficult, and may eventually prove to be impossible, to make predictions about non-test performance in the real life target situation on the basis of a single test performance, no matter how true to real-life the test tasks might be. This is so because language use, even in highly restricted domains, such as taxi-driving, accounting, welding, biochemistry, or waiting tables, is so complex and unpredictable that coverage, or sampling of tasks, will be inadequate. Skehan (1984: 208), for example, writing about the problems of testing English for specific purposes, notes that "Merely making an interaction "authentic" does not guarantee that the sampling of language involved will be sufficient, or the basis for wide ranging and powerful predictions of language behavior in other situations." Spolsky (1986: 150) agrees, and reminds us that how speech acts are realized is the result of a complex interaction among many contextual variables, and although we might study pragmatic values and sociolinguistic probabilities of various forms appearing in different contextual environments, "the complexity is such that we cannot expect ever to come up with anything like a complete list from which sampling is possible." Bachman (1990) points out that now that it has become commonplace to recognize that language use takes place in contexts, and must be interpreted with reference to the context, and since the domain of language use consists of a potentially infinite number of unique instances, the assumption that we will be able to predict future communicative performances on the basis of a single test performance becomes untenable. He offers an example of attempting to produce a test of English proficiency for taxi-drivers in Bangkok by making lists of actual utterances the drivers might be expected to control. It soon became clear that the complexity involved in negotiating meaning even in this relatively narrowly defined context meant that "there was probably an infinite variety of conversational exchanges that might take place" (Bachman, 1990: 312). Skehan hypothesizes a similar problem in another domain, that of a waiter in a restaurant:

> Although at first sight 'waiter behavior' might seem to be a straightforward affair, we soon need to ask questions like: what range of customers needs to be dealt with? What range of food is to be served? Once one probes a little, the well-defined and restricted language associated with any role is revealed to be variable, and requiring a range of language skills. (Skehan, 1984: 216)

Tests developed in the real-life mold, which equate language ability with a specific language performance, are analogous to the training courses criticized by Widdowson, above, as failing to test the ability of the learners to deal with new, unexpected, unique communication problems. This is a real problem for specific purpose language testing. Tests might contain tasks that mirror faithfully those of the target situation, and these tasks might meaningfully engage the test takers' language ability, and yet the test overall might not be truly representative of the target situation, since there are simply too many possible variations of target situation to cover adequately in a test.

This problem is considered in more detail in the discussion of authenticity below, but for the moment let us agree that what is required in LSP testing is not the holistic replication of a specific purpose domain, but rather the use of features or characteristics of tasks in specific purpose language use situations in the construction of test tasks. This leads us to a view of LSP testing in which test tasks are developed on the basis of an analysis of characteristics of context and tasks in target language use situations. It is this analysis of target language use task characteristics which will allow us to make inferences about language ability in the specific purpose domain. The distinction between "ability" and "performance" is an essential one in the approach to language testing advocated in this book. The interaction between ability and task characteristics leads to authenticity, which can be interpreted as the extent to which the test does in fact engage the test takers in tasks characteristic of the target language use situation.

It should be clear from this discussion, too, that language tests are not either specific purpose or general; rather, there are degrees of specificity, which can be described along two dimensions: (a) the amount of content or background knowledge required for carrying out test tasks, and (b) the narrowness of interpretations which may be made on the basis of test performance about language use in real-life contexts. In other words, language tests will be more or less specific purpose in relation to the degree to which they require the engagement of specific purpose content knowledge in responding to the test tasks and the degree to which they allow generalizations about language use in specific situations. For example, the Test of English as a Foreign Language (TOEFL) is a test intended to measure English proficiency broadly interpreted, without the engagement of any special background knowledge or specific reference to use (Educational Testing Service, 1965). It would thus be considered a more general purpose language test. On the other hand, a test such as the Proficiency Test in English

Language for Air Traffic Controllers (PELA), a language test for trainee air traffic controllers in Europe (Institute of Air Navigation Services, 1994), requires a large amount of specialized knowledge about air traffic control, and interpretations of language use are specifically limited to the work of air traffic control officers. The PELA is, therefore, a prototypical example of a highly specific purpose language test. Between these two extremes is a test such as the Test of English for Educational Purposes (TEEP), intended as a test of academic English skills (Associated Examining Board, 1984), though not related to any specific field, and the Taped Evaluation of Assistants' Classroom Handling (TEACH), a test of instructors' ability to present information to students in specific academic fields (Abraham and Plakans, 1988). On these grounds, Douglas (2000) emphasizes that test developers must always take both test purpose and task characteristics into account when setting out to measure communicative language ability.

5.4. CRT and NRT

A very important concept for specific purpose language testing is that of criterion-referenced (CR) testing. Usually contrasted with norm-referenced (NR) language testing, CR testing differs from NR tests both in design and in the interpretation we make of performance on them. NR tests are designed to maximize distinctions among test takers so as to rank them with respect to the ability being tested. CR tests, on the other hand, are designed to represent levels of ability or domains of content, and performance on them is interpreted with reference to the criterion level (Bachman, 1990). In other words, on NR tests, passing would be determined by relative ranking within the population of test takers, while on CR tests, test takers might all achieve the criterion and so pass. As an example of CR test use, suppose we wanted to test prospective candidates for certification as scuba divers to see whether they knew certain essential diving terminology, such as "be jacket," "pony bottle," and "regulator." Scuba divers use this terminology in pre-dive checks, so it is essential to know whether candidates can carry out a pre-dive check using the appropriate vocabulary. We are not interested in discovering who knows the most terms; we want to know which candidates know all the essential terms. Our performance criterion, then, is whether the candidate can use all the terms appropriately. On the other hand, continuing with the scuba diving example, as a way of motivating students to learn the various concepts associated with diving, such as the relationship between depth and pressure, and the different types of

equipment, an instructor might offer a weekly prize to the five students who get the highest scores on quizzes. In this case, the aim would be to rank the students against each other so that the top five could be identified—a norm-referenced use of the tests. Both types of test uses are relevant to LSP testing; however, the development process associated with CR testing, which involves a detailed analysis of the target language use situation, is of most direct relevance in LSP testing, particularly with regard to a fundamental concept in specific purpose testing—authenticity.

It is important, therefore, to note that specific purpose language tests might be developed as either CR or NR tests. However, CR testing offers an important perspective to LSP testing: the necessity of specifying precisely the level of ability or the domain of content that is to be the criterion for performance. Thus, the process of developing a CR test, requiring as it does the precise, detailed specification of not only the features of the specific purpose target language use situation, but also the criteria for evaluating performance, is extremely useful in LSP test development. In fact, the development of evaluation criteria, or rating scales, is perhaps the most important, and also the most vexing, problem in LSP test development (McNamara 1990a, 1996). Until very recently, the task of developing assessment scales has been left to test developers and other applied linguists, and, not surprisingly, the scales they have come up with reflect a linguistic orientation, so that such categories as grammar, cohesion, vocabulary, fluency, intelligibility or comprehension are commonly employed. As such, a precise definition of assessment criteria is an essential part of the LSP test development process, and CRT procedures offer a systematic approach to specifying these criteria.

5.5. Authenticity

Since authenticity is such an important concept in specific purpose language testing, it is necessary to consider its meaning in some detail and with some precision. Kramsch (1993) points out that the term has been used to indicate a reaction against the often artificial language of language textbooks and tests; it refers to the way language is used in non-pedagogic, non-test, natural communication. Since the publication of Widdowson's *Explorations in applied linguistics* (1979), many language teachers and testers have come to view authenticity as a property not of spoken and written texts themselves, but of the uses people put them to:

> It is probably better to consider authenticity not as a quality residing in instances of language but as a quality which is bestowed upon them, created by the response of the receiver. Authenticity in this view is a function of the interaction between the reader/hearer and the text which incorporates the intentions of the writer/speaker … Authenticity has to do with appropriate response. (Widdowson, 1979: 166)

In other words, a text is said to be authentic when it is being used for the purpose intended by its author. Therefore, a key concept in Widdowson's formulation above is that of interaction between the language user and the text. Douglas (2000) makes use of this notion in his characterization of authenticity in specific purpose language testing.

Widdowson (1979, 1983) notes that there is often a confusion between the use of 'authentic' to refer to examples of language actually produced by users in a communicative situation versus reference to the activities and procedures that language users engage in, in association with the forms of language produced. He suggests a distinction between the terms "authentic" and "genuine." The former refers to activities or processes associated with instances of language use while the latter denotes the actual spoken or written texts produced by the users. Thus, our use of a set of instructions from a lab manual for the purpose of testing an instructor's ability to understand and use imperatives would be the use of a "genuine" text for a purpose other than that for which it was intended. Bachman (1991) reminds us of Widdowson's point, quoted above, that authenticity is a function of an interaction between a language user and a discourse, and proposes two aspects of authenticity: "situational authenticity" and "interactional authenticity." The first aspect is composed of authentic characteristics derived from an analysis of tasks in the target language use (TLU) situation, the features of which are realized as test task characteristics. Thus, situational authenticity can be demonstrated by making the relationship between the test task characteristics and the features of tasks in the target language use situation explicit. The second aspect of authenticity, interactional, is closely related to Widdowson's definition above, and involves the interaction of the test taker's specific purpose language ability with the test task. The extent to which the test taker is engaged in the task, by responding to the features of the target language use situation embodied in the test task characteristics, is a measure of interactional authenticity. It is important in specific purpose language tests that both these aspects of authenticity are present. It is quite possible, for example, that a test task may be perceived by test takers as having nothing whatever to do with their field of study, but which they nevertheless find quite interesting and which engages their

communicative language ability interactively. Performance on the task would be interpretable as evidence of their communicative language ability, but not in the context of the target language use situation. By the same token, a test task may contain all the contextual attributes of the target situation and yet fail to engage the test taker meaningfully in communicative language use. Mere emulation of a target situation in the test is not sufficient to guarantee communicative language use, and, as Lewkowicz (1997) has pointed out, the focus on the interaction between the test taker's language ability and the situational characteristics of the test task is a strength of this dichotomous view of authenticity in specific purpose language tests.

Drawing on Bachman's (1990) conception of authenticity, Douglas (2000) proposes to employ his dual notion of authenticity in specific purpose language testing. According to Douglas, in LSP test development, what we must do is first describe a target language use situation in terms of features of context and task. We must then specify how these characteristics will be realized in the test so as to engage the test taker in test tasks, performance on which can be interpreted as evidence of communicative language ability with reference to the target situation. Building on the work of Bachman and Palmer (1996), Bachman, *et al.* (1991), and Davidson and Lynch (1993), Douglas (2000) develops a "means of classifying test tasks on the basis of dimensions … that we abstract from authentic language use" (Bachman 1990: 317) in the construction of specific purpose language tests.

5.6. Inference

According to Douglas (2000), a central goal in language testing is making judgements about test takers on the basis of their performance on a test. That is, testers give tests to elicit performances that they can observe so that they can make inferences about qualities of test takers that they cannot observe. A fundamental question involves what testers wish to make inferences about. They might want to make inferences just about a person's language ability; for example, this candidate is able to write business letters in English, using correct syntax, vocabulary, and spelling. Alternatively, they might want to make a statement not only about language ability, but also, particularly in LSP testing, about specific purpose background knowledge; for example, this candidate is able to write business letters in English, incorporating appropriate types and amounts of information from material provided.

Inferences of the latter type would be more complicated, since testers would need not only a measure of language ability, but also one of background knowledge. This would enable them to disentangle the two types of knowledge and understand, for example, whether a candidate's failure to incorporate appropriate types and amounts of information was due to a lack of language ability or a lack of background knowledge. Finally, testers might wish to make inferences about a candidate's specific purpose language ability, in which case, language and background knowledge would be left intertwined. The type of inference testers want to make would depend on the purpose for which they are giving the test, but in specific purpose language testing, the first type of inference, that about a decontextualized language knowledge, is probably not very useful. The second type, where they want to separate out language knowledge from background knowledge, would be most useful when, for example, the test takers were trainees in the specific purpose field. This is even more crucial when testers need to know if the "low test performance" of test takers was the result of problems with the language or a lack of background knowledge, so they could offer appropriate remedial instruction. The final possibility, inferences about the dual component specific purpose language ability, would be most useful in situations where testers could take the test takers' specific purpose background knowledge for granted, as in the case of qualified doctors who wish to demonstrate their language ability for purposes of licensure.

Making appropriate inferences is a crucial aspect of specific purpose language testing. McNamara (1989, 1996) has distinguished between making inferences on the basis of LSP test performance about ability to do future tasks or jobs in the target language use situation, on the one hand, versus making inferences about ability to use language in specific future tasks or jobs, on the other. This seems a subtle distinction, but it is of extreme importance for the theoretical foundations of LSP testing. McNamara cautions against the first type of inference since job performance is influenced by a number of factors, such as personality characteristics, that are independent of language ability. He makes a theoretical distinction between a strong performance hypothesis, about an individual's ability to perform target tasks successfully, and a weak performance hypothesis, about ability to use language in the target situation, and prefers the latter. Douglas (2000) argues that we are not attempting to measure communicative success in LSP tests, but rather the knowledge and abilities that underlie communicative performances, and that this point is related to McNamara's. According to Douglas, it is practically impossible not only to sample the

variety of tasks in a target domain adequately, but also even to list all the possible types of communicative events that language users must cope with. Thus, he agrees with McNamara that we should restrict ourselves in LSP testing to making inferences about language ability and not about job performance. However, Douglas also tried to establish a case for making inferences about specific purpose language ability, a construct defined on the basis of an interaction between language knowledge and specific purpose background knowledge. This departs somewhat from McNamara, who would include job-related background knowledge in his list of factors essentially unrelated to language knowledge. Douglas also argues that language knowledge must be interpreted differently from one domain of use to another. As Chapelle (in press) points out, context constrains language choice, and if testers are interested in making inferences about test takers' abilities to use language in specific situations, then background knowledge associated with those situations must be a part of the construct they wish to measure with their tests.

Bachman and Palmer (1996) suggest that there are three possibilities for defining the relationship between language ability and background knowledge (which they call topical knowledge). The first one is making inferences only about language knowledge in situations where test takers vary widely in background knowledge. The second one includes both language knowledge and background knowledge in situations where there is minimal variety among test takers. The last one defines language knowledge and background knowledge as separate constructs in cases where the test developers and score users are uncertain about the relative strength of test takers' background knowledge (Bachman and Palmer, 1996: Chapter 6). Douglas has taken the view in his book (2000) that the first scenario is not relevant to the LSP testing enterprise. Moreover, he discusses both of the remaining possibilities in relation to test purpose.

6. LSP Tests

Specific purpose language tests are by definition communicative. The definition of communicative language testing employs a number of key terms: communicative language ability, specific contexts of use, and test constraints. As a special case of communicative language testing, specific purpose language testing encompasses these concepts. Drawing on these characteristics, Douglas (2000) proposes a more precise definition of specific purpose language tests.

> A specific purpose language test is one in which test content and methods are derived from an analysis of a specific purpose target language use situation, so that test tasks and content are authentically representative of tasks in the target situation, allowing for an interaction between the test taker's language ability and specific purpose content knowledge, on the one hand, and the test tasks on the other. Such a test allows us to make inferences about a test taker's capacity to use language in the specific purpose domain. (Douglas, 2000: 19)

Douglas discusses reasons for wishing to develop "specific purpose language" tests, and notes that language performance varies with both context and test task. He argues that our interpretations of test takers' language ability must vary from situation to situation. He claims that technical language—that used in any academic, professional or technical field, including cooking, law, physics, chemistry, air traffic control, scuba diving, religion, or stamp collecting—has specific purpose characteristics that people who work in the field must control. In other words, what is commonly referred to, often disparagingly, as jargon has a specific communicative function within that field, namely precision.

In considering how specific purpose language testing is related to other types of, and approaches to language testing, Douglas discussed the distinction between so-called general tests and LSP tests. He notices that while all tests have purposes, in LSP testing, the notion of purpose is typically more narrowly focused than in more general language testing. He notices that there is a problem inherent in this focus, however, since there is in principle no way of determining how specific 'specific' needs to be. A criticism of specific purpose testing has been an assumption that if a test taker could perform the real-life test task, he or she would be able to perform in the target language use situation. However, there are serious problems in demonstrating this to be the case. It is impossible, except in the most restricted language use situations, to specify with any completeness the range of language forms that will be required. This is so because language use, even in relatively specific domains, is so complex and unpredictable that coverage, or sampling of tasks, will be inadequate.

As a way out of the dilemma of never-ending specificity on the one hand and non-generalizability on the other, Douglas referred to context and task characteristics, which are drawn from an analysis of a target language use situation. These will allow test developers to make inferences about language ability in the specific purpose domain. LSP test development, what must be done is to first describe a target language use situation in terms of characteristics of context and task. Then, the LSP test developer

should specify how these characteristics will be realized in the test so as to engage the test taker in test tasks, performance on which can be interpreted as evidence of language ability with reference to the target situation. In this connection, attempts have been made in the field of language assessment to reconcile language testing with TLU situation. One such attempt was the introduction of performance assessment.

6.1. Performance Assessment (PA)

In the past two decades, there has been a major shift in language testing towards the development and use of performance tests. The basis for this shift is the expectation that such tests would assess a more valid construct of what it really means to know a language. In the following sections, I will review the topic of performance testing by focusing on its definitions, theory, development, and research. I will begin with a review of the different definitions of performance testing and provide examples of the types of performance tests that have been developed and used. I will then examine the extent to which performance tests have drawn upon the theoretical discussions of competence and performance. The next section will describe the research that has been carried out on performance tests. My discussion of performance assessment will end with an agenda for development and research on the many unanswered questions concerning performance testing

6.1.1. Background and Definitions

Language testing has always followed linguistic theories of the time. Thus, the communicative era in the 1970s generated a wave of criticism of the traditional non-communicative tests. These tests were seen as being limited in their concept and as producing artificial language, as opposed to the language normally produced by human beings. For example, the kind of tests used for testing oral language included mostly mechanical repetition of words and sentences and the supplying of pattern answers to pattern questions. In subsequent years there was a shift in language testing towards the development and use of tests that resembled features of real language use and that required test takers to perform language that was authentic, direct, communicative, and performance-based. Such tests, it was believed, would reflect better ‘real life’ language use as they would tap a broader construct of ‘what it means to know a language.’ A number of terms were used along with these types of tests. Clark (1975) referred to “direct tests” in which both the testing format and the procedure duplicate, as closely as possible, “the setting and operation of real life” situations in which language proficiency

is normally demonstrated. Jones (1977) proposed performance tests in which test takers provide information on functional language ability. Morrow (1977) recommended tests that would offer test takers the opportunity for spontaneous language use in authentic settings and activities which the candidate would recognize as relevant. Canale and Swain (1980) referred to performance-based communicative tests which required test takers to perform language while considering criteria such as saying the right thing, at the right time, to the right person. The Foreign Service Institute (FSI) Oral Interview (OI) test was the most relevant example of such a direct, performance-based test (Clark, 1975; Jones, 1977), requiring test takers to use language in a face-to-face oral interaction. The tester asked questions on a variety of topics, and the test taker provided the oral language sample which was then evaluated by the tester with the aid of a rating scale.

In this way, 'performance' became one feature among a number of others, such as 'direct,' 'functional,' and 'authentic,' all of which characterized communicative tests of that era. The unique aspect of the 'performance' feature was that test-takers were expected to replicate, as much as possible, the type of language used in non-testing situations (Bachman, 1990; Bailey, 1985). Thus, performance testing referred to tests where a test taker is tested on what s/he can do in the second language in situations similar to 'real life.' Jones (1985) specified that such tests also required the application of prior learning experiences in an actual or simulated setting where either the test stimulus, the desired response, or both were intended to lend a high degree of realism to the test situation. The above description characterized features of performance tests in the 1970s. In the 1980s, performance testing became associated more with specific tasks and contexts of professional preparation and certification, mostly in the workplace (Wesche, 1992). In this context, performance testing borrowed from the field of vocational testing in which a test taker needs to carry out realistic tasks applying language skills in actual or simulated settings (Carroll and Hall, 1985). The criteria used to evaluate the performance was an approximation of the way performance would be judged in the specific and actual target circumstances, including adequate fulfillment of tasks. Wesche (1992) notes that these tests tap both second language ability and the ability to fulfill nonlinguistic requirements of the given tasks. With these types of tests, the main psychometric feature is that of predictive validity; the tests predict how well a test taker will perform under real conditions in a specific context (Jones, 1985). The underlying assumptions with those type of performance tests is that nonlinguistic factors

are present in any language performance; consequently, it is important to understand their role and channel their influence on language performance. In this regard, McNamara (1996) has proposed a distinction between strong and weak hypotheses on performance tests. In the strong sense, knowledge of the second language is a necessary but not sufficient condition for success on the performance-test tasks: success is measured in terms of performance on the task, and not only in terms of knowledge of language. In the weak sense, knowledge of the second language is the most important, and sometimes the one factor, relevant for success on the test. The specific contexts in which performance testing is used involves a clientele (students, employees, etc.) with certain shared second language needs that can be identified and described, and that can subsequently be translated into test tasks and overall test design. Performance testing, therefore, is associated with a specific context and its strongest requirement will be a detailed description of that context and the language performances associated with it (Sajavaara, 1992; Wesche, 1992).

Jones (1985) distinguished among three types of performance tests according to the degrees that the tasks require actual 'performances.' In a 'direct' assessment, the examinee is placed in the actual target context, and the second language performance is assessed in response to the naturally evolving situation. In the "work sample" type, there is a realistic task which is generally set in the target context: this type enables control of the elicitation task and a comparison of the performance of different examinees while simultaneously retaining contextual realism. The 'simulation' type creates simulation settings and tasks in such a way that they represent what are thought to be pertinent aspects of the real-life context. 'Role playing' is frequently used as a simulation technique where both the examiner and the examinee play roles. There have also been a number of efforts to use devices such as video, audio recorders, and telephones. For all these types, however, it should be clear that it is never possible to satisfy all the conditions of performance communication and contextual grounding since testing is not really a normal activity. Recognizing this fact, more recent techniques utilize a variety of non-testing procedures that reflect the real performance context: these include record reviews, portfolios, self assessment, participant and non-participant observations, and external indicators.

Wesche (1992) differentiated between performance testing in the work-place and in the instructional context. In the work-place context, tests are used for job certification and for prediction of post-training behavior. In the instructional context, tests are used

for washback, diagnostic feedback, and increasing students' motivation. Early introduction of performance tests can help communicate to learners the importance of language objectives, instructors expectations, and criteria for judging performances. Texts and tasks which are used in performance testing also make very good instructional tasks, and ratings obtained from performance tests can be translated to diagnostic feedback in the form of profile scores. Thus, performance tests can actually be introduced in the pre-instruction phase for placement, formative diagnosis, and achievement purposes; during the program itself, these tests can be used for achievement purposes, for summative testing at the end of a program, and for certification purposes. In instructional situations where the goals are based on an analysis of large language needs, there is a place in the curriculum for an evaluation system which includes performance-type tasks.

In constructing a performance test, a need analysis is conducted in order to provide a detailed description of the specific context and tasks which learners will need to perform, the specific conditions under which these tasks will be performed, and the criteria against which the performance can be judged. Then, the learners' performances can be judged over a range of tasks that need to be sampled, using a variety of instruments and procedures. The needs analysis will specify the context of the second language use, the type of interactions foreseen, the roles, discourse types, and language functions to be performed, and the basis on which successful fulfillment of the second language tasks is to be judged. It is with respect to these needs that the performance test is designed, texts and tasks are selected, and evaluation criteria are determined. These are then translated into appropriate test objectives and tasks, and later into actual test design and scoring. Performance tests are generally assessed with the aid of rating scales which describe what a person can do with the language in specific situations.

There are a number of questions that need to be addressed in constructing performance tests: How can the evaluation criteria reflect the kinds of judgments and consequences that the performance would entail? What relative weighting should be given to the different criteria? How can the scoring information be interpreted and presented so as to give maximum information back to the test users? There are also questions more generally related to the criteria by which the performance should be judged: What is the proportion of 'language' vs. 'domain knowledge' to be assessed? Who should be the judge of the performance—a native speaker, a domain specialist, or a teacher? Although most performance tests do use the native speaker as the top level of

the scale (ACTFL, 1986; Emmett, 1985), this issue has been a topic of debate in the language testing literature for many years (Alderson, 1980; Bachman, 1990). Hamilton, *et al.* (1993) claim that performance on a test involves factors other than straight second language proficiency that cause an overlap in the performance of native and non-native speakers. Therefore, the reference to native speaker performance is unwarranted.

In the past few years, performance testing has become a common form of assessment in the educational research context. It is associated with any procedure not employing paper-and-pencil multiple choice items, and it includes a variety of assessment alternatives such as open ended responses, constructed responses, problem solving tasks, essays, hands-on science problems, computer simulations of real world problems, exhibits, and portfolios of students' work. (Linn, Baker, and Dunbar, 1991)

6.1.2. Examples of PA

Over the years, a wide variety of performance tests have been developed in the language testing field. Below are descriptions of relevant performance tests, although the following examples are by no means an exhaustive list.

The early English for Special Purpose (ESP) test (Carroll, 1980) included specific performances expected of students in academic situations. It was a significant development as the test was based on a needs analysis (Munby, 1978) and test specifications assessing whether prospective students at British universities would be able to cope with the language demands of their studies. The tests were task- and performance-based and were linked to performances that needed to be predicted. These specifications then became the English Language Testing Service (ELTS) test with two parts (Alderson and Hughes, 1981; Weir, 1988). The first part assessed more general, but academically related, language skills of reading and listening. The second part involved candidates choosing content areas relevant to their language abilities. Similar types of performance tests include the Test of English for Educational Purposes (TEEP) (Emmett, 1985; Weir, 1988) and the International English Language Testing System test (IELTS; British Council/UCLES, 1989). The IELTS replaced the earlier British ELTS test and is an EAP performance test in winch most tasks simulate those encountered by students in university settings.

Cole and Neufeld (1991) report on a test used by the Canadian Public Service to certify the second language skills of employees in bilingual positions. It includes performance tasks from a vast bank of work-related role-plays reflecting a range of

situations that might occur in each of the major public service job classifications. Contextualized interviews are introduced into each assessment according to the situations in which employees might need to use the second language. These tasks are based upon a detailed analysis of second language needs in bilingual government positions (Cole and Neufeld, 1991). The writing tests include, at the lower proficiency levels, tasks such as filling out forms and writing short work-related messages; at higher levels the tests include preparation of correspondence and executive summaries.

Todesco, Cahill, and Frey (1983) report on the Canadian Public Service's Communicative English at Work instructional program for Francophone government employees in bilingual positions. This is an intensive training program in which job-related performance testing elements are integrated into all evaluation activities. Each test includes work-related role-plays and self assessment elements that reflect daily work situations in the Canadian Public Service. The weight given to performance elements at different points in the training varies. The testing system prepares students for the Canadian Public Service Bilingual Certification tests.

Sajavaara (1992) describes the Finnish Foreign Language Diploma for Professional Purposes that offers information for employers to use in screening new professionals in terms of foreign language skills. The test provides certificates for various workplace situations, and it serves business sectors that require ability to communicate in foreign languages. The test is being offered in English, German, and Swedish; French and Russian are planned for a later phase as are workplace domains such as business administration services and engineering industries. Further plans are also being made to establish similar testing for management, public administration, research development and technology, secretarial services, and service industries. The Diploma provides a certification and a nationally accepted assessment of English proficiency in professional contexts. Employers can acquire a nonpartisan evaluation of job candidates in areas that may be crucial in terms of professional competence. The battery includes criteria that pertain to the common core of language use as well as criteria that pertain to more specific needs arising from jobs and professions.

Perhaps what is most interesting about this Finnish test is that it is tailor-made according to job-specific orientations. A detailed questionnaire is administered to the candidates and to the employers before the decision is made on the specifics of each individual test. The questionnaires allow the examination to be flexible, relevant, and accountable in terms of each candidate's actual needs. For example, managers and

secretaries will be exposed to different language-related tasks; their language needs may overlap to a degree but the precise tasks are not similar. Mastery of professional practices, job-related experiences, and

knowledge about subject matter and processes in the profession concerned are all addressed since they are important elements in the attempt to cope with the language tasks. Thus, a salesperson may be asked to give a product demonstration or perform a service based on a written and/or oral brief; a secretary may be asked to draft a letter on the basis of prior correspondence and a simulated telephone message. Demands on accuracy and fluency vary from one testee category to another and from one task to another. Each examination lasts about five hours and consists of reading comprehension and writing, listening comprehension and speaking, face to face oral interaction, and vocabulary and grammar. Content validity is achieved by deriving most of the tasks from the needs analyses conducted with the candidates before the examination.

McNamara (1990a) reports on a performance test used for health professionals in Australia (the Occupational English Test—OET) which is used for immigrants and refugees who want to be medical practitioners (nurses, physiotherapists, occupational therapists, dentists, speech pathologists, veterinarians, and surgeons). It is administered along with other measures developed by the relevant professional examining body. The OET test contains a listening test.

In this listening test, the test takers answer questions related to a talk on a professionally relevant subject as well as a consultation between a general practitioner and a patient. The reading part is multiple-choice, based on articles from professional journals or magazines. The speaking test uses profession-specific content within a common format that consists of two role-plays and a short interview. The assessment scales for the speaking test contain six categories of the semantic differential type that indicate overall communicative effectiveness, intelligibility, fluency, comprehension, appropriateness of language, and resources of grammar and expression. The writing test is also profession-specific and requires writing a letter of referral based on case notes or extracts from medical records.

A large number of contextualized forms of assessment utilizing a variety of creative procedures have recently been developed for general educational, school contexts. Moss (1994) reports on assessment procedures in which students prepare “papers” or “exhibits” that are evaluated by a committee of teachers, parents, other students, and members of the community. In another example, teachers meet regularly to discuss

progress of individual students or curriculum issues in much the same way that physicians conduct case conferences. In another project, committees of teachers, supervisors, and others at the school level engage in periodic audits of the individual portfolios, and committees at higher levels of the system review the procedures of the school level committees to ensure that appropriate standards are being followed. On the system level, Moss reports on diagnostic and contextualized forms of assessment in which committees of teachers design a district-wide portfolio-assessment system and invite educators from outside the district to audit the portfolio-evaluation process.

6.1.3. History of PA

Language testers, in their attempt to develop appropriate measures of language ability, have devoted much time and effort to describing the construct of language knowledge. The rationale is that a clear definition and identification of the structure of language will enable language testers to design tests to match such descriptions, and these will have direct consequences for the validity and design of language assessment. This section provides a short review of some of those construct descriptions, and examines the extent to which they have contributed to performance assessment.

Chomsky's (1965) distinction between competence and performance provided the basis for differentiating between a 'slate' and 'actual performance.' Later expansion by Hymes (1972) introduced the difference between linguistic and communicative competence versus linguistic and communicative performance, referring to the interaction between grammatical and sociolinguistic competence, on the one hand, and performance on the other. Hymes included a new type of ability, the 'ability for use,' which is the individual's underlying potential to realize a possible, feasible, and appropriate speech act, and not just the actual performance. Communicative performance signifies the realization of the user's underlying communicative competence.

The language testing literature of the 1970s devoted much space to the issue of the structure of language, mostly around the notion of what constitutes language ability. Oller (1976; 1979) promoted the notion of language as a unitary factor rather than as a divisible construct. He specified underlying language behavior as being based on learners' pragmatic grammar of expectancy—a psychological representation of the language users' ability to map utterances onto contexts; This ability was to be operationalized through integrative tests such as cloze and dictation since, in these tests,

"learners had to mobilize their linguistic and extra-linguistic knowledge to reconstitute the meaning of a text" (Brindley, 1986). Oller was not specific about whether his hypothesis implied competence or performance, but he did refer to an underlying competence. Yet, it was not these aspects in Oller's writings that drew attention; rather, it was the notion of a unitary factor, manifested through cloze testing, that received attention and criticism (Vollmer and Sang, 1983).

Bachman and Palmer (1982) demonstrated that language was not composed only of one general factor, but of a higher order general factor plus two trait factors which they called grammatical and pragmatic competence. Canale and Swain (1980) addressed a broader concept of the components of language which was not focused on the concept of a well-formed sentence but on the appropriate use of language in a specific context. Canale and Swain's approach adopted Hymes' notion of communicative competence; their argument was that linguistic competence should be viewed as part of communicative competence since rules of grammar are meaningless without rules of use. Accordingly, communicative competence consisted of grammatical, sociolinguistic and discourse competence; they later added a strategic component, implying that a valid measure of language ability needs to include these four components.

Canale and Swain included 'ability for use' in what they called "communicative performance"—the realization of these competencies and their interaction in the actual production and comprehension of utterances, and subsequently the actual demonstration of this knowledge in real second language situations and for authentic communicative purposes. McNamara (1996) notes that their definition of performance referred to 'actual use,' implying that their model lacked a notion of 'potential' for use or underlying skill. Yet, two components of their model (i.e., discourse competence and strategic competence) involved 'ability for use' rather than merely 'knowledge.' Discourse competence, for example, includes cohesion and coherence, and it is not clear that the ability to perceive and create coherence in discourse is entirely a matter of knowledge. Strategic competence also seems to involve a similar compounding of knowledge with an ability or skill; this is exemplified by such a strategy as 'how to paraphrase grammatical forms that one has not mastered or cannot recall momentarily.' While the Canale and Swain framework broadened the scope of language testing theory—introducing new sub-components within communicative competence and giving preference to the communicative components over linguistic ones—they did not address the competence-performance distinction in a clear and coherent way.

> Indeed, a few testers did try to get away from the distinction and to claim that any linguistic behavior constitutes instances of performance (Rae, 1985). This broader view does away with the communicative-performance division since competence can only be inferred through performance which is therefore all that can be directly observed and hence assessed. Since no communicative theoretical model made a clear distinction between competence and performance, pragmatic considerations and operational definitions should guide the development of language tests. With no underlying theory of performance, actual performances, translated into tasks and actions, became the de facto theory. The tests, then, were communicative, functional, authentic, and direct, with a special focus on performance and ignoring the notion of competence. The main criteria for determining what it means to know a language was in performing tasks. These performance definitions have since dictated the process of test development: The purpose and context for a test are defined based on a needs analysis; samples of the 'behavior' in that context are defined, actual performance or simulation tasks that elicit the performance are selected; tasks are performed by the test-taker (in simulated or real situations); the language samples are elicited; and the language samples are assessed, usually by means of rating scales which define criteria for successful performance. Language testers, then, have turned to behavioral definitions in which language is described in performance terms rather than in implicational terms. Performance tests also have gained high face validity, high washback, and high client acceptability. Competence has not been the focus of these language tests, as only what could be observed was measured, and covert mental acts were totally overlooked. (Shohamy, 1995: 196)

A number of factors, therefore, explain the emergence of performance-type assessment:

1) The de-emphasis between competence and performance in the various communication models and its minimal effect on test development;
2) The broad acceptance of communicative performance in the domain of language teaching;
3) The limited repertoire of competence tasks that could be used;
4) The societal need for demonstrating 'face validity' and the need of language-testers to show decision-makers that tests are, in fact, what they are expected to test and can therefore be trusted;
5) The strong influence of Hymes' communicative views, the distancing of formal linguistics from application, and the upgrading of notional syllabuses, pragmatics, speech act theory, etc.;
6) The rise of communicative teaching trends which meant that the test could be expected to create washback effects;
7) The strong appeal of rating scales; and
8) The direct relationship between test performance based on needs analysis and criteria.

While the absence of 'competence' in performance testing seemed logical, the question remains whether there is evidence for construct validity of performance tests when they

are based on a performance *per se* (i.e., whether overlooking competence still results in construct-valid tests).

Messick (1994) distinguishes between constructs and tasks. Constructs, according to Messick, refer to theories of competence knowledge, communication, and skills underlying performance; tasks refer to performance. Thus, Messick claims that there is a need to establish construct validity through empirical evidence. Messick notes the distinction between competence and performance as follows: Although competence must be inferred from observations of performance of behaviors, these inferences are not often straightforward, particularly those inferences about lack of competence from poor performance. Similarly, in using performance tests in educational contexts, Messick makes the following claim: "We are rarely concerned just with the particular performance *per se* but also with the knowledge, skill, and other attributes that enable both the given performance and a range of other performances engaging the same knowledge and skills" (Messick, 1994:16). This suggests that constructs like relevant knowledge and skills, rather than domain-relevant tasks and performance ought to drive the development, scoring, and interpretation of performance assessment. Messick claims that there is a need to ponder the potential consequence of adopting a task-centered approach as opposed to a construct-centered approach in performance assessment. The nature of the construct should guide the selection and design of the relevant tasks as well as the rationale for scoring criteria and rubrics. Focusing on constructs also illuminates construct-irrelevant variables which might distort task performance, or scoring, or both, and threaten test validity. Adopting a task-oriented approach in performance testing, when it is not driven by a sound theory of performance, is insufficient. Such an approach may be viewed as simplistic and narrow as long as it does not include evidence of construct validity. Thus, there is a need to develop a theory of performance which emphasizes the many variables that interact in that construct.

McNamara (1996) claims that, in constructing a theory of performance, it is important to outline variables that relate language to other cognitive and affective areas (e.g., subject matter knowledge, personality, gender, attitudes, and beliefs), as well as define a whole set of non-linguistic variables that integrate language knowledge with communication skills. Also, there is a need to introduce aspects of performance in the first language in order to make the conception of communicative competence broader than is found in Hymes' model. Thus, a serious question arises regarding the validity of

performance tests: These tests are not based on a fuller construct theory but rather on a narrow view of communication not including many of these additional components. McNamara also states that there is a need for a comprehensive model of performance based on an understanding of factors underlying 'ability for use' which will guide the selection of tasks for tests. Such a model may include elements of first language ability but they should be second-language-communication based and take into consideration the many variables relevant to the communicative act. The model, as a multidimensional construct, should include resources from various domains: psychology, sociology, communication, subject-matter knowledge, etc. Clearly, grammar will also be included in such a construct, but this could be tested empirically as part of the process of establishing the validity of the construct. The model needs to be rich enough to enable conceptualization of all the significant issues involved in second language communicative-performance situations, and there is no limit to the dimensionality of such a model as it should be as complete as possible. Having constructed a model, it will then be necessary to determine what is appropriate and possible to assess in a given test situation. Explicitness of the model, and reference to its dimensions, will be critical in any discussion of the construct validity of a given assessment procedure.

Thus, there is a need for an expanded theory of performance testing. To date, a number of steps have been taken in this direction. The most important one is the emergence of the Communicative Language Ability model (Bachman, 1990). In this model, language ability concentrates around organizational and pragmatic competencies. Organizational competence consists of grammatical and textual competencies while pragmatic competence consists on illocutionary and sociolinguistic competencies According to Bachman (1990), the model consists of both knowledge, that is. competence, and the capacity for implementing, or executing, that competence in appropriate contextualized communicative language. Bachman claims that an attempt is made to characterize the processes by which the various components interact with each other as well as with the context in which language use occurs. However, to date the model has not been construct-validated, and it is often claimed to be too complex and difficult to apply (Spolsky, 1989).

In spite of the wide use of performance testing, there has been limited research on its nature and its psychometric properties. The past few years have witnessed an important trend in this direction, and the 1993 Language Testing Research Colloquium (LTRC) was devoted primarily to research on performance testing. Research conducted on

performance testing can be a good source for understanding and constructing a performance theory as well as a means for validating performance testing constructs. A review of a number of studies which have investigated various dimensions of performance testing will indicate possible future directions for performance-test validation.

6.1.4. Review of Research on PA

A number of validation studies on performance tests were conducted by McNamara (1990b; 1991) using the Rasch Item Response Theory (IRT) to investigate various aspects of content and construct validity. In the analysis of data from an ESP listening test for health professionals, the Occupational English Test (OET), developed on behalf of the Australian Government, he showed how Rasch IRT can be used in the validation of the listening subtest (McNamara, 1991). He demonstrated that, despite the fact that the two-part test involved rather different kinds of listening tasks, statistical tests confirmed the finding of the partial credit analysis itself, that it is possible to construct a single dimension using these items for measuring listening ability in health professional contexts. It also showed that the kinds of listening tasks required of candidates in the two parts of the test represent significantly different tasks in terms of the level of ability required to perform successfully. McNamara argues for the usefulness of IRT as a tool in the exploration of test constructs, and discusses the implications of the empirical analysis for the validity of language-performance tests involving the skills of speaking and writing. In another study, McNamara (1990a) showed how the different parts of the OET test could be validated.

He demonstrated that certain assessment criteria on the OET had a disproportionate role in the allocation of test scores in a way that called into question the communicative orientation of the test; candidates were to some extent measured on selected features which were important to the raters independently of the design of the test.

A few studies have examined the predictive validity of other types of performance tests. Elder (1993) examined the relationship between the language proficiency of overseas students as measured by the ELTS test and subsequent performances by these students in teacher education studies. The findings did not offer conclusive evidence about the value of the ELTS as a predictor of performance, but it confirmed evidence from previous studies that language makes an important difference at low proficiency levels. Elder claims that, at other proficiency levels, many other factors such as subject

knowledge, scholastic aptitude, and cultural adaptability are likely to interact with language ability in determining progress; these factors should be taken into account in making initial selection for language assessment measures.

A number of studies have focused on various aspects of the rating systems used in performance tests. McNamara (1990a) showed that grammar and appropriateness played a major role in the rating systems of performance tests. He found that, as time went by, raters appeared to be evaluating the candidate's production more in purely linguistic terms, that is, less on the role of communicative partner and more on the role of language analysis. The rater becomes less influenced with what s/he is hearing because of the ongoing grammatical inspection of output, marking the candidate down accordingly. In the final analysis then, it is features of accuracy which count more heavily in the rater's mind. McNamara's research also showed that some aspects of performance measurement were made as a result of an interaction between the behavior of candidates and the orientation of raters: some of these outcomes were not necessarily related to the test.

A number of additional studies have compared different types of elicitation procedures on performance tests. Stansfield and Kenyon (1988) examined the concurrent validity of direct versus semi-direct tests in a number of languages. They found a high correlation between the two types of tests and recommended the use of semi-direct tests as valid and practical substitutes for direct tests. Wigglesworth and O'Loughlin (1993) investigated the comparability of two versions of an oral interaction test (a direct 'live interview' version and a semi-direct 'tape-based' version as part of the ACCESS test, the Australian assessment of communicative English skill test administered to intending migrants to Australia. They showed that the two versions were highly comparable.

Shohamy (1994) researched the validity of direct versus semi-direct tests using both qualitative and quantitative procedures. While the concurrent validity of the two tests, as obtained from correlations was high (Shohamy, Gordon, Kenyon, and Stansfield, 1989; Shohamy and Stansfield, 1991), qualitative analyses of the two tests revealed that they differed in a number of aspects. Differences were found in the number of functions and topics used in the elicitation tasks and in the communicative strategies used (more shifts to L1 resources on the direct test; more paraphrasing and self correction on the semi-direct test). The discourse obtained from the two tests differed in terms of a number of features such as rhetorical functions and structures, genre expectations, communicative

properties, discourse strategies, prosodic paralinguistic features, and speech functions and discourse markers. Shohamy, Donitze-Schmidt, and Waizer (1993) examined the types of discourse obtained from five different elicitation procedures but using identical tasks. Some of the procedures required interactions with humans, others with different types of machines (telephone, video recorder, and tape-recorder). Results showed that the discourse obtained from the different procedures, in identical tasks, differed in a number of respects. The language obtained from the machine-induced procedures was more direct and did not involve pragmatic devices. The language elicited from the human interactions was more elaborated and indirect, and involved a large number of pragmatic and social devices. Differences were also revealed by various linguistic and discourse features. The discourse which was most elaborated, and also included the highest number of pragmatic devices, was that which was elicited from the telephone.

With the growing interest in conversational and discourse analysis, many recent studies have focused on these aspects of performance tests. These studies analyze the type of oral discourse elicited, the type of interaction, and the context in which the discourse is evaluated. This research brings together a number of fields such as conversational analysis, discourse analysis, cross-cultural pragmatics, and language-variation studies. Such an interdisciplinary approach will provide a better understanding of the oral language performances obtained from performance tests; it also has practical implications for designing effective means of performance assessment.

Van-Lier (1989) considered the criterion of conversational ‘naturalness’ in the oral interview. In particular, he examined issues of interview turn-taking, symmetry of topic nomination and maintenance, and characteristics of the process of interviewing which thwart the purpose of the interview as a test to simulate communication. Weiyun He (1994) reported on a study of language proficiency interviews obtained from the University of Cambridge First Certificate in English. The research focused on the discourse practice of elaborating an answer to a sequentially preceding question when different perceptions were held by the interviewer and interviewee concerning the purpose of the interview.

Lazaraton (1994) conducted a number of qualitative studies on various aspects of the Cambridge Assessment of Spoken English (the CASE). In one study, she monitored examiner conduct in the CASE test: she described and analyzed the language used by native speakers in communicative situations, the ways that native speakers accommodate situations, and the ways that native speakers accommodate their less-

than-proficient interlocutors. She also focused on the role that questions play in this accommodation process, and especially on native speakers' question-behaviors in the language proficiency interviews. She analyzed course-placement interviews in a number of tests.

The analysis of course-placement interviews by Lazaraton (1994) showed that interviewers routinely modify their question prompts in response to some perceived trouble for the nonnative interlocutors. Using conversation analysis she described a variety of these question-turn modifications as they occurred in these interview data and showed that these sample modifications differed in a variety of testing situations as well as in ordinary conversations.

Similarly, Ross (1992) examined the variations in questions posed by interviewers at key junctures in the interview process. By analyzing the discourse of full length oral proficiency interviews, he found that perceptions of oral proficiency are reflected in the extent of accommodation in interviewer questioning. The extent of accommodation provides a strong indicator in determining oral proficiency. He also noticed that, at the upper end of the proficiency continuum, where accommodation diminishes, issues of register, style, and rhetorical skill become more important; at the lower end of the continuum, where the majority of second language learners are located, the degree of interviewer accommodation can provide a useful metric for assessment.

In a more recent study, Ross (1994) analyzed twelve Foreign Service Institute interviews to examine systematic procedures-scripts for conducting the interview. Results suggest that the interviews are conducted according to a procedural script which provides an underlying structural schema for the interviewer and that formulaic speech plays a crucial role in helping the interviewer frame new topics in major portions of the interview. A further finding was that interviewers may resort to time-sharing procedures in order to dedicate some of their limited attentional resources to the task of monitoring and evaluating the interviewee's speech. Language interviews are not subject to audience-design constraints which are characteristics of ordinary conversations.

Another research topic has involved the criteria for judging the quality of the language obtained on performance tests. The question of whether to use native speaker performance on rating scales has been the subject of long debates. Most rating scales have used the native speaker as the top level of the performance (Emmett, 1985; Hughes, 1989). Yet, in a series of studies conducted by Hamilton, Topes, McNamara, and Sheridan (1993), they found great variability in the performance of native speakers

on the IELTS test. Specifically, they found that performance by native speakers was related to their educational level and work experience. Sheridan (1991) found similar results in the writing sub-test of the IELTS. They noted that the results provide support to the distinction made by McNamara (1996), between the strong and weak interpretation of the term "performance." In the stronger sense, a likely source of the results in these studies, success is measured in terms of performance on the task, not only in terms of knowledge of language. The issue of native speakers comes up also in relation to how native speakers actually evaluate oral proficiency. Barnwell (1989a) found that the ACTFL oral proficiency scales did not provide a sufficiently transparent set of criteria to allow non-specialist native raters to use them consistently. Barnwell (1989b) consequently questions the reliability of native-speaker intuition as a suitable reference point for defining the criteria for proficiency rating.

The above set of studies is not a comprehensive description of all research taking place on performance tests; yet it provides indications as to the direction in which research on performance testing is currently moving. The following section will examine a number of these directions for further research and development in the area of performance testing.

6.1.5. PA: Research Agenda

The most important future direction for research is in the development of theoretical models and constructs for performance tests. As was noted above, the construct of performance as used today is very limited and narrow; it does not refer to a range of cognitive and affective factors involved in the performance of communicative tasks such as general reasoning power, emotional states, personality factors, knowledge of the L1, etc. (McNamara, 1996). There is a need, therefore, to expand current models to incorporate aspects beyond linguistic components since performance tests combine both knowledge of the domain and linguistic knowledge. Once expanded models are developed, there is a need for them to be validated in various ways. Thus, the tasks which are developed for performance tests must be open to criticism and examination so that they do not either under-represent or over-represent given constructs. Despite the intuitive appeal of performance tests as representing 'real language use,' there is no reason to suppose that a person's performance on these tests will tell us everything we want to know about his/her proficiency in specific contexts.

There are many other unanswered questions regarding aspects of performance tests. Most performance tests use rating scales for assessing the language samples. However, there are many questions regarding the appropriacy of rating scales in performance testing. Rating scales are often constructed on the assumption of a hierarchy. Since rating scales are assumed to represent the construct of what is being assessed, there is a need to examine the validity of different types of rating scales in light of a clear definition of the construct. It is, therefore, important to validate the different scales which are commonly used to examine the validity of the hierarchies, and to examine whether certain hierarchies that exist in one domain can be transferred to another. The situation gets very complex when language interacts with domain knowledge. For example, Lumley (1993) found that bands on rating scales have a considerable degree of overlap, and he notes that "it would be surprising if one were to claim to show that one skill had to be fully acquired before the next could be mastered" (1993:49). There are also many specific issues regarding the use of rating scales for different types of performances assessed on tests. These include:

1) Definitions of 'successful' or 'acceptable' performances;
2) The choice of components to be included;
3) The extent of the analytic framework for rating scales;
4) The extent to which rating scale should emphasize the domain knowledge or the language knowledge, or both;
5) The background of the judges;
6) The desirability of using homogeneous groups of judges for performance tests, or using heterogeneous groups representing a variety of different backgrounds in terms of education, linguistic ability, professional background, educational background, etc.

Other areas that require research include the variety of types of elicitation procedures and tasks. Lumley (1993) mentions the need to know more about how far changes in text and context affect the estimated level of difficulty of different sub-skills: there is also a need to know under what circumstances particular micro-skills cluster together. Such information may be very useful for the selection of tasks of varied difficulty levels. Related to this concern is the use of different simulation devices, alternative assessment procedures, or authentic/ethnographic observation approaches in real life situations at the work place. Questions such as "What do we mean by authentic tests?" and "How authentic is authentic?" need to be examined. If, indeed, it is not possible to replicate real life in testing situations, then there is a need to determine the

validity of different simulation procedures and the extent to which they represent actual performances.

Along these lines, it is important to experiment with a variety of alternative procedures that are becoming common nowadays, for example, portfolios which include a variety of work-sample performances, computer technologies, videos, etc. Examination of the validity of these procedures and their relative contributions to the performances is also needed. Following a person around as they perform in real life situations, may seem to be a difficult alternative, but a variety of such ethnographic and qualitative approaches to assessment must be examined as well. Real life observations, as well as other types of external indicators, are commonly used nowadays in a variety of school contexts.

Information is urgently needed regarding the various issues of sampling. Questions such as the following will require further investigation:

1) How many performances are needed in order to arrive at valid conclusions?
2) How many tasks and how many different procedures are needed?
3) For what duration should performance be sampled?
4) How often should a person's language be sampled?
5) At what points in a person's career or in his/her language development, and over what period of times, should performance be sampled?

These issues need further study so that researchers can draw valid conclusions regarding the proficiency of the test taker. There is also a need to employ a variety of theories related to work performance which specify hierarchy ordering and weighting of skills in order to set up sampling frames and establish content validity (Skehan, 1984).

Although one of the most important rationales for performance tests is predictive validity, it is surprising how few studies are available in this area. Skehan (1989) claims that gains in predictive validity, although very important, are often achieved at the expense of narrowness and poor generalizability. Also, questions of this type relate to the proportion of subject-matter or domain knowledge versus language knowledge within given performance tasks and tests. In addition, studies should be directed towards the performance-testing process, the types of interactions taking place, the types of discourse produced, and the extent to which that discourse is 'test discourse' or 'a real life' discourse. Some of the qualitative studies reported above are good initiatives in that direction. Another issue of great importance regarding performance testing is that of practicality. Performance tests take considerable time to administer and are very

costly to rate and evaluate. Also, because performance tests require an initial needs analysis, they serve a more limited population than do tests of general language knowledge. In addition, they may require individualized administration and scoring, and they can be significantly more expensive than other alternatives. This is one reason why testing institutions hesitate to use performance tests on a wider basis.

Thus, one important direction in research is to explore procedures which are less costly, yet obtain valid results and can be administered to a large segment of the population. The use of semi-direct tests is an example of such work as these tests are more practical to use. In these situations, it is important to examine the cost-benefits of using such tests compared to more authentic ones. In spite of the practical limitations, some institutions consider performance testing to be a sufficiently critical issue to justify the cost. Thus, the availability of expertise and financial resources must be weighed against the importance of the decisions to be based on the test results, as well as the issue of whether significantly better decisions can be expected from a performance-test procedure (Wesche, 1992). This cost-benefit analysis is particularly important when performance tests are used for certification purposes; the representativeness of content and the reliability of the test are crucial when major decisions rest upon a single test performance.

There are also a variety of technical issues that need to be examined in performance testing: these issues include the speed of presentation of lectures, the capacity to understand large quantities of visually presented information, the ability to relate to different stimuli, the effect of cultural background, and the familiarity with the domain in which language is being tested. All of these issues are especially relevant to overseas students and immigrants (Skehan, 1989). Another issue to be studied further is the use of performance testing across different levels of proficiency. Nowadays, performance tests are generally reserved for the certification of relatively advanced learners who are nearing their goal, but it may very well be that different types of performance tests, using different types of simulation techniques, can be appropriate for different levels of proficiency. Second language acquisition theory as yet provides no principled way of assessing interlanguage abilities at early stages of acquisition in relation to the requirements of complex real-world verbal tasks. A better understanding of “foreigner talk” (systematic linguistic and interactional accommodation made by native speakers communicating with less proficient non-native speakers) may eventually provide such guidance (Wesche, 1987). This guidance would be in the form of principles by which

linguistic and cognitive demands may be purposefully constrained in a given language testing situation to allow performance assessment of lower-proficiency language learners. At present, the intuition of native-speaker testers interacting with non-native examinees is the only mechanism of this sort available. A related issue is the need to examine the composition of performance tests (Should they be all performance-based or should they use a combination of performance testing and general proficiency testing?)

The wide interest in performance testing in the general educational literature implies that much of the research in that domain will benefit language testing. There is ample work taking place today on interpreting and contextualizing results from alternative assessment and on examining performance tests with new tools in the larger educational context. Moss (1994) raises questions related to reliability and challenges the applicability of the classic psychometric paradigms to performance testing. She notes that in traditional psychometric theories, reliability has been taken for granted as a necessary but insufficient condition for validity, yet she doubts whether the classic definitions of reliability and validity are applicable to the new forms of assessment. Moss introduces the hermeneutic approach which emphasizes holistic and integrative interpretations of human phenomena. In the classic psychometric approach to assessment, each performance is scored independently by judges who have no additional knowledge about the test taker or about, the judgements of other readers. Inferences about achievement, competence, or growth are based upon composite scores, aggregated from independent observations across readers and performances, and referenced to relevant criteria or norm groups. In contrast, the hermeneutic approach involves holistic interpretations of collected performances that privilege readers who are most knowledgeable about the context in which the assessment occurs and that situate the textual and contextual evidence available in a rational debate among the interpreters. The interpretation might be warranted by criteria like a reader's extensive knowledge of the learning context, multiple and varied sources of evidence, and the ethic of disciplined, collaborative inquiry that encourages challenges and revisions to initial interpretations.

Thus, in the area of reliability, the earlier common practice was to look for agreement among judges. Yet, when it comes to judging performances which involve a variety of 'knowledge' of different kinds, there may be a need for judges who represent those different backgrounds, and they are not expected to agree with each other. Some testers have proposed a phase of discussion, or interpretation of the data, through a

dialogue to discuss the findings of the different judges. This process of including different judges who come from different backgrounds in discussions and interpretations can provide more comprehensive insights into the construct that is being measured (Delandshere and Petrosky, 1994; Moss, 1992, 1994). Moss claims that attention to reliability actually works against critical dialogue; it leads to procedures that attempt to exclude the values and contextualized knowledge of the reader, and it forecloses a dialogue among readers about the specific performances being evaluated. Moss is not excluding the notion of reliability; rather, she advocates that it be considered as one alternative that should always be justified in critical dialogue and in confrontation with other possible means of warranting knowledge claims. Such approaches need to be experimented with in the language domain.

Performance testing opens a new era of assessment where language is being assessed in context along with other skills, not in an isolated, decontextualized manner. Yet, there are still many unresolved issues that need to be examined in theory, research, and test development. No doubt, the next decade will be devoted to attempts at answering such questions using advanced statistical procedures, advanced technologies, and qualitative and quantitative analytic procedures. Performance testing will also bring together a number of related disciplines, in addition to language, and specifically those areas related to the very contexts where performance tests are to be implemented.

In addition to performance assessment, many different LSP tests have also been developed throughout the world over the past two decades. In the following sections, a number of LSP tests will be reviewed. The particular tests which are discussed were selected for inclusion in this literature review because they illustrate a variety of LSP assessment techniques and illustrate the qualities of good testing practice. In considering the ways in which various teams of testers have tackled the problems of LSP test development, this short review hopes to bring the reader to an appreciation of the art of LSP testing in all its variety and creativity. Last but not least, an account of portfolio assessment will also be presented.

6.2. LSP Tests for Admission Purposes

One category of LSP tests are those which are administered for “admission” purposes. These tests are usually used as a gateway for people who try to enter universities to pass through. However, some of these tests are administered for

admission purposes not to educational but to occupational programs. The most famous members of this class of LSP tests are the IELTS, CAEL, and OET.

6.2.1. IELTS

As Alderson and Clapham (1992) rightly point out, in the early 1960s, the British Council administered a set of procedures called the British Council Subjective Assessments (BCSA), in which local British Council officers assessed listening, speaking, reading, and writing skills according to a five-point scale. Examiners were not provided with test materials for this, but they were given some advice about the kinds of tests they might use. This test could be said to have belonged to the traditional era of language testing when it was not considered necessary to link a test to a particular theory of language acquisition, and when examiners were not overtly concerned with the consistency of results (Moller, 1981). As the 1960s progressed and language testers sought to write tests which not only reflected the structuralist language concerns of the time (Lado, 1961), but which were also reliable, the BCSA began to lose credibility and was gradually replaced by the British Council's English Proficiency Test Battery (EPTB), often known as the Davies Test.

The Davies Test contained measures of grammar, sound discrimination, and correct identification of suprasegmental features—all in discrete-point format—as well as rather more integrative tests of reading (the intrusive word test, referred to in recent publications as the "cloze elide" test (Manning, 1987), and a modified cloze test intended to test grammatical skills. For several years the test was admired. It was also considered to be successful at identifying those students whose English was good enough for educational courses in Britain. However, by the mid 1970s, when Hymes (1972) had stressed the need for a sociocultural element to be introduced into theories of language, and functional and notional syllabuses were contributing to the attempt to teach communicative competence, the Davies Test was considered to be outdated. The linguistic theory on which it was based was thought to be inadequate to account for the facts of language use.

The British Council, therefore, approached the University of Cambridge Local Examinations Syndicate (UCLES) for assistance in the design and administration of a new test, the English Language Testing Service (ELTS) test, which was introduced worldwide in 1980. This was an English for Specific Purposes (ESP) test which was based on a "needs analysis" approach. It contained tests in different subject areas, so

that students in different disciplines could take tests which were related to their own fields of study. The test gradually gained support from test users, and was for some time considered to be accurate. However, six years after its inception, the feeling that ELTS itself was based on an outmoded theory of language proficiency led to the establishment of the ELTS Revision Project, charged among other things with updating the test's construct. The International Development Program of Australian Universities and Colleges combined efforts with the University of Cambridge Local Examinations Syndicate (UCLES) and the British Council to produce a more internationally orientated test and this was launched in 1989 as the International English Language Testing System (IELTS).

6.2.1.1. The ELTS Test

When the ELTS test was first mooted in the late 1970s, the work of John Munby was becoming well-known in English language teaching circles. His work was to a large extent itself derived from or influenced by writings in applied linguistics on the nature of communicative competence, by the work of Wilkins (1976), the Council of Europe (Trim, 1973), and others on notional/functional syllabuses, and by research by sociolinguists into the nature of communication in a variety of settings, and was thus very much part of the sociolinguistic/communicative emphasis of its time. Munby's (1978) Communicative Needs Processor (CNP) in which 'the appropriate specification of communicative competence is processed from a profile of language communication needs' (Munby, 1978) rapidly became influential in syllabus design, especially in the field of English for Specific Purposes (ESP), where the need for accurate syllabuses was arguably most acute. The analysis of students' language needs in terms of the "linguistic requirements of the target situation" became commonplace in such circles, and the Munby model was frequently used as a basis for such an analysis.

Similarly in language proficiency testing, where test designers have to decide upon a syllabus for their test before they can begin to construct items, Munby's work was influential. The ELTS test was not the only proficiency test to apply the Munby model: the Associated Examining Board's Test of English for Educational Purposes (TEEP) was constructed on the basis of results from a needs analysis conducted by Weir (1983) explicitly using Munby's CNP as the framework for the design of data collection instruments. In both cases, it was argued that such a model was appropriate, since the tests were intended for specific purposes: to screen overseas students wishing to study

academic subjects in English-medium countries. In the case of the ELTS test, the specific purpose approach of the designers resulted in a battery of tests with a variety of modules in specific areas: Life Sciences, Medical Sciences, Technology, and so on. According to Carroll (1981), the new test was based upon a set of specifications which attempted directly to apply the Munby model to a definition of test purpose and test content.

In fact the Munby model had been attacked even in the 1970s. Widdowson (1977) pointed out that a definition of language needs in terms of target situations of language use was inadequate as a basis for language teaching, since a teacher—and arguably a syllabus designer—has to consider not only the goals of language teaching, but also the means by which learners are to be led to those goals. Reviews by Davies (1981), Mead (1982), and Skehan (1984) were critical of the impracticality, lack of comprehensiveness, and theoretical implausibility of the model.

In addition, however, the implementation of the Munby model in the case of the ELTS test was severely criticized by language testers. Criper (1981) and Clapham (1981) take the author of the "Test Specifications" to task for the exaggerated claims made for the validity of the specifications. In particular, they point out that no empirical data was ever gathered on the basis of which one might have justified the claims, and that the statistical justifications for the test are thus completely unfounded. Nevertheless, Clapham (1996) at least considered the Munby model to be in principle an interesting and useful way to think about test construction.

6.2.1.1.1. The ELTS Validation Study

In 1986, Criper and Davies completed a four-year validation study of the ELTS test (Criper and Davies, 1988). In this study, they point out that ELTS is 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'. (Criper and Davies, 1988:9-10)

They suggest that their factor analyses show that ELTS is in fact unifactorial on a principle components analysis. They also argue that the rotation of factors suggests a

dominant first general factor followed by a second (reading) factor and a third (listening) factor—no specific factors emerge (Criper and Davies, 1988: 11). In their summary of the results of the Edinburgh Validation Study of the ELTS, they argue that ELTS draws on the methodology of a needs analysis design albeit imperfectly and in an imbalanced way. This reveals the inadequacies of needs analysis models. ELTS, therefore, reveals the drawbacks of English for Specific Purposes (ESP) models but at the same time shows considerable success as an English for Specific Purposes (ESP) test.

A conference was held in 1986 to discuss the first draft of the Edinburgh Validation Study of the ELTS. According to Hughes, Porter, and Weir (1988), it is pointed out, in the report of that conference, that considerably more work needs to be done on defining the complexity of constructs that underlie the ELTS test in order to make possible a proper examination of the degree to which the test addresses them.

One issue that emerged during the conference was ‘what should the construct of a test like ELTS be?’ Unfortunately, no answers were forthcoming during the conference, other than a general agreement that the Munby model was considered to be less relevant as a definition of language proficiency than had been thought in the 1970s. It was also agreed that a revised ELTS test must be based upon an appropriate and acceptable view of language and language use. In 1987, the ELTS Revision Project started a three-year program to revise the ELTS test, and among other tasks was charged with identifying and defining a suitable construct for a revised ELTS test.

6.2.1.1.2. The ELTS Revision Project

The ELTS Revision Project considered what view of language proficiency ought to underlie the new test from a variety of perspectives. Perhaps the most important approach was to review the literature on testing for academic and educational purposes, both from a theoretical and practical point of view. A survey of the analyses of the linguistic needs of comparable target populations was also made. A series of meetings were held with language testers and with experienced teachers of English for Specific or Academic Purposes (EAP) at which discussions focused on what content ought to be included in a revised ELTS test. Language testing researchers were then invited to a conference at which all the data gathered during this initial Data Collection Phase was reviewed, and recommendations were solicited as to test content and format.

After this conference, teams of experienced teachers and teachers of English for Academic Purposes (EAP) were invited to devise their own specifications and sample test items for each of the test components. As a check on the content validity of these materials, the draft specifications and sample test items were shown to academic subject specialists who were invited to comment on the suitability of the texts and tasks for their students. They were also shown to English for Academic Purposes (EAP) teachers and to specialists in academic reading and writing. The feedback from all these activities produced widely varying opinions which inevitably had to be sifted, adjusted, and rejected as well as accepted.

Inevitably the revised test was to be subject to important constraints which were partly administrative and financial, but which also took account of the high face validity of the existing ELTS. The revisions, therefore, had to balance the often opposing claims of testing theory and test feasibility. The British Council insisted on the following non-negotiable constraints:

1) The revised test had to be cheaper to administer than its predecessor, and had to be shorter and less cumbersome.
2) It had to be capable of being administered overseas, with rapid reporting of results.
3) It had to maintain continuity with the existing test. Specifically there had to be a degree of subject-specific modularization, all four macro-skills had to be tested and reported in a profile scale, and the scores had to be reported on band scales equivalent to the current ELTS scales.

A good number of applied linguists were asked to help in the ELTS Revision Project. Although many of their recommendations were outside of the scope of the ELTS Revision Project, some common themes and controversies did emerge. The first of these was the issue of commonality. Some applied linguists insisted that the revised version of the ELTS should take account of what language users have in common (i.e., underlying competence), rather than what differentiates them (Alderson and Clapham, 1992). A more common theme than the notion of underlying competence was that of variability, in various disguises. This assertion, that a valid test must take account of degree of learner control is similar to Bialystock's (1982) argument, to the sense that we should test not only quantity of knowledge but also quality of knowledge. Moreover, more than half of the applied linguists who helped the ELTS Revision Project wanted candidates to be given tasks which were as similar as possible to those they would meet during their future courses. Finally, some of the applied linguists argued that

proficiency was divisible by skill, and there are thus tests of the four macro-skills: reading, writing, listening, and speaking.

6.2.1.2. The Emergence and Development of IELTS

The ELTS Revision Project resulted in the emergence of the IELTS. In 1988, Criper and Davies undertook a study of the practicality, validity, and reliability of ELTS and recommended some simplification and shortening of the test. A compromise was sought between practicality and maximum predictive power (Alderson and Clapham, 1992). The number of subject specific modules was reduced from six to four, and the Non-Academic test was replaced by the General Training Module. In addition, 'International' was taken as a prefix to acknowledge the involvement from 1989 of the International Development Program Education Australia (IDPEA), who joined the British Council and the University of Cambridge Local Examinations Syndicate (UCLES) in managing the test. One important advantage of this aspect of the test's management was that it helped to prevent any perception of Eurocentric bias and instead ensured a fully international perspective. Charge and Taylor (1997) notice that, over recent years, the employment of teams of trained test writers based in both Australia and the UK, together with the regular exchange of materials between UK and Australian Chief Examiners, has ensured that the right balance is found between what is country-specific (and hence inappropriate in an international test) and what is part of the non-country-specific culture which underlies the English language and which it is legitimate to expect candidates taking IELTS to know.

IELTS went live in 1989, with each candidate taking two general subtests, Listening and Speaking, and two specialized subtests, Reading and Writing. The general subtests tested general English while the specialized subtests (Modules A, B, and C) were intended to test skills in particular areas suited to the candidate's chosen course of study. From 1989 the number of people taking the test grew steadily, and by 1994 there were well over 35,000 candidates taking IELTS in 200 test centers around the world. After appropriate consultation and research, it was decided that further modifications to the test should be introduced in April 1995. These improvements were made in seven main areas.

6.2.1.3. Subject-Specific Subtests

Prior to April 1995 IELTS included three subject-specific subtests—Academic Modules A, B, and C (for Reading and Writing)—which were designed to meet the

needs of candidates in three broad discipline areas: physical sciences and technology, life and medical sciences, and arts and social sciences. Despite its attractiveness on face validity grounds, this subdivision of the test into three subject-specific subtests caused some administrative concern because test centers and receiving institutions were often unclear about the appropriate subtests for different courses. In addition, it was not always clear whether it would be better to match a candidate to a subject-specific subtest on the basis of their previous or their intended discipline area. The International Editing Committee, therefore, recommended that a research study be undertaken to investigate the effectiveness of a one-module approach. The results of this project, together with results from important research into second language reading and English for Specific Purposes (ESP) testing by Clapham (1993, 1995, 1996), showed that one test for all academic candidates did not discriminate for or against candidates of any discipline area. For this reason, the three subject-specific subtests in IELTS were replaced in April 1995 by one Academic Reading Module and one Academic Writing Module.

In addition, the thematic link between the Reading and Writing Modules (for Academic and General Training) was removed. A major consideration in the revision of the Reading and Writing Modules related to issues of construct validation. It was recognized that the thematic link of the original test design, although desirable in some senses, nevertheless increased the potential for confusing the assessment of reading ability with the assessment of writing ability. Research confirmed that candidates are not disadvantaged by the removal of the link between the Reading and Writing Modules. The removal of this link also makes it easier to ensure comparability of task difficulty across the different versions of each Reading or Writing Module.

6.2.1.4. General Training Banding

Until April 1995 General Training was restricted to assessment over six Bands, since the General Training Reading and Writing Modules did not have the discriminating characteristics to operate over the same range as the Academic Modules. This in itself was inconsistent since General Training candidates took the Listening and Speaking Modules which were assessed over nine Bands. General Training was mainly used in the Australian context for entry to Technical and Further Education colleges and secondary schools, and there was a perceived need to bring the General Training

Reading and Writing Modules into line with the Academic Reading and Writing Modules and score them over nine Bands. This was duly effected in April 1995.

6.2.1.5. Window for the Speaking Module

Before April 1995 several test centers were finding it difficult to allocate time and space for all candidates to take Speaking Modules on the same day as the rest of the test. One practical result of this was that the number of test dates was rising unnecessarily because of the constraints on the Speaking Module, which in turn had security implications. To address this problem, the window for the Speaking Module was increased to include two days immediately after the day of the test administration. As a result, the number of IELTS administrations can now be better controlled, and greater flexibility is ensured for test centers. Given the increasing number of IELTS candidates overall, this has proved especially beneficial for larger centers.

6.2.1.6. Fairness and Subjective Marking

It has always been important to maintain adequate reliability in both the objectively and the subjectively marked modules of IELTS. A rigorous process of test production has produced Reading and Listening versions with an average Cronbach Alpha of 00.88, calculated from the performance of over 90,000 candidates on thirteen reading and listening versions. The reliability of Speaking and Writing Modules cannot be estimated in the same way, but quality is assured through a comprehensive program of training, certification, and monitoring of examiners. Performance in the Speaking Module is recorded onto cassette and the recorded interviews, together with scripts from the Writing Module, are kept by the test center for a minimum of two months. All IELTS results are routinely checked prior to release, and automatic remarking is required where consistent profile scores across the four skill areas are identified. In addition, a formal procedure now allows candidates to query their results within one month of the results being issued.

6.2.1.7. Fairness and Data Collection

For some candidates success or failure on IELTS naturally has serious implications, so decisions based upon the test must take into account the whole context of each individual candidate and how their language ability relates to the demands of a particular course of study or training. Candidates are now given time to transfer their answers at the end of the Listening Module to an Optical Mark Reader Answer Sheet

similar to that which was already used for Reading Module. In addition, all candidates complete a Candidate Information Sheet (CIS) which also serves as the test center registration form. The Candidate Information Sheet and the Listening and Reading Answer Sheets are returned to University of Cambridge Local Examinations Syndicate (UCLES) on a regular basis for processing and validation research. Routine data collection and analysis of this sort enables test use and user to be closely monitored so that the test can continue to develop appropriately. This is particularly important for an internationally available test where there is a need to be sensitive to the broad range of cultural, social, and educational contexts in which candidates live and work. An ongoing program of IELTS-related validation research investigates the relationship between candidates demographic characteristics, background affective attitudes, cognitive and metacognitive strategies, and test performance. A validation section is included in IELTS Annual Report.

6.2.1.8. Dispatch of Live Materials

Security of test material is of course of the utmost importance, and the re-use of materials cannot be allowed to compromise test security in any way. The relationship between security and re-use of IELTS test material depends in part on the number of module versions available. Since April 1995 more test versions have been made available for centers to choose from over a six-month period. “The decoupling of Reading and Writing Modules allows test centers greater flexibility in combining the various modules for each test administration, and has been a significant step in enhancing the security aspect of the test.” (Charge and Taylor, 1997: 378). Test material generally remains in use for a maximum of one year, after which it is withdrawn from circulation.

6.2.1.9. Computerized Administration

Maintaining the service of IELTS test centers with flexible administrations and a quick turn-round of results generates an administrative burden in terms of organization and information handling. An increasing IELTS candidature would suggest an increasing administrative burden, and this needs to be managed effectively and efficiently to avoid any deterioration in the quality of service. To make this administrative process easier, the University of Cambridge Local Examinations Syndicate (UCLES), the British Council, and the IDPEA have developed a software package with which the IELTS centers are now equipped. The centers can also

download much of the information that University of Cambridge Local Examinations Syndicate (UCLES) requires for central administration and validation from the Internet.

6.2.1.10. The Current Status of IELTS

The International English Language Testing System (IELTS) (its recent version of 2000) is widely recognized as a reliable means of assessing whether candidates are ready to study or train in the medium of English. IELTS is owned by three partners, The University of Cambridge Local Examinations Syndicate, the British Council and IDP Education Australia (through its subsidiary company IELTS Australia Pty Limited).

The year-2000 version of IELTS consists of six modules. All candidates take the same listening and speaking modules. There is a choice of reading and writing modules according to whether a candidate is taking the academic or general training version of the test. The academic version of the reading and the writing modules is especially designed for candidates taking the test for entry to undergraduate or postgraduate studies or for professional reasons. The general training version of these modules, on the other hand, has been designed for candidates taking the test for entering to vocational or training programs not at degree level, for admission to secondary school, and for immigration purposes.

The test modules are taken in the following order. First, the candidates spend 30 minutes to answer the 40 test items which accompany the four sections of the listening module. Second, depending on their purpose, the candidates take either the academic or the general training reading module. The reading module consists of three sections and forty items in either version. The candidates spend 60 minutes on this module. Third, the writing module also consists of the academic and the general training versions. Each version includes two writing tasks and the candidates are expected to spend 60 minutes on these tasks. The final module is the speaking module. This consists of a conversation between the candidate and an examiner and takes between ten and fifteen minutes. There are five sections in this module. As such, the total test time is two hours and forty-five minutes.

The listening module is in four sections, each with ten questions. The first two sections are concerned with social needs. There is a conversation between two speakers and then a monologue. The final two sections are concerned with situations related to educational or training contexts. There is a conversation between up to four people and

then a monologue. A variety of question types is used, including: multiple choice, short-answer questions, sentence completion, notes/chart/table completion, labeling a diagram, classification, and matching. Candidates hear the recording only once and answer the questions as they listen. Ten minutes are allowed at the end to transfer answers to the answer sheet.

The reading module has two versions: academic and general training. In the academic version, there are three reading passages, of increasing difficulty, on topics of general interest and candidates have to answer 40 questions. The passages are taken from magazines, journals, books and newspapers. At least one text contains detailed logical argument. A variety of question types is used, including: multiple choice, short-answer questions, sentence completion, notes/chart/table completion, labeling a diagram, classification, matching lists/phrases, choosing suitable paragraph headings from a list, identification of writers' views/attitudes—yes, no, not given.

The general training reading module also consists of three sections with increasing difficulty and 40 questions. The texts in this module are taken from notices, advertisements, leaflets, newspapers, instruction manuals, books and magazines. The first section contains texts relevant to basic linguistic survival in English, with tasks mainly concerned with providing factual information. The second section focuses on the training context and involves texts of more complex language. The third section involves reading more extended texts, with a more complex structure but with the emphasis on descriptive and instructive rather than argumentative texts. A variety of question types is used, including: multiple choice, short-answer questions, sentence completion, notes/chart/table completion, labeling a diagram, classification, matching lists/phrases, choosing suitable paragraph headings from a list, identification of writers' views/attitudes—yes, no, not given, or true, false, not given.

The academic writing module consists of two writing tasks, and it is suggested that candidates spend about 20 minutes on task 1, which requires them to write at least 150 words, and 40 minutes on task 2—250 words. The assessment of task 2 carries more weight in marking than task 1. In task 1 candidates are asked to look at a diagram or table and to present the information in their own words. They are assessed on their ability to organize, present, and possibly compare data, describe the stages of a process, describe an object or event, or explain how something works. In task 2, candidates are presented with a point of view, argument or problem. They are assessed on their ability

to present a solution to the problem, present and justify an opinion, compare and contrast evidence and opinions, evaluate and challenge ideas, evidence or arguments. Candidates are also judged on their ability to write in an appropriate style.

The general training writing module also consists of two tasks. It is suggested that candidates spend about 20 minutes on task 1, which requires them to write at least 150 words, and 40 minutes on task 2—250 words. The assessment of task 2 carries more weight in marking than task 1. In task 1, candidates are asked to respond to a given problem with a letter requesting information or explaining a situation. They are assessed on their ability to engage in personal correspondence, elicit and provide general factual information, express needs, wants, likes and dislikes, express opinions, complaints, etc. in task 2, candidates are presented with a point of view, argument or problem. They are assessed on their ability to provide general factual information, outline a problem and present a solution, present and justify an opinion, evaluate and challenge ideas, evidence or arguments. Candidates are also judged on their ability to write in an appropriate style.

The speaking module has five sections. First, in the introduction section, the examiner and candidate introduce themselves and the candidate is encouraged to talk briefly about their life, home, work and interests. Second, in the extended discourse section, the candidate is encouraged to speak at length about some familiar topic of general interest or of relevance to their culture, place of living or country of origin. This will involve explanation, description or narration. Third, in the elicitation section, the candidate is given a task card with some information on it and is encouraged to take the initiative and ask questions either to elicit information or to solve a problem. Next, in the speculation and attitudes section, the candidate is encouraged to talk about their future plans and proposed course of study. Alternatively the examiner may choose to return to a topic raised earlier. Finally, in the conclusion section, the interview is concluded. Candidates are assessed on their ability to communicate effectively with native speakers of English. The assessment takes into account evidence of communicative strategies and appropriate use of grammar and vocabulary.

IELTS results are reported on a nine-band scale. In addition to the score for overall language ability, IELTS provides a score, in the form of a profile, for each of the four skills. These scores are also reported on a nine-band scale. Each Overall Band Score (OBS) corresponds to a descriptive statement which gives a summary of the English language ability of a candidate classified at that level.

6.2.1.11. Research into IELTS

Results of the International English Language Testing System (IELTS) battery trials in Australia were reported by Griffin (1990). In this report, Griffin argues that the IELTS tests of productive language skills use direct assessment strategies and subjective scoring according to detailed guidelines. The receptive skills tests use indirect assessment strategies and clerical scoring procedures. Component tests in reading, writing, listening, speaking, and grammar and vocabulary were developed by international teams for use in measuring English language competence and identifying suitable candidates for study in English-language-medium programs. The report describes the trial subject sample and test component characteristics, and presents and discusses detailed statistical results for each test item, reliability statistics, and data on inter-test correlations and interrater reliability. The grammar and vocabulary component was removed from the test, and some item deletions are noted.

The nature and development of the International English Language Testing System (IELTS) instrument were described by Ingram (1990). Ingram noticed that the test was the result of a joint Australian-British project to develop a new test for use with foreign students planning to study in English-speaking countries. It was expected, Ingram goes on, that the modular instrument would become the principal test taken by overseas students intending to study or train in Australia or Britain, and that comprehensive test administration procedures would be established in most parts of the world. According to Ingram, IELTS would assess both general proficiency and English for academic and other special purposes in one of four broad discipline areas (general training, arts and social sciences, physical sciences and technology, and life and medical sciences). An introductory section of Ingram’s paper offers background information on the test construction process, and subsequent sections describe in greater detail the test format, subtests and specifications, efforts to make the test international in nature and to limit cultural bias, and issues in test management and administration.

In another paper, Ingram (1991) claims that the International English Language Testing System (IELTS) assesses proficiency in English both generally and for special purposes of non-native English speakers studying, training, or learning English in English-speaking countries. According to Ingram, The Speaking subtest of the IELTS measures a candidate's general proficiency in speaking in everyday situations via a structured, five-phase oral interview. During the interview, the interviewer and candidate exchange introductions, the candidate speaks at length about a familiar topic,

queries the interviewer, expresses ideas and attitudes, and the interviewer concludes the session. Maximum speaking proficiency is then rated against the Band Scale. Ingram also notices that documentation for the speaking test includes training and administration manuals for interviewers, a sample test, and the Speaking Band Scale. Issues to be addressed in future research on the IELTS Speaking subtest include validity and reliability of the Speaking test, its relationship with other subtests of the IELTS and with other tests, rater consistency, the use of lay persons as interviewers, and effect of the interview situation.

Buell (1992) discusses research conducted in the spring of 1991 that measured the relationship of reading subtest scores to teacher ratings of students' reading abilities. Sixty-eight advanced-level students in an intensive English program took an institutional version of the Test of English as a Foreign Language (TOEFL) and a specimen reading module of the International English Language Testing System (IELTS) battery. The students' reading abilities were assessed by their teachers, using a scale devised by the American Council on the Teaching of Foreign Languages (ACTFL). Reliability estimates were obtained and correlations were run. The research tested the hypothesis that the results of a reading test for academic purposes, based on current theories of test design and construction, would correlate better with teacher observations than would results of a more traditional test (the TOEFL). Data analysis indicated instead that each of the tests correlated moderately well with teacher observations. Although somewhat different patterns of correlations occurred with graduate versus undergraduate students, and with natural sciences majors versus arts and social sciences majors, most differences were not statistically significant.

IELTS has been used for the validation of the instruments used in some research projects over the past few years. One example is the study by Lynch (1994). He reports on a two-part evaluation of the Test of English at Matriculation (TEAM) in use at the University of Edinburgh. TEAM has been used since 1987 to identify entering non-native speakers of English who are likely to be at risk linguistically and who should receive English language support. Separate samples of candidates' scores were used to assess: (1) TEAM's concurrent validity with other measures of English language proficiency, such as the English Language Proficiency Test Battery (EPTB) and the International English Language Testing System (IELTS); and (2) TEAM's predictive validity in relation to academic outcome. The results indicate strong correlations between TEAM and existing proficiency tests, particularly with EPTB. The findings

also suggest that TEAM performs predictively as well as other measures, with scores on the TEAM listening subtest being especially indicative.

Geranpayeh (1994) reports on a study conducted to determine if comparisons between scores on the Test of English as a Foreign Language (TOEFL) and the International English Language Testing System (IELTS) are justifiable. The test scores of 216 Iranian graduate students who took the TOEFL and IELTS, as well as the Iranian Ministry of Culture and Higher Education Test of English Proficiency (MCHE), from 1990-92 were compared. The study found high to moderate correlations between TOEFL and IELTS scores. Comparisons indicate that a score of 6 on IELTS is roughly equated with 600 on TOEFL, the minimum requirement for non-native speakers to gain admittance to most English-language graduate schools. A score of 6.5 on IELTS is roughly equated with 600 on TOEFL, the minimum requirement for non-native speakers to gain admittance to a linguistics department in most English-language graduate schools. The scores of the most proficient subjects on the two tests were found to be less comparable than the scores of less proficient subjects.

Charge and Taylor (1997) describe the nature of recent changes in the International English Language Testing System (IELTS) and provide a rationale for their introduction. They note that IELTS is widely accepted as a reliable means of assessing whether candidates are ready to study or train in the medium of English. Barett-Lennard (1997) suggests that foreign students in Australia may not be getting preparation needed to integrate successfully into university study. Barett-Lennard reviews research that indicates these students not only need help preparing for the International English Language Testing System (IELTS) but also with learning at an Australian university. Academic preparation courses that attempt to address both needs are also discussed. Wallace (1997) questions the removal of a link between reading and writing tasks in the International English Language Testing System (IELTS) examinations on two grounds: that this removal is prejudicial to those students whose native cultures may not provide the appropriate schemata to effectively write; and that it is unrealistic in terms of the measurement of study/linguistic skills required in university study.

6.2.2. CAEL

The Carleton academic English Language (CAEL) Assessment is a criterion-referenced test which is part of a collaborative testing system developed at Carleton University, Ottawa, Canada. It combines the logic of performance testing with the ethics

of gradual admission. It is logical that students should be tested on language in use for academic purposes if that is precisely the way in which they intend to use language. It is ethical to recognize degrees of proficiency, and allow students to begin study in their academic field on a limited basis, based on their degree of proficiency in English, with the support of one or more English for Academic Purposes (EAP) courses. The CAEL testing system may provide a model for other universities, colleges, and institutions that wish to generate positive washback effects from language testing for purposes of admission to academic programs and link language testing with learning.

Initially, the testing committee decided that the Ontario Test of English as a Second Language (OTESL) (Wesche and Cray, 1984), produced by the Province of Ontario, might provide the answer. It was a carefully constructed, extensively examined performance test, and a number of the contributing members to OTESL worked at Carleton University. OTESL provided an example of direct testing of academic language proficiency. "Testing is said to be 'direct' when it requires the candidate to perform precisely the skill which we wish to measure" (Hughes, 1989). Obviously, because candidates know they are in a test situation, the tasks cannot be considered completely authentic, but every attempt was made in OTESL to simulate the academic contexts in which students would use their language.

There were, however, only two versions of OTESL. If language proficiency testing was to be a permanent and on-going activity of the University, it was essential that multiple, parallel versions be available for test security. Of equal importance, having involved faculty from all disciplines, the test committee concluded that the development of an in-house test(s) would generate continued cooperation. One of the hallmarks of the CAEL model is its collaborative, generative potential for positive washback: the linking of SSL student learning, EAP classrooms and the academy.

Thus, early in 1988, a group of EAP teachers in the Center for Applied Language Studies began to work with professors in the faculties of Science, Engineering, Social Science and Arts, collecting information about actual language performance requirements in introductory first-year classes. A number of language activities were consistent across all faculties. For example, students need to be able to take information from lectures. Although students exhibit a multitude of different note-taking strategies during lectures, when asked by professors to fill in diagrams, flow charts, or label items on a map, they exhibit very similar responses. Such consistent responses also emerge when professors provide their own definitions of key terms, and when professors signal

important information with discourse markers such as, "ok ... and these are really the three key-points …" or "so the main thing I want you to remember is ...." It was possible, in this way to identify and catalogue the actual language performance requirements of students, across faculties at the university, and to record lectures in progress. A selective sample of students from the lectures were also asked to bring their notes to follow-up small group discussion sessions and photocopies were made of actual student notes and analyzed to identify how students use their language skills to accomplish the tasks required in lectures.

As a result of this initial analysis, and using OTESL as a model, several versions of the CAEL Assessment were developed for pilot testing by September 1988. These versions consisted of actual lecture segments taken from introductory lectures at the first-year level from the faculties of Science (biology) and Social Science (geography). The topics of the lectures were general (food, weather), but the information provided by the professor was specific to that professor's particular emphasis in the course. In all cases the professor's agenda was to introduce to his/her students information about a general topic, which the professor assumed the students knew nothing about. Students were asked to fill in diagrams, flow charts, etc., which were actually part of the source lecture. They were asked to read about the same topic, and answer questions about the reading in much the same way as the students in the first-year class had been asked to do. The reading and the lecture provided a "scaffold" (Bruner, 1986) for the students' written response, which was the natural outcome (as it had been in the source classes) of the information provided.

These versions of the CAEL Assessment were pilot tested with students in a number of different settings: in a local high school, in ESL pre-university level classes, in credit ESL classes, and in the context of Introductory Psychology classes with fully admitted first-year students who represented the university as a whole, both in their range of ability and in their faculty of study. Today, CAEL Assessment is an important example of LSP testing used for purposes of admission.

6.2.3. OET

The Occupational English Test (OET) was originally administered by the National Office for Overseas Skills Recognition of the Australian government Department of Employment, Education, and Training. In 1991, administrative responsibility for the OET was transferred to the Language Testing Research Center of the National Language and Literacy Institute of Australia (NLLIA), at the University of Melbourne

(McNamara, 1990a, 1996). The OET was developed to assess the English language ability of immigrant health professionals seeking admission to Australian training and licensure examinations, and is available throughout Australia and at some 50 Australian embassies and consulates overseas. Eleven medical professions are represented in the various forms of the test: dentists, dieticians, nurses, occupational therapists, pharmacists, physicians, physiotherapists, podiatrists, radiographers, speech pathologists, and veterinarians. In 1994, of approximately 1200 candidates, about one-third were physicians, the largest single group. The test consists of four components: profession specific writing and speaking sub-tests, and reading and listening sub-tests, both of which are non-profession specific. The test taker's performance is assessed by means of a rating scale. The scale ranges from 1, representing the lowest level of proficiency, to 6, representing native or near-native proficiency.

This test has the potential for eliciting quite a rich language performance in terms of LSP ability. Only textual knowledge is not included in either the input or the scoring criteria. Grammatical knowledge is fairly traditionally represented in the scoring grid, with intelligibility referring to accent or pronunciation, and resources of grammar and expression referring to knowledge of morphology, syntax, and vocabulary forms (McNamara, 1996). Functional knowledge is explicitly called for in the prompt card, where the test taker is instructed to talk to the patient about the advantages and disadvantages of the three pieces of equipment: this task thus combines ideational and heuristic functions. Moreover, there is implicit in the interaction a requirement for the manipulative function in that the test taker must recommend a particular piece of equipment and persuade the patient to accept it. It is also implicit in the task that standard Australian English is the dialect that must be comprehended by the test taker, including cultural references, idioms, and figures of speech, and that the candidate should produce language that a person unfamiliar with medical terminology could understand. For example, it was felt by the informants consulted by the test developers that a problem for overseas trained medical professionals was a lack of 'ordinary colloquial language to refer for example to “bowel movements” or of not recognizing that for the patient "stomach" might mean a quite different and far more extensive anatomical area than an anatomy textbook might suggest' (McNamara 1996: 192). This notion is intended to be captured in the evaluation category appropriateness of language.

The construct of LSP ability also includes appropriate background knowledge to accomplish the test task, although this knowledge is not assessed directly in the OET. It

is clear that the test takers will have to know something about wheelchairs, walking frames, and sticks in order to talk to the patient about their relative advantages and disadvantages in the particular case at hand. Furthermore, the candidates will need to communicate information about various forms of therapeutic physical exercises in order to convince the patient that they are not as pointless as he or she apparently thinks. Finally, the test taker must use strategic competence to engage an appropriate physical therapy discourse domain to mediate between his or her language knowledge and background knowledge in executing a complex performance, interacting with the interlocutor on-line to achieve a communicative goal, or recommending a particular line of treatment and convincing the client to follow it.

6.2.4. UETESOL

The Northern Examinations and Assessment Board, a consortium of British universities in the north of England, introduced the University Entrance Test in English for Speakers of Other Languages (UETESOL) in 1990. UETESOL is a revised version of the Joint Matriculation Board's University Entrance Test in English, which had been introduced in 1966. The objective of the UETESOL is to assess the English language skills (listening, reading, writing, and speaking) of candidates, whose first language is not English, for admission to British universities. The test is intended to measure language skills considered to be common to the fields of science, engineering, social sciences, and business studies. The UETESOL contains five sections: writing, editing, reading, speaking, and listening.

6.3. Tests for Certification Purposes

A second category of LSP tests includes the tests that are normally used for “certification as teachers” purposes. In spite of their recency, the tests that fall within this category have been able to gain the interest of many LSP test developers. Some of the most significant members of this class of LSP tests are the TEACH, the ELSA, and the Italian proficiency test for language teachers.

6.3.1. Proficiency Test for Language Teachers: Italian

The Proficiency Test for Language Teachers was introduced in 1993, having been developed by staff at the National Language and Literacy Institute of Australia, Language Testing Research Center, at the University of Melbourne. The test has two main functions. (1) It serves as a benchmark for teacher education by making the language requirements of the foreign language teacher explicit. (2) It is used to certify

language teachers by helping determine whether those applying for employment in primary schools are proficient enough in the language to perform their duties as language teachers (Elder, 1993b). The Proficiency Test for Language Teachers: Italian consists of five sub-tests: listening, text editing, reading, writing, and speaking.

6.3.2. TEACH

The second LSP test of the language of teaching the Taped Evaluation of Assistants' Classroom Handling (TEACH). The TEACH is a performance test developed in 1985 at Iowa State University, in the United States, to provide evidence of the oral English proficiency of prospective teaching assistants in a classroom in their own field of study. The term teaching assistant, in the US university context, refers to postgraduate students who have been offered employment as instructors in undergraduate classes as a way of financing their studies. In many scientific and technical disciplines in US universities, a significant proportion of teaching assistants are international students whose first language is other than English. The TEACH is used in addition to the Speaking Proficiency English Assessment Kit (SPEAK), a general purpose test of oral ability (Educational Testing Service, 1986). The entire test consists of three parts: (1) meet the class, (2) introduce the topic, and (3) answer students' questions.

6.3.3. ELSA

The English Language Skills Assessment (ELSA) was developed by the New South Wales Department of School Education and the Adult Migrant English Service in 1991-92. It was meant for the purpose of assessing the English language ability of applicants for employment as teachers who had qualified outside Australia and who came from a non-English-speaking background (McDowell, 1995). The developers of ELSA decided to treat reading and writing as integrated, rather than independent, skills and to test them together in a single sub-test. The texts were taken from genuine school documents, newsletters to parents, newspaper articles on education, and academic papers. There are seven sections in the reading and writing paper, each containing different tasks. The reading tasks include multiple-choice comprehension items, short-answer questions, matching tasks, and error correction tasks. There are two writing tasks: a short-answer, limited production task and a longer direct writing task. Both are based on written input.

6.4. Tests for Vocational Purposes

Some LSP tests are used outside academic settings. They are normally used in occupational settings that do not require an academic degree. A waiter, for instance,

should know how to use appropriate language while waiting on the guests in a restaurant. Recently some LSP tests have been developed for this and similar vocational purposes. The most important class of non-academic LSP tests is the one that includes tests normally used for vocational purposes.

6.4.1. CEIBT

The Certificate in English for International Business and Trade (CEIBT) was introduced in 1990 by the University of Cambridge Local Examinations Syndicate. It is intended to allow candidates whose first language is not English to demonstrate an ability to function efficiently in an office or business where English is used. There are three sub-tests: listening, oral interaction, and reading and writing. The tests are all set in the context of one international company, and the prompts and input data are all thematically related to that company.

6.4.2. Japanese Language Tests for Tour Guides

This is another of the tests produced by the NLLIA Language Testing Research Center at the University of Melbourne, and has a dual purpose: to indicate to employers the language proficiency of applicants for positions as Japanese-speaking tour guides, and to provide a selection criterion for applicants to tour guide training courses. Thus, prospective test takers are of two types: those with some experience and/or training as tour guides, and those with no training or experience (Brown, 1995). The test was produced in consultation with experienced tour guides, and the assessment criteria were based on their judgements about the necessary features of quality tour guide communication as well as more linguistically oriented features (Brown, 1993). Raters are drawn as far as possible from within the tour guide industry but teachers of Japanese language are also trained as raters. Moreover, both native speakers and non-native speakers of Japanese are used as raters. A study of the effect of rater background (Brown, 1995) found no significant differences in ratings produced by tour guide versus non-tour guide raters or between native versus non-native speakers of Japanese, given that the raters were adequately trained. Ratings are made on a six-point scale.

6.4.3. PELA

The Proficiency test in English Language for Air traffic controllers (PELA) is a test designed to measure the English language proficiency of trainee air traffic controllers in Europe. The PELA was developed through the European Organization for the Safety of Air Navigation (EUROCONTROL) Institute of Air Navigation Services (IANS) in

Luxembourg, between January 1992 and December 1993. It has the potential of being used to measure the specific purpose English language abilities of student air traffic controllers in the 38 countries of the European Civil Aviation Conference (Institute of Air Navigation Services, 1994). This is a high stakes test to define and assess a “criterial level in ATC English for newly qualifying ATC trainees” (Teasdale, 1993: 143). The language of air traffic control is an extremely well defined field specific domain, owing to the fact that there are established conventions for communication, published by the International Civil Aviation Organization (ICAO), Rules of the air and air traffic services (International Civil Aviation Organization, 1985). This document “defines the content, form, and ordering of elements of utterances, as well as specifying the circumstances in which specific phrases are to be used” (Teasdale 1996: 2). In the PELA, the problems of only a single aircraft at a time are considered in the test tasks, a situation somewhat problematic for candidates in pretesting, according to Teasdale. Nevertheless, because of practical considerations, this constraint remains a part of the test and simply has to be taken into account when performances are interpreted. The PELA consists of a listening sub-test with a number of written limited production tasks, and a speaking sub-test with two production tasks.

6.4.4. CBLC

The Royal Society of Arts Certificate in Business Language Competence (CBLC) is intended to assess communicative skills within a business context in languages other than English, including French, German, Italian, and Spanish. There are five levels of certification: Basic, Survival, Threshold, Operational, and Advanced. At each level there are five or six sections, called elements, covering such activities as listening, reading, telephoning, conversing, making presentations, and writing. There are a number of tasks in each section, and the input materials are what the test developers call 'simulated authentic' (Royal Society of Arts Examinations Board, 1994a: 47). Indeed the materials do have a visually authentic look about them. Results in each section are rated either pass or fail and candidates must pass all sections to receive a full certificate.

6.4.5. TOEIC

Test of English for International Communication (TOEIC) was developed in 1979 by Educational Testing Service (ETS) in the US in response to requests from clients for a standardized test of English listening and reading in the context of international business (Educational Testing Service, 1996). The test is now administered by an

independent commercial subsidiary of ETS, the Chauncey Group. TOEIC is given World-Wide, although the majority of test takers and score users are located in Asia, particularly Japan and Korea, and it is taken by over a million candidates a year. It consists of 100 multiple-choice listening comprehension items and 100 multiple-choice reading comprehension items. The reading sub-test has three sections: sentence completion, error recognition, and comprehension of short texts. Candidates are given 75 minutes to complete all three parts.

The producers of TOEIC say that the test is intended to measure the everyday skills of people working in an international environment. The scores indicate how well people can communicate with others in business, commerce, and industry. The test "does not require specialized knowledge or vocabulary" (Educational Testing Service, 1996: 1). The scores are based on the number correct out of 100 items, and are converted to a scale from 5 to 495 points.

6.4.6. OIBEC

The Oxford International Business English Certificate (OIBEC) was developed by the University of Oxford Delegacy of Local Examinations (1990). It is aimed at business men and women working in international commerce who wish to obtain a certificate of competence in English language skills for purposes of promotion or changing employment (University of Oxford Delegacy of Local Examinations, 1990). The test is given at two levels: First Level, a basic qualification, and Executive Level, an advanced qualification. Available in some 44 countries world-wide, the test takes about two and a quarter hours, comprising a 20-minute listening component, a 20-minute speaking component, and a one hour and 35-minute reading and writing component. A significant feature of the OIBEC is a case study booklet, which provides extensive information about a problem that forms the context of the test. The candidates are given three days to study this information, which consists of narrative, tables, letters, memos, and other printed input, and may take it and any notes they might make about it into the examination room, along with a dictionary.

6.5. Other LSP Tests

In addition to the above tests, a number of other LSP tests, with limited scope, also exist. Business English Performance Test (BEPT) was developed at the English Language Center at Drexel University, in Philadelphia. It was designed to test the English skills needed by international students in a Master's degree program in business

administration to participate in cooperative work experiences in OS workplaces. The CEELT (Cambridge examination in English for language teachers) is another LSP test introduced in 1987 by the University of Cambridge Local Examinations Syndicate (UCLES) to provide a means of certifying the English language competence of teachers of English whose first language is other than English. Other examples of LSP tests include Diploma in Public Service Interpreting (DPSI), Listening Summary Translation Examination (LSTE), etc. (Scott, *et al*., 1996).

The LSP tests that were discussed here comprise only a limited fraction of the rich repertoire of LSP tests available. For purposes of brevity, however, I refrain from discussing the remaining LSP tests. The most recent development in the field of LSP testing is that of portfolio assessment discussed below.

6.6. Portfolio Assessment

Another aspect of the assessment of specific purpose writing (and other skills as well) is that of portfolio assessment. The use of portfolios, particularly in the assessment of reading, writing, and speaking, is a growing trend and certainly has potential applications in LSP assessment. Portfolio assessment is one means of alternative assessment and refers to the "purposeful, selective collection of learner work and reflective self-assessment that is used to document progress and achievement over time with regard to specific criteria" (Kohonen, 1997: 15). Already, portfolio assessment is used in many specific purpose areas, for example in mathematics (Asturias, 1994), chemistry (Phelps, 1997), physics (Slater, 1994), teacher training (Dubetz, *et al.*, 1997), and English for academic purposes (Spath, Hirschmann, and Traversa, 1997).

Collaboration between LSP practitioners and instructors in the specific purpose content areas is a productive approach to the assessment of LSP development in a specific discipline. The same portfolio of work could be used for the dual purposes of assessing learning in the specific purpose field and progress in the acquisition of field specific language ability. Criteria for assessing LSP use could be developed in cooperation with the field specific content instructor and thus reflect the indigenous criteria established within that discipline. The most usual practice in portfolio assessment involves the learner in preparing his or her own portfolio, sometimes in collaboration with the instructor, sometimes not, placing in it examples of various types of language performances, including drafts and revisions as well as finished products (Spath, Hirschmann, and Traversa, 1997). A speaking portfolio would contain taped samples of a learner's presentations and interactions, as well as any documentation

associated with the spoken performances. In vocational training and internship programs, a portfolio of samples of actual work undertaken might be used to document progress and readiness for employment or promotion. As Cohen (1994) points out, a portfolio produces, in effect, a set of multiple measures of writing, providing a depth and breadth of coverage not usually possible with conventional tests. Moya and O'Malley (1994) suggest five positive characteristics of portfolio assessment procedures: (a) comprehensive, (b) predetermined and systematic, (c) informative, (d) tailored, and (e) authentic.

Kohonen (1997: 14) compares standardized testing with alternative assessment practices along ten dimensions and concludes that alternative assessment, including portfolio assessment, "entails a movement towards a culture of evaluation in the service of learning." Kohonen sees the portfolio as an interface between learning and evaluation. Perceived advantages of portfolio assessment include the potential for a more comprehensive, process-oriented assessment of long-term progress in writing, and learners becoming more self-critical and reflective about their own work (Savitch and Serling, 1997), and more publicly accountable for their own progress (Herter, 1991).

Portfolio assessment is not exempt from a concern for the qualities of good testing practice. Problems with portfolio assessment include the fact that there is very little research evidence to support the claims made by its proponents and the acknowledgement that there are difficulties in maintaining consistency in assessments across individuals and over time (Madaus, *et al.*, 1997; Brown and Hudson, 1998). Indeed, Hamp-Lyons and Condon (1993) report that assessors need training to standardize assessment criteria just as raters do in more traditional testing. Hamp-Lyons (1993) points out in this regard that grades assigned to portfolios tend to cluster close together and thus lose much discriminatory value in showing differences between more and less proficient writers (cited in Cohen, 1994). Demonstrating the validity of interpretations of portfolio assessment is also a problem: the criteria used to determine validity must reflect the holistic nature of language development, must be sensitive to individual student differences, and must accurately reflect student progress (Moya and O'Malley, 1994; Brown and Hudson, 1998). Another problem is that portfolio assessment can be very time-consuming for assessors. In addition, given the time involved, portfolio assessment samples a less representative portion of performances than can a form of assessment that elicits a greater variety of tasks (Madaus *et al.*, 1997). In spite of these problems, however, portfolio remains a popular form of

alternative assessment, and, for certain purposes, the advantages appear to outweigh the disadvantages. As such, portfolio assessment techniques offer a powerful means of LSP assessment that can have the additional benefit of empowering learners and second language users (Pollari, 1997). It must, however, be remembered that portfolio, and other alternative means of assessment, must reflect the qualities of good testing practice in the same way that other assessment instruments must (Brown and Hudson, 1998; Weigle, 2001).

7. Research into LSP Testing

Since 1981 there has been some response to Alderson's plea for more research into LSP testing, and there have been several studies into the effect of background knowledge on EAP test performance.

Three articles by Alderson and Urquhart (1983, 1985a, and 1985b) aroused considerable interest and led to several follow-up studies. These articles described three studies carried out with students attending English classes in Britain in preparation for going to British universities. In each, Alderson and Urquhart compared students' scores on reading texts related to their own field of study with those on texts in other subject areas. In the third study, three groups of students in different disciplines (Business and Economics, Science and Engineering, and Liberal Arts) took the Social Studies and Technology Modules of the ELTS test. The students' scores on the modules were somewhat contradictory. On the one hand, for example, Science and Engineering students taking the Technology module did better than the Business and Economics students who took the same test, and as well as the Liberal Arts students, although their language proficiency was lower. On the other hand, the Business and Economics students did no better than the Science and Engineering group on the Social Studies module. Since the authors used pairwise tests to assess the results, they were not able to test for the interaction between students and tests. However, they concluded that background knowledge had some effect on test scores, but that this was not consistent, and that future studies should take account of linguistic proficiency and other factors as well.

Koh (1985) had somewhat similar results with three groups of students (two in Science and one in Business Studies at Singapore University). Using analysis of variance to estimate the effect of background knowledge on cloze test results, she found that there was an interaction between student group and test, but that students did not always do best in their own subject areas. The Business students, for example, had their

highest scores on the Science text. However, it turned out that half these students had studied Science previously, so it could be that prior knowledge was affecting their scores. The group with the highest language proficiency (one of the Science groups) did consistently better than the other two on all the texts, which were on Business, History, Politics and Science. She concluded that prior knowledge did affect test scores but that ignorance of the subject matter could be compensated for by high linguistic proficiency.

Shoham, Peretz and Vorhaus (1987) rejected the use of cloze in studies of the role of background effect, and used comprehension, referent and vocabulary-in-context questions for their study at Ben Gurion University. They used three-way analysis of variance to analyze their results but once again the results were inconclusive. While students in the Biological and Physical Sciences did better at the scientific texts, the Humanities and Social Science students did not do better on the test in their own subject area. (Peretz and Shoham, 1990 had similar results.) The authors' explanation for this was that the texts were only indirectly related to the students' specialized fields of study, and suggested that this might support Lipson's suggestion (1984) that 'a totally unfamiliar text is often easier to comprehend than a text with a partially familiar content' (Shoham, Peretz, and Vorhaus, 1987: 86). This contention of Lipson's is indeed radical. If it were supported by further research, there would be an almost unassailable reason for dropping ESP testing from university proficiency tests, since until each student has his or her own tailor-made test, ESP tests will have to be focused on fairly general subject areas, which will inevitably be only partially familiar to post-graduates in many highly specialized fields. If Lipson's idea were taken to its logical conclusion, of course, proficiency tests would have to contain material outside any candidate's experience. The Joint Matriculation Board (JMB) University Test in English for Speakers of Other Languages follows just such an approach, with passages in esoteric subjects such as silver markings and heraldic devices. Item writers have difficulty finding suitable texts and the ensuing materials are often excessively dull.

Hale (1988) commented on the inconclusive results of some of the above studies, and on the small sample sizes of some of them. For his study, he looked at all candidates taking the Test of English as a Foreign Language (TOEFL) over four administrations to see whether there was a more consistent interaction between students' major field area and text content with the larger sample sizes. The reading passages were all aimed at the general reader but were based on a wide range of topics in the arts and sciences. The numbers of candidates in the four sessions ranged from

approximately six to ten thousands. For the purposes of the main study, subjects were divided into two groups (Humanities and Social Sciences in one, and Biological and Physical Sciences in the other). Hale used analysis of variance and found that for three of the four test forms the effect of subject area was significant at .001. Students' reading performance was affected by a combination of their major-field area and the nature of the passages, but the effect was not large. Hale says this was possibly because the texts were taken from general sources rather than from subject specific textbooks. Hale's reason for using such large sample sizes was to provide a greater opportunity for any statistically significant effects to be detected. Using large sample sizes certainly has this effect, and once the number of subjects becomes really large almost anything can appear significant. It seems, therefore, that Hale's huge sample sizes may militate against his significant results being as informative as he might have hoped.

A question related to that of the effect of prior knowledge on test results is the question of whether tests in students' own subject areas are better predictors than more general ones. Tan (1990) used regression analysis to see whether familiarity with test content or level of language proficiency was the best predictor of ability in reading comprehension. Undergraduates at the University of Malaya were given 'prior knowledge tests compiled by their own subject teachers, along with discipline-related cloze reading tests and a form of a 'general' proficiency test, the English Proficiency Test Battery (EPTB). In all subject areas under study (Medicine, Law, and Economics) she found that comprehension of a discipline-related text could be predicted by both knowledge of the subject area and by language level, but that language level was the better predictor.

Researchers vary in the amount of effect which they think readers' background knowledge has on comprehension. Clarke and Silberstein (1977) say that readers bring more to the text than writers because of their "formidable amount of information and ideas, attitudes and beliefs" (1977: 49), but Eskey (1988) considers that this effect is exaggerated. He accepts that successful reading depends on relating received information to prior knowledge, but says that such reading cannot be carried out without successful decoding, and that at least some understanding of the text can take place without the activation of related schemata.

Whatever the truth may be, many researchers have studied the effect of prior knowledge on comprehension, some comparing the performance of students from

different cultural or educational backgrounds, and others comparing students with different amounts of knowledge of a topic.

In a study into the effect of background culture on interpretation, Tannen (1979) showed young American and Greek women a six minute film about a boy stealing pears and then asked them to re-tell the story. In their recalls, the subjects showed that there were many ways in which they organized and changed the content of the film to fit their expectations, and Tannen found that many of these seemed to be culturally determined. For example, many of the Americans showed awareness of the requirements of media productions, and commented on the film's technique, strange color and unusual sound effects. None of the Greeks commented on the technical side of the film, or criticized it in any way; they concentrated on interpreting the meaning of the story.

Steffensen, Joag-Dev, and Anderson (1979) asked Indians and Americans to read and then recall two letters, one describing an Indian wedding, and the other an American one. The subjects read the letter relating to their own culture faster, and produced appropriate elaboration to the story, that is they added facts which were not in the original letter, but which made sense. When they recalled the other passage they misinterpreted it, adding culturally based distortions. The authors concluded that cultural schemata showed a pervasive influence on comprehension and memory.

Anderson, *et al.*, (1977a) presented students in weight-lifting or music classes with two texts, each of which could be read in two ways. The first could be understood as a wrestling match or an escape from prison, and the second as a game of cards or a woodwind rehearsal. The weight-lifting students took it for granted that the first text concerned a wrestling match and that the second one was about a game of cards. The music students thought the first text described someone escaping from prison and the second one was about a music rehearsal. Although either reading of the texts led to some anomalies, most of the students managed to fit these into their understanding of the texts, and a startling 80% never realized that there might be different interpretations.

In a rather different approach, Spilich, *et al.*, (1979) chose students who had a high (HK) or a low (LK) level of knowledge about a topic, baseball, and gave them a very technical account of a baseball match. The HK students recalled the passage more coherently than the LK ones, and scored higher on a set of questions about the passage. They also produced more elaboration of the input and gave graphic accounts of the game. The LK subjects gave very short accounts which were often out of order, and included many irrelevancies. From the results the authors assumed that the subjects

matched input with their knowledge structures, and that because the HK subjects had more knowledge structures related to baseball they were able to process the information more readily.

Vesonder (1979) had similar results in a study based on Kintsch's (1974) theory of text processing. He presented science and non-science students with two scientific texts, and found that the science students recalled the texts better than the non-scientists, and were more accurate at recognizing statements which had appeared in the texts. Vesonder found that his Kintsch-based grammar, which did not take background knowledge into account, was useful as a structural processing mechanism, but was unable to explain the differences between the two groups of students.

Freebody and Anderson (1983) gave students explanations of a familiar and an unfamiliar game. The texts were carefully matched to contain similar topics, and the grammar and the vocabulary were almost identical. There were easy and difficult versions of the texts. In the difficult versions one third of the content words were exchanged with less common ones. The students were asked to summarize the texts, recall them freely and answer sentence recognition questions. Topic familiarity accounted for three times as much of the score variance as vocabulary difficulty, and so the researchers assumed that prior knowledge was more important than text difficulty.

Symons and Pressley (1993) asked groups of students at different stages in a psychology course to search through a textbook for material that would be relevant to a given topic. The students were given ten low-inference questions which could be answered by reference to explicit statements in the text. Students did progressively better at the task as their course progressed and they became more familiar with the subject matter. They did not improve in the same way when they were given similar tasks related to earth sciences. The authors presumed firstly that, as Pichert and Anderson (1977) say, the incoming information fills slots into an activated schema and is processed more easily than is information that does not fit the schema, and secondly, that attention is directed towards information which is considered relevant and important. They also found that information which is consistent with prior knowledge has a storage advantage over unfamiliar information during encoding and is thus more likely to be recalled.

One educational researcher who, without benefit of schema theory, had a strong influence on research into the effect of prior knowledge on reading comprehension, was Ausubel (1960, 1963). He described how learners drew on previously acquired concepts

when they were trying to understand new information (Ausubel, 1963). He introduced 'advance organizer' which is now widely used in education. Ausubel argued that by providing students with advance organizers, teachers could help them to assimilate new information better. In order to test whether an advance organizer helped comprehension, Ausubel (1960) chose students from eight different academic disciplines and gave them a multiple-choice test on the contents of a passage on metallurgy. Five days before this test, half the students had been given an introductory text on the composition of metals, and the other half a text on the history of the processing of iron and steel. Ausubel's expectation was that the text on the composition of metals would work as an advance organizer and that students who had read it would score higher in the metallurgy test than those who had only read the historical passage. From Ausubel's point of view the results of the study were disappointing as the students who read the introductory script did little better than the other group. However, the results are interesting because they appear to have been partly affected by the students' major field of study, that is, their prior subject knowledge. In a later experiment, in which the text was on a subject unknown to any of the students (endocrinology), there was again no significant difference between the experimental and control groups overall, but there was a significant difference ($p = .01$) among those students who had the lowest marks in a verbal ability score. Ausubel concluded that students at a higher level of language ability could spontaneously organize new material, whereas those at a lower level could not.

One point which has not yet been mentioned is the fact that readers, even in their first language, have different levels of reading ability, so that some can loosely be described as 'good' and some as 'bad' readers. Perfetti and Lesgold (1977) say that three of the sources of these individual differences are variation in people's speed of verbal coding, differences in short term memory capacity, and varying sensitivity to discourse structures. Interestingly, different theorists have produced diametrically opposed views on the effect of level of reading ability on students' dependence on prior knowledge. For example, Perfetti and Lesgold (1977) argue that whereas good readers can take advantage of background knowledge and context, readers who are slow at decoding symbols and words overburden their short-term memory and cannot call up the appropriate schemata. However, Stanovich (1991) says that the quicker word recognition skills of the better reader are not due to superior context skills, since in many reaction-time studies it has been shown that poor readers often use context more

than good ones. Stanovich thinks that as word recognition efficiency improves, the effects of background knowledge and context dependency decrease.

Jensen and Hansen (1995) used multiple regression analysis to compare the effects of prior knowledge and listening ability on university students' performance on academic listening tasks. They collected data over six administrations of a listening test based on an academic lecture. At each administration the students listened to a lecture on either a technical subject (for example, chemistry or biology) or a non-technical one (for example history or social anthropology). The choice of lecture depended on students' prior knowledge, which was assumed to be appropriate if they said 'yes' to a question asking whether they had studied the topic of the lecture before. A different pair of tests was used for each administration. Jensen and Hansen found that prior knowledge had a significant effect on test scores on all but one of the technical passages, but on only one of the non-technical ones, and that in all cases the students' level of listening ability had a stronger effect than prior knowledge. There was no interaction between level of listening ability and the effect of prior knowledge. We do not know, however, how subject-specific the lectures were, although, since they came from introductions to courses, they may have been fairly general in content. Nor do we know how much variation there was in the students' levels of listening ability: if the spread was only small, we would not expect the effect of background knowledge to differ between high and low scoring students. It is interesting that all but one of the significant subject effects related to the technical tests, and this suggests that science students are better able to cope with non-science texts than non-scientists are with scientific ones.

Kattan (1990: 3) wanted to see whether it was worth giving university students at Bethlehem University ESP proficiency tests based on the subject in which they were majoring, or "whether a measure of a more generalized competence would do just as well." She compared the predictive validity of a 'neutral' reading test. with that of two subject specific tests (one for students majoring in English and one for those majoring in Nursing Studies). She correlated students' scores on both the neutral test and their own subject test with their grade point average (GPA) over a period of eighteen months, and found a correlation of 00.71 (n=54) between the results of the test for English majors and their GPA. The correlation between the students' neutral test scores and their GPA was only 00.36. The nurses' correlations (based on only 19 students) were, on the other hand, not significant. The size of the correlation between the English

students' subject specific test and their GPA is surprising, since the GPA is presumably based at least as much on subject knowledge as on linguistic proficiency. However, the test contained only sixteen items and its Cronbach alpha reliability index was 00.54. Since low reliability reduces the potential of a test to correlate with other measures, this suggests that the predictive validity of this English test was either exceptionally good, or that the English major at Bethlehem University places a stronger emphasis on English proficiency than do other majors, and that the students' GPA is partially, therefore, a reflection of their ability to use English. As such, the English GPAs, therefore, might not be similar in composition to those of other subject majors.

It would be interesting to see whether a repeat study produced comparable results. Another way of assessing the effect of subject area on test performance is to use bias analysis to see whether test items discriminate against students who are not familiar with the subject area of a text. O'Neill, Steffen, and Broch (1994) used Differential Item Function (DIF) to see whether the proportion of correct items among students taking TOEFL reading tests was higher when the content of the texts was based on 'home-field' rather than 'non-home-field' subject matter. The researchers used the Educational Testing Service (ETS) DIP program, which is based on the Mantel-Haenszel statistical technique, and compares the odds of two groups answering an item correctly when ability levels are taken into account. The results agreed with those of several of the other studies reported above: students in biological and physical science performed better than the other students on the science-based texts, but humanities majors did no better than the scientists on humanities-based subjects. However, it must be remembered that all TOEFL reading passages are designed to be appropriate to all students, regardless of their field of study, and so a strong subject area effect would not be expected. Moreover, any differences might be too subtle to be detected by bias analysis. The fact, therefore, that, in spite of this, two groups of students did perform significantly better in their own subject area is interesting.

Several points emerge from the above studies. Firstly, language proficiency levels seem to play at least as important a role as background knowledge in the comprehension of reading texts. Secondly, background knowledge itself is not easily assessed: a student who is in Business Studies may well have previously worked in another discipline such as Science, or may have scientific interests in his or her spare time. Thirdly, although the above studies were in many ways inconclusive, there did seem to be a tendency for science students to perform better than other students at science-based tests, but to

perform as well as the humanities students on humanities based ones. Finally, the level of specificity of the subject-based texts probably varied widely in the different studies, but this was not fully taken into account in the studies.

In the most thorough study of the relationship between SP background knowledge and language ability to date. Clapham (1996) studied performance on the reading modules of the International English Language Testing System (IELTS). Although Clapham's study focused on reading tests, her results have important implications for all types of tests in which background knowledge is intended to play a role in test performance and in the interpretation of scores.

In her study, Clapham made a number of important findings. First, students achieved significantly higher scores on the reading sub-test in their own subject area than on the sub-test outside it. However, in a pilot study (Clapham, 1993), using a different set of passages, no significant differences were found. This suggests that the passages in her two studies varied significantly in their degree of specificity, but that if passages were sufficiently specific, test takers did better at tests in their own subject area.

Second, Clapham found that there was no significant subject area effect for the undergraduate students in her study, but there was a subject effect for the postgraduates in her test population. This finding indicates that the test takers' level of SP background knowledge may have had an effect on their language performance. Additionally, Clapham found that test takers with scores of less than 60% on a grammar sub-test did not appear to profit from their background knowledge: there were no significant subject area effects. Students with grammar scores above 60% did show highly significant subject area effects. Thus, it appears that level of language knowledge, and specifically structural knowledge, had an influence on the effect of background knowledge on test performance. At the same time, however, there was no steady increase in the effect of background knowledge as students' level of proficiency rose. Rather, there seemed to be a threshold below which students were not able to make use of this knowledge and above which they were.

Third, Clapham found that when scores were analyzed on the reading sub-tests containing texts of widely varying specificity, language proficiency accounted for 44% of the variance while the addition of background knowledge variables added only 1%. This finding suggests that what really counted in the test takers' performance was their language ability and not their background knowledge. This result agrees with that of

Tan (1990), who found that comprehension of a discipline-related text could be predicted both by knowledge of the subject area and by language level, but that language level was the better predictor. Clapham went a step further, however: she removed the least field specific texts from her data and carried out her analysis a second time. When the less specific texts were excluded, the contribution of language proficiency, though still strong, was less marked: 26% of the variance was due to language ability; adding background knowledge raised the figure to 38%. It thus seems likely that as the modules become more subject specific, background knowledge will have a proportionately stronger effect on test scores.

Finally, Clapham looked at the performance of those test takers who scored very highly on the grammar module. She found, in addition to the threshold between low and intermediate proficiency readers, that readers above 80% on the grammar test were less affected by subject area than were the intermediate readers. Thus, it seemed that readers with a high level of language competence were so proficient that they could compensate for a certain lack of background knowledge by making fuller use of their language resources.

Much more research is needed before we can understand clearly the relationship between language knowledge and SP background knowledge, but Clapham's study suggests that, at least for intermediate level students, background knowledge did make a difference in their reading test performance. One might argue that, if general language knowledge were high enough, as indicated, for example, by a test of general grammatical knowledge, then SP testing would be unnecessary since language knowledge would compensate for a lack of background knowledge. However, it seems likely, given Clapham's finding regarding the stronger effect on test performance of the more field specific texts, that highly specific texts would have a significant background knowledge effect even among the most highly proficient test takers, but we certainly need more research directed specifically at this question.

An important question for LSP test development is, given that the specificity of texts varies so widely, what are the factors that contribute to it? Clapham found that the amount of field specific vocabulary did not affect the degree of specificity of the text so much as whether the vocabulary was explained or not. She also found that the source of the text did not seem to matter as much as the rhetorical functions of various sections of that text. The more academic a passage was, the more highly specific its subject matter tended to be, as it was aimed at more specialized audiences. However, it is not clear that

there is any agreement about what academic means. Furthermore, Clapham found that cohesive devices, such as referring to a ship variously as the vessel, the craft, or she, in the more highly field specific texts tended to be lexical rather than explanatory, making this aspect of comprehension more difficult for the less nautical readers. The specificity of a text was also likely to depend on the extent to which comprehension of the text required knowledge of subject specific concepts which were not explained in the text. It thus appears that the amount of context-embedded information in a text or prompt affected field specificity.

There is clearly a problem here for LSP testers: texts, and even parts of texts, vary greatly in their specificity and this variation is not necessarily obvious to test developers. In fact, Lewkowicz (1997) found that it was not easy for either native speakers or non-native speakers of English to distinguish real-life texts from testing texts. Therefore, the LSP test writer must employ what Selinker (1979) has called "Subject Specialist Informant" (SSI) techniques, which entail the involvement of practitioners from the SP field in question to work with the test development team on the selection and use of appropriate texts. The immediate lesson is one that echoes advice given by Davidson and Lynch (1993) that test development is best carried out as a collaborative effort involving a wide range of people for whom the test is important. In the case of LSP testing, as Rea-Dickins (1987) argues, this range includes experts in the vocational, technical, professional, and academic fields that are the TLU domains of interest.

8. Final Remarks

Notwithstanding the fact that LSP testing is still in its early days, it has attracted the attention of many researchers. More and more universities are offering LSP courses. The upsurge of interest in LSP teaching and testing has produced a noticeable and rich literature. Many LSP practitioners are publishing books devoted entirely to the subject of LSP testing and teaching. It is, therefore, justifiable to assume that the literature reviewed in this chapter does not show the whole picture of LSP testing. Rather, it only opens a new window onto the field.

It is noteworthy here that, due to practical limitations, the literature reviewed in this chapter has a limited scope. It only includes the most salient instances from among the rich repertoire available. For more information about any of the aspects of LSP teaching and testing, interested readers are invited to refer to the sources listed in the 'reference' section of this study.

CHAPTER THREE
METHODOLOGY

1. Introduction

We now come to the methodology that functioned as the pedestal for the present study. This chapter will describe the steps that were taken by the investigator in the process of carrying out the study. Specific sections will be devoted to a discussion of the subjects, instruments, and procedures that were employed in this investigation.

2. Subjects

The population from which the subjects of the present study were drawn included junior and senior students majoring in electronics at three Iranian universities: University of Shiraz, Shahid Bahonar University of Kerman, and Azad University of Bushehr. These students took the sample version of the IELTS (University of Cambridge Local Examinations Syndicate, 2000). They were then classified into four proficiency groups: proficient, fairly proficient, semi-proficient, and non-proficient. Table III.1 shows the frequency analysis of these proficiency levels.

Table III.1: *frequency of subjects in each proficiency level*

	Frequency	Valid Percent	Cumulative Percent
Proficient	93.00	17.20	17.20
Fairly proficient	186.00	34.40	51.60
Semi proficient	164.00	30.30	81.90
non proficient	98.00	18.10	100.00
Total	541.00	100.0	

The mean and the standard deviation of the IELTS were used as the criterion for the classification of subjects. Subjects who had scored higher than ‘mean-plus-one’

standard deviation were assigned to the top proficiency level (i.e., the proficient group). A total of 93 subjects were found to be the members of this group. 17.2% of the subjects were proficient. Through the same procedure, subjects who stood within the 'mean-plus-one' standard deviation range were assigned to the second proficiency level or the fairly proficient group. This group included 186 members and accounted for 34.4% of the sample for the present study. The third group, the semi-proficient class, included 164 members, 30.3% of the sample. This group included the subjects whose scores on the IELTS fell within the mean-minus-one standard deviation range. The last group which included 98 subjects (i.e., 18.1% of the sample) was called the non-proficient group. The subjects who had scored below the mean-minus-one standard deviation range were assigned to this group.

The justification for such a classfication is three-fold. First, this strategy has been adopted by a large number of scholars throughout the world (See Clapham, 1996). In addition, when the number of subjects in a group is bigger than the minimum number required for Z-distribution (not smaller than 60), the differences in subject frequency across different subject groups stop being statistically significant. Finally, SPSS automatically rotates the data and meets the homogeniety assumption when performing data analyses (See Bryman and Cramer, 1999).

One of the independent variables of the present study was text familiarity. A definition of text familiarity has already been provided in the first chapter (See, chapter 1, section 11). In order to obtain dependable data, the subjects of the study were expected to be totally familiar with the content of the texts that appeared in the Electronics Module of the Task-Based Reading Test (TBRT-EM), partially familiar with the content of those that appeared in the General Module of the test (TBRT-GM), and totally unfamiliar with the passages that appeared in the Accounting Module of the test (TBRT-AM) (For more information about the Task-Based Reading Test (TBRT) and its different modules (EM, GM, and AM), see the sections on 'instruments' and 'procedures' below). Therefore, some of the subjects (37 people) who reported unqualified on the basis of the 'text familiarity' criterion were discarded from the study.

3. Instruments

Three different instruments were used in the present study: (1) The sample version of the IELTS (UCLES, 2000), (2) the Self-report Questionnaire developed by the

investigator, and (3) the TBRT (AM, EM, and GM Modules) developed by the investigator. A description of these instruments appears below.

3.1. The IELTS

One of the steps of the present study was to assess the subjects' level of proficiency. The investigator had to decide whether the subjects belonged in the Limited English Proficient (LEP) group or the Non-Limited English Proficient (Non-LEP) group. A further problem was that the subjects' "reading comprehension" ability was in the focus of the study. In other words, the job of the investigator was not only to identify the subjects' level of proficiency but to do so on the basis of their reading comprehension ability. Moreover, practical considerations (like the subjects' unwillingness to cooperate) made it even more urgent to hit the two goals with one shot. It was, therefore, decided that the IELTS (UCLES, 2000) be administered. In this connection, it is noteworthy that the IELTS was "the most suitable instrument due to its 'modularity' claims" (See Chapter 2, section 6.2.1). In other words, the assumption behind the IELTS is that the 'link between the reading and writing modules has been lifted' since 1995 (See chapter two, section 6.2.1.3). Appendix A presents the IELTS General Training Reading Module used in this study.

In addition to its importance in the classification of the subjects of the study into different proficiency levels, the IELTS was also used for the validation of the main instrument of the present research, the TBRT. The correlation between each module within the TBRT test and the IELTS was used as the validity index of that module (See section 4.3. below).

3.2. Task-Based Reading Test (TBRT)

The major instrument used in the present study was a Task-Based Reading Test (TBRT) with three modules: (a) the electronics module (TBRT-EM), (b) the accounting module (TBRT-AM), and (c) the general module (TBRT-GM). Each module consisted of 40 items that measured subjects' performance on five reading tasks as delineated in chapter one: true-false, sentence-completion, outlining, writer's-view, and skimming. Each module consisted of five passages of varying lengths, textual complexity, and readability indexes. However, the texts that appeared in the different modules were chosen in such a way as to ensure maximum correspondence to the IELTS General Training Reading Module (UCLES, 2000) in terms of such textual features as

readability, structural complexity, etc. Table III.2 presents the readability statistics for the IELTS General Training Reading module.

Table III.2: *Readability statistics for the IELTS General Training Reading Module*

PROPERTIES		PASSAGES				
		1	2	3	4	5
Counts	Words	155	237	379	442	826
	Characters	795	1244	1867	2286	3930
	Paragraphs	5	11	7	8	7
	Sentences	7	16	18	24	36
Averages	Sentence per paragraph	1.4	1.2	3.6	3	5.1
	Words per sentence	21	13.5	20.6	17.8	22.8
	Characters per word	5	5	4.7	5.1	4.6
Readability	Passive sentences	28%	6%	0%	0%	19%
	Flesch reading ease	37.6	53.4	50	44.8	49.4
	Flesch-Kincaid grade level	12	9.1	11.4	11.4	11.1

In addition to readability analysis, nine university instructors who are experienced teachers of ESP courses at the University of Shiraz, Shahid Bahonar University of Kerman, and Azad University of Bushehr were asked to judge whether the texts were of the suitable level of difficulty for the prospective subjects.

The texts that appeared in the TBRT-EM (See Appendix B) were all taken from the content areas that junior and senior university students majoring in electronics had already studied as part of their academic courses. They included five topics: (a) magnetic flux, (b) vacuum tube diodes, (c) bridge circuits, (d) incandescent lamps, and (e) digital and analog computers. Since the subjects of the present study were all majoring in electronics, the passages within this module would be totally familiar for them. Table III.3 presents the readability statistics for the TBRT-EM module.

Table III.3: *Readability statistics for the TBRT-EM module*

PROPERTIES		PASSAGES				
		1	2	3	4	5
Counts	Words	155	237	379	443	826
	Characters	796	1197	1836	2290	3931
	Paragraphs	2	2	5	8	7
	Sentences	9	17	18	25	42
Averages	Sentence per paragraph	4.5	8.5	3.6	3.1	6
	Words per sentence	17.2	13.9	21	17.7	19.6
	Characters per word	5	4.9	4.7	5	4.6
Readability	Passive sentences	22%	5%	0%	0%	19%
	Flesch reading ease	37.6	53.4	50.1	44.8	49.4
	Flesch-Kincaid grade level	12	9.2	11.4	11.4	11.1

In the same vein, the TBRT-AM module (See Appendix D) included five texts. These texts were selected from the materials that were part of the academic courses of university students majoring in accounting. They included the following five topics: (a) chain stores, (b) interest, (c) clearinghouses, (d) assets and liabilities, and (e) corporate finance. It is noteworthy that, since the subjects of the present study were all majoring in electronics, the texts within the TBRT-AM module were judged totally unfamiliar for

them. As was mentioned earlier (See section 2 above), those subjects who reported somehow familiar with the texts in the TBRT-AM module were discarded from the data. Table III.4 presents the readability statistics for the TBRT-AM module.

Table III.4: *Readability statistics for the TBRT-AM module*

PROPERTIES		PASSAGES				
		1	2	3	4	5
Counts	Words	155	238	379	442	826
	Characters	926	1160	1967	2275	4131
	Paragraphs	2	2	5	8	7
	Sentences	9	16	17	23	43
Averages	Sentence per paragraph	4.5	8	3.4	2.8	6.1
	Words per sentence	17.2	14.8	22.2	19.2	19.2
	Characters per word	5.8	4.7	5.1	5	4.9
Readability	Passive sentences	22%	6%	0%	0%	20%
	Flesch reading ease	37.6	53.4	50.1	44.7	49.4
	Flesch-Kincaid grade level	12	9.4	11.4	11.4	11.1

The same procedures were used in the selection of the passages that appeared in the TBRT-GM module (See Appendix C). Unlike the two other modules, the texts within this module were expected to contain propositional content with which the subjects of the present study reported partially familiar. Five passages were selected from the *Encyclopedia Encarta* computer package. These texts included such general-digest topics as (a) natural hazards, (b) national parks and sanctuaries, (c) the sensory system of sharks, (d) classification of airplanes, and (e) mission to moon. Table III.5 presents the readability statistics for the TBRT-GM module.

Table III.5: *Readability statistics for the TBRT-GM module*

PROPERTIES		PASSAGES				
		1	2	3	4	5
Counts	Words	155	237	379	442	826
	Characters	827	1287	1927	2431	4023
	Paragraphs	2	2	5	8	7
	Sentences	9	17	18	25	44
Averages	Sentence per paragraph	4.5	8.5	3.6	3.1	6.2
	Words per sentence	17.2	13.9	21	17.6	18.7
	Characters per word	5.2	5.2	4.9	5.3	4.7
Readability	Passive sentences	22%	5%	0%	0%	18%
	Flesch reading ease	37.6	53.4	50.1	44.7	49.4
	Flesch-Kincaid grade level	12	9.2	11.4	11.4	11

After the selection of the texts, each TBRT module was developed in such a way as to resemble the IELTS General Training Reading Module.

3.3. Self-report Questionnaire

As it was mentioned earlier, text familiarity was one of the "independent variables" of the present study. The investigator had to determine whether the subjects had any prior familiarity with the propositional content of the texts that appeared in the different modules (AM, EM, and GM) of the TBRT developed for purposes of data collection.

To this end, two steps were taken: (a) administration of a Self-report Questionnaire through which the subjects indicated their degree of text familiarity with each text, and (b) selection of the texts for inclusion in the TBRT on the basis of "text familiarity cline." The investigator, therefore, developed and administered the Self-report Questionnaire to determine subjects' distribution over the text familiarity cline. This questionnaire was a Likert Scale composed of 20 items (four similar items for each text in each module) through which the subjects indicated their degree of text familiarity with the five passages that appeared in each of the TBRT modules. To ensure maximum understanding, the questionnaire was written in the subjects' native language—Farsi. Appendix E presents the Self-report Questionnaire.

4. Procedures

The IELTS was used as the pedestal for item construction for the TBRT. The content analysis of the 40 items that appeared in the sample version of the IELTS General Training Reading Module (UCLES, 2000) was at the heart of the process of item construction for the TBRT modules. It was decided that each module within the TBRT should include no more than 40 items, the same number of items as appeared in the IELTS General Training Reading Module. Moreover, the items were supposed to measure the performance of the subjects on five different tasks. The first group of items that measured subjects' performance on true-false tasks included twelve items. Each item was followed by three answers: true, false, and not given. The subjects were expected to read the corresponding passages and to decide whether the propositions expressed in the true-false items were given in the passage or not, and if yes, to make their own choice whether the items were true or false.

The second set of items in each module aimed at measuring the subjects' performance on sentence completion tasks. This set included eight items. The items in this set were eight open-ended sentences which could be completed in two ways. Following this set of items was a list of possible endings. The subjects' job was to read the corresponding passage and, on the basis of the information present in the passage, to choose two possible endings from the list to complete each item.

A third group of items measured the subjects' performance on outlining tasks. This category included six items. The subjects were expected to read a passage. Each paragraph within the passage was labeled with a letter from the English alphabet. The subjects were expected to choose from among a list of summaries the one that best

represented the propositions expressed in each paragraph. They would then match the summary for each paragraph with the label that signified that paragraph.

Subjects' performance on the task of 'identifying the writer's view' was also measured. Five multiple-choice items followed a passage in each module. Each item had three choices: yes, no, not given. The subjects were expected to read the passage and to decide whether the propositions expressed in these five items were given in the passage or not, and if yes, whether they represented the views of the writer of the passage or not.

The last set of items measured the subjects' performance on skimming tasks. The nine items of this category asked the subjects to skim the reading passage for two types of information: dates and proper nouns. The former included five items while the latter included four items. The subjects' job was to skim the reading passage and to identify the date or the proper noun that was questioned in the item.

In order to ensure maximum correspondence between the modules, the rubrics and instructions that were used in the three modules were exactly the same. The general formats of the three modules (spacing, margins, use of italics, etc.) were also the same. Maximum care was also taken, through the pretesting process, to ensure that the item facility and discrimination indexes of the items did not exhibit any meaningful difference across the different modules.

4.1. Pilot Administration of TBRT Pre-test Modules

In order to determine whether the items that appeared in the different modules of the TBRT were effective, malfunctioning or non-functioning, it was vital that the modules be administered in a pre-test administration. Since the purpose of this process was to screen the items so that the most suitable ones would be included in the final version of the TBRT, more items were included in the pre-test version of the TBRT. The investigator included 80 items in each module, twice as many as were necessary for the final version of the TBRT.

The test was then administered to a group of 36 university students majoring in electronics at the Azad University of Bushehr. All of the subjects took the IELTS and the TBRT-GM pre-test module first. Then the subjects were randomly assigned to two half-groups. The first half-group took the TBRT-EM pre-test module followed by the TBRT-AM pre-test module while the second half-group took the TBRT-AM pre-test

module followed by the TBRT-EM pre-test module. This procedure was necessary to control any practice effect which might introduce error variance into the results.

The results of the pre-test administration of the pre-test version of the TBRT modules were then input to item analyses procedures. After item analysis, from among the 80 items that appeared in each trial module, the 40 items that had the best item facility and item discrimination indexes (within the 0.3 to 0.9 range, and resembling those of the IELTS) were chosen for inclusion in the final version of the corresponding module of the TBRT. It is noteworthy that more than 40 items were found to be effective. However, since the investigator had to include only 40 items in each module, the same number of items as appeared in the IELTS General Training Reading Module, some effective items were also discareded along with the malfunctioning and non-functioning items.

4.2. Trial Administration of the TBRT Modules

After the development of the final version of the TBRT, in order to determine whether the TBRT reading modules were suitable for data collection, the modules were evaluated through a trial administration. The modules, along with the IELTS General Training Reading module (UCLES, 2000) were, therefore, administered to a group of 20 senior university students majoring in electronics at Azad University of Bushehr. All these students took the IELTS General Training Reading and the TBRT-GM modules in one administration session, and the TBRT-EM and TBRT-AM modules in another session.

A two week time interval was allowed between the two administrations. Moreover, to control the practice effect, a counter-balanced design was used in each administration. That is, ten subjects were randomly assigned to the first-half and the ten remaining subjects to the second-half groups. In the first session, the first-half group took the IELTS module first followed by the TBRT-GM module whereas the second-half group took the TBRT-GM module first followed by the IELTS module. In the second administration, the first-half group took the TBRT-AM followed by the TBRT-EM modules while the second-half group took the TBRT-EM followed by the TBRT-AM modules.

The investigator then sought to determine the reliability and the validity indexes of the TBRT modules. The SPSS and MINITAB statistics packages were employed for these analyses. For purposes of measuring the concurrent validity of the TBRT

modules, the correlation coefficients between the different modules of the TBRT and the IELTS General Training Reading Module were calculated. The smallest correlation coefficient, found between the TBRT-AM module and the IELTS, was 00.87. The reliability index of each module within the TBRT was also calculated. Here again, the smallest reliability index belonged to the TBRT-AM module (i.e., 00.89, Cronbach Alpha).

4.3. Final Administration of the TBRT Modules

For purposes of data collection, the final version of the TBRT was administered to a total of 578 junior and senior university students majoring in electronics who took the IELTS General Training Reading Module, TBRT-AM, TBRT-GM, and TBRT-EM over a four-week period. These subjects took the tests in three different universities: University of Shiraz, Shahid Bahonar University of Kerman, and Azad University of Bushehr. The procedure for the final administration of the tests was similar to that of the pilot administration. Here again, for purposes of minimizing the practice effect, a counter-balanced design was used for test administration (See section 4.1. above).

In addition to these tests, the subjects also answered the items that appeared in the Self-report Questionnaire. The assumption behind the TBRT-AM module was that the subjects should not be familiar with the content of the passages within this module. Moreover, they were expected to be only partially familiar with the content of the passages within the TBRT-GM module, and totally familiar with the content of the texts in the TBRT-EM module. On the basis of their responses to the Self-report Questionnaire, 37 of these subjects were discarded from the data since they did not meet the text-familiarity assumption of the study. The remaining 541 subjects were used as the sample for the present study.

The results, after analysis, revealed that the modules had satisfactory reliability and validity indexes. Tables III.6 and III.7 report the validity and the reliability indexes of the IELTS General Training Reading Module, the TBRT-GM module, the TBRT-AM module, and the TBRT-EM module respectively.

Table III.6: *Validity analysis of the different modules*

CORRELATION MATRIX				
	IELTS	TBRT-GM	TBRT-AM	TBRT-EM
IELTS	01.0000			
TBRT-GM	00.9397	01.0000		
TBRT-AM	00.9188	00.9252	01.0000	
TBRT-EM	00.9477	00.9447	00.9017	01.0000

Table III.7: *Reliability analysis of the different modules*

RELIABILITY COEFFICIENTS				
	IELTS	TBRT-GM	TBRT-AM	TBRT-EM
Number of cases	541	541	541	541
Number of Items	40	40	40	40
Alpha	00.8355	00.8628	00.8527	00.8527

5. Final Remarks

It was possible for the investigator to administer the sample of the IELTS General Training Reading module (UCLES, 2000) as it appeared in its original format published by the University of Cambridge Local Examinations Syndicate. However, for purposes of gaining maximum face validity, the IELTS General Training Reading module was also typed using the same typeset, format, and fonts as were used in the TBRT modules. In addition, the rubrics and instructions of the TBRT modules were exactly the same as those of the IELTS module (See appendixes A, B, C, and D).

CHAPTER FOUR
RESULTS AND DISCUSSION

1. Introduction

As chapters 1 and 2 showed, evidence concerning the effect of subject area on Language for Specific Purposes (LSP) test performance is mixed (See, for example, Alderson and Urquhart, 1985b; Koh, 1985; and Lipson, 1984). It does not always seem to be the case that students achieve higher reading comprehension scores on tests that are based on familiar subject matter. One of the major aims of the present study was to examine whether students taking LSP tests were at an advantage if they took the reading module in their own academic field of study. A second major aim of the study was to determine whether students' proficiency level affected their performance of LSP test in any meaningful way. This chapter reports the empirical findings of the study and provides answers to the questions that lie at the heart of the investigation (See chapter 1 section 10).

2. Analysis and Results

Two major types of analyses were performed on the data: descriptive analyses, and inferential analyses. The results of data analysis can be classified under three major headings: (a) frequency analysis, (b) test performance analysis, and (c) task performance analysis. The following sections will provide an in-depth account of the findings of the study. While section 2.1. provides the results of the application of descritive statistics (i.e., frequency analysis), section 2.2. reports the findings of the application of inferential statistics (e.g., ANOVA, Regression Analysis, etc.).

2.1. Frequency Analysis

The first set of analyses included the frequency analysis of the subjects' test and task performance. In this set, two types of analyses were conducted: (a) frequency analysis of subjects' test performance on the IELTS General Training Reading module, TBRT-GM module, TBRT-AM module, and TBRT-EM module, and (b) frequency analysis of subjects' task performance of true-false, sentence-completion, outlining, writer's-view, and skimming tasks.

Table IV.1 shows the frequency analysis of subjects' test performance. The subjects had their lowest mean score on TBRT-AM (15.67), and their highest mean score on TBRT-EM (24.71). This indicates that the subjects were more able to do the test within their own subject area than the test outside their area of expertise. Since the standard deviations of the tests were roughly the same, the variety of subjects' test performances can be related to other variables. Figure 1 represents the mean plot for subjects' test performance.

Table IV.1: *Frequency analysis of subjects' test performance*

DESCRIPTIVE		IELTS	TBRT-GM	TBRT-AM	TBRT-EM
N	Valid	541	541	541	541
	Missing	0	0	0	0
Mean		20.9741	20.9020	15.6765	24.7116
Median		21.0000	22.0000	16.0000	25.0000
Mode		14.00	13.00	8.00	16.00
SD		7.1216	7.7225	7.3940	7.3185
Variance		50.7178	59.6367	54.6711	53.5611
Minimum		7.00	7.00	2.00	10.00
Maximum		37.00	38.00	34.00	39.00

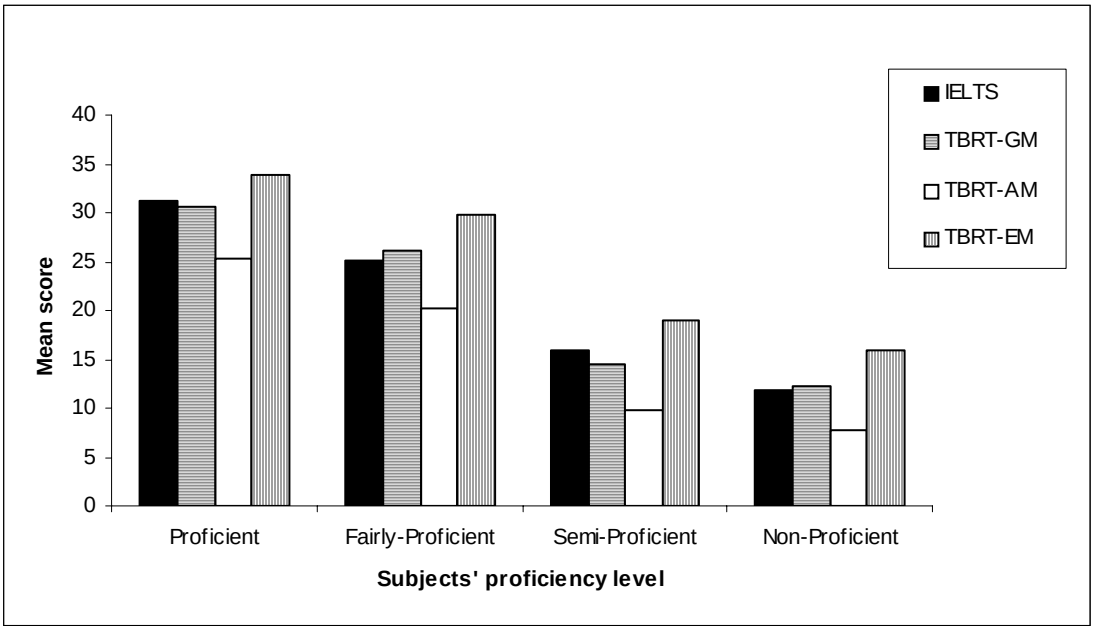

Figure 1: *Mean plot for subjects' test performance*

Table IV.2 is the tabular representation of the frequency analysis of subjects' performance on the true-false task. Here again, the greatest mean belonged to TBRT-EM (7.81) and the smallest mean to TBRT-AM (4.92). The subjects' mean scores on the IELTS and the TBRT-GM were very close (6.89 and 6.87, respectively). Figure 2 shows the mean plot for subjects' performances on the true-false task.

Table IV.2: *Frequency analysis of subjects' true-false task performance*

DESCRIPTIVE		IELTS	TBRT-GM	TBRT-AM	TBRT-EM
N	Valid	541	541	541	541
	Missing	0	0	0	0
Mean		6.8965	6.8780	4.9205	7.8152
Median		7.0000	7.0000	5.0000	8.0000
Mode		6.00	9.00	3.00	10.00
SD		2.6957	2.9596	2.7980	2.7847
Variance		7.2670	8.7592	7.8289	7.7547
Minimum		1.00	1.00	.00	1.00
Maximum		12.00	12.00	12.00	12.00

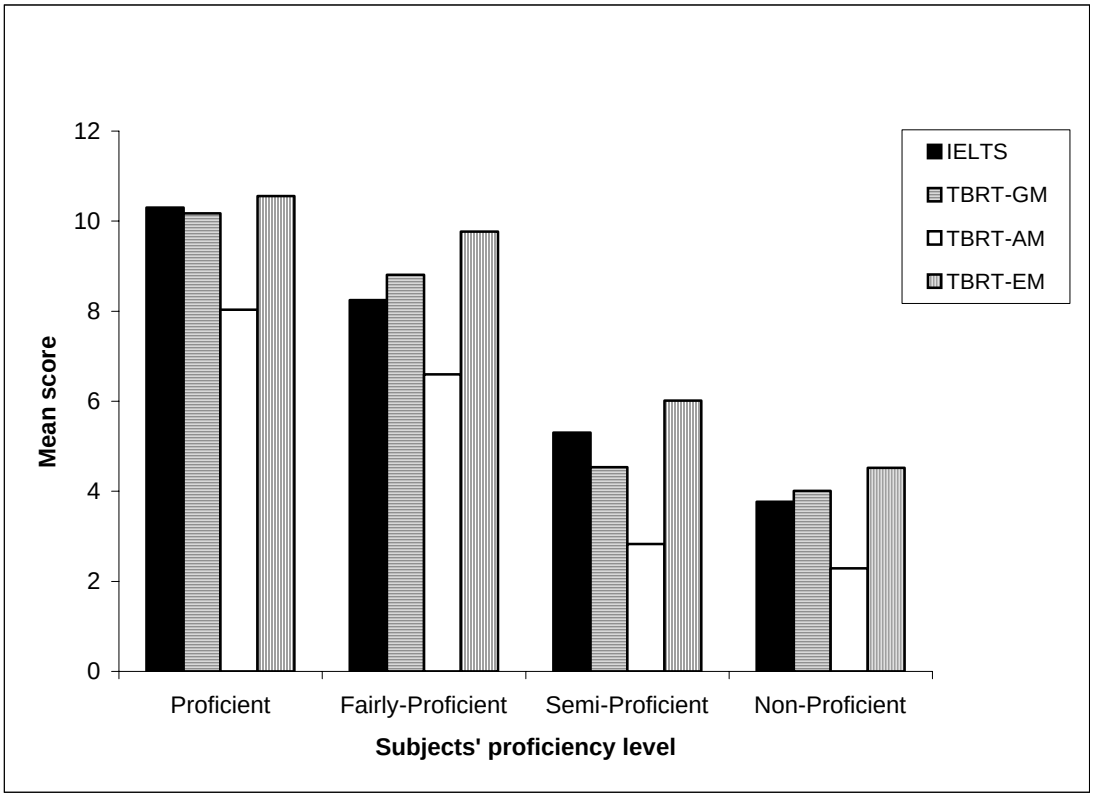

Figure 2. *Mean plot for subjects' true-false task performance*

Table IV.3 represents the frequency analysis for subjects' performance on the sentence-completion task. The mean for the IELTS and the TBRT-GM were somewhat close (2.28 and 2.38, respectively). However, once more, the greatest mean score belonged to subjects' performances on the TBRT-EM (3.32) and the smallest one to the

TBRT-AM (1.79). Figure 3 represents the mean plot for subjects' sentence-completion task performance.

Table IV.3: *Frequency analysis of subjects' sentence completion task performance*

DESCRIPTIVE		IELTS	TBRT-GM	TBRT-AM	TBRT-EM
N	Valid	541	541	541	541
	Missing	0	0	0	0
Mean		2.2847	2.3882	1.7911	3.3290
Median		2.0000	2.0000	1.0000	3.0000
Mode		1.00	1.00	1.00	3.00
SD		1.6137	1.7408	1.5925	1.8516
Variance		2.6040	3.0305	2.5359	3.4286
Minimum		.00	.00	.00	.00
Maximum		7.00	8.00	7.00	8.00

Note: Multiple modes exist; the smallest value is shown

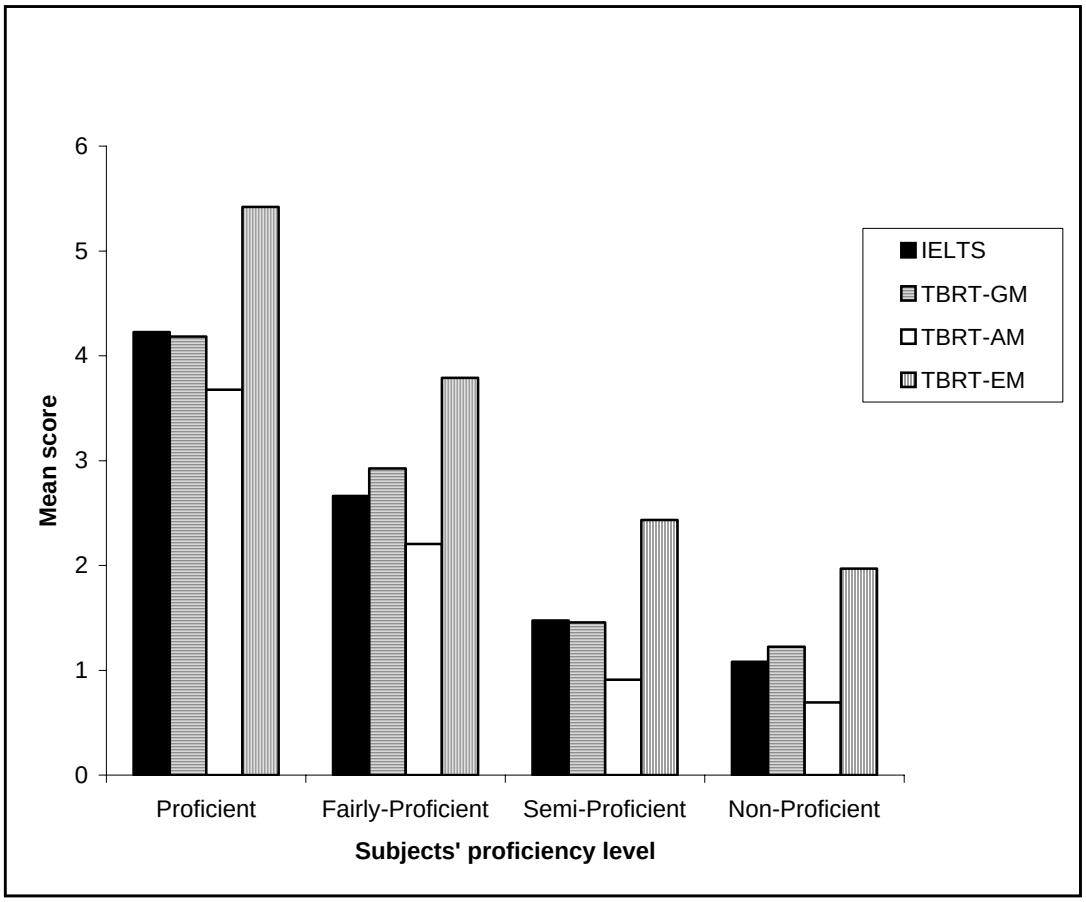

Figure 3. *Mean plot for subjects' sentence-completion task performance*

A third reading task that was under investigation in the study was the outlining task. The subjects' job was to read a few paragraphs and choose suitable headings for them from among a list of possible headings. Table IV.4 shows the frequency analysis of subjects' outlining task performance. As can be seen from the table, here again, the greatest mean belonged to the TBRT-EM (3.97) and the smallest mean to the TBRT-

AM (2.76). The means of the IELTS and the TBRT-GM were somewhat close to each other (3.39 and 3.49, respectively). Figure 4 represents the mean plot for subjects' outlining task performance.

Table IV.4: *Frequency analysis of subjects' outlining task performance*

DESCRIPTIVE		IELTS	TBRT-GM	TBRT-AM	TBRT-EM
N	Valid	541	541	541	541
	Missing	0	0	0	0
Mean		3.3937	3.4935	2.7006	3.9723
Median		3.0000	3.0000	3.0000	4.0000
Mode		4.00	3.00	2.00	4.00
SD		1.5013	1.5487	1.3981	1.4911
Variance		2.2540	2.3986	1.9546	2.2233
Minimum		.00	.00	.00	.00
Maximum		6.00	6.00	6.00	6.00

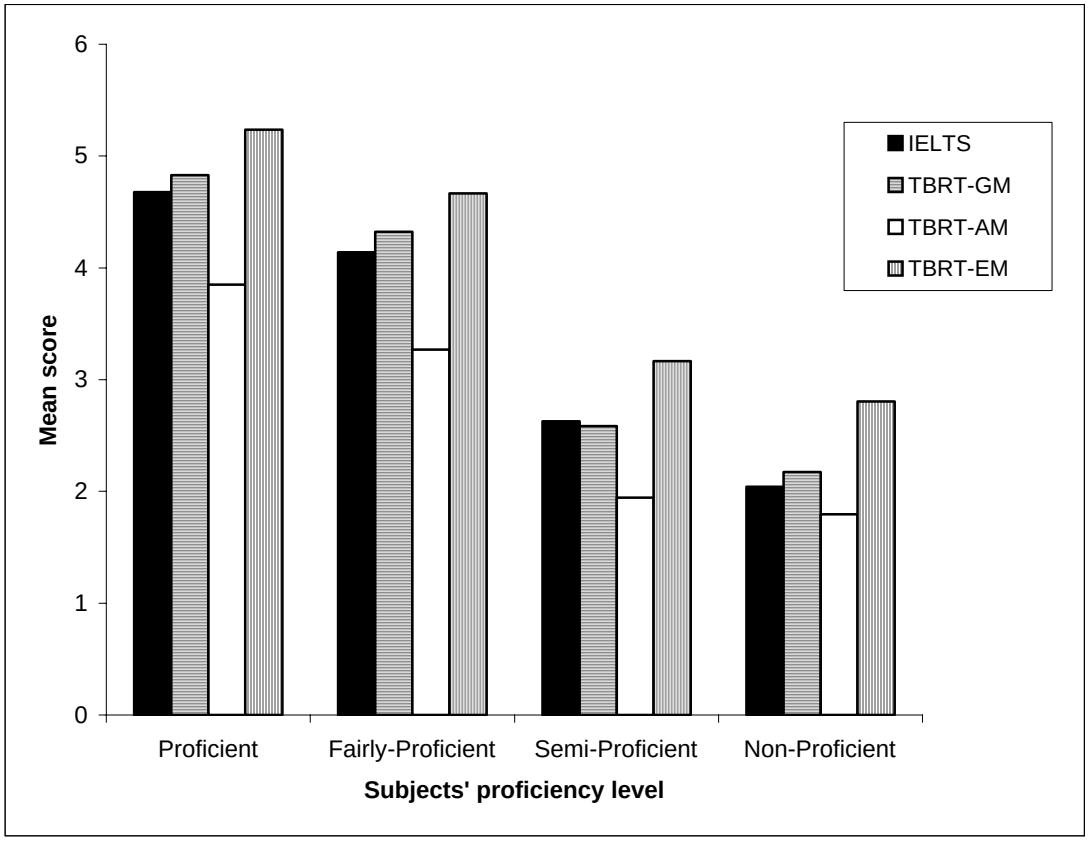

Figure 4. *Mean plot for subjects' outlining task performance*

The subjects' performance on the writer's-view task was also in the focus of the study. The subjects were expected to read a passage and indicate whether the propositions expressed in test items were given in the text or not, and if yes, whether they were part of the claims of the writer of the text or not. Table IV.5 is the tabular representation of the frequency analysis of subjects' performance on this task. As it can be understood from the table, the greatest mean belonged to the TBRT-EM (3.38). The smallest mean was found to belong to the TBRT-AM (2.35). The mean of the IELTS

General Training Reading module was 3.00. The TBRT-GM mean was 2.91. Figure 5 represents the mean plot for the subjects' performance on the writer's-view task.

Table IV.5: *Frequency analysis of subjects' writer's-view task performance*

DESCRIPTIVE		IELTS	TBRT-GM	TBER-AM	TBRT-EM
N	Valid	541	541	541	541
	Missing	0	0	0	0
Mean		3.0092	2.9131	2.3235	3.3845
Median		3.0000	3.0000	2.0000	4.0000
Mode		3.00	3.00	3.00	4.00
SD		1.2859	1.3059	1.3013	1.2751
Variance		1.6536	1.7054	1.6933	1.6260
Minimum		.00	.00	.00	.00
Maximum		5.00	5.00	5.00	5.00

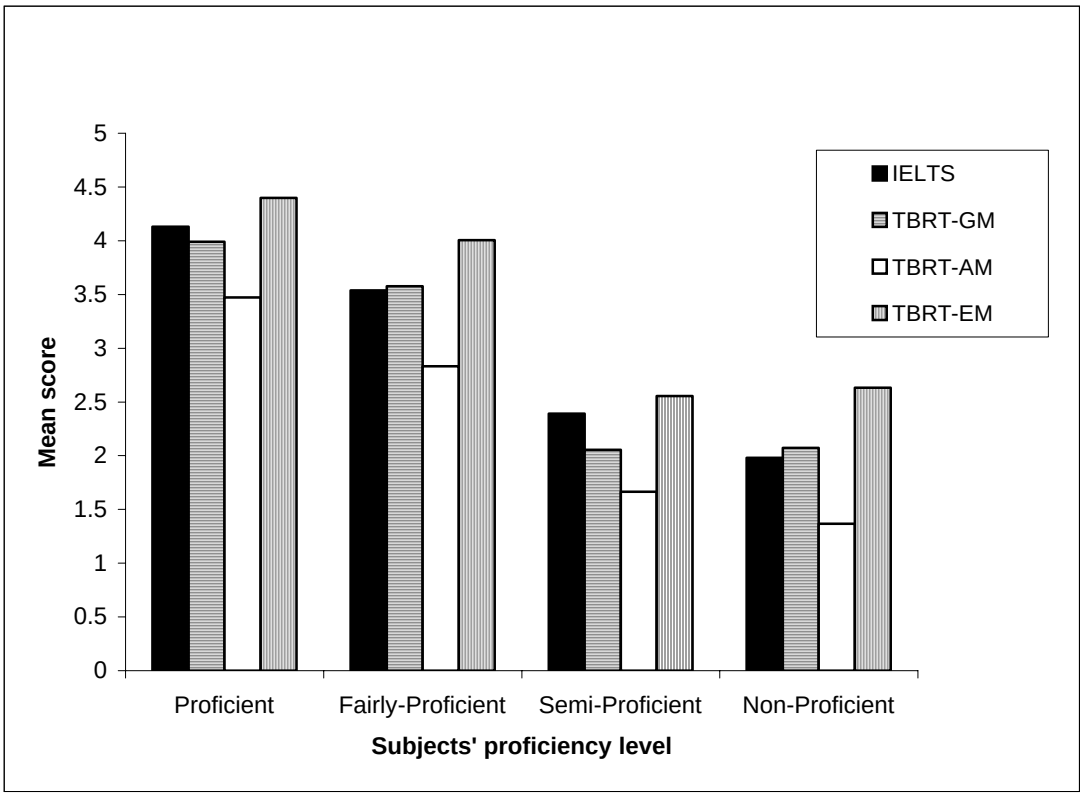

Figure 5. *Mean plot for subjects' writer's-view task performance*

The last reading task that the subjects were asked to perform was skimming. The subjects' job was to skim a reading passage and answer a few questions. Some of these questions asked them to skim for dates and some others for proper nouns. Table IV.6 shows the frequency analysis for the subjects' performances on the skimming task. As it is evident from the table, the greatest mean, like in other tasks, belonged to the TBRT-EM module (6.21). Like in other tasks, the smallest mean belonged to the TBRT-AM (3.94). The means of the IELTS General training Reading module and the TBRT-GM

module were somewhat close (5.39 and 5.22, respectively). Figure 6 displays the mean plot for the subjects' performance on the skimming task.

Table IV.6: *Frequency analysis of subjects' skimming task performance*

DESCRIPTIVE		IELTS	TBRT-GM	TBRT-AM	TBRT-EM
N	Valid	541	541	541	541
	Missing	0	0	0	0
Mean		5.3900	5.2292	3.9409	6.2107
Median		5.0000	5.0000	4.0000	6.0000
Mode		5.00	6.00	5.00	9.00
SD		2.1971	2.2059	2.3278	2.1396
Variance		4.8272	4.8659	5.4187	4.5777
Minimum		1.00	.00	.00	1.00
Maximum		9.00	9.00	9.00	9.00

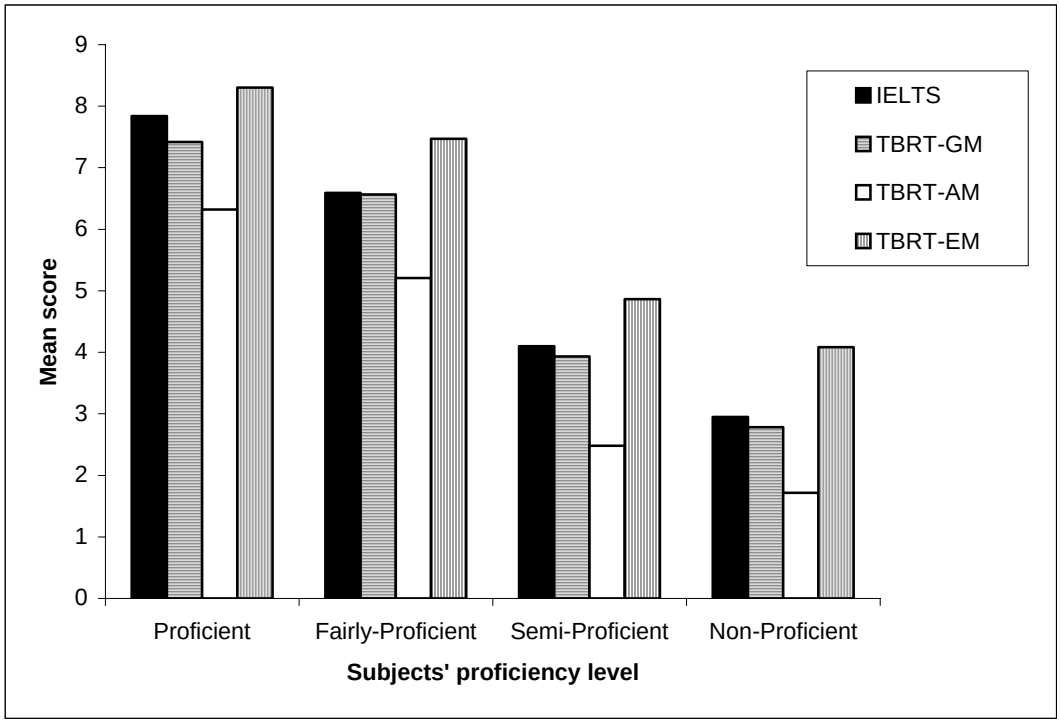

Figure 6. *Mean plot for subjects' skimming task performance*

The figures and tables presented up to this point report the subjects' task performance across different modules. In order to see which task(s) had been performed significantly better, subjects' scores on the same task across different modules were added together, after being scaled, and were recorded as subjects' overall task performance score. Table IV.7 shows the frequency analysis for the subjects' overall task performance. As the table shows, the smallest mean belongs to the sentence-completion task (122.41), true-false task is next (220.91), outlining is third (226.00),

skimming fourth (230.78), and writer's-view last (232.60). Figure 7 represents the mean plot for subjects' overall task performance.

Table IV.7: *Frequency analysis of subjects' overall task performance*

DESCRIPTIVE		True-false	Sentence completion	outlining	Writer's-view	Skimming
N	Valid	541	541	541	541	541
	Missing	0	0	0	0	0
Mean		220.9181	122.4122	226.0012	232.6063	230.7866
Median		225.0000	112.5000	216.6667	220.0000	233.3333
Mode		275.00	62.50	133.33	160.00	155.56
SD		85.7288	69.9802	83.6700	80.6275	88.4293
Variance		7349.43	4897.22	7000.66	6500.78	7819.74
Minimum		58.33	.00	.00	20.00	44.44
Maximum		391.67	362.50	400.00	400.00	400.00

Note: a Multiple modes exist. The smallest value is shown

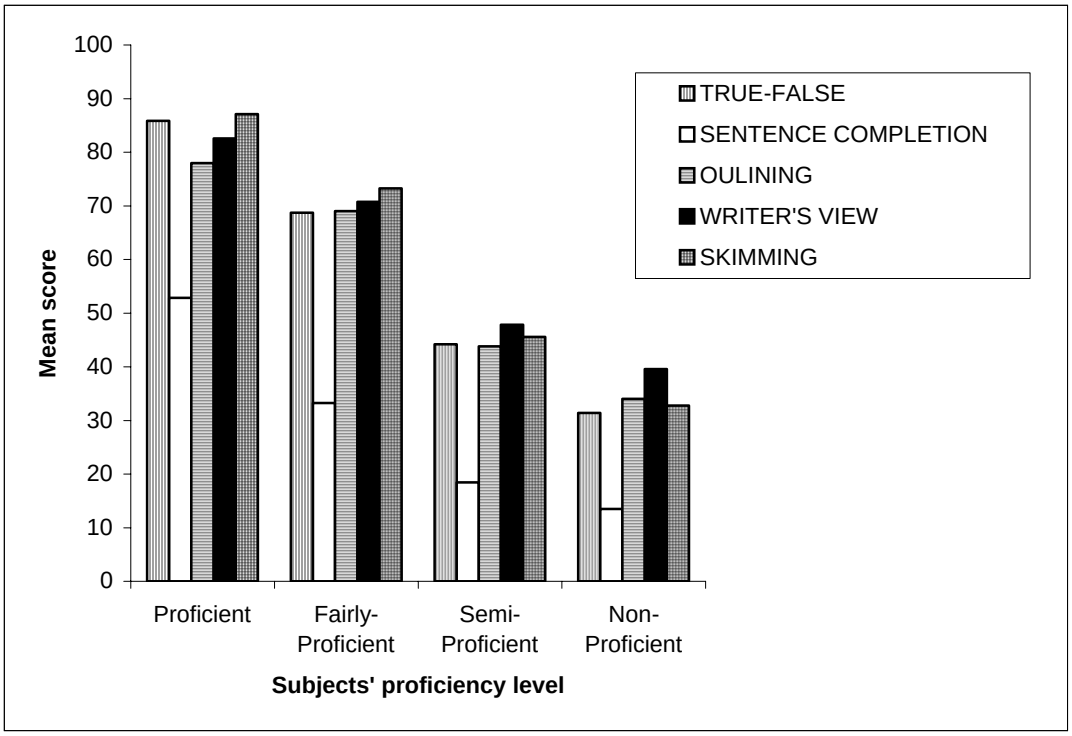

Figure 7. *Mean plot for subjects' overall task performance*

2.2. Other Statistical Analyses

As was mentioned before, the present study aimed at providing answers to the research questions and hypotheses listed in Chapter 1, section 10. On the whole, the study aimed at examining the relationships between task type, subjects' level of proficiency, and text familiarity, and subjects' differential as well as overall task and test performances. To this end, a set of inferential statistics were conducted over the data. The following sections report the results of the data analysis and discuss the empirical findings of the study.

After data collection, the investigator sought to analyze the data. In a personal communication with professor David C. Howell, he said:

> … you will want to run the overall analysis, and then break them down into more refined simple effects. I would suggest that, because you have repeated measures, you base each analysis on just the data in question. You don't want to include data that are not relevant to the particular statistical hypotheses, because that is likely to give you problems with compound symmetry. (David C. Howell, E-mail communication, 12 June 2002)

The data were analyzed by means of appropriate statistics on the basis of the hypotheses in question. Specifically, four statistics were used: (1) One-Way ANOVA comparison of means, (2) Univariate General Linear Model, (3) Multivariate General Linear Model, and (4) Multiple Regression Analysis. The investigator chose not to use the Repeated Measures General Linear Model because "… they require the assumption that the correlations among pairs of levels of the repeated variable are constant" (Howell, 1995: 348). Moreover, the Scheffé test was used because "… with the Scheffé test we drop the requirement that the overall F from the analysis of variance be significant. We can run multiple comparisons with the Scheffé test regardless of whether or not the overall F is significant" (Howell, 1995: 309).

2.2.1. The Effect of Proficiency

The first aim of the study was to determine if subjects' level of language proficiency affected their LSP test performance at a given level of text familiarity. To this end, the performance of subjects across all proficiency levels (i.e., proficient, fairly proficient, semi-proficient, and non-proficient) were compared for significant differences. The results indicated that subjects from different proficiency levels performed differently on tests with totally familiar, partially familiar, and totally unfamilar propositional content. In other words, no matter whether the propositional content of the test was familiar, unfamiliar, or partially familiar, subjects at a given proficiency level performed significantly different from those at any other proficiency levels. Table IV.8 reports the results of the post hoc Scheffé test for subjects' performance on tests with varying degrees of familiar propositional content.

The overall performances of the subjects on the TBRT were also studied. The sum of subjects' scores on the three TBRT modules (TBRT-EM, TBRT-GM, and TBRT-AM) indicated their total TBRT score. In order to determine if subjects' proficiency level had an effect on their overall test performance, the main effect analysis of variance was conducted. The results indicated that subjects' overall performance on the TBRT at each

of the proficiency levels differed significantly from subjects' performance at any of the the other proficiency levels (See table IV.9 below).

Table IV.8: *Scheffé test results for subjects' performance on tests with varying degrees of familiar propositional content*

						95% Confidence Interval	
Propositional Content of Test	Subjects' Proficiency Level		Mean Difference	Std. Error	Sig.	Lower Bound	Upper Bound
Totally Familiar	Proficient	Fairly proficient	10.5242*	.8332	.000	8.1877	12.8607
		semi proficient	37.2087*	.8516	.000	34.8205	39.5969
		non proficient	44.7594*	.9497	.000	42.0961	47.4228
	Fairly proficient	semi proficient	26.6845*	.7027	.000	24.7138	28.6552
		non proficient	34.2352*	.8189	.000	31.9388	36.5317
	Semi proficient	non proficient	7.5507*	.8376	.000	5.2017	9.8997
Partially Familiar	Proficient	Fairly proficient	10.9946*	.9060	.000	8.4538	13.5355
		Semi proficient	40.0608*	.9261	.000	37.4638	42.6578
		Non proficient	45.8152*	1.0328	.000	42.9190	48.7115
	Fairly proficient	Semi proficient	29.0662*	.7642	.000	26.9232	31.2092
		Non proficient	34.8206*	.8905	.000	32.3234	37.3179
	Semi proficient	Non proficient	5.7544*	.9109	.000	3.2000	8.3088
Totally Unfamiliar	Proficient	Fairly proficient	13.1048*	1.0208	.000	10.2421	15.9676
		Semi proficient	38.8139*	1.0434	.000	35.8879	41.7400
		Non proficient	43.7442*	1.1636	.000	40.4811	47.0074
	Fairly Proficient	Semi proficient	25.7091*	.8610	.000	23.2945	28.1236
		Non proficient	30.6394*	1.0033	.000	27.8258	33.4530
	Semi proficient	Non proficient	4.9303*	1.0263	.000	2.0523	7.8083

*The mean difference is significant at the .05 level.

Table IV.9: *Scheffé test for subjects' overall performance on the TBRT across different levels of language proficiency*

Subjects' Level of Proficiency		Mean Difference	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Proficient	Fairly proficient	4.6165*	.2132	.000	4.0198	5.2132
	Semi proficient	15.4778*	.2179	.000	14.8679	16.0877
	Non proficient	17.9092*	.2430	.000	17.2291	18.5893
Fairly proficient	Semi proficient	10.8613*	.1798	.000	10.3581	11.3646
	Non proficient	13.2927*	.2096	.000	12.7063	13.8791
Semi proficient	Non proficient	2.4314*	.2144	.000	1.8315	3.0312

Based on observed means.
*The mean difference is significant at the .05 level.

A second aim of the study was to determine if subjects' level of language proficiency affected their performance on a given task at different levels of the text familiarity cline. As it was mentioned earlier (see Chapter 3, section 4), the present study set out to measure subjects' performance on five reading tasks: true-false judgments, sentence completion, outlining, identifying writer's view, and skimming. Subjects' performance on each of these tasks was compared across different levels of proficiency at the three points on the text familiarity cline: total familiarity, partial familiarity, and total unfamiliarity.

The first task studied in this series was the true-false task. The investigator set out to determine if the performance of subjects on the true-false task at each level of text familiarity varied as a result of their proficiency levels. The results indicated that, in the

context of a test with totally familiar propositional content, the performance differences between subjects at each proficiency level and those at any of the other levels were highly significant. However, the performance difference between proficient and fairly proficient subjects was not significant at the 0.001 level. In the context of tests with partially familiar propositional content, the performance differences between subjects at each proficiency level with those at any of the other levels were significant except for that of semi-proficient versus non-proficient subjects. A similar finding was observed in the context of tests with totally unfamiliar propositional content. Here again, except for the performance difference between semi- and non-proficient subjects, the performance of subjects at any other proficiency level revealed to be significantly different in comparison to those of the subjects at other proficiency levels (See table IV.10).

Table IV.10: *Scheffé test results for subjects' true/false task performance on tests with varying degrees of familiar propositional content*

						95% Confidence Interval	
Propositional Content of Test	Subjects' Proficiency Level		Mean Difference	Std. Error	Sig.	Lower Bound	Upper Bound
Totally Familiar	Proficient	Fairly proficient	6.5860*	1.5691	.001	2.1856	10.9864
		Semi proficient	37.8912*	1.6038	.000	33.3935	42.3889
		Non proficient	50.3228*	1.7886	.000	45.3069	55.3386
	Fairly proficient	Semi proficient	31.3052*	1.3234	.000	27.5938	35.0166
		Non proficient	43.7367*	1.5422	.000	39.4119	48.0616
	Semi proficient	Non proficient	12.4316*	1.5775	.000	8.0077	16.8554
Partially Familiar	Proficient	Fairly proficient	11.3799*	1.6618	.000	6.7195	16.0403
		Semi proficient	46.9621*	1.6986	.000	42.1987	51.7256
		Non proficient	51.3487*	1.8943	.000	46.0364	56.6609
	Fairly proficient	Semi proficient	35.5822*	1.4016	.000	31.6515	39.5130
		Non proficient	39.9687*	1.6333	.000	35.3883	44.5492
	Semi proficient	Non proficient	4.3865	1.6707	.077	-.2988	9.0718
Totally Unfamiliar	Proficient	Fairly proficient	11.9624*	1.7130	.000	7.1584	16.7663
		Semi proficient	43.3582*	1.7509	.000	38.4480	48.2684
		Non proficient	47.8879*	1.9526	.000	42.4119	53.3638
	Fairly proficient	Semi proficient	31.3959*	1.4448	.000	27.3440	35.4477
		Non proficient	35.9255*	1.6836	.000	31.2039	40.6471
	Semi proficient	Non proficient	4.5296	1.7222	.076	-.3000	9.3592

*The mean difference is significant at the .05 level.

The second task studied in this series was the sentence-completion task. Subjects' performance on this task across different proficiency levels at each point on the text familiarity cline was analyzed for significant differences. The results indicated that the performance of subjects at each proficiency level was significantly different from those of subjects at any of the the other levels except for that of semi-proficient versus non-proficient subjects at each and every level of the text familiarity cline (See Table IV.11).

Table IV.11: *Scheffé test for subjects' sentence-completion task performance on tests with varying degrees of familiar propositional content*

						95% Confidence Interval	
Propositional Content of Test	Subjects' Proficiency Level		Mean Difference	Std. Error	Sig.	Lower Bound	Upper Bound
Totally Familiar	Proficient	Fairly proficient	20.3629*	2.2645	.000	14.0125	26.7133
		Semi proficient	37.3304*	2.3145	.000	30.8395	43.8212
		Non proficient	43.1246*	2.5812	.000	35.8859	50.3633
	Fairly proficient	Semi proficient	16.9674*	1.9099	.000	11.6113	22.3236
		Non proficient	22.7617*	2.2256	.000	16.5202	29.0031
	Semi proficient	Non proficient	5.7942	2.2765	.092	-.5900	12.1785
Partially Familiar	Proficient	Fairly proficient	15.7258*	2.1756	.000	9.6245	21.8271
		Semi proficient	34.0685*	2.2237	.000	27.8323	40.3047
		Non proficient	36.9788*	2.4800	.000	30.0241	43.9336
	Fairly proficient	Semi proficient	18.3427*	1.8350	.000	13.1966	23.4887
		Non proficient	21.2530*	2.1383	.000	15.2564	27.2497
	Semi proficient	Non proficient	2.9103	2.1872	.622	-3.2235	9.0442
Totally Unfamiliar	Proficient	Fairly proficient	18.4140*	1.8898	.000	13.1142	23.7137
		Semi proficient	34.6110*	1.9316	.000	29.1941	40.0280
		Non proficient	37.2943*	2.1542	.000	31.2532	43.3353
	Fairly proficient	Semi proficient	16.1971*	1.5939	.000	11.7271	20.6671
		Non proficient	18.8803*	1.8574	.000	13.6714	24.0891
	Semi proficient	Non proficient	2.6832	1.8999	.574	-2.6448	8.0113

*The mean difference is significant at the .05 level.

Analyses were also performed to determine if subjects' level of proficiency affected their performance on outlining tasks at a given point on the text familiarity cline. The results indicated that, as in the case of the sentence-completion task, subjects in the non-proficient and semi-proficient groups did not show any significant difference in performance on this task in tets with totally unfamiliar and totally familiar propositional content. In the context of tests with partially familiar propositional content, the same subjects showed a somewhat significant difference in performance.

Table IV.12: *Scheffé test for subjects' outlining task performance on tests with varying degrees of familiar propositional content*

						95% Confidence Interval	
Propositional Content of Test	Subjects' Proficiency Level		Mean Difference	Std. Error	Sig.	Lower Bound	Upper Bound
Totally Familiar	Proficient	Fairly proficient	9.4982*	2.4543	.002	2.6153	16.3811
		Semi proficient	34.5321*	2.5086	.000	27.4970	41.5672
		Non proficient	40.5073*	2.7976	.000	32.6616	48.3529
	Fairly proficient	Semi proficient	25.0339*	2.0701	.000	19.2286	30.8391
		Non proficient	31.0091*	2.4122	.000	24.2443	37.7739
	Semi proficient	Non proficient	5.9752	2.4674	.120	-.9444	12.8948
Partially Familiar	Proficient	Fairly proficient	8.4229*	2.4093	.007	1.6663	15.1795
		Semi proficient	37.3765*	2.4626	.000	30.4705	44.2825
		Non proficient	44.2415*	2.7463	.000	36.5398	51.9431
	Fairly proficient	Semi proficient	28.9536*	2.0321	.000	23.2548	34.6523
		Non proficient	35.8185*	2.3680	.000	29.1778	42.4592
	Semi proficient	Non proficient	6.8649*	2.4222	.046	7.229E-02	13.6576
Totally Unfamiliar	Proficient	Fairly proficient	9.6774*	2.4147	.001	2.9057	16.4492
		Semi proficient	31.7390*	2.4681	.000	24.8175	38.6605
		Non proficient	34.2257*	2.7525	.000	26.5068	41.9447
	Fairly proficient	Semi proficient	22.0616*	2.0367	.000	16.3501	27.7731
		Non proficient	24.5483*	2.3733	.000	17.8927	31.2039
	Semi proficient	Non proficient	2.4867	2.4276	.789	-4.3212	9.2946

*The mean difference is significant at the .05 level.

However, this difference was so small that it could be neglected. In other words, a trend towards insignificance was observable. Moreover, in the context of tests with totally familiar propositional content, the difference in performance between proficient and fairly proficient subjects, though significant at the 0.05 level, was not significant at the 0.002 level. These subjects did not show any significant difference in their performance on the outlining task in the context of tests with partially familiar propositional content at the 0.007 level, and on tests with totally unfamiliar propositioal content at the .001 level. Table IV.12 above reports the results of the post hoc Scheffé test for subjects' outlining task performance on tests with varying degrees of familiar propositional content.

Along the same lines, data analysis was also performed to determine if subjects from different proficiency levels performed differently on the writer's view task at any given point on the text familiarity cline (See table IV.13).

Table IV.13. *Scheffe test for subjects' "Writer's View" task performance on tests with varying degrees of familiar propositional content*

						95% Confidence Interval	
Propositional Content of Test	Subjects' Proficiency Level		Mean Difference	Std. Error	Sig.	Lower Bound	Upper Bound
Totally Familiar	Proficient	Fairly proficient	7.8495*	2.5531	.025	.6896	15.0093
		Semi proficient	36.8594*	2.6096	.000	29.5412	44.1776
		Non proficient	35.3039*	2.9102	.000	27.1426	43.4653
	Fairly proficient	Semi proficient	29.0100*	2.1534	.000	22.9711	35.0488
		Non proficient	27.4545*	2.5093	.000	20.4174	34.4915
	Semi proficient	Non proficient	-1.5555	2.5667	.947	-8.7536	5.6426
Partially Familiar	Proficient	Fairly proficient	8.2796*	2.5505	.015	1.1271	15.4321
		Semi proficient	38.6874*	2.6069	.000	31.3767	45.9981
		Non proficient	38.3564*	2.9072	.000	30.2034	46.5094
	Fairly proficient	Semi proficient	30.4078*	2.1512	.000	24.3751	36.4405
		Non proficient	30.0768*	2.5067	.000	23.0470	37.1066
	Semi proficient	Non proficient	-.3310	2.5641	.999	-7.5217	6.8597
Totally Unfamiliar	Proficient	Fairly proficient	12.7957*	2.6451	.000	5.3779	20.2135
		Semi proficient	36.1697*	2.7036	.000	28.5879	43.7515
		Non proficient	42.1154*	3.0151	.000	33.6601	50.5708
	Fairly proficient	Semi proficient	23.3740*	2.2309	.000	17.1176	29.6304
		Non proficient	29.3197*	2.5997	.000	22.0292	36.6103
	Semi proficient	Non proficient	5.9457	2.6592	.173	-1.5116	13.4031

*The mean difference is significant at the .05 level.

As it is observable from table IV.13, the ANOVA results revealed that subjects at each proficiency level did perform significantly different from subjects at any of the the other proficiency levels. However, there were two exceptions. First, the difference between the performance of non-proficient and semi-proficient subjects was not significant in the context of tests with varying degrees of familiar propositional content. Second, although proficient and fairly proficient subjects showed significant performance differences on

tests with partially and totally familiar propositional content at the 0.05 level, these differences were not significant at the 0.015 level in the context of partial text familiarity, and at the 0.025 level in the context of total text familiarity.

Finally, analyses were also conducted to see if subjects' proficiency level affected their skimming task performance in any statistically significant way at any given level of text familiarity. The results of these analyses have been reported in table IV.14. As it can be understood from the table, subjects at each proficiency level did perform differently from subjects at any of the the other proficiency levels at all the three levels of text familiarity. However, the mean difference between semi-proficient and non-proficient subjects in the context of tests with totally unfamiliar propositional content was not significant at the 0.002 level.

Table IV.14: *Scheffé test for subjects' skimming task performance on tests with varying degrees of familiar propositional content*

						95% Confidence Interval	
Propositional Content of Test	Subjects' Proficiency Level		Mean Difference	Std. Error	Sig.	Lower Bound	Upper Bound
Totally Familiar	Proficient	Fairly proficient	9.1995*	1.9527	.000	3.7234	14.6757
		Semi proficient	38.1691*	1.9959	.000	32.5719	43.7664
		Non proficient	46.8827*	2.2259	.000	40.6405	53.1249
	Fairly proficient	Semi proficient	28.9696*	1.6470	.000	24.3508	33.5884
		Non proficient	37.6832*	1.9192	.000	32.3009	43.0654
	Semi proficient	Non proficient	8.7136*	1.9631	.000	3.2082	14.2190
Partially Familiar	Proficient	Fairly proficient	9.4982*	1.9151	.000	4.1274	14.8690
		Semi proficient	38.7381*	1.9575	.000	33.2485	44.2276
		Non proficient	51.4849*	2.1830	.000	45.3629	57.6069
	Fairly proficient	Semi proficient	29.2399*	1.6153	.000	24.7100	33.7698
		Non proficient	41.9867*	1.8823	.000	36.7080	47.2653
	Semi proficient	Non proficient	12.7468*	1.9254	.000	7.3474	18.1463
Totally Unfamiliar	Proficient	Fairly proficient	12.3656*	2.1650	.000	6.2940	18.4371
		Semi proficient	42.6764*	2.2129	.000	36.4706	48.8822
		Non proficient	51.2033*	2.4679	.000	44.2825	58.1241
	Fairly proficient	Semi proficient	30.3108*	1.8261	.000	25.1898	35.4317
		Non proficient	38.8377*	2.1279	.000	32.8703	44.8051
	Semi proficient	Non proficient	8.5269*	2.1766	.002	2.4230	14.6309

*The mean difference is significant at the .05 level.

Another set of analyses performed on the data concerned the effect of subjects' proficiency level on their overall task performance. It can be recalled from the preceding sections that five tasks lie at the heart of the present study: true-false, sentence-completion, outlining, writer's-view, and skimming. The sum of subjects' scores on each task across the three TBRT modules (EM, GM, and AM) indicated their total score for that task. The investigator then carried out a main effect analysis of variance on the data to see if subjects' proficiency level influenced their task performance in any meaningful way.

The results indicated that, in the case of the true-false task, subjects' performance at each proficiency level differed significantly from those at any and all the other proficiency levels. In addition, the investigator compared subjects' performance on the sentence-completion task across different proficiency levels for significant differences. The results indicated that the performance of subjects at each proficiency level significantly differed from those at any of the other levels. In case of the non-proficient versus semi-proficient subjects, the mean difference was significant at the 0.05 level.

Table IV.15: *Scheffé test for subjects' overall task performance*

						95% Confidence Interval	
Task Type	Subjects' Level of Proficiency		Mean Difference	Std. Error	Sig.	Lower Bound	Upper Bound
True-False	Proficient	Fairly proficient	9.9761*	.9521	.000	7.3117	12.6405
		Semi proficient	42.7372*	.9731	.000	40.0139	45.4605
		Non proficient	49.8531*	1.0853	.000	46.8160	52.8901
	Fairly proficient	Semi proficient	32.7611*	.8030	.000	30.5139	35.0083
		Non proficient	39.8770*	.9358	.000	37.2583	42.4957
	Semi proficient	Non proficient	7.1159*	.9572	.000	4.4373	9.7945
Sentence Completion	Proficient	Fairly proficient	18.1676*	1.2217	.000	14.7487	21.5864
		Semi proficient	35.3366*	1.2487	.000	31.8422	38.8310
		Non proficient	39.1326*	1.3926	.000	35.2355	43.0296
	Fairly proficient	Semi proficient	17.1691*	1.0304	.000	14.2855	20.0526
		Non proficient	20.9650*	1.2007	.000	17.6048	24.3252
	Semi proficient	Non proficient	3.7959*	1.2282	.023	.3589	7.2330
Outlining	Proficient	Fairly proficient	9.1995*	1.4008	.000	5.2796	13.1195
		Semi proficient	34.5492*	1.4317	.000	30.5425	38.5559
		Non proficient	39.6582*	1.5967	.000	35.1899	44.1264
	Fairly proficient	Semi proficient	25.3497*	1.1815	.000	22.0434	28.6559
		Non proficient	30.4586*	1.3767	.000	26.6059	34.3114
	Semi proficient	Non proficient	5.1090*	1.4082	.004	1.1681	9.0498
Writer's View	Proficient	Fairly proficient	9.6416*	1.4914	.000	5.4678	13.8153
		Semi proficient	37.2388*	1.5244	.000	32.9728	41.5049
		Non proficient	38.5919*	1.7001	.000	33.8344	43.3494
	Fairly proficient	Semi proficient	27.5973*	1.2579	.000	24.0770	31.1175
		Non proficient	28.9503*	1.4659	.000	24.8482	33.0525
	Semi proficient	Non proficient	1.3531	1.4994	.846	-2.8429	5.5491
Skimming	Proficient	Fairly proficient	10.3544*	1.1628	.000	7.1005	13.6084
		Semi proficient	39.8612*	1.1885	.000	36.5353	43.1871
		Non proficient	49.8570*	1.3254	.000	46.1479	53.5660
	Fairly proficient	Semi proficient	29.5068*	.9807	.000	26.7623	32.2512
		Non proficient	39.5025*	1.1428	.000	36.3044	42.7006
	Semi proficient	Non proficient	9.9958*	1.1690	.000	6.7245	13.2671

*The mean difference is significant at the .05 level.

However, at the 0.023 level, this difference was no longer significant. The third task under question in the present study was the outlining task. Subjects' performance on this task resembled their performance on the sentence-completion task. Except for the slight similarity in performance between semi-proficient and non-proficient subjects, subjects' performance at other proficiency levels were totally different. In other words, subjects' performance at each proficiency level differed significantly from those at any of the other levels. As for semi-proficient versus non-proficient subjects, the mean

difference was significant at the 0.05 level but not at the 0.004 level. The present study was also concerned with subjects' overall "writer's-view" task performance. ANOVA results for this task indicated that subjects' overall performance on the writer's-view task at each proficiency level was significantly different from those at any of the other proficiency levels except between semi-proficient and non-proficient subjects. In other words, the performance difference between semi-proficient versus non-proficient subjects was not significant. The last task that the subjects were asked to perform in the present study was skimming. Here again, the results indicated that subjects' overall skimming task performance at each level was significantly different from those at any of the other proficiency levels (See table IV.15 above).

2.2.2. The Effect of Text Familiarity

In addition to language proficiency, subjects' performance is thought to be under the influence of the degree to which subjects are familiar with the propositional and textual features of the text. The review of the literature (See chapter 2) reports a good number of studies that sought to determine the influence of text familiarity on reading comprehension. For purposes of the present study, subjects' familiarity with the propositional content of texts was considered as an independent variable that might affect their test and task performance. In this section, the findings of the study as they applied to text familiarity are reported.

In the first place, the investigator sought to determine if text familiarity affected subjects' overall test performance. The data were submitted to the SPSS statistics package and were analyzed by means of the main effect ANOVA statistic. The results indicated that text familirity affected test performance, and that the performance differences among subjects at the three levels of text familirity were statistically significant. In other words, subjects' test performance at each level on the text-familiarity cline differed significantly from their performance at any of the other text-familiarity levels. Table IV.16 illustrates the results of the post hoc Scheffé test for subjects' overall test performance as it relates to the degree of text familiarity.

Table IV.16: *Scheffé test for subjects' test performance across different levels of text familiarity*

Propositional Content of the Test		Mean Difference	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Familiar	Partially Familiar	3.8096*	.1768	.000	3.3764	4.2428
	Unfamiliar	9.0351*	.1768	.000	8.6020	9.4683
Partially Familiar	Unfamiliar	5.2255*	.1768	.000	4.7923	5.6587

Based on observed means
*The mean difference is significant at the .05 level.

In addition to test performance, subjects' task performance at each text-familiarity level was also analyzed. The type of analysis used in this connection was the post hoc Scheffé test due to its stringent characteristics. Subjects' true-false task performance across different levels of text familiarity was studied first. The results of the analysis of variance (ANOVA) revealed a significant difference in subjects' performance on this task at each text-familiarity level as compared to any of the other levels. The second task studied in the context of text familiarity was the sentence-completion task. A comparison of subjects' performance on this task at each text-familiarity level with the other levels indicated a significant difference. A comparison of subjects' performance on outlining task across the different text-familiarity levels also indicated the existence of significant differences. In other words, subjects' performance on this task at each text-familiarity level significantly differed from their performance on the same task at each of the other levels of text familiarity. The main effect of text familiarity on subjects' performance on the writer's-view task was also studied. The results, after analysis, indicated that subjects' performance at each text-familiarity level significantly differed from their performance at each of the other levels. Finally, the skimming task was also studied in the context of text familiarity. Here again, the investigator calculated the main effect of text familiarity on subjects' performance on the skimming task. The results, after analysis, showed that subjects' performance on the skimming task at each text-familiarity level was significantly different from their performance on the same task at each of the other levels of text familiarity. Table IV.17, on the next page, illustrates ANOVA results for the influence of text familiarity on subjects' performance on reading tasks at different levels of text familiarity.

Table IV.17: *Scheffé test for subjects' task performance across different levels of text familiarity*

						95% Confidence Interval	
Task Type	Propositional Content of the Test		Mean Difference	Std. Error	Sig.	Lower Bound	Upper Bound
True-False	Familiar	Partial	7.8096*	.7895	.000	5.8753	9.7439
		Unfamiliar	24.1220*	.7895	.000	22.1877	26.0563
	Partial	Unfamiliar	16.3124*	.7895	.000	14.3781	18.2467
Sentence Completion	Familiar	Partial	11.7606*	1.0131	.000	9.2786	14.2426
		Unfamiliar	19.2237*	1.0131	.000	16.7417	21.7057
	Partial	Unfamiliar	7.4630*	1.0131	.000	4.9810	9.9450
Outlining	Familiar	Partial	7.9791*	1.1616	.000	5.1332	10.8249
		Unfamiliar	21.1953*	1.1616	.000	18.3495	24.0412
	Partial	Unfamiliar	13.2163*	1.1616	.000	10.3704	16.0621
Writer's View	Familiar	Partial	9.4270*	1.2367	.000	6.3969	12.4570
		Unfamiliar	21.2200*	1.2367	.000	18.1899	24.2500
	Partial	Unfamiliar	11.7930*	1.2367	.000	8.7629	14.8230
Skimming	Familiar	Partial	10.9057*	.9642	.000	8.5434	13.2680
		Unfamiliar	25.2208*	.9642	.000	22.8585	27.5831
	Partial	Unfamiliar	14.3151*	.9642	.000	11.9527	16.6774

*The mean difference is significant at the .05 level.

2.2.3. The Effect of Task Type

The investigator also conducted another set of analyses to determine if task type influenced subjects' performance in the context of text familiarity in any meaningful way. The aim of these analyses was to determine if subjects' performance on different tasks was influenced differently when the tasks appeared in tests with (a) totally familiar propositional content (i.e., TBRT-EM), (b) partially familiar propositional content (i.e., TBRT-GM), and (c) totally unfamiliar propositional content (i.e., TBRT-AM).

In the first place, subjects' performances on different tasks in the context of tests with totally familiar propositional content were compared for significant differences. The results, after analysis, indicated that only subjects' performance on the sentence-completion task differed significantly from their performance on the remaining tasks (true-false, outlining, writer's-view, and skimming). The investigator also compared subjects' performance on different tasks in the context of tests with partially familiar propositional content for significant differences.

Table IV.18: *Scheffé test for subjects' differential task performance at a given text familiarity level*

						95% Confidence Interval	
Propositional Content of the Test	Task Type		Mean Diff.	Std. Error	Sig.	Lower Bound	Upper Bound
Totally Familiar	True-false	Sentence-completion	23.5136*	1.4662	.000	18.9943	28.0328
		Outlining	-1.0783	1.4662	.969	-5.5975	3.4410
		Writer's-view	-2.5632	1.4662	.549	-7.0824	1.9561
		Skimming	-3.8817	1.4662	.136	-8.4009	.6375
	Sentence-completion	Outlining	-24.5918*	1.4662	.000	-29.1110	-20.0726
		Writer's-view	-26.0767*	1.4662	.000	-30.5959	-21.5575
		Skimming	-27.3953*	1.4662	.000	-31.9145	-22.8760
	Outlining	Writer's-view	-1.4849	1.4662	.906	-6.0041	3.0343
		Skimming	-2.8035	1.4662	.455	-7.3227	1.7158
	Writer's-view	Skimming	-1.3185	1.4662	.937	-5.8378	3.2007
Partially Familiar	True-false	Sentence-completion	27.4646*	1.4970	.000	22.8503	32.0788
		Outlining	-.9088	1.4970	.985	-5.5231	3.7055
		Writer's-view	-.9458	1.4970	.983	-5.5601	3.6685
		Skimming	-.7856	1.4970	.991	-5.3999	3.8287
	Sentence-completion	Outlining	-28.3734*	1.4970	.000	-32.9877	-23.7591
		Writer's-view	-28.4104*	1.4970	.000	-33.0246	-23.7961
		Skimming	-28.2502*	1.4970	.000	-32.8644	-23.6359
	Outlining	Writer's-view	-3.6969E-02	1.4970	1.000	-4.6512	4.5773
		Skimming	.1232	1.4970	1.000	-4.4910	4.7375
	Writer's-view	Skimming	.1602	1.4970	1.000	-4.4541	4.7745
Totally Unfamiliar	True-false	Sentence-completion	18.6152*	1.4463	.000	14.1572	23.0732
		Outlining	-4.0049	1.4463	.105	-8.4630	.4531
		Writer's-view	-5.4652*	1.4463	.007	-9.9232	-1.0072
		Skimming	-2.7829	1.4463	.448	-7.2409	1.6751
	Sentence-completion	Outlining	-22.6201*	1.4463	.000	-27.0782	-18.1621
		Writer's-view	-24.0804*	1.4463	.000	-28.5384	-19.6224
		Skimming	-21.3981*	1.4463	.000	-25.8562	-16.9401
	Outlining	Writer's-view	-1.4603	1.4463	.907	-5.9183	2.9978
		Skimming	1.2220	1.4463	.950	-3.2360	5.6800
	Writer's-view	Skimming	2.6823	1.4463	.487	-1.7757	7.1403

*The mean difference is significant at the .05 level.

The results revealed the same pattern of relationships as was found in the context of tests with totally familiar propositional content. In other words, only the performance of subjects on the sentence-completion task differed significantly from their performance on each of the other tasks. No significant difference was observed among the remaining tasks. Along the same lines, subjects' performances on different tasks in the context of tests with totally unfamiliar propositional content were compared for meaningful differences.

The results, after analysis, revealed that subjects' performance on the sentence-completion task differed significantly from their performance on each of the other tasks. Moreover, the difference between subjects' performance on true-false versus writer's-view tasks, though significant at the 0.05 level, was not significant at the 0.007 level. No significant differences were observed among the remaining tasks. Table IV.18 above reports ANOVA results for subjects' performance on different tasks in the context of tests with varying degrees of familiar propositional content.

In addition to the above analyses, subjects' performances on different tasks over the whole text familiarity cline were also compared for any significat difference (See table IV.19). As the table shows, subjects' performance on the sentence-completion task differed significantly from their performance on each of the other tasks. The difference between subjects' performance on the true-false task and their performance on the skimming task, though significant at the 0.05 level, was not significant at the 0.002 level. Moreover, the difference between subjects' performance on the true-false task and the outlining task, though significant at the 0.05 level, was not significant at the 0.027 level. The differences among subjects' performances on the remaining tasks were not significant.

Table IV.19: *Scheffé test fo subjects' overall TBRT task performance*

					95% Confidence Interval	
Task Type		Mean Difference	Std. Error	Sig.	Lower Bound	Upper Bound
True-false	Sentence-completion	23.1978*	.6032	.000	21.3393	25.0562
	Outlining	-1.9973*	.6032	.027	-3.8558	-.1389
	Writer's-view	-2.9914*	.6032	.000	-4.8498	-1.1329
	Skimming	-2.4834*	.6032	.002	-4.3419	-.6249
Sentence-completion	Outlining	-25.1951*	.6032	.000	-27.0536	-23.3367
	Writer's-view	-26.1892*	.6032	.000	-28.0476	-24.3307
	Skimming	-25.6812*	.6032	.000	-27.5396	-23.8227
Outlining	Writer's-view	-.9940	.6032	.607	-2.8525	.8644
	Skimming	-.4861	.6032	.957	-2.3445	1.3724
Writer's-view	Skimming	.5080	.6032	.950	-1.3505	2.3664

*The mean difference is significant at the .05 level.

2.2.4. Interaction Analyses

In addition to the effects of each of the independent variables on subjects' test and task performance, it was also hypothesized that the interaction bewteen any given pair and also all of the independent variables under study might be a source of variance. Therefore, a number of analyses were performed to determine if the interactions between the independent variables of the study (i.e., language proficiency, text-familiarity, and task type) were responsible for variation in subjects' performance of individual tasks on the one hand, and their overall TBRT test performance on the other. As it can be understood from tables IV.20 and IV.21, the results of interction analysis led to the following conclsions:

(1) The interaction between text familiarity and proficiency level leads to a significant difference in subjects' performance on the true-false task.

(2) The interaction between text familiarity and proficiency level leads to a significant difference in subjects' performance on the outlining task at the 0.05 level but not at the 0.029 level.

(3) The interaction between text familiarity and proficiency level does not lead to any significant difference in subjects' performance on the sentence-completion, writer's-view, and skimming tasks.

(4) The interaction between task type and text familiarity leads to a significant difference in subjects' overall test performance.

(5) The interaction between task type and proficiency level leads to a significant difference in subjects' overall test performance.

(6) The interaction between proficiency level and text familiarity leads to a significant difference in subjects' overall test performance at the 0.05 level, but not at the 0.009 level.

(7) The interaction between text familiarity, task type, and proficiency level leads to a significant difference in subjects' overall test performance at the 0.05 level, but not at the 0.027 level.

Table IV.20: Multivariate Test: *Interaction analysis for subjects' task performance*

		Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	Total TBRT	99289.918*	11	9026.356	1067.506	.000
	True-False	805143.769*	11	73194.888	434.122	.000
	Sentence completion	413358.555*	11	37578.050	135.364	.000
	Outlining	522514.228*	11	47501.293	130.155	.000
	Writer's view	541105.077*	11	49191.371	118.894	.000
	Skimming	758802.310*	11	68982.028	274.307	.000
Intercept	Total TBRT	619126.861*	1	619126.861	73221.342	.000
	True-False	4357068.122*	1	4357068.122	25841.980	.000
	Sentence completion	1531972.403*	1	1531972.403	5518.487	.000
	Outlining	4716119.822*	1	4716119.822	12922.276	.000
	Writer's view	4927619.755*	1	4927619.755	11909.900	.000
	Skimming	4745318.464*	1	4745318.464	18869.756	.000
Text Familiarity	Total TBRT	19586.237*	2	9793.119	1158.188	.000
	True-False	139364.085*	2	69682.042	413.288	.000
	Sentence completion	92307.952*	2	46153.976	166.256	.000
	Outlining	109376.768*	2	54688.384	149.847	.000
	Writer's view	111934.149*	2	55967.075	135.271	.000
	Skimming	154528.727*	2	77264.363	307.242	.000
Proficiency Level	Total TBRT	76857.626*	3	25619.209	3029.868	.000
	True-False	636428.556*	3	212142.852	1258.229	.000
	Sentence completion	310237.813*	3	103412.604	372.514	.000
	Outlining	393359.785*	3	131119.928	359.272	.000
	Writer's view	415181.452*	3	138393.817	334.493	.000
	Skimming	584204.278*	3	194734.759	774.363	.000
Interaction Between Proficiency level and text Familiarity	Total TBRT	169.696*	6	28.283	3.345	.003
	True-False	4800.395*	6	800.066	4.745	.000
	Sentence completion	1492.392	6	248.732	.896	.497
	Outlining	5161.495*	6	860.249	2.357	.029
	Writer's view	3616.293	6	602.716	1.457	.189
	Skimming	1488.183	6	248.031	.986	.433

*Results are significant at the 0.05 level

Table IV.21: Multivariate Test: *Interaction analysis for subjects' overall test performance*

	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	3865055.097*	59	65509.408	221.857	.000
Intercept	19582007.976*	1	19582007.976	66317.145	.000
Task type (1)	696090.589*	4	174022.647	589.351	.000
Text familiarity (2)	594180.767*	2	297090.383	1006.137	.000
Proficiency level (3)	2270504.143*	3	756834.714	2563.124	.000
Interaction (1) + (2)	13330.914*	8	1666.364	5.643	.000
Interaction (1) + (3)	68907.740*	12	5742.312	19.447	.000
Interaction (2) + (3)	5023.990*	6	837.332	2.836	.009
Interaction (1) + (2) + (3)	11534.770*	24	480.615	1.628	.027

*Results are significant at the 0.05 level

2.2.5. Regression Analyses

All the analyses reported up to this point only reveal the existence of a meaningful difference in subjects' test and task performance due to the impact of the independent variables in question (i.e., task type, language proficiency, and text familiarity). A more important aim of the study, however, was the determination of the relative impact of each of these independent vriables. In other words, it was of greater importance to determine which independent variable contributes more to LSP student's task and test performace scores. It was hypothesized in chapter one (section 10) that text familiarity was in charge of the greatest share of variance. The results of data analysis, however,

rejected this hypothesis and revealed that language proficiency had by far the greatest share of variance. The second greatest share of variance belonged to task type. The smallest portion of variance was accounted for by text familiarity.

To come to these conclusions, a few multiple regression analyses were conducted. The model used for each analysis was the "enter" model. The first analysis compared the relative impact of text familiarity and language proficieny on subjects' overall test performance. In this case, language proficiency accounted for 79.5% of the variance whereas text familiarity only accounted for 18.6% of the variance. Moreover, the exclusion of text-familiarity did not affect the relative importance of language proficiency. In addition, the tolerances for proficiency and text familiarity were 01.00 and 01.00 respectively, suggesting that multi-collinearity is unlikely. In other words, the findings were not sample-specific. Table IV.22 compares unstandardized and standardized regression coefficients for 'test performance' as the dependent variable.

Table IV.22: *Regression analysis for overall test performance as the dependent variable*

Independent Variables	Unstandardized Coefficients	Std. Error	Standardized Coefficients	t	Sig.	Tolerance
Proficiency	6.778	.129	.795	52.691*	.000	1.000
Proficiency	6.778	.122	.795	55.349*	.000	1.000
Text Familiarity	1.905	.147	.186	12.989*	.000	1.000

The second regression analysis took subjects' performance on tests with different degrees of familiar propositional content as its dependent variable. In this case, too, language proficiency was shown to have by far the strongest relationship with the results. In the context of tests with totally familiar propositional content, it accounted for 61.2% of the variance in comparison to task type (another independent variable of the study) which accounted for only 18.3% of the variance. Here again, the exclusion of 'task type' did not affect the impact of proficiency. Moreover, no evidence of multi-collinearity was observed. In the context of tests with partially familiar propositional content, language proficiency and task type were found to take care of 61.2% and 15.7% of the variance respectively. No fluctuation in the impact of language proficiency was observed due to the exclusion of task type from analysis. Here again, the tolerances for language proficiency and task type were 01.00 and 01.00 respectively, indicating the lack of multi-collinearity. In the context of tests with totally unfamiliar propositional content, too, the greatest share of variance belonged to language proficiency. While task type accounted for only 16.5% of the variance, language proficiency was found to be in charge of 61.2% of the variance. In addition, the impact of language proficiency did not

fluctuate due to the exclusion of task type. No evidence of multi-collinearity was observed either (See table IV.23).

Table IV.23: *Regression analysis for text familiarity as the dependent variable*

Dependent Variable	Independent Variables	Unstandardized Coefficients	Std. Error	Standardized Coefficients	t	Sig.	Tolerance
Total Text Familiarity	Proficiency	16.375	.407	.612	40.191*	.000	1.000
	Proficiency	16.375	.396	.612	41.301*	.000	1.000
	Task Type	3.384	.274	.183	12.345*	.000	1.000
Partial text Familiarity	Proficiency	17.094	.418	.618	40.857*	.000	1.000
	Proficiency	17.094	.410	.618	41.686*	.000	1.000
	Task Type	2.998	.284	.157	10.575*	.000	1.000
Total Text Unfamiliarity	Proficiency	15.885	.394	.612	40.277*	.000	1.000
	Proficiency	15.885	.386	.612	41.180*	.000	1.000
	Task Type	2.965	.267	.165	11.116*	.000	1.000

The relative impacts of text familiarity, task type, and language proficiency on subjects' task performance were also studied. Once more, the greatest share of variance belonged to langauge proficiency. It accounted for 58% of the variance. Task type and text familiarity accounted for 15.9% and 14.1% of the variance respectively. In this connection it is noteworthy that neither the exclusion of any of the 'task type' and 'text familiarity' variables nor the exclusion of both of them affected the relative importance of language proficiency on subjects' task performance. Even more interesting than this was the finding that task type had a greater share of varianve than text familiarity. The results also indicated that there was no evidence of multi-collinearity. This is important since it showes that the results are not sample-specific. The tolerances for language proficiency, task type, and text familiarity were 01.00, 01.00, and 01.00, respectively. Table IV.24 reports the comparison of unstandardized and standardized regression coefficients for 'task performance' as the dependent variable.

Table IV.24: *Regression analysis for task performance as the dependent variable*

Independent Variables	Unstandardized Coefficients	Std. Error	Standardized Coefficients	t	Sig.	Tolerance
Proficiency	16.451	.257	.580	64.118*	.000	1.000
Proficiency	16.451	.252	.580	65.369*	.000	1.000
Task Type	3.116	.174	.159	17.905*	.000	1.000
Proficiency	16.451	.248	.580	66.406*	.000	1.000
Task Type	3.116	.171	.159	18.190*	.000	1.000
Text Familiarity	4.788	.297	.141	16.140*	.000	1.000

The relative impacts of text familiarity and language proficiency on subjects' performance of each task were also studied. Again, it was found that language proficiency had by far the strongest relationship with the results. In relation to the true-

false task, language proficiency accounted for 73.4% of the variance while the sahre of text familiarity was not bigger than 12.4% of the variance. In addition, in relation to the sentence-completion task, language proficiency was responsible for 57.8% of the variance while text familiarity accounted for only 20.9% of the variance. In connection to the outlining task, language proficiency was found to be in charge of 57% of the variance while text familiarity had a share of only 12.5% of the total variance. Along the same lines, language proficiency accounted for 54.8% of the variance in the context of the writer's-view task. The share of text familiarity in this context was 14.1%. Finally, language proficiency accounted for 68.6% of the variance in relation to the skimming task whereas text familiarity accounted for only 16.6% of the variance. These findings indicated that language proficiency had the greatest share of variance when the true-false task was taken into account as the dependent variable, and the smallest share with writer's-view as the dependent variable. Text familiarity, on the other hand, left its maximum influence on the sentence-completion task and its minimum influence on the true-false task. It should also be noted that the results of regression analyses for individual reading tasks did not indicate the existence of multi-collinearity. The tolerances for text familiarity and language proficiency in the context of each reading task were 01.00 and 01.00, respectively. This indicates that the findings were not specific to the sample under investigation. Table IV.25 reports the comparison of unstandardized and standardized regression coefficients for true-false, sentence-completion, outlining, writer's-view, and skimming tasks respectively.

Table IV.25: *Regression analysis for task type as the dependent variable*

Dependent Variables	Independent Variables	Unstandardized Coefficients	Std. Error	Standardized Coefficients	t	Sig.	Tolerance
True-False Task	Proficiency	19.322	.445	.734	43.449*	.000	1.000
	Proficiency	19.322	.437	.734	44.174*	.000	1.000
	Text Familiarity	3.905	.524	.124	7.455*	.000	1.000
Sentence-Completion Task	Proficiency	13.620	.477	.578	28.543*	.000	1.000
	Proficiency	13.620	.461	.578	29.514*	.000	1.000
	Text Familiarity	5.880	.553	.209	10.641*	.000	1.000
Outlining Task	Proficiency	15.235	.546	.570	27.895*	.000	1.000
	Proficiency	15.235	.540	.570	28.212*	.000	1.000
	Text Familiarity	3.990	.647	.125	6.169*	.000	1.000
Writer's View Task	Proficiency	15.284	.580	.548	26.368*	.000	1.000
	Proficiency	15.284	.572	.548	26.743*	.000	1.000
	Text Familiarity	4.713	.684	.141	6.887*	.000	1.000
Skimming Task	Proficiency	18.794	.495	.686	37.983*	.000	1.000
	Proficiency	18.794	.482	.686	39.004*	.000	1.000
	Text Familiarity	5.453	.577	.166	9.450*	.000	1.000

3. Discussion

The present study set out to find answers to a number of questions delineated in section 10 of chapter one. However, the results of data analysis reported in the previous sections can provide answers to a good number of other questions as well. On the whole, the findings of this investigation can be re-classified into a minimum of thirteen classes of aims. In this section, the findings of the study as they relate to these classes are discussed.

One of the major aims of the study was to determine if subjects' level of proficiency introduced any difference in their task performance at each specific point along the text-familiarity cline. The findings of the study indicated that subjects' performances of the true-false and skimming tasks when the tasks appeared in a test with totally familiar propositional content were a function of their level of proficiency. In the same context, the performance of only the semi-proficient subjects compared to the non-proficient subjects did not show any meaningful difference on sentence-completion, outlining, and writer's-view tasks. In the context of a reading test with partially familiar propositional content, only the performance difference observed between semi-proficient and non-proficient subjects when performing true-false, sentence-completion, writer's-view, and outlining tasks was not significant. Moreover, in the context of a reading text with totally unfamiliar propositional content, only the performance difference observed between semi-proficient versus non-proficient subjects when performing true-false, sentence-completion, outlining, and writer's-view tasks was statistically significant.

A second aim of the study was to determine whether there was any meaningful relationship between subjects' level of proficiency and their test performance in the context of text-familiarity cline. Literature is full of reports that envisage the existence of such a relationship. However, the existance of this kind of relationship cannot be taken for granted in the context of LSP testing. The results of the present study indicated that subjects' test performance was a function of their level of proficiency, no matter whether the propositional content of the test was totally familiar, partially familiar, or totally unfamiliar. In other words, at all points on the text-familiarity cline, proficiency affects performance differentially.

Moreover, the study aimed at finding out whether subjects' level of proficiency affected their test performance regardless of the probable effect of text familiarity. The literature is also full of reports that maintain the existence of such an impact. The results of the present study also supported this contention. The level of proficiency of the

subjects affected their test performance when the test consisted of a combination of totally familiar, partially familiar, and totally unfamiliar types of propositional content.

The study also aimed at determining the probable impact of subjects' degree of text-familiarity on their test performance. The results of the study indicated that such an influence did exist. Background knoweldge of the propositional content of reading tests affected performance positively. The subjects of the study performed significantly better on tests with totally-familiar propositional content. This finding lends credence to the existence of a text-familiarity cline. Moreover, the results indicated that the performance of subjects at each point on this cline differed from their performance at each of the other points. The results also revealed that not only complete text-familiarity but also degrees of it serve as an advantage for subjects taking a reading test. This finding takes sides with the claims of Alderson and Urquhart (1985), and Clapham (1996).

Another aim of the study was to find out if there was any meaningful relationship between text familiarity and task performance. In other words, the study aimed at determining if subjects performed the same task with any significant variation across different levels of the text-familiarity cline. The results of the study supported this contention, and indicated that subjects' performance on a particular task at any given point on the text-familiarity cline differed significantly from their performance on the same task at any other point on the text-familiarity cline. This finding is also in line with Clapham's (1996) claims.

A somewhat different aim of the study was to explain how subjects' level of proficiency differentially affected their performance on a given task across different text-familiarity levels. The results, after analysis, indicated that subjects' performance on the true-false, outlining, and skimming tasks varied in accordance to their level of proficiency when these tasks appeared in tests with totally familiar, partially familiar, or totally unfamiliar propositional content. However, the difference observed between semi-proficient and non-proficient subjects when performing sentence-completion and writer's-view tasks on tests of varying degrees of familiar propositional content was not statistically significant. This supports the "reading threshold" hypothesis. In other words, in order to be able to draw on prior knowledge (that is, to activate schemata), readers need to have already reached a specific level of language proficiency (a threshold level) to be able to disentangle themselves from the web of formal and structural features of the text.

The impact of task type on subjects' test performance was also studied in the context of text familiarity. The purpose of this probe was to determine if subjects' performance on one task was comparable to their performance on other tasks at the same text-familiarity level. The difference between the sentence-completion task and all the other tasks (true-false, outlining, writer's-view, and skimming) revealed significant when these tasks appeared in tests with varying degrees of familiar propositiaonl content. In addition, in tests with totally unfamiliar propositional content, the difference between the true-false task and the writer's-view task was also meaningful

The impact of task type on subjects' test performance was also studied in the context of subjects' overall test performance (i.e., regardless of the text-familiarity cline). The difference between the sentence-completion and true-false tasks, on the one hand, and all the other tasks, on the other, turned out to be significant. The one-to-one comparison of the remaining tasks also afforded significant results. However, there were three exceptions: (a) outlining versus writer's-view, (b) outlining versus skimming, and (c) writer's-view versus skimming. These comparisons afforded no significant results.

Another step taken in the study was to determine if the interaction between two or more of the independent variables (i.e., subjects' proficiency level, task type, and text-familiarity cline) introduced any meaningful difference in subjects' test and task performance. Subjects' task performance was studied in the context of the interaction between subjects' degree of text-familiarity and level of proficiency. The results indicated that this interaction only affected subjects' performance of the true-false and outlining tasks in a significant way. The writer's-view, sentence-completion, and skimming tasks were not influenced in a meaningful way by this interaction. As for subjects' overall test performance, the interaction between text familiarity and task type appeared significant. Subjects' overall test performance was also affected by the interaction between text familiarity and language proficiency in a meaningful way. Moreover, the interaction between task type and language proficiency caused a meaningful difference in subjects' overall test performance. Finally, the interaction among text familiarity, task type, and language proficiency was an important source of variance in subjects' overall test performance.

A comparison of the results of regression analyses reported in this study with the findings of Clapham's (1996) study is intriguing indeed. While Clapham attaches greater importance to text familiarity (accounting for 38% of the variance) in comparison to language proficiency (accounting for 26% of the variance), the present

investigation came up with a somewhat different result. As intriguing as it may seem, in none of the comparisons made between any given pair of the independent variables under study in relation to subjects' overall as well as differential test and task performances did language proficiency account for less than 50% of the variance. Moreover, the very high tolerance indexes reported in this study reject any chance for multi-collinearity to occur. This indicates that the findings of the present study are far from being sample-dependent. A quick look at the tolerance indexes reported in the regression tables above reveals that, in each case, the collinearity statistic was equal to 01.00 which signifies the lack of multi-collinearity. Moreover, the effect of text familiarity on task performance was found to be smaller than the effect of task type. On these grounds, it can safely be argued that perhaps the development and use of LSP tests is out of consideration. As such, the results of this study are somewhat close to Lipson's (1984) contention that LSP testing is not really justified. The greater impact of task type, in comparison to text familiarity, on subjects' performance, however, stands against Lipson's claims. The findings of the study indicate that, instead of giving students passages with esoteric propositional content, it is better to give them a rich variety of reading tasks, and measure their performances on them.

4. Final Remarks

The findings of the present study are all based on the 95% confidence interval. In other words, all of the significant findings reported in the previous sections were significant at the 0.05 level (i.e., Alpha=0.05). Some of these findings are not significant if we modulate the confidence interval. The reader's attention is specifically drawn to the following considerations:

(1) The interaction between text familiarity and language proficiency does not cause meaninful variation in subjects' overall test performance at the 0.009 level.

(2) The interaction among text familiarity, language performance, and task type does not cause significant variation in subjects' overall test performance at the 0.027 level.

(3) The interaction between text familiarity and subjects' proficiency level is not a source of significant variation in subjects' overall performance on the outlining task at the 0.029 level.

(4) The difference observed in the true-false task performance of proficient versus fairly proficient subjects on tests with totally familiar propositional content is not significant at the 0.001 level.

(5) The difference observed in the outlining task performance of proficient versus fairly proficient subjects on tests with partially familiar propositional content is not significant at the 0.007 level.

(6) The difference observed in the outlining task performance of proficient versus fairly proficient subjects on tests with totally unfamiliar propositional content is not significant at the 0.001 level.

(7) The difference observed in the outlining task performance of proficient versus fairly proficient subjects on tests with totally familiar propositional content is not significant at the 0.002 level.

(8) The difference observed in the writer's-view task performance of proficient versus fairly proficient subjects on tests with partially familiar propositional content is not significant at the 0.015 level.

(9) The difference observed in the writer's-view task performance of proficient versus fairly proficient subjects on tests with totally familiar propositional content is not significant at the 0.025 level.

(10) The difference observed in the skimming task performance of non-proficient versus semi-proficient subjects on tests with totally unfamiliar propositional content is not significant at the 0.002 level.

(11) The difference observed in the overall sentence-completion task performance of non-proficient versus semi-proficient subjects is not significant at the 0.023 level.

(12) The difference observed in the overall outlining task performance of non-proficient versus semi-proficient subjects is not significant at the 0.004 level.

(13) The difference observed in the subjects' overall performance of the writer's-view and true-false tasks is not significant at the 0.007 level.

Needless to say, any interpretation of the findings of the present study should heed these points. In addition, the findings of this investigation are based on the performance of the subjects who took part in the study and should not be overgeneralized to the

whole population of Iranian university students majoring in electronics unless after a generalizability study on the findings.

CHAPTER FIVE
CONCLUSION

1. Introduction

This study had three main purposes: (a) to investigate the value of giving academic students task-based reading modules in different subject areas, (b) to determine how language proficiency affects academic students LSP test performance, and (c) to find out more about the effect of background knowledge on reading. This concluding chapter will begin by commenting on the research questions which were introduced in Chapter 1. Then, a discussion of the implications of the research findings will follow. The findings will also be related to the issues of LSP proficiency testing. The chapter shall conclude with a brief description of the areas in which further research is needed.

2. Brief Overview of the Findings

Do students in a specific academic subject area achieve significantly higher scores in a reading test within their own subject area than they do in a test outside their subject areas? Do students with partial familiarity with the propositional content of a reading test achieve significantly higher scores on the test than students who do not have that knowledge? Which contributes to students' reading scores: background knowledge, task type, or level of proficiency? Do reading tasks affect students' reading scores differentially? Is ESP testing really needed? These and many other questions stimulated the investigator to undertake the present study. In order to make the study practically executable, the researcher had to break these general questions into readily researchable ones. (See Chapter 1, section 10).

The investigation began with the hope of finding answers to a number of questions. These questions attempted to (1) probe the probable effect of background knowledge on LSP reading test performance, (2) query the motivation for including reading tasks in LSP reading tests, and (3) investigate the impact of language proficiency on reading comprehension. In addition, the study aimed at illuminating whether the importance attached to "text specificity" in LSP reading tests was empirically justifiable.

The study showed that students achieve significantly higher scores on reading tests in their own subject area than on reading tests outside it. For purposes of this study, 'own subject area' referred to the students' present or past subject area, and not to their future one. Students were asked in the self-report questionnaire (See Appendix E) to indicate their degree of text familiarity (total familiarity, partial familiarity, or total unfamiliarity). Their responses to the questionnaire were used for a classification of the students according to their background knowledge. The results of the study supported the claim that the students' scores were highest on the test within their subject area, higher on the test partially familiar to them, and lowest on the test outside their subject area.

The results also indicated that the inclusion of reading tasks in a reading test is empirically supported. As it can be recalled from chapter 4, the only task that stood at odds with the other tasks in the study was the sentence completion task. By definition, sentence completion requires greater structural knowledge. It can, thus, be considered more of a writing task than of a reading task. Since subjects did not show any differential performance on the remaining tasks even in tests outside their subject area, the compartmentalization of reading tests into reading tasks should be approached with caution. That is, before the inclusion of any task in a reading test, we must come to an agreement as to what comprises a reading task.

The results also indicated that reading test performance was a function of language proficiency. The findings of the study revealed that the classification of students into LEP (Limited English Proficient) and Non-LEP (Non-Limited English Proficient) was a wise strategy. The interesting point in this connection was the observation that the effect of language proficiency took the same form in reading tests both inside and outside the students' subject area. Students from different proficiency levels performed consistently better in tests within their subject area than in tests outside it. Although text familiarity had an impact on students' performance, its share of variance was much smaller than that of language proficiency.

3. Implications of the Study

It is clear from this study that background knowledge plays a key role in the reading process, but that this role is not as important as the role of either language proficiency or task type. Anyway, the findings of the study have some implications for the validity of at least the following notions: Schema theory, Bachman's notion of Levels of Contextualization, reading Threshold Hypothesis, and Given and New strategy.

We do not yet know enough about the cognitive processes of the brain to be able to tell whether we store and retrieve knowledge in the ways proposed by schema theorists. Whatever the actual processes involved in reading comprehension may be, the value of schema theory to applied linguists is that it proposes formal structures for the acquisition and retrieval of knowledge, and thus gives some form to the amorphous notion of background knowledge. An area of schema theory which might produce useful insights into the findings of this study relates to 'formal' schemata. Many reading researchers believe that people have formal schemata of the structure of different kinds of writing, and that if they can activate the appropriate formal schemata when reading, their comprehension and memory of that text will be enhanced. If this is the case, a student who has the formal schemata relating to the reading passages in his or her test might find the passages easier to read than one who does not. The findings of the present study also have some implications for Bachman's notions of Levels of Contextualization (See Bachman, 1990 :132). Although not explicitly related to it, the Given-new strategy ties in with Levels of Contextualization. It accounts for some of the findings of the present study too. Applying the given-new idea to a reading passage, one can say that the author of the text would have written it for readers with an expected amount of Given knowledge in the chosen field, and that this knowledge would be necessary for a complete understanding of the passage. For a reader outside that field, without the required prior knowledge, the information would not be Given, and in cases where this reader had to make inferences and go beyond the passage for comprehension, he might fail to understand the text fully unless he is proficient enough. If, however, the passage contained a high proportion of Given material, then prior knowledge about the subject area might not be necessary. It would be interesting to relate these ideas to students' reading test introspections.

Perfetti and Lesgold (1977) suggest that bad readers are so slow at decoding symbols that they cannot retain materials in their short-term memory long enough to be able to call up the relevant background knowledge. This is likely to be the case to an even

greater extent with low proficiency L2 readers. This 'threshold hypothesis' is supported in the present research: the subject area of the reading modules had negligible effect on the performance of low proficiency students. This does not mean that such learners do not attempt to use background knowledge, but that they are not able to incorporate it effectively to their reading. For advanced learners, too, there may be a further threshold above which they are able to use bottom-up reading processes as automatically as native speakers. Above this threshold readers are so proficient linguistically that they can compensate for a certain lack of background knowledge by making full use of their language resources. However, the validity of the existence of such a threshold requires further research.

Although it may be accepted that readers' levels of language proficiency will affect their use of prior knowledge, there has been disagreement about the relative importance of these two factors in reading comprehension. In some of the early research based on schema theory it was claimed that background knowledge is more important than level of language proficiency. However, this does not accord with the findings of the present study. As we know, once a text is highly specialized, and is based on complex concepts which are familiar to only a limited group of readers, good language proficiency is no longer sufficient for text comprehension. The effect of background knowledge on reading comprehension, therefore, depends not only on the proficiency level of the students, but also on the specificity of the reading passages. This brings us directly to two central problems in research into the effect of background knowledge on reading: firstly, the difficulty in assessing that background knowledge, and secondly, the difficulty in ensuring that the reading passages are suitably specific for the purposes of the research. These issues, too, have some implications for the findings of the present research.

4. Relating the Study to LSP Testing

We now return to the outstanding research issue which was formulated in Chapter 1: If it is accepted that background knowledge does have some effect on reading comprehension, should this be explicitly taken into account when LSP tests are devised? Should students intending to study in different academic areas be given reading tests in these different subject areas, so that they are not disadvantaged by a lack of appropriate background knowledge? The answer to this question is naturally affected by the fact that the present study was based on three text-familiarity levels. The results might have

been different if the research had been based on a more detailed conception of text familiarity. What can safely be said in connection to the above questions is that, as subject areas become narrower, it becomes more difficult to allocate students to the appropriate text familiarity levels, and even with more specific tests, such as Engineering, there is a risk that some students will not find that the reading passages relate to their own branch of the subject.

According to Clapham (1996), the question of which subject areas should be tested in an LSP test has not yet been satisfactorily answered, and if we look at the case of the IELTS we can see some of the problems. Because there were no research findings to help decide which subject areas the IELTS test battery should contain, the IELTS project committee decided that the modules should be based on three subject areas. This decision was made partly at the suggestion of some receiving institutions and partly because approximately equal numbers of students were registering for courses in three subject areas. However, this division of science into three subject areas has turned out to be neither necessary nor advisable. The fact that many or most students study one subject area during their university course and are therefore text familiar suggests that a subject-area-based reading test would be appropriate for students in that subject area. It might therefore be possible to reduce the size of reading tests. However, care must be taken since this smears the distinction between LSP and knowledge tests.

Returning back to our question, it is really difficult to make sure that ESP texts are specific to their designated subject area, and that complex procedures have to be followed to ensure that the texts are specific. The inclusion of a general reading passage in what is supposedly a subject specific test may not matter to the students as long as it does not place them at a disadvantage, but it destroys the whole purpose of having tests in different subject areas. All the effort of producing subject specific tests is wasted if such tests turn out to consist of general reading passages. However, even if the texts are specific or highly specific, it is not clear from the present study how many students would profit or suffer from taking reading modules in different subject areas. It therefore seems advisable not to give academic students subject specific reading modules, but to give them an English for General Academic Purposes (EGAP) reading test instead. If it is accepted that LSP students should all take the same reading module, a decision has to be made about the types of reading passages on which the test should be based. Should they be non-academic texts? Or, should they be academic texts taken from a variety of disciplines?

5. Suggestions for Further Research

Some of the suggestions made here are related to the effect of background knowledge on reading comprehension, and some to LSP testing. However, the findings of most of them are likely to be interrelated.

The first suggestion is that a qualitative study be carried out into students' reading processes. Students should be asked to introspect as they take the reading modules within and outside their own study area. They should focus on the use they make of background knowledge as they describe what goes through their mind as they answer each test item. This will be a worthwhile research since its results can help us define the construct of background knowledge.

In the present study the range of subjects was not wide enough for the investigator to get a full idea of the stages at which students at varying levels of proficiency start to use different reading processes. The students who took the TBRT test were either junior or senior university students. A large study should be undertaken in which L2 students over the whole spectrum of language ability from new beginners to those with the highest levels of proficiency are tested for differences in their use of top-down and bottom-up processes in general, and the use they make of background knowledge in particular. Such a study should perhaps be based on a homogeneous group of students sharing the same first language, since variations in the reading processes of different speech communities might otherwise confound the results.

Research into the comparative effects of background knowledge and language proficiency on reading comprehension suggests that when reading materials are not highly specific for their readers, language proficiency will be a better predictor of reading scores than background knowledge, but that as the texts become more specific, background knowledge becomes more important. It would be interesting to run multiple regression analyses on students' scores on a range of subtests of varying levels of specificity, to see whether this supposition is correct. If it is, regressions on the partially familiar reading tests might show that background knowledge fails to account for any of the score variance, whereas on the totally familiar tests, background knowledge might account for more of the variance than language proficiency.

It would be interesting to classify the students according to narrower disciplines such as engineering, law and physics, or yet narrower ones such as civil engineering and maritime law, or quantum physics and to compare their test performances with those from the broader set of classifications although there are too few students in some of the

disciplines for this to be worthwhile. With a larger and better balanced student sample, such a comparison will make a valuable study since broad groupings must mask important differences between the disciplines within a subject area.

There is some uncertainty as to what an 'academic' text might be. If it is decided that an LSP test should include reading passages based on academic materials, more research is needed into what distinguishes an academic text from a non-academic one.

To get more idea of the students' background knowledge, the present study used their responses to the self-report questionnaire. No attempt was made to measure other, possibly untestable, factors such as memory or general knowledge. However, because of the variation in people's memory and the range of their general knowledge, it has to be accepted that any survey, however thorough, would be incomplete. Yet, it would be a worthwhile research.

It was mentioned earlier that the present study aimed at evaluating the difference in performance (a) of LEP subjects across different test tasks on the same module, and (b) of non-LEP subjects across different tasks on the same module. These are essentially reliability considerations, since they look at how well the different tasks measure the ability that underlies the module (be it reading, or ESP reading). Thus, another suggestion for further research will be to use the present data to calculate coefficient alpha for the 5 tasks within each module. Then, to calculate the relative effect of tasks, a single facet g-study can be conducted, using task as the facet. These analyses can be done first by group, and then by aggregating across groups.

It was also mentioned earlier that another aim of the present study was to evaluate performance difference (c) of LEP subjects across different modules on the same task, and (d) of non-LEP subjects across different modules on the same task. It is possible to address these two issues together with the (a) and (b) issues mentioned in the previous paragraph in a single analysis. To this end, a 2-facet g-study should be conducted with tasks and modules as facets, with tasks nested within modules.

Rasch analysis can be carried out to see whether any items or sets of items showed evidence of bias against students who take modules which are not in their own subject area.

Bias analyses are designed to identify variations in item performance caused by gross differences in the groups being compared. Students in one subject areas are not sufficiently disparate. This possible lack of disparity raises important questions in any research based on a limited range of subjects. A worthwhile research would be to use

disparate subjects and subject areas to replicate the present study and to carry out bias analysis on its data.

Another study can be carried out into whether students are able to complete the appropriate module more quickly than the inappropriate one. Such a study may turn out to show that tests which contain the most highly specific subtests, are finished by proportionately more students who are totally familiar with that specific subject area.

Research can also address the issue of whether the reading passages in LSP reading modules are specific to the appropriate subject area? An investigation into the LSP modules' Test Method Facets may reveal no variation in the specificity of the test items. In this case, it seems likely that any differences in the subtests' specificity are due to variations in the reading passages.

Is it possible to identify some characteristics of the reading passages which make them either more or less specific to their chosen subject areas. The source of a reading passage does not necessarily determine its level of specificity. The rhetorical function of reading passages rather than their sources can affect their specificity. The subject specificity of a text is also likely to depend on the extent to which comprehension of that text requires knowledge of subject specific concepts which are not explained in the text. This can be the subject of another research.

The literature on the effects of background knowledge shows that background knowledge does affect the reading comprehension of ESL/EFL readers, but it also shows that comparatively little is known about the manner in which it does so. The rival effects of content and formal knowledge, for example, need more study, and so does the effect of familiar as compared to 'salient' passages.

If it is accepted that background knowledge does have some effect on reading comprehension, should this be explicitly taken into account when LSP proficiency tests are devised? Should students intending to study in different academic areas be given reading tests in these different subject areas, so that they are not disadvantaged by a lack of appropriate background knowledge? This can be the topic of another research.

Another issue in connection to the effect of background knowledge is the question that arises from the fact that it is not clear whether students with low levels of language proficiency depend more on background knowledge for their interpretation of written texts than do high level students. Many researchers think that, at the lower levels, students are unable to decode written language sufficiently to be able to bring top-down processing to bear. Some others have found that low level students used background

knowledge to help them make sense of incomprehensible material. Some scholars also suspect that background knowledge becomes less important as students become more linguistically proficient. A worthwhile study will, therefore, try to see at which levels of proficiency students seem to be most affected by their background knowledge.

It is true that students with different linguistic backgrounds may process reading in different ways, and it is true that if members of these different backgrounds are grouped together these differences may obscure important reading effects. It is certainly important, therefore, that at least some second language research should be carried out on linguistically homogeneous groups. However, the choice of sample must depend on the purpose of the research. If its aim is to enquire into aspects of the reading process which are considered to be universal, it makes perfectly good sense to use heterogeneous samples, as it does if the purpose of the research relates to the testing of multinational LSP students. However, since variation in the subjects' first language and culture may obscure findings, it is desirable, where possible, to study one or two subsets of linguistically homogeneous subjects as well.

The use of anomalous, artificially designed texts cannot give us full knowledge of how readers process 'real' texts, but the choice of text type must depend on the purpose of the research. If the research is specifically concerned with the comprehension of natural texts, then it should be based on such texts. A study can compare the inclusion of authentic versus artificial texts in LSP tests.

Since there is no perfect measure of reading comprehension, and since different testing techniques affect results, it is only sensible that, where possible, multiple measures should be collected from each subject involved in a reading comprehension study. A combination of techniques like recall protocols, grammatical achievement tests, responsive writing and retrospective interviews can be suggested. This technique can be used, in a replication of the present study, for the classification of subjects' into different proficiency levels. The results of that study should then be compared with the findings of the present study.

As was pointed out earlier, Douglas (2000) claims that there is such a thing as specific purpose language knowledge (LSP ability), and that the nature of language knowledge may be different from one domain to another. However, this is a strong claim, and a contentious one. It was not the purpose of this study to become embroiled in the issue of whether there are multiple competencies, each associated with a

particular context, or whether there is one competence, variably drawn upon in different contexts. This can be the topic of a worthwhile exploratory research.

In addition to the suggestions made up to this point, questions such as the following will require further investigation: How many performances are needed in order to arrive at valid conclusions in LSP testing? How many tasks and how many different procedures are needed in LSP tests? For what duration should LSP performance be sampled? How often should a person's LSP language be sampled? At what points in a person's career or in his/her language development, and over what period of times, should his/her LSP performance be sampled? These issues need further study so that researchers can draw valid conclusions regarding the proficiency of the LSP test taker. There is also a need to employ a variety of theories related to work performance which specify hierarchy ordering and weighting of skills in order to set up sampling frames and establish content validity for LSP tests.

6. Final Remarks

Although the students in the present study represented a wide range of language proficiency, wide enough for some research to be carried out into reading at different levels of language ability, none of the students were beginners, and few, if any, were very highly proficient, since few of the most linguistically able students would be attending the English classes at which the TBRT modules were administered. This means that the findings are inevitably limited. It is possible that with a wider sample there will be stronger evidence of what reading precisely is. The results will also be more informative if the reading modules are more subject-specific. A study is needed in which the student sample covers the complete range of L2 speakers from new beginners to top proficiency learners.

REFERENCES

Abraham, R., & Plakans, B. (1988). Evaluating a screening/training program for nonnative speaking teaching assistants. *TESOL Quarterly*, 22, 505-508.

Alderson, J. C. (1980). Native and non-native speaker performance on cloze tests. *Language Learning*, 30, 59-76.

Alderson, J. C. (1981). Report of the discussion on testing English for specific purposes. In J. C. Alderson, & A. Hughes (Eds.).

Alderson, J. C. (1988b). Testing and its administration in ESP. In D. Chamberlain, & R. J. Baumgardner (Eds.).

Alderson, J. C. (1991). Language testing in the 1990s: how far have we come? How much farther have we to go? In S. Anivan (Ed.), *Current developments in language testing* (pp. 1-26). Singapore: Regional Language Centre.

Alderson, J. C., & Clapham, C. (1992). Applied linguistics and language testing: a case study of the ELTS Test. *Applied Linguistics*, 13, 149-67.

Alderson, J. C., & Hughes A. (Eds.). (1981). *Issues in language testing*. London: British Council.

Alderson, J. C., & Urquhart, A. H. (1983). The effect of student background discipline on comprehension: a pilot study. In A. Hughes, & D. Porter (Eds.).

Alderson, J. C., & Urquhart, A. H. (Eds.). (1984). *Reading in a foreign language*. London: Longman.

Alderson, J. C., & Urquhart, A. H. (1985a). The effect of students' academic discipline on their performance on ESP reading tests. *Language Testing*, 2, 192-204.

Alderson, J. C., & Urquhart, A. H. (1985b). This test is unfair: I'm not an economist. In P. C. Hauptman, R. Le Blanc, & M. B. Wesche (Eds.).

Allen, J. P. B., & Widdowson, H. G. (1974). Teaching the communicative use of English. *International Review of Applied Linguistics*. XII(I).

Allwright, R. L. (1982). Perceiving and pursuing learners' needs. In M. Geddes, & G. Coffey, B. (1984). ESP-English for specific purposes. *Language Teaching*, 17(1).

American Council on the Teaching of Foreign Languages. (1986). *ACTFL proficiency guidelines.* Hastings-on-Hudson, NY: American Council on the Teaching of Foreign Languages.

Anderson, R. C., *et al*., (1977a). Frameworks for comprehending discourse. *American Educational Research Journal,* 14, 367-381.

Anthony, L. (1997). ESP: What does it mean? Retrieved from the World Wide Web on April 6, 2000, on CUE.

Asturias, H. (1994). Using students' portfolios to assess mathematical understanding. *Mathematics Teacher*, 87(9), 698-701.

Ausubel, D. P. 1960. The use of advance organizers in the learning and retention of meaningful material. *Journal of Educational Psychology*, 51, 267-72.

Ausubel, D. P. (1963). *The psychology of meaningful verbal learning*. New York, NY: Grune & Stration.

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., Davidson F., & Milanovic, M. (1991). The use of test method characteristics in the content analysis and design of English proficiency tests. Paper presented at the Language Testing Research Colloquium, Princeton, NJ.

Bachman, L. F. & Palmer, A. S. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly*, 16, 449-465.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Bailey, K. (1985). If I had known then what I know now: Performance testing of foreign teaching assistants. In P. Hauptman, R. LeBlanc, & M. Wesche (Eds.), *Second language performance testing* (pp. 153-180). Ottawa: University of Ottawa Press.

Barnwell, D. (1989a.) Naive native speakers and judgements of oral proficiency in Spanish. *Language Testing*, 6,152-63.

Barnwell, D. (1989b). Proficiency and the native speaker. *ADFL Bulletin*, 20, 42-46.

Barrett-Lennard, S. (1997). Encouraging autonomy and preparing for IELTS: Mutually exclusive goals? *Prospect*, 12 (3), 29-40.

Bialystock, E. (1982). On the relationship between knowing and using linguistic forms. *Applied Linguistics*, 3(3), 181-206.

Brindley, G. (1986). *The assessment of second language proficiency*: *Issues and Approaches*. Adelaide: National Curriculum Resource Center.

Brown, A. (1993). The role of test-taker feedback in the test development process: Test-takers' reactions to a tape-mediated test of proficiency in spoken Japanese. *Language Testing*, 10(3), 277-303.

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing*, 12(1), 1-15.

Brown, J. D., & Hudson, T. (1998). The alternatives in language assessment. *TESOL Quarterly*, 32(4), 653-676.

Bruner, J. (1986). Play, thought and language. *Prospects*: *Quarterly Review of Education*, 16(1), 77-83.

Bryman, A., & Cramer, D. (1999). *Quantitative data analysis with SPSS Release 8 for Windows: A guide for social scientists*. London: Routledge.

Buell, James G. (1992). TOEFL and IELTS as measures of academic reading ability: An exploratory study. Eric No. ED354758.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 6, 67-84.

Candlin, C. N. (1987). Towards task-based language learning. In C. N. Candlin, & D. F. Murphy (Eds.), *Language learning tasks* (pp. 5-22.).

Candlin, C. N., Leather, J. H., & Bruton, C. J. (1976). Doctors in casualty: Applying communicative competence to components of specialist course design. *International Review of Applied Linguistics*,14, 246-72.

Carroll, B. J. (1980). *Testing communicative performance*. Oxford: Pergamon Press.

Carroll, B. J. (1981). Specifications for an English language testing service. In J. C. Alderson, & A. Hughes (Eds.).

Carroll, B. J., & Hall, P. (1985). *Make your own language tests*. Oxford: Pergamon.

Carter, David G. (1981). Competency testing: Assessment or politics? Paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles, CA, April.

Carver, D. (1983). Some propositions about ESP. *English* for *Specific Purposes,* 2, 131-137.

Chambers, F. (1980). A re-evaluation of needs analysis. *English for Specific Purposes*, 1(1), 25-33.

Chapelle, C. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman, & A. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32-70). Cambridge: Cambridge University Press.

Chapelle, C., & Douglas, D. (1993). Interpreting second language performance data. Paper presented at Second Language Research Forum, Pittsburgh, PA.

Charge, N., & Taylor, L. B. (1997). Recent developments in IELTS. *ELT Journal*, 51(4), 374-380.

Chitravelu, N. (1980). Introduction: English for special purposes project. In British Council, *The University of Malaya English for Special Purposes Project* (V-XVI).

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Clapham, C. (1981). Reaction to the Carroll paper (1). In J. C. Alderson, & A. Hughes (Eds.).

Clapham, C. (1993). Is ESP testing justified? In D. Douglas & C. Chapelle (Eds.), *A new decade of language testing research* (pp. 257-271). Alexandria, VA: TESOL Publications.

Clapham, C. (1995). What makes an ESP reading test appropriate for its candidates? In A. Cumming, & R. Berwick (Eds.), *Validation in language testing*. Clevedon, North Somerset: Multilingual Matters.

Clapham, C. (1996). *The development of IELTS*: *a study of the effect of background knowledge on reading comprehension*. Cambridge: Cambridge University Press.

Clark, I. (1972). *Foreign language testing*: *theory and practice*. Philadelphia, PA: Center for Curriculum Development, Inc.

Clark, J. L. D. (1975). *Direct testing of speaking proficiency*: *Theory and practice*. Princeton, NJ: Educational Testing Service.

Clarke, M. A., & Silberstein, S. (1977). Toward a realization of psycholinguistic principles in the ESL reading class. *Language Learning*, 27, 135-54.

Close, R. A. (1992). *A teachers' grammar*: *The central problem of English*. Hove: Language Teaching Publications.

Coffey, B. (1984). ESP-English for specific purposes. *Language Teaching*, 17(1).

Cohen, A. (1994). *Assessing language ability in the classroom*. Boston, MA: Heinle and Heinle.

Cole, G., & Neufeld, D. (1991). Tests for evaluating the second language of public officials in Canada. *Actes du Colloque Bulletin*: *Association quebequoise des enseignants du francais langue seconde* (AQEFLS).Ottawa: University of Ottawa.

Criper, C. (1981). Reaction to the Carroll paper (2). In J. C. Alderson, & A. Hughes (Eds.).

Criper, C., & Davies, A. (1988). *ELTS validation project report.* ELTS Research Report 1(I). London and Cambridge: British Council and University of Cambridge Local Examinations Syndicate.

Cummins, J. (1979). Cognitive/academic language proficiency, linguistic interdependence, the optimum age question and some other matters. *Working Papers on Bilingualism*, 19, 121-129.

Davidson, F., & Lynch, B. (1993). Criterion-referenced language test development: a prolegomenon. In A. Huhta, K. Sajavaara, & S. Takala (Eds.), *language testing*: *new openings* (pp. 73-89). Jyvaskyla, Finland: Institute for Educational Research, University of Jyvaskyla.

Davies, A. (1981). Review of Munby's *Communicative syllabus design*. *TESOL Quarterly*, 15(3), 332-336.

Davies, A. (1990). *Principles of language testing*. Oxford: Blackwell.

Delandshere, G., & Petrosky, A. (1994). Capturing teachers' knowledge: Performance assessment a) and post-structural epistemology, b) from a poststrucruralist perspective, c) and post-structuralism, and d) none of the above. *Educational Researcher*. 23, 5 11-18.

Douglas, D. (2000). *Assessing language for specific purposes*. Cambridge: Cambridge University Press.

Douglas, D., & Chapelle, C. (Eds.). (1993). *A new decade of language testing research: Collaboration and cooperation.* Washington, DC: TESOL.

Dubetz, N., Turley, S., & Erickson, M. (1997). Dilemmas of assessment and evaluation in preservice teacher education. In A. Goodwin (Ed.), *Assessment for equity and inclusion* (pp. 197-210). London: Routledge.

Dudley-Evans, T. (1987). Introduction to genre analysis and ESP. *ELR Journal*, 1, 1-9.

Dudley-Evans T. (1997). *Developments in English for Specific Purposes*: *A Multi-disciplinary Approach*. Cambridge: Cambridge University Press.

Dudley-Evans, T. (1998). An Overview of ESP in the 1990s. Paper presented at The Japan Conference on English for Specific Purposes, Fukushima.

Dudley-Evans, T., & Henderson, W. (1990). *The language of economics: The analysis of economics discourse.* London: Modern English Publications.

Dudley-Evans, A., & St. John, M. J. (1998). *Developments in ESP*: *a multi-disciplinary approach*. Cambridge: Cambridge University Press.

Educational Testing Service. (1965). *Test of English as a Foreign Language.* Princeton, NJ: Educational Testing Service.

Educational Testing Service (1996). *TOEIC examinee handbook*. Princeton, NJ: Educational Testing Service.

Elder, C. (1993). Language proficiency as a predictor of performance in teacher education. *Melbourne Papers in Applied Linguistics*, 2, 68-89

Elder, C. (1993a). How do subject specialists construe classroom language proficiency? *Language Testing*,10(3), 235-254.

Elder, C. (1993b). The Proficiency Test for Language Teachers: Italian, Volume 1: Final report on the test development process. Melbourne: NLLIA Language Testing Centre, University of Melbourne.

Elder, C. (1993c). The Proficiency Test for Language Teachers: Italian, Volume 2: appendices. Melbourne: NLLIA Language Testing Centre, University of Melbourne.

Emmett, A. (1985). The Associated Examining Board's Test in English for Educational Purposes (TEEP) In P. C. Hauptman, R. LeBlanc, & M. B. Wesche (Eds.), *Second language performance testing* (pp. 131-151). Ottawa: Ottawa University Press.

Eskey, D. E. (1988). Holding in the bottom: An interactive approach to the language problems of second language readers. In P. L. Carrell, J. Devine, & D. E. Eskey (Eds.).

Ewer, J. R., & Hughes-Davies, E. (1971, 1972). Further notes on developing an ELT program for students of science and technology. *ELT Journal*, 26(1), 65-70.

Ewer, J. R., & Latorre, G. (1969). *A course in basic scientific English*. London: Longman.

Freebody, P., & Anderson, R. C. (1983). Effects of vocabulary difficulty, text cohesion and schema availability on reading comprehension. *Reading Research Quarterly*, 18, 272-93.

General Medical Council. (1995). Changes to the PLAB test. London: General Medical Council.

Geranpayeh, A. (1994). Are score comparisons across language proficiency test batteries justified?: An IELTS-TOEFL comparability study. *Edinburgh Working Papers in Applied Linguistics*, 5, 50-65.

Grellet, F. (1981). *Developing reading skills*. Cambridge: Cambridge University Press.

Griffin, P. (1990). Characteristics of the test components of the IELTS battery: Australian trial data. Paper presented at the Regional English Language Centre, Singapore.

Hale, G. A. (1988). Student major field and text content: Interaction effects on reading comprehension in the test of English as a foreign language. *Language Testing*, 5, 49-61.

Halliday, M. A. K., McIntosch, A., & Strevens, P. (1964). *The linguistic science and language teaching*. London: Longman.

Hamilton, J., Lopes, M., McNamara, T. F., & Sheridan, E. (1993). Rating scales and native speaker performance on a communicatively oriented EAP test. *Melbourne Papers in Language Testing*, 2, 1-24.

Hamp-Lyons, L. (1993). Components of portfolio evaluation: ESL data. Paper presented at the annual AAAL Conference, Atlanta.

Hamp-Lyons, L., & Condon, W. (1993). Questioning assumptions about portfolio-based evaluation. *College Composition and Communication*, 44(2), 176-190.

Harley, B., Allen, P., Cummins, J., & Swam, M. (Eds.). (1990). *The development of second language proficiency*. New York: Cambridge University Press.

Henning, G., & Cascallar, E. (1992). *A preliminary study of the nature of communicative competence*. Princeton, NJ: Educational Testing Service.

Herbert, A. J. (1965). *The structure of technical English*. London: Longman.

Herter, R. (1991). Writing portfolios: Alternatives to testing. *English Journal*, January, 90-91.

Hopstock, P. J., Bucaro, B. J., Fleischman, H. L., Zehler, A. M., & Eu, H. (1993). *Descriptive study of services to limited English proficient students*. *Volume II: Survey results.* Report to the U.S. Department of Education, Office of Policy and Planning. Arlington, VA: Development Associates, Inc.

Hornberger, N. H. (1989). Tramites and Transportes: the acquisition of second language communicative competence for one speech event in Puno, Peru. *Applied Linguistics*, 10, 214-230.

Howell, David C. (1995). *Fundamental statistics for the behavioral sciences*. Belmont, California: International Thomson Publishing.

Hughes, A. (1989). *Testing for language teachers*. Cambridge: Cambridge University Press.

Hughes, A., Porter, D., & Weir, C. J. (Eds.). (1988). *ELTS validation project: Proceedings of a Conference Held to Consider the ELTS Validation Project report.* ELTS Research Report l(II). London and Cambridge: British Council and University of Cambridge Local Examinations Syndicate.

Hymes, D. (1971). Competence and performance in linguistic theory. In R. Huxley, & E. Ingram (Eds.), *Language acquisition*: *models and methods* (pp. 3-24). London: Academic Press.

Hymes, D. (1972). On communicative competence. In J. B. Pride, & J. Holmes (Eds.), *sociolinguistics* (pp. 269-292). Harmondsworth, UK: Penguin Books.

Hymes, D. (1974). *Foundations in sociolinguistics*: *an ethnographic approach*. Philadelphia: University of Pennsylvania Press.

Hutchinson, T., & Waters, A. (1987). *English for specific purposes*: *A learning centered approach*. Cambridge: Cambridge University Press.

Ingram, D. E. (1990). The International English Language Testing System (*IELTS*): Its nature and development. Eric No. ED342211.

Ingram, D. E. (1991). The International English Language Testing System: the speaking test. *Working Papers of the National Languages Institute of Australia*, 1(1), 101-114.

Institute of Air Navigation Services. (1994). *PELA*: *a test in the Proficiency in English Language for Air Traffic Control*. Luxembourg: Institute of Air Navigation Services.

International Civil Aviation Organization (ICAO). (1985). *Rules of the air and air traffic services*. Doc. 444-RAC/501.12.

Iowa State University, Graduate College (no date (a)). *Instructions for preparing for the TEACH test.* Xerox; Ames, IA: Iowa State University.

Jensen, C., & Hansen, C. (1995). The effect of prior knowledge on EAP listening-test performance. *Language Testing*, 12, 99-119.

Johns, A. M. (1990). ESP in the USA: State of the art. Talk given at the second Latin American ESP Colloquium, Santiago de Chile.

Johns, A. M. (1991). English for specific purposes (ESP): Its history, contributions, and future. In Celce-Murcia, M. (Ed.), *Teaching English as a foreign language*. New York: Newbury House.

Johns, A., & Dudley-Evans, T. (1991). English for Specific Purposes: International in scope, specific in purpose. *TESOL Quarterly*, 25, 297-314.

Johnson, D. W., Johnson, R. T., & Holubec, E. J. (1993). *Circles of learning*. Edina, MN: Interaction Book Company.

Johnson, K. (1995). *Language teaching and skill learning*. Oxford: Basil Blackwell.

Jones, R. L. (1977). Testing a vital connection. In J. Phillips (Ed.), *The language connection*: *From the classroom to the world* (pp. 237-265). Skokie, IL: National Textbook Company.

Jones, R. L. (1985). Second language performance testing. In P. C. Hauptman, R. LeBlanc & M. B. Wesche (Eds.), *Second language performance testing* (pp. 15-24). Ottawa: University of Ottawa Press. 15-24.

Jones, G. M. (1990). ESP textbooks: Do they really exist? *English for Specific Purposes*, 9(1), 89-93.

Kattan, J. (1990). *The construction and validation of an EAP test for second year English and nursing majors at Bethlehem University.* Unpublished Ph.D. Dissertation, Lancaster University, Lancaster.

Kintsch, W. (1974). *The representation of meaning in memory*. Hillsdale, NJ: Lawrence Erlbaum.

Koh, M. Y. (1985). The role of prior knowledge in reading comprehension. *Reading in a foreign language*, 3, 375-380.

Kohonen, V. (1997). Authentic assessment as an integration of language learning, teaching, evaluation and the teacher's professional growth. In A. Huhta, V. Kohonen, L. Kurki-Suonio, & S. Luoma (Eds.), *Current developments and*

alternatives in language assessment: *proceedings of LTRC 96* (pp. 7-22). Jyvaskyla, Finland: University of Jyvaskyla.

Kramsch, C. (1993). *Context and culture in language teaching*. Oxford: Oxford University Press.

Kunnan, A. J. (1998). *Validation in language assessment*: *Selected papers from the 17th testing research colloquium*. London: Lawrence Erlbaum Associates, Publishers.

Lado, R. (1961). *Language testing: The construction and use of foreign language tests. A teacher's book.* New York, N.Y: McGraw-Hill Book Company.

Lazaraton, A. (1994). Question turn modification in language proficiency interviews. Paper presented at the annual American Association for Applied Linguistics Conference Baltimore, MD.

Lee, P., & Stuart, M. (1997). Making a video. *Journal of Geography in Higher Education*, 21 (1), 127-34.

Lewkowicz, J. (1997). Investigating authenticity in language testing. Unpublished doctoral thesis, University of Lancaster.

Linn, R. L., Baker, E., & Dunbar, S. B. (1991). Complex performance-based assessment: expectations and validation criteria. *Educational Researcher*, 20(8), 15-24.

Lipson, M. Y. (1984). Some unexpected issues in prior knowledge and comprehension. *The reading teacher*, April, 760-764.

Lockwood, J. (2001). ELT curriculum design in Hong Kong workplaces: Interdisciplinary implications for ESP teacher education. *Hong Kong Journal of Applied Linguistics*, 5(2), 83-93.

Lumley, T. (1993). Reading comprehension sub-skills: Teacher's perceptions of content in an EAP test. *Melbourne Papers in Applied Linguistics*, 2, 25-55.

Lumley, T. (1993). The notion of subskills in reading comprehension tests: an EAP example. *Language Testing*, 10, 211-34.

Lynch, T. (1994). The University of Edinburgh test of English at matriculation: Validation report. *Edinburgh Working Papers in Applied Linguistics*, 5, 66-77.

Lynch, T., & Anderson, B. (1991). *Study speaking*. Cambridge: Cambridge University Press.

Mackay, R., & Mountford, A. (Eds.). (1978). *English for specific purposes*. London: Longman.

Madaus, G., Raczek, A., & Clarke, M. (1997). The historical and policy foundations of the assessment movement. In A. Goodwin (Ed.), *Assessment for equity and inclusion* (pp. 1-34). London: Routledge.

Manning, W. (1987). *Development of cloze elide Tests of English as a second language*. Princeton, NJ: Educational Testing Service.

Mason, D. (1989). An examination of authentic dialogues for use in the ESP classroom. *English* for *Specific Purposes*, 8(1), 85-92.

McDowell, C. (1995). Assessing the language proficiency of overseas-qualified teachers: the English Language Skills Assessment (ELSA). In G. Brindley (Ed.), *Language assessment in action* (pp. 11-29). Sydney: NCELTR.

McNamara, T. F. (1989). ESP testing: general and particular. In C. Candlin, & T. McNamara (Eds.), *Language, learning, and community* (pp. 125-142). Sydney, NSW: National Centre for English Language Teaching and Research, Macquarie University.

McNamara, T. F. (1990a). Assessing the second language proficiency of health professionals. Unpublished Doctoral Dissertation, Department of Linguistics and Language Studies, University of Melbourne.

McNamara, T. F. (1990b). Item response theory and the validation of an ESP test for health professionals. *Language Testing*, 7, 52-76.

McNamara, T. F. (1991). Test dimensionality: IRT analysis of an ESP listening lest. *Language Testing*, 8, 139-159.

McNamara, T. F. (1996). *Measuring second language performance*. London: Longman.

McNamara, T. F. (1996). *Second language performance assessment*: *Theory and research*. London: Longman.

Mead, R. (1982). Review of Munby: *Communicative syllabus design. Applied Linguistics*, 3, 70-78.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23(2), 13-23.

Moller, A. (1981). *A study in the validation of proficiency tests of English as a foreign language*. Unpublished Ph.D. Dissertation. University of Edinburgh, Edinburgh.

Morrow, K. (1977). Authentic tests and ESP. In S. Holden (Ed.), *English for specific purposes*. London: Modern English Publications.

Morrow, K. (1980). *Skills for reading*. Oxford: Oxford University Press.

Moss, P. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research*, 62, 229-258.

Moss, P. (1994). Can there be validity without reliability? *Educational Researcher*, 23(2), 5-12.

Moya, S., & O'Malley, J. M. (1994). A portfolio assessment model for ESL. *The journal of Educational Issues of Language Minority Students*, 13, 13-36.

Munby, J. (1978). *Communicative syllabus design*. Cambridge: Cambridge University Press.

National Languages and Literacy Institute of Australia Language Testing Centre. (1992). *The Japanese Language Test for Tour Guides*. *Handbook for candidates*. Melbourne: National Languages and Literacy Institute of Australia.

National Languages and Literacy Institute of Australia Language Testing Centre. (1993). *Proficiency Test for Teachers*: *Italian. Draft handbook for candidates*. Melbourne: National Languages and Literacy Institute of Australia.

North, B. (1994). *Scales of language proficiency, a survey of some existing systems.* Strasbourg: Council of Europe CC-LANG (94) 24.

Northern Examinations and Assessment Board (NEAB). (1996a). *University Entrance Test in English for Speakers of Other Languages. Listening Skills.* Audiotape, 19 June. Manchester: NEAB.

Northern Examinations and Assessment Board (NEAB). (1996e). *University Entrance Test in English for Speakers of Other Languages. Written English.* Question Paper, 19 June. Manchester: NEAB.

Nuttall, C. (1982). *Teaching reading skills in a foreign language*. London: Heinemann.

Oller, J. W. (1976). Evidence for a general language proficiency factor: An expectancy grammar. *Die Neueren Sprachen*, 2, 165-74.

Oller, J. W. (1979). *Language tests at school*. London: Longman.

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.

Peretz, A. S., & Shoham, M. (1990). Testing reading comprehension in LSP. *Reading in a Foreign Language*. 7, 447-55.

Perfetti, C. A., & Lesgold, A. M. (1977). Discourse comprehension and sources of individual differences. In M. A. Just, & P. A. Carpenter (Eds.).

Perren, G. (1974). Forward in Teaching languages to adults for special purposes. *CILT Reports and Papers*, 11, London: CILT.

Phelps, A. (1997). Portfolio assessment in high school chemistry: one teacher's guidelines. *Journal of Chemical Education*, 74(5), 528-531.

Pichert, J. W., & Anderson, R. C. (1977). Taking different perspectives on a story. *Journal of Educational Psychology*, 69, 309-15.

Pollari, P. (1997). Could portfolio assessment empower ESL learners? Portfolios in the teaching of English as a foreign language in Finnish upper secondary school. In A. Huhta, V. Kohonen, L. Kurki-Suonio, & S. Luoma (Eds.), *Current developments and alternatives in language assessment: proceedings of LTRC 96* (pp. 37-54). Jyvaskyla, Finland: University of Jyvaskyla.

Rae, P. (1985). Language testing and the communicative language teaching curriculum. In Y. Lee, C. Fok, & G. Low (Eds.), *New directions in language testing* (pp. 15-32). Oxford: Pergamon.

Rea-Dickins, P. (1987). Testing doctors' written communicative competence: an experimental technique in English for specialist purposes. *Quantitative Linguistics*, 34,185-218.

Robinson, P. (1991). *ESP today: A practitioner's guide.* New York, NY: Prentice Hall.

Rodgers, C. (1969). *Freedom to learn*. Columbus, OH: Charles Merrill.

Ross, S. (1992). Accommodative questions in oral proficiency interviews. *Language Testing*, 9, 173-186.

Ross, S. (1994). Formulaic speech in language proficiency interviews. Paper presented at the annual American Association for Applied Linguistics Conference. Baltimore, MD.

Royal Society of Arts (RSA) Examinations Board. (1994a). *CBLC-Certificate in Business Language Competence*: *syllabus*. Coventry, UK: RSA.

Sajavaara, K. (1992). Designing tests to match the needs of the workplace. In E. Shohamy, & A. Walton (Eds.), *Language assessment for feedback*: *testing and other strategies* (pp. 123-144). Dubuque, IA: Kendall/Hunt.

Savitch, J., & Serling, L. (1997). I wouldn't know I was smart if I didn't come to this class. In A. Goodwin (Ed.), *Assessment for equity and inclusion* (pp. 141-162). London: Routledge.

Scott, M., Stansfield, C., & Kenyon, U. (1996). Examining validity in a performance test: the Listening Summary Translation Exam (LSTE)–Spanish version. *Language Testing*, 13(1), 83-109.

Selinker, L. (1979). Interlanguage. *International Review of Applied Linguistics,* 10, 209-231.

Selinker, L., *et al.*, (1970). Grammar and technical English. In R. C. Lugton (Ed.), *English Teaching Forum*, October, 1972, 3-14.

Selinker, L., Trimble, T., & Trimble, L. (1976). Presuppositional information in EST discourse. *TESOL Quarterly*, 10(3), 281-290.

Selinker, L., & Trimble, L. (1976). Scientific and technical writing: The choice of tense. *English Teaching Forum*, 14(4).

Sheridan, E. (1991). *A comparison of native/non-native speaker performance on a communicative test of writing ability (IELTS).* Melbourne: University of Melbourne. (Unpublished MA thesis).

Shoham, M., Peretz, A. S., & Vorhaus, R. (1987). Reading comprehension tests: General or subject specific? *System*, 15, 81-8.

Shohamy, E. (1994). The validity of direct versus semi-direct oral tests. *Language Testing*, 11, 99-124.

Shohamy, E. (1995). Performance assessment in language testing. *Annual Review of Applied Linguistics*, 15, 188-211.

Shohamy, E., Donitze-Schmidt, S., & Waizer, R. (1983). The effect of the elicitation method on the language samples obtained on oral tests. Paper presented at the annual Language Testing Research Colloquium. Cambridge, 1983.

Shohamy, E., Gordon, C., Kenyon, D., & Stansfield, C. (1989). The development and validation of a semi-direct test for assessing oral proficiency in Hebrew. *Bulletin of Higher Hebrew Education*, 4.

Shohamy, E., & Stansfield, C. (1991). The Hebrew oral test: An example of international cooperation. *A1LA Bulletin*, 7, 79-90.

Slater, T. (1994). Portfolios for learning and assessment in physics. *Physics Teacher*, 32(6), 370-373.

Skehan, P. (1984). Issues in the testing of English for specific purposes. *Language Testing*, 1, 211-220.

Skehan, P. (1989). Language Testing Part I & Part II. *Language Teaching*, 22,1-13.

Spath Hirschmann, S., & Traversa, A. (1997). When a portfolio program adopts the portfolio. In A. Huhta, V. Kohonen, L. Kurki-Suonio, & S. Luoma (Eds.), *Current developments and alternatives in language assessment*: *proceedings of LTRC 96* (pp. 55-70). Jyvaskyla, Finland: University of Jyvaskyla.

Spilich, G. J., Vesonder, G. T., Chiesi, H. L., & Voss, J. F. (1979), Text processing of domain-related information for individuals with high and low domain knowledge. *Journal of Verbal Learning and Verbal Behavior*, 18, 275-290.

Spolsky, B. (1973). What does it mean to know a language? Or, how do you get someone to perform his competence? In J. W. Oller, & J. Richards (Eds.), *Focus on the learner*: *pragmatic perspectives for the language teacher* (pp. 164-176). Rowley, MA: Newbury House.

Spolsky, B. (1986). A multiple choice for language testers. *Language Testing*, 3, 147-158.

Spolsky, B. (1989). Communicative competence, language proficiency, and beyond. *Applied Linguistics*, 1(1), 138-156.

Spolsky, R. (1995). *Measured words*: *the development of objective language testing*. Oxford: Oxford University Press.

Stanovich, K. E. (1991). Word recognition: Changing perspectives. In R. Barr, M. Kamil, P. Mosenthal, & P. D. Pearson (Eds.).

Stansfield, C., & Kenyon, D. (1988). *Development of the Portuguese speaking test.* Washington, DC: Center for Applied Linguistics.

Steffensen, M. S., Joag-Dev, C. & Anderson, R. C. (1979). A cross cultural perspective on reading comprehension. *Reading Research Quarterly*, 15, 10-29.

Strevens, P. (1977). *New orientations in the teaching of English.* Oxford: Oxford University Press.

Strevens, P. (1980). Functional English (ESP). In *Teaching English as an international language*: *From practice to principle* (pp. 105-121). Oxford: Pergamon Press.

Strevens, P. (1988). The learner and teacher of ESP. In Chamberlain & Baumgardner, 6, 39-44.

Strevens, P. (1988). ESP after twenty years: A re-appraisal. In M. Tickoo (Ed.), *ESP: State of the Art* (pp. 1-13). Singapore: SEAMEO Regional Centre.

Strevens, P., & Johnson, E. (1983). SEASPEAK: A project in applied linguistics, language engineering, and eventually ESP for sailors. *English for Specific Purposes*, 2(2), 123-29.

Swain, M., & Canale, M. (1982). The role of grammar in a communicative approach to second language teaching and testing. In *Issues of Language Assessment: Foundations and Research*. Proceedings of the Annual Language Assessment Institute, Illinois.

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. Gass, & C. Madden (Eds.), *Input in second language acquisition*. Rowley, MA: Newbury House.

Swales, J. (1971). *Writing scientific English*. Walton-on-Thames: Nelson.

Swales, (1985). *Episodes in ESP*. Oxford: Pergamon Institute of English.

Swales, J. M. (1989). Service English program design and opportunity cost. In Johnson (Ed.), *The second language curriculum* (pp. 79-90).

Swales, J. M. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

Swales, J. (1995). The role of the textbook in EAP writing research. *English* for *Specific Purposes*, 14(1), 3-18.

Symons, S., & Pressley, M. (1993). Prior knowledge affects text search success and extraction of information. *Reading Research Quarterly*, 28, 251-9.

Tan, S. (1990). The role of prior knowledge and language proficiency as predictors of reading comprehension among undergraduates. In J. De Jong, & D. Stevenson (Eds.), *Individualizing the assessment of language abilities* (pp. 214-224). Clevedon, UK: Multilingual Matters.

Tannen, D. (1979). What's in a frame? In R. O. Freedle (Ed.).

Teasdale, A. (1993). Authenticity, validity, and task design for tests of well-defined LSP domains. *Thames Valley University Working Papers In English Language* Teaching, 2, 137-150.

Teasdale, A. (1996). Content validity in tests for well-defined LSP domains: An approach to defining what is to be tested. In M. Milanovic, & N. Saville (Eds.), *Studies in language testing*. Cambridge: Cambridge University Press.

Tickoo, M. L. (Ed.). (1988). *ESP*: *State of the art*. Singapore: SEAMEO Regional English Language Centre.

Todesco, A., Cahill, W., & Frey, G. (1983). *A system of evaluation for the new English program*. Ottawa: Language Training Canada.

University of Cambridge Local Examinations Syndicate (UCLES). (1990). *International English Language Testing System. Module A. Specimen version.* Cambridge: UCLES.

University of Cambridge Local Examinations Syndicate (UCLES). (1994). *An Introduction to IELTS.* Cambridge: UCLES.

University of Cambridge Local Examinations Syndicate (UCLES). (1995a). *Certificate of Proficiency In English.* Cambridge: UCLES.

University of Cambridge Local Examinations Syndicate (UCLES). (1995b). *International English Language Testing System: specimen materials. Academic reading.* Cambridge: UCLES.

University of Cambridge Local Examinations Syndicate (UCLES). (1995c). *International English Language Testing System: specimen materials. Academic writing.* Cambridge: UCLES.

University of Cambridge Local Examinations Syndicate (UCLES). (1995d). *International English language Testing System: specimen materials. Listening. Text and Audiotape*. Cambridge: UCLES.

University of Cambridge Local Examinations Syndicate (UCLES). (no date). *CEIBT content and administrative information.* Cambridge: UCLES.

University of Oxford Delegacy of Local Examinations. (1990). *OIBEC: executive level syllabus and specimen material.* Oxford: University of Oxford.

Upshur, J. (1971). Productive communication testing: a progress report. In G. Perren, & J. Trim (Eds.), *Applications in linguistics* (pp. 435-442). Cambridge: Cambridge University Press.

Ur, P. (1996). *A course in language teaching*: *Practice and theory*. Cambridge: Cambridge University Press.

Urquhart, A. H. (1984). The effect of rhetorical ordering on readability. In J. C. Alderson, & A. H. Urquhart (Eds.).

Van-Lier, L. (1989). Reeling, writhing, drawling, stretching and fainting in coils: Oral proficiency interviews as conversations. *TESOL Quarterly*, 23, 489- 508.

Vesonder, G. T. (1979). *The role of knowledge in processing experimental reports.* Unpublished Ph.D. Dissertation, University of Pittsburgh, PA.

Vollmer, H. J. (1981). Why are we interested in 'General Language Proficiency'? In J. C. Alderson, & A. Hughes. (Eds.).

Vollmer, H. J., & Sang, F. (1983). Competing hypotheses about second language ability: A plea for caution. In J. W. Oller (Ed.), *Issues in language testing research* (29-75). Rowley, MA: Newbury House.

Wallace, C. (1997). Global implications of curriculum and materials design. *ELT Journal*, 51 (4), 370-73.

Waters, A. (Ed.). (1982). *Issues in ESP*. Oxford: Pergamon Press.

Waters, M., & Waters, A. (1992). Study skills and study competence: Getting the priorities right. *ELT Journal*, 46(3), 264-73.

Weigle, S. C. (2001). *Assessing writing*. Cambridge: Cambridge University Press.

Weir, C. J. (1983). *Identifying the language problems of overseas students in tertiary education in the United Kingdom*. Unpublished Ph.D. Dissertation, University of London, London.

Weir, C. (1988). The specification, realization and validation of an English language proficiency test. In A. Hughes (Ed.), *Testing English for university study* (45-110). London: Modern English Publications.

Weir, C. (1990). *Communicative language testing.* New York: Prentice Hall.

Weir, C. (1993). *Understanding and developing language tests.* New York: Prentice Hall.

Weiyun He, A. (1994). Elicited versus volunteered elaboration: Talk and task in language proficiency interviews. Paper presented at the annual American Association for Applied Linguistics Conference. Baltimore, MD.

Wesche, M. (1987). Second language performance testing: The Ontario Test of ESL as an example. *Language Testing*, 4, 28-47.

Wesche, M. (1992). Performance testing for work-related second language assessment. In E. Shohamy, & R. Walton (Eds.), *Language assessment for feedback*: *Testing and other strategies* (103-122). Kendall/Hunt Publishing Company.

Wesche, M., & Cray, E. (1984). The Ontario Test of English as a Second Language (OTESL): A Progress Report. *TESL Talk*, 15(3), 17-21.

Widdowson, H. G. (1977a). The communicative approach and its application. In Moody & Moore (Eds.), *English for specific purposes: An international seminar*. Bogota: The British Council.

Widdowson, H. G. (1977b). Approaches to discourse. In C. Gutknecht (Ed.), *Grundbegriffe und Hauptstromungen der linguistik*, Hamburg: Hoffmann und Campe.

Widdowson, H. G. (1978). *Teaching language as communication.* Oxford: Oxford University Press.

Widdowson, H. G. (1979). *Explorations in applied linguistics.* Oxford: Oxford University Press.

Widdowson, H. G. (1983). *Learning purpose and language use*. Oxford: Oxford University Press.

Wigglesworth, G., & O'Loughlin, K. (1993). An investigation into the comparability of direct and semi-direct versions of an oral interaction test in English. *Melbourne Papers in Applied Linguistics*, 2, 56-67.

Williams, M. (1988). Language taught for meetings and language used in meetings: Is there anything in common? *Applied Linguistics*, 9(1), 45-58.

Wilkins, D. (1976). *Notional syllabuses*. Oxford: Oxford University Press.

IELTS

SECTION 1 QUESTIONS 1-13

QUESTIONS 1-5

Look at the information below about the use of vehicles in the University grounds.

In boxes 1-5 on your answer sheet write

TRUE *if the statement is true*

FALSE *if the statement is false*

NOT GIVEN *if the information is not given in the passage*

EXAMPLE	Answer
The campus roads are not open to general members of the public.	TRUE

1 University employees do not need to pay for their parking permits.

2 Parking in Halls of Residence is handled by the Wardens of the Halls.

3 Having a University permit does not allow staff to park at Halls.

4 Parking permits cost £20 a year.

5 Students living in Hall do not need permission to park in Hall car parks.

USE OF UNIVERSITY GROUNDS BY VEHICULAR TRAFFIC

The University grounds are private.

The University authorities only allow authorised members of the University, visitors and drivers of vehicles servicing the University to enter the grounds.

Members of staff who have paid the requisite fee and display the appropriate permit may bring a vehicle into the grounds. A University permit does not entitle them to park in Hall car parks however, unless authorised by the Warden of the Hall concerned.

Students may not bring vehicles into the grounds during the working day unless they have been given special permission by the Security Officer and have paid for and are displaying an appropriate entry permit. Students living in Halls of Residence must obtain permission from the Warden to keep a motor vehicle at their residence.

Students are reminded that if they park a motor vehicle on University premises without a valid permit they will be fined £20.

QUESTIONS 6-13

Look at the patient information leaflet below.
*Match each of the following sentences with **TWO** possible endings **A-M** from the box below.*
*Write the appropriate letters **A-M** in boxes 6-13 on your answer sheet.*

EXAMPLE	Answer
Borodine tablets should not be given to....	**A and M**

Questions 6 and 7

Borodine tablets might be used to treat ...

Questions 8 and 9

You must ask your doctor **before** taking Borodine tablets if you are already being treated for ...

Questions 10 and 11

You do not need to consult your doctor immediately if Borodine tablets give you ...

Questions 12 and 13

You must consult your doctor at once if you find Borodine tablets cause ...

Possible Endings

A	children under 12 years of age.
B	a headache.
C	an uncomfortable feeling in your stomach.
D	symptoms similar to a cold.
E	a change in your skin color.
F	anything treated by a prescription medicine.
G	a kidney complaint.
H	a whitening of the eye.
I	sore or broken skin.
J	a fungal infection.
K	a feeling of sadness.
L	shortness of breath.
M	a woman expecting a child.

PATIENT INFORMATION LEAFLET

The name of your medicine is ***Borodine*** tablets

WHAT ARE *Borodine* TABLETS USED FOR?

Borodine tablets are used to help relieve hay fever and conditions due to allergies, in particular skin reactions and a runny nose.

It is not recommended that ***Borodine*** tablets are given to children under 12 years of age or pregnant or breastfeeding women.

BEFORE YOU TAKE ***Borodine*** TABLETS

In some circumstances it is very important not to take ***Borodine*** tablets. If you ignore these instructions, this medicine could affect your heart rhythm.

Are you taking oral medicines for fungal infections?

Have you suffered a reaction to medicines containing ***Borodine*** before?

Do you suffer from any liver, kidney, or heart disease?

If the answer to any of these questions is YES, do not take ***Borodine*** tablets before consulting your doctor.

AFTER TAKING ***Borodine*** TABLETS

Borodine tablets, like many other medicines, may cause side-effects in some people.

If you faint, stop taking *Borodine* tablets and tell your doctor immediately.

In addition ***Borodine*** tablets may cause problems with your vision, hair loss, depression or confusion, yellowing of your skin or your eyes.

If you have these effects whilst taking *Borodine* tablets tell your doctor immediately.

Other side-effects are dizziness or headaches, and indigestion or stomach ache. However, these effects are often mild and usually wear off after a few days' treatment. If they last for more than a few days, tell your doctor.

SECTION 2 QUESTIONS 14-20

Questions 14-20

*Look at the introduction to West Thames College below and at the following statements **(Questions 14-20)**.*

In boxes 14-20 on your answer sheet write

TRUE *if the statement is true*

FALSE *if the statement is false*

NOT GIVEN *if the information is not given in the passage*

14 Chiswick Polytechnic was closed at the same time West Thames College was opened.

15 Most of the students at the college come from outside the local area.

16 The college changed its name to West Thames College in 1993.

17 There are currently 6000 students over the age of 19 attending the college.

18 Students under the age of 16 cannot attend any of the courses offered by the college.

19 The college offers a more mature environment in which to learn than a school.

20 There are fewer subjects to study in the sixth form of a school than at the college.

WEST THAMES COLLEGE
BACKGROUND INFORMATION FOR CANDIDATES

West Thames College (initially known as Hounslow Borough College) came into existence in 1976 following the merger of Isleworth Polytechnic with part of Chiswick Polytechnic. Both parent colleges, in various guises, enjoyed a long tradition of service to the community dating back to the 1890s.

The college is located at London Road, Isleworth, on a site occupied by the Victorian house of the Pears family, Spring Grove House. An earlier house of the same name on this site had been the home of Sir Joseph Banks, the botanist who named Botany Bay with Captain Cook in 1770. Later he founded Kew Gardens. Situated at the heart of West London, West Thames College is ideally placed to serve the training and education needs of local industry and local people. But its influence reaches much further than the immediate locality.

Under its former name, Hounslow Borough College, it had already established a regional, national, and international reputation for excellence. In fact, about eight per cent of its students come from continental Europe and further afield, whilst a further 52 per cent are from outside the immediate area. Since 1 April 1993, when it became independent of the local authority and adopted its new title, West Thames College has continued to build on that first class reputation.

These days there is no such thing as a typical student. More than half of West Thames College's 6000 students are over 19 years old. Some of these will be attending college part-time under their employers' training schemes. Others will want to learn new skills purely out of interest, or out of a desire to improve their promotion chances, or they may want a change in career.

The college is also very popular with 16-18 year olds, who see it as a practical alternative to a further two years at school. They want to study in the more adult atmosphere the college provides. They can choose from a far wider range of subjects than it would be practical for a sixth form to offer. If they want to go straight into employment they can still study at college to gain qualifications relevant to the job, either on a day-release basis or through network or the Modern Apprenticeship Scheme.

Questions 21-26

*Look at the West Thames College's Service for Students below. Each paragraph **A-H** describes a different service provided by the college. From the following list (**i-xi**) choose the most suitable summaries for the paragraphs **A, C,** and **E-H.***

*Write the appropriate numbers (**i-xi**) in boxes 21-26 on your answer sheet.*

***NB** There are more summaries than paragraphs, so you will not use them all.*

i	A shop for the books and the stationery needed to study
ii	Counseling and welfare willing to listen, offer advice or arrange a referral
iii	An Examinations Office arranging exams and issuing certificates
iv	A Registrar's Office handling all fee payments and related enquiries
v	A Medical Service offering on-site assistance with health-related problems
vi	A tutorial system for regular one-to-one guidance, support and feedback
vii	Careers Advice helping students into employment
viii	An Admission Service providing assistance in choosing and applying for higher education courses
ix	A Student Union representing students on college committees
x	Clubs and societies for students' free time
xi	A Learning Support Service supporting students in studying, presenting information and handling numbers

21 Paragraph **A**

EXAMPLE	Answer
Paragraph **B**	**xi**

22 Paragraph **C**

EXAMPLE	Answer
Paragraph **D**	**i**

23 Paragraph **E**

24 Paragraph **F**

25 Paragraph **G**

26 Paragraph **H**

WEST THAMES COLLEGE SERVICES FOR STUDENTS

A
As a full-time student at West Thames College you will have your own personal mentor who will see you each week to guide you through your studies, and discuss any problems which may arise. We take a cooperative approach to the assessment of your work and encourage you to contribute to discussion.

B
This service provides specialist assistance and courses for those who need help to improve their writing, oral and numeracy skills for the successful completion of their college courses. Help with basic skills is also available.

C
This service is available to anyone who is undecided as to which course to follow. It is very much a service for the individual, whatever your age, helping you to select the best option to suit your circumstances. The service includes educational advice, guidance and support, including a facility for accrediting your previous experience—the Accreditation of Prior Learning (APL). The Admission Office is open Monday to Friday 9.00 am to 5.00 pm. All interviews are confidential and conducted in a relaxed and friendly atmosphere. Evening appointments are available on request.

D
The College Bookshop stocks a wide range of books, covering aspects of all courses, together with a good selection of stationery. It also supplies stamps, phone cards, blank videos and computer disks. The shop is open at times specified in the Student Handbook in the mornings, afternoons and evenings.

E
When students are weary from study and want the chance to relax and enjoy themselves with friends, they can participate in a number of recreational activities. Depending on demand, we offer a range of sporting activities including football, badminton, basketball, table tennis, volleyball, weight training and aerobics. For the non-sporting students we offer a debating society, video club, hair and beauty sessions, as well as a range of creative activities. Suggestions for activities from students are always welcome.

F
This confidential service is available if you have practical or personal difficulties during your course of study, whether of a financial or personal nature. Our student advisors can help you directly or put you in touch with someone else who can give you the help you need.

G
The College Nurses are there for general medical advice and for treatment of illness or injury. All visits are confidential. First aid boxes and fully-trained First Aiders are also on hand at various locations around the college.

H
West London employers have a permanent base in the center of college, with access to a database of more than 24,000 jobs available locally and in Central London. They will also help you with job applications and interview techniques.

SECTION 3 Questions 27-40

Read the following passage and answer questions 27-40.

The Discovery of Uranus

Someone once put forward an attractive though unlikely theory. Throughout the Earth's annual revolution around the Sun there is one point of space always hidden from our eyes. This point is the opposite part of the Earth's orbit, which is always hidden by the sun. Could there be another planet there, essentially similar to our own, but always invisible?

If a space probe today sent back evidence that such a world existed it would cause not much more sensation than Sir William Herschel's discovery of a new planet, Uranus, in 1781.

Herschel was an extraordinary man—no other astronomer has ever covered so vast a field of work—and his career deserves study. He was born in Hanover in Germany in 1738, left the German army in 1757, and arrived in England the same year with no money but quite exceptional music ability. He played the violin and oboe and at one time was organist in the Octagon Chapel in the city of Bath. Herschel's was an active mind, and deep inside he was conscious that music was not his destiny; he therefore read widely in science and the arts, but not until 1772 did he come across a book on astronomy. He was then 34, middle-aged by the standards of the time, but without hesitation he embarked on his new career, financing it by his professional work as a musician. He spent years mastering the art of telescope construction, and even by present-day standards his instruments are comparable with the best.

Serious observation began in 1774. He set himself the astonishing task of 'reviewing the heavens', in other words, pointing his telescope to every accessible part of the sky and recording what he saw. The first reviewing was made in 1775; the second, and most momentous, in 1780-81. It was during the later part of this that he discovered Uranus. Afterwards, supported by the royal grant in recognition of his work, he was able to devote himself entirely to astronomy. His final achievements spread from the Sun and Moon to remote galaxies (of which he discovered hundreds), and papers flooded from his pen until his death in 1822.

Among these there was one sent to the Royal Society in 1781, entitled *An Account of a Comet*. In his own words: On Tuesday the 13th of March, between 10 and 11 in the evening, while I was examining the small stars in the neighborhood of H Geminorum, I perceived one that appeared visibly larger than the rest; being struck with its uncommon magnitude, I compared it to H Geminorum and the small star in the Quartile between Aurigi and Gemini, and finding it to be much larger than either of them, suspected it to be a comet.

Herschel's care was the hallmark of a great observer; he was not prepared to jump to any conclusions. Also, to be fair, the discovery of a new planet was the last thought in anybody's mind. But further observation by other astronomers besides Herschel revealed two curious facts. For a comet, it showed a remarkably sharp disc; furthermore, it was moving so slowly that it was thought to be a great distance from the sun, and comets are only normally visible in the immediate vicinity of the sun. as its orbit came to be worked out the truth dawned that it was a new plant far beyond Saturn's realm, and that the 'reviewer of the heavens' has stumbled across an unprecedented prize. Herschel wanted to call it georgium sidus (Star of George) in honor of his royal patron King George III of Great Britain. The planet was later for a time called Herschel in honor of its discoverer. The name Uranus, which was first proposed by the German astronomer Johann Elert Bode, was in use by the late 19th century.

Uranus is a giant in construction, but not so much in size; its diameter compares unfavorably with that of Jupiter and Saturn, though on the terrestrial scale it is still colossal. Uranus' atmosphere consists largely of hydrogen and helium, with a trace of methane. Through a telescope the planet appears as a small bluish-green disc with a faint green periphery. In 1977, while recording the occultation of a star behind the planet, the American astronomer James L.

Elliot discovered the presence of five rings encircling the equator of Uranus. Four more rings were discovered in January 1986 during the exploratory flight of *Voyager2*. In addition to its rings, Uranus has 15 satellites ('moons'), the last ten discovered by *Voyager2* on the same flight; all revolve about its equator and move with the planet in an East-West direction. The two largest moons, Titania and Oberon, were discovered by Herschel in 1787. The next two, Umbriel and Ariel, were found in 1851 by the British astronomer William Lassell. Miranda thought before 1986 to be the innermost moon, was discovered in 1948 by the American astronomer Gerard Peter Quiper.

Questions 27-31

Complete the table below.

Write a date for each answer.

Write your answers in boxes 27-31 on your answer sheet.

EVENT	DATE
Example William Herschel was born	**Answer** 1738
Herschel began investigating astronomy	(27) ……
Discovery of the planet Uranus	(28) ……
Discovery of the moons Titania and Oberon	(29) ……
First discovery of Uranus' rings	(30) ……
Discovery of the last 10 moons of Uranus	(31) ……

Questions 32-36

Do the following statements reflect the claims of the writer of the Reading Passage?

In boxes 32-36 on your answer sheet write

YES *if the statement is true*

NO *if the statement is false*

NOT GIVEN *if the information is not given in the passage*

EXAMPLE	Answer
Herschel was multi-talented	**YES**

32 It is impossible that there is a planet hidden behind the Sun.

33 Herschel knew immediately that he had found a new planet.

34 Herschel collaborated with other astronomers of his time.

35 Herschel's newly discovered object was considered to be too far from the sun to be a comet.

36 Herschel's discovery was the most important find of the last three hundred years.

Questions 37-40

*Complete each of the following statements (**Questions 37-40**) with a name from the Reading Passage.*

Write your answers in boxes 37-40 on your answer sheet.

The suggested names of the new planet started with … **(37)** …, then … **(38)** …, before finally settling on Uranus.

The first five rings around Uranus were discovered by … **(39)** ….

From 1948 until 1986, the moon … **(40)** … was believed to be the moon closest to the surface of Uranus.

TBRT- EM

SECTION 1 QUESTIONS 1-13

QUESTIONS 1-5

Look at the information below about MAGNETIC FLUX.

In boxes 1-5 on your answer sheet write

TRUE *if the statement is true*

FALSE *if the statement is false*

NOT GIVEN *if the information is not given in the passage*

EXAMPLE	Answer
Particles in the vicinity of a magnet will be affected by magnetic flux.	TRUE

1 Tesla is used to measure large magnetic flux densities.

2 The CGS system of units does not make use of tesla for measuring magnetic flux.

3 The Earth has a magnetic field as big as 10000 teslas.

4 The magnetic field of a synchroton is equal to that of the Earth.

5 Superconducting magnets have the largest magnetic flux possible.

MAGNETIC FLUX

Magnetic flux is the force that a magnet or electromagnetic source leaves on other magnets or particles in its vicinity. In the International System of Units, magnetic flux density is measured in teslas whereas Gauss, which is the unit of measurement of magnetic flux density in the centimeter-gram-second (cgs) system of units, is a much smaller unit than the tesla. One gauss is the same as 10^{-4} (0.0001) tesla. Thus, it is normally used to describe very small magnetic flux densities.

The magnetic field of the Earth is about one gauss. The magnetic field between the poles of a typical horseshoe-type magnet is about 300 gauss. The strongest fields that can be created with permanent magnets reach a few thousand gauss. A synchroton, which is a very large machine used for accelerating subatomic particles, uses electromagnets with fields measuring around 10,000 gauss. A superconducting magnet can have a magnetic field of 500,000 gauss or more.

QUESTIONS 6-13

Look at VACCUM TUBE DIODES below.
*Match each of the following sentences with **TWO** possible endings **A-M** from the box below.*
*Write the appropriate letters **A-M** in boxes 6-13 on your answer sheet.*

EXAMPLE	Answer
In their structure, vacuum tube diodes employ ...	**A and M**

Questions 6 and 7

The capacity of triodes to amplify depends on ...

Questions 8 and 9

The normally-negative electrode in a vacuum tube diode, the cathode, is ...

Questions 10 and 11

The cathode and the anode have extensively been used in ...

Questions 12 and 13

In triodes, the grid is responsible for controlling ...

Possible Endings

A	a plate and a negative electrode.
B	the potential of the grid.
C	the rectification of alternating current.
D	the introduction of a third electrode.
E	the electrode that emits electrons.
F	a negatively-charged plate.
G	the flow of current between the cathode and the anode.
H	the small changes in the voltage between the grid and the cathode.
I	a heated filament or a small, heated, metal tube.
J	the positive halves of the cycle.
K	the structure of triodes.
L	the number of electrons that reach the anode.
M	a crystal or steel vacuum capsule.

VACCUM TUBE DIODES

Vacuum Tube diodes consist of a crystal or steel vacuum capsule and two or more electrodes between which electrons can move freely. They usually contain two electrodes. The cathode is a heated filament or a small, heated, metal tube that emits electrons by thermionic emission. In contrast, the anode, or plate, is the element that collects electrons. In diodes, the plate attracts the electrons emitted by the cathode only when the latter is positive with respect to the cathode. A negatively-charged plate disallows the flow of current through the tube. If you apply an alternating potential to the anode, the tube passes current only during the positive halves of the cycle. Thus, it acts as a rectifier. Scientist use diodes extensively in the rectification of alternating current.

Triodes are formed by the introduction of a third electrode, a grid, interposed between the cathode and the anode. A triode was the basic tube used for amplifying current for many years. The function of the grid is to control the current flow. The grid repels electrons. It can impede the flow of electrons between the cathode and the anode at a certain negative potential. At lower negative potentials, the electron flow depends on the grid potential. The capacity of triodes to amplify depends on the small changes in the voltage between the grid and the cathode. These changes cause larger changes in the number of electrons reaching the anode.

SECTION 2 QUESTIONS 14-20

Questions 14-20

Look at BRIDBE CIRCUITS below and at the following statements ***(Questions 14-20)****.*

In boxes 14-20 on your answer sheet write

TRUE *if the statement is true*

FALSE *if the statement is false*

NOT GIVEN *if the information is not given in the passage*

14 The ability of an electronic device to create electric current by moving in magnetic fields can be measured by means of a bridge.

15 Sir Charles Wheatstone was the first person who measured the ability of a bridge to resist the flow of electric current.

16 Scientists use a galvanometer to measure the ability of a bridge to store electricity.

17 A loop is a Wheatstone bridge that capacitates the user to control the value of resistors.

18 When you apply electric current to a loop, the voltage breaks into two equal halves and flows through the four arms of the loop.

19 In order to change one type of energy into another, the electric current must flow through a transducer.

20 One of the components of a bridge is an output transducer.

BRIDHGE CIRCUITS

Bridge is an electric circuit suitable for measuring such electrical values as resistance, capacitance and/or inductance. Resistance refers to the ability to resist the flow of electric current. Capacitance is the ability to store electricity. Inductance is the ability to create electric current by moving in magnetic fields. The British Physicist Sir Charles Wheatstone made the first bridge circuit to measure the value of an unknown resistance through juxtaposing it to three known resistances.

Almost all electrical bridges are variations of the infamous Wheatstone bridge. The Wheatstone bridge consists of four connected resistances that form four arms of a loop. The user knows the values of three of these four resistances (R_1, R_2, and R_3) and does not know that of one (R_x). R_1, R_2, and R_3 are variable resistors controlled by the user. The arrangement of resistors is such that electric current splits at one corner of the diamond and flows through R_1 and R_3 on one side and R_2 and R_x on the other side.

A galvanometer, a meter that measures electric current, connects the point of the loop between R_1 and R_3 to the point of the loop between R_2 and R_x. Current flows out of the diamond at the point between R_3 and R_x. The user varies the resistance of R_1, R_2, and R_3 until the galvanometer reads zero. The ratio of the resistance of R_1 to the resistance of R_3 then equals the ratio of the resistance of R_2 to the resistance of R_x, and the user can easily estimate the value of R_x because he knows the values of R_1, R_2, and R_3.

Ohmmeters are devices that use an internal Wheatstone bridge to measure the value of a resistor in an electric circuit. Other electrical bridges use techniques similar to the Wheatstone bridge to estimate the value of capacitors and inductors. Transducers—devices that convert one type of energy into another—often employ Bridge circuits.

The transducers used with bridge circuits convert types of energy such as heat, light, or sound into electrical energy. When the output of a transducer forms an element of a bridge circuit, changes in the level of the energy input to the transducer result in dramatic and easily detectable changes in the output of the bridge circuit.

Questions 21-26

*Look at INCANDESCENT LAMPS below. From the following list (**i-xi**) choose the most suitable summaries for the paragraphs **A, C,** and **E-H.***

*Write the appropriate numbers (**i-xi**) in boxes 21-26 on your answer sheet.*

***NB** There are more summaries than paragraphs, so you will not use them all.*

i	The history and the use of household lightbulbs
ii	Early and recent developments in the structure of lightbulbs
iii	The role of American inventors in the invention of incandescent lamps
iv	The general characteristics of incandescent lamps produced these days
v	How ductile drawn tungsten filaments produce light
vi	What scientists mean by the principle of incandescence
vii	The use of carbon-arc lamp and the way it works
viii	Who first invented the incandescent lamp
ix	The general structure of the incandescent lamp
x	Non-electrical lamps, their use, and the way they produce light
xi	How the inert gas reacts in the presence of current to produce light

21 Paragraph **A**

EXAMPLE	Answer
Paragraph **B**	**Xi**

22 Paragraph **C**

EXAMPLE	Answer
Paragraph **D**	**i**

23 Paragraph **E**

24 Paragraph **F**

25 Paragraph **G**

26 Paragraph **H**

INCANDESCENT LAMPS

A
Incandescent Lamps produce light by heating a material to a high temperature. The most familiar example of an incandescent lamp is the common household lightbulb. It consists of a stretched or coiled filament of tungsten metal sealed inside a bulb filled with a gas that does not react with the tungsten or the bulb.

B
This inert gas is a combination of nitrogen and argon in a proportion suitable to the wattage, or brightness, of the bulb. When electric current flows through the filament, it heats the filament to a temperature—of about 3000 °C—which causes the filament to glow and emit light.

C
The principle of incandescence is at the heart of the incandescent lamp. The principle simply tells us that solids and gases give off visible light when burning or when an electric current heats them to a sufficiently high temperature, and that each material gives off light in a color characteristic of that material.

D
The invention of vacuum pumps made it possible to implement incandescent lamps for regular lighting. In 1878 British scientist Sir Joseph Wilson Swan invented the modern lightbulb. It uses carbon filaments in evacuated glass bulbs. The infamous American inventor Thomas Alva Edison also invented the lightbulb.

E
From the time of Edison's work, scientists have improved the lightbulb in several ways. One of the most significant changes was the introduction in 1911 of lamps made with filaments of tungsten, which has the highest melting point of any metal. William David Coolidge, an American engineer working for General Electric Research Laboratory, was the man to make such an enhancement. In 1908 Coolidge had developed a process to make tungsten ductile—to draw tungsten into a wire without breaking.

F
Today, ductile drawn tungsten filaments are the backbone of most lightbulbs. Other improvements have added to the efficiency of the lightbulbs. These improvements include the addition of coiled filaments and the use of inert gas in bulbs. Lightbulbs with an inside coating of hydrofluoric acid offer softer light than coated bulbs.

G
Besides the common lightbulb, a variety of other incandescent lamps exist. One is the carbon-arc lamp, used for spotlights and motion-picture projection. This lamp provides light by heating two carbon electrodes that have an arc of high-current electricity passing between them and from the ionized gases in the arc.

H
The gas-mantle lamp is a nonelectric incandescent lamp that provides light by heating a lattice of metal oxides to the point of glowing. Another instance of a nonelectric incandescent lamp is the limelight used in theatrical lighting until the turn of the century. It provides light by heating a block of calcium oxide in a flame fueled by oxygen and hydrogen.

SECTION 3 Questions 27-40

Read the following passage and answer questions 27-40.

DIGITAL AND ANALOG COMPUTERS

Computers can be either digital or analog. Digital refers to the processes in computers that manipulate binary numbers, which represent switches that are turned on or off by electrical current. Analog refers to numerical values that have a wide range. As an example, consider a desk lamp. If it has a simple on/off switch, then it is digital, because the lamp either produces light at a given moment or it does not. If a dimmer replaces the on/off switch, then the lamp is analog, because the amount of light can vary continuously from on to off and all intensities in between. Analog computer systems were the first type of computers ever made. A popular analog computer used in the 20th century was the slide rule. The English scientist, William Oughtred, invented it in 1621. The slide rule was in wide use until the invention of calculators in 1976. Recently, many people have shown new interest in analog computers, particularly in fields such as neural networks that respond to continuous electrical signals. Most modern computers, however, are digital machines whose components have a finite number of states—on or off of bits. These bits can be combined to show data such as numbers, letters, graphics, and program instructions.

Digital logic is at the heart of the operational function for all modern digital computers. Digital Logic, also called binary logic in computer science is a strict set of rules for showing the relationships and interactions among numbers, words, symbols, and other data stored or entered in the memory of a computer. The system uses binary arithmetic, in which a set of 1s and 0s (called bits) represent a number. These bits go together in meaningful ways through the operation of digital logic and physically describe electrical voltage states in a computer's circuitry. Digital logic uses the bit value 1 to represent a transistor with electric current flowing through it and the bit value 0 to represent one with no such current flowing through it.

Machine code is the technical term to refer to the instructions in the form of a set of binary digits that direct a computer's function. These binary digits, or bits, switch specific groups of transistors, called gates, on or off. There are three basic logic states, or functions, for logic gates: AND, OR, and NOT. An AND gate takes the value of two input bits and tests them to see if they are both equal to 1. If they are, the output from the AND gate is a 1, or true. If they are not, the AND gate will output a 0, or false. An OR gate tests two input bits to see if either of the bits is equal to 1. If either input bit is equal to 1, the gate outputs a 1; if both input bits are 0, it outputs a 0. A NOT gate negates the input bit, so an input of 1 results in an output of 0 and vice versa.

Combinations of logic gates in open or closed states can be used to represent and execute operations on data. A series of logic gates together form a logic circuit, the output of which can provide input to another logic circuit or produce the result of a function. It is possible to perform highly complex operations using combinations of the AND, OR, and NOT functions.

Binary logic was first proposed by 19th-century British logician and scientist George Boole, who in 1847 invented a two-valued system of algebra that represented logical relationships and operations, which was called Boolean Algebra, used by German engineer Konrad Zuse in the 1930s for his Z1 calculating machine.

The American physicist John Atanasoff and his graduate student Clifford Berry also used it in the design of the first digital computer in the late 1930s. During 1944 and 1945 American scientist John von Neumann suggested the use of the binary arithmetic system for storing programs in computers. In 1936, British scientist Alan Turing also recognized how binary logic was fit for the development of digital computers, a concept adopted by American scientist Claude Shannon in 1948.

From then on, digital Sampling is used for the transformation of an analog signal into one based on discrete units—a digital signal. In the creation of taped music, sampling enables the composer, producer, or remix engineer to borrow discrete vocal or synthetic parts from other taped work. Live sound may also be sampled. A telephone handset changes sound waves into an analog signal that moves up and down like a wave. In the digitizing stage the waveform is sampled thousands of times per second. Each part of the sampled wave is given a binary code sign (made up of clusters of the digits 0 and 1) related to the height of the wave at that point. The information is then sent along the telephone line. Using digital signals, messages can be transmitted quickly, accurately, and economically.

Questions 27-31

Complete the table below.

Write a date for each answer.

Write your answers in boxes 27-31 on your answer sheet.

EVENT	DATE
Example The design of the first digital computer	**Answer** 1930s
Turing noticed the fitness of binary logic for digital PCs	(27) ……
Boole proposed his system of algebra	(28) ……
The invention of the electronic calculator	(29) ……
Claude Shannon agreed with what Turing had claimed	(30) ……
The invention of slide rule	(31) ……

Questions 32-36

Do the following statements reflect the claims of the writer of the Reading Passage?

In boxes 32-36 on your answer sheet write

YES *if the statement is true*

NO *if the statement is false*

NOT GIVEN *if the information is not given in the passage*

EXAMPLE	Answer
Analog signals are not discrete.	**YES**

32 Binary logic is based on Boolean algebra.

33 Digital computers cannot process continuous electrical signals.

34 The invention of electronic calculators resulted in people's lack of interest in the slide rule.

35 To be able to process binary information computers must be equipped with transistors.

36 Analog computers are not capacitated to manipulate numbers bigger than two.

Questions 37-40

*Complete each of the following statements (**Questions 37-40**) with a name from the Reading Passage.*

Write your answers in boxes 37-40 on your answer sheet.

Binary logic was used by the … **(37)** … and his student … **(38)** … in the design of digital computers.

The oldest computer invented by … **(39)** … in the seventeenth century was in wide use until the 1970s.

A calculating machine based on Boolean logic was invented by … **(40)** … who called it Z_1.

TBRT-GM

SECTION 1 QUESTIONS 1-13

QUESTIONS 1-5

Look at the information below about NATURAL HAZARDS.

In boxes 1-5 on your answer sheet write

TRUE *if the statement is true*

FALSE *if the statement is false*

NOT GIVEN *if the information is not given in the passage*

EXAMPLE	Answer
Natural hazards are responsible for the death of some people.	TRUE

1 Volcanoes, hurricanes and earthquakes are chronic and continuous in nature.

2 The impact of an earthquake on buildings can be predicted before the occurrence of the quake.

3 Mount Pinatubo is a volcanic mountain that stands in the Philippines.

4 According to the passage, geologists are scientists who study the behavior of volcanoes.

5 Water storage and embankments are two engineering solutions that can prevent floods from occurring.

NATURAL HAZARDS

Continuously occurring or chronic natural hazards are often unrecognized as such and difficult to identify. They may adversely affect lots of people, animals and plants. Many natural hazards, such as earthquakes, volcanoes, and hurricanes are unavoidable, but measures can be taken to lower their impact. Thus buildings can be designed to withstand quakes, and ways are also being developed to predict their occurrence.

In 1991 the volcano Mount Pinatubo in the Philippines erupted. Fortunately a team of geologists from the United States Geological Survey were present, who predicted the eruption and saved the lives of many people who were evacuated. Flood impacts can be reduced by engineering solutions. These solutions may involve water storage and embankments, and warnings and advice given to the public in advance of major storms. When Hurricane Andrew struck Florida in 1992 it caused $12 billion of damage, but due to the advance warning of the storm only 50 people died.

QUESTIONS 6-13

Look at NATIONAL PARKS AND SANCTUARIES below.
*Match each of the following sentences with **TWO** possible endings **A-M** from the box below.*
*Write the appropriate letters **A-M** in boxes 6-13 on your answer sheet.*

EXAMPLE	Answer
National parks are possessed by governments in ...	**A and M**

Questions 6 and 7

Forlandet National Park is ...

Questions 8 and 9

Lake District National Park in the United Kingdom encompasses ...

Questions 10 and 11

Spain's Doana National Park provides ...

Questions 12 and 13

National parks and sanctuaries in Britain are not supervised and controlled ...

Possible Endings

A	South America.
B	England's tallest mountain.
C	by the government.
D	residence for red deer.
E	active farms.
F	food for migrating birds on their way to Africa.
G	a narrow island and sanctuary located in Norway.
H	for purposes of recreation.
I	throughout the park.
J	ancient settlements.
K	an important breeding ground for guillemot ducklike auks.
L	habitat for the last surviving wild lynxes.
M	the United States.

NATIONAL PARKS AND SANCTUARIES

Governments posses sanctuaries in North and South America. However, the government does not entirely posses them in Britain. Nor are they supervised and managed primarily for purposes of recreation and wildlife. Their residents possess many of these sanctuaries which encompass ancient settlements. Many ancient towns and villages exist within 2331 square km Lake District National Park. Active farms, stone quarries, and ancient mines are scattered throughout the park. These ranches provide residence for red deer, fox, swans, and trout. The hundreds of lakes in the park inspired the park's name. England's tallest mountain, Scafell Pike, also stands here. Stone and Earthen monuments together with burial mounds of England's Stone, Iron, and Bronze Ages are sheltered in 1437 square km Peak National Park and 694 square km Exmoor National Park.

Similarly, national parks in Norway encapsulate colonies of seabirds, walrus, and reindeer herds. Forlandet National Park is a narrow island. Several small glaciers cling to its high peaks. The 640 square km island lies along the northernmost reach of the ocean stream from the Gulf of Mexico, which creates a mild climate, making this an important breeding ground for guillemot ducklike auks. Seals, eider ducks, and geese also depend on its habitats. Spain's Doana National Park provides a 507 square km wildlife refuge where birds that nest in northern Europe feed while migrating to Africa. The last surviving wild lynxes in southern Europe find sanctuary there as well.

SECTION 2 QUESTIONS 14-20

Questions 14-20

Look at THE SENSORY SYSTEMS OF SHARKS below and at the following statements ***(Questions 14-20)****.*

In boxes 14-20 on your answer sheet write

TRUE *if the statement is true*

FALSE *if the statement is false*

NOT GIVEN *if the information is not given in the passage*

14 Sharks use their developed sensory systems for finding food.

15 The largest section of the brain of every shark is devoted to its sense of smell.

16 Nocturnal animals have well-developed sense of smell called tapetum lucidum.

17 A narrow strip of sensory cells can be found along the sides of the body of a shark which enables it to see its prey in extremely dark waters.

18 Almost all species of sharks are color blind.

19 A special reception system called clusters of ampullae of Lorenzini capacitates sharks to find prey swimming at distances over 1 meter.

20 Sharks usually use their electrosensors for purposes of attacking other animals which are in the final stages of feeding.

THE SENSORY SYSTEMS OF SHARKS

The well-developed sensory systems of sharks capacitate them with unmatched advantages—in comparison to almost every other animal—when hunting or feeding. The sense of smell comprises almost one-third of a shark's brain. A shark's sense of smell is so powerful that it can detect perfumes and odors in the water hundreds of meters from their source. Sharks can detect as little as one part per million of substances in the water, such as blood, body fluids, and chemical substances produced by animals under stress. Some sharks can detect as few as ten drops of liquid tuna in the volume of water it takes to fill an average swimming pool.

Sharks' eyes detect and capture virtually small movements and they can sense in gloomy conditions, making them effective hunters in virtually dark depths. Like cats and other nocturnal hunters, sharks have a reflective layer in the back of their eyes, called the tapetum lucidum, which magnifies low levels of light. In clear water, sharks see their prey when it is about 20 to 30 meters away.

Sharks' eyes also contain specific cells that detect color, and behavioral studies suggest that sharks can see colors as well as black, white, and shades of gray. These studies also revealed that luminous and glimmering objects and bright colors, such as yellow and orange, may attract sharks.

Sharks employ an extra sensory system—which scientists call the lateral line—to detect vibrations in the water which fish, boats, surfers or even swimmers often create. A narrow strip of sensory cells running along the sides of the body and into the shark's head comprises the lateral line. This sensory system is especially sensitive to sounds in the low-frequency ranges, such as those which struggling wounded fish or other animals emit.

Additionally, the functioning of neurons and muscles in living animals create electrical currents which sharks sense in no time. The shark's electrosensors—the clusters of ampullae of Lorenzini—exist over the shark's head of all sharks. This reception system is effective only over distances of less than 1 meter. It may aid sharks in the final stages of feeding or attack. Scientists also concede that this system may somehow capacitate sharks to detect the feeble electromagnetic fields of the Earth, ushering them in migration.

Questions 21-26

*Look at CLASSIFICATION OF AIRPLANES below. From the following list (**i-xi**) choose the most suitable summaries for the paragraphs **A, C,** and **E-H.***

*Write the appropriate numbers (**i-xi**) in boxes 21-26 on your answer sheet.*

***NB** There are more summaries than paragraphs, so you will not use them all.*

i	The general structure and the design of sea planes
ii	Technical characteristics of amphibian planes
iii	The way a space craft takes off and lands
iv	Take off and landing characteristics of carrier-based airplanes
v	General characteristics of helicopters
vi	Technical features of short-range airplanes
vii	The sophisticated under-carriage system of pontoon planes
viii	Major classes of airplanes
ix	Take off and cruise characteristics of Vertical Take-off planes
x	The skis some planes use in the Arctic and Antarctic regions
xi	Gear systems of land planes and the runways they can use

21 Paragraph **A**

EXAMPLE	Answer
Paragraph **B**	**Xi**

22 Paragraph **C**

EXAMPLE	Answer
Paragraph **D**	**I**

23 Paragraph **E**

24 Paragraph **F**

25 Paragraph **G**

26 Paragraph **H**

CLASSIFICATION OF AIRPLANES

A
Airplanes are classifiable into various classes including land planes, carrier-based airplanes, seaplanes, amphibians, vertical takeoff and landing, short takeoff and landing, and space shuttles.

B
Designers usually design land planes to operate from a paved surface, typically a runway, and equip some of them to operate from grass or other unfinished surfaces. Land planes usually have wheels. Some specialized aircrafts operating in the Arctic or Antarctic regions have skis instead of wheels.

C
As a modified type of land planes which can takeoff from and land aboard naval aircraft bases, carrier-based airplanes have a strengthened structure. A landing gear helps them handle the stresses of catapult-assisted takeoff, in which steam-driven catapults launch the craft. They also make arrested landings by using hooks attached to the underside of their tails.

D
Pontoon planes are technically-modified land planes with floats in place of wheels so they can operate from water. Their designers have designed a number of seaplanes from scratch to operate only from water bases. Pontoon planes may have small floats connected to their outer wing panels to help steady them at low speeds on the water, but the plane's floating hull usually bears the weight of the plane.

E
Amphibians operate from both water and land headquarters. Very often, an amphibian is an extraordinary Pontoon, with a boat-like hull and the addition of specifically designed under-carriage system. When extended, it can capacitate the airplane to taxi right out of the water onto land headquarters. Historically, some Pontoons possessed a beaching gear, a system of cradles on wheels positioned under the floating aircraft. It allowed the pilot to roll the aircraft onto land.

F
Vertical Takeoff and Landing airplanes typically implement the jet thrust from their turbines, pointed down at the Earth, to take off and land straight up and down. After taking off, the airplane usually transitions to wing-borne flight in order to cover a longer distance or carry a significant load. A helicopter is a typical example of such an aircraft.

G
Short-Takeoff-and-Landing aircrafts are able to function on relatively short runways. Their designs usually employ optimized wings and high-lift instruments on the wings for optimum performance throughout takeoff and landing as distinguished from an airplane that has a wing optimized for high-speed cruise at high altitude. These airplanes are usually cargo airplanes. Some serve in a passenger-carrying capacity as well.

H
A NASA space shuttle is an aircraft unprecedented by any other because it flies as a fixed-wing airplane within the atmosphere and as a spacecraft in outer space. After rising from the launching base, the space shuttle flies like a rocket out of the atmosphere. During landing, the shuttle becomes the world's most sophisticated engine-less glider.

SECTION 3 Questions 27-40

Read the following passage and answer questions 27-40.

MISSION TO MOON

In 1958, the United States and the USSR were both working hard to be the pioneer to send a satellite to the Moon. Their early probes failed. On October 11, 1958, Pioneer 1 was launched by the United States to orbit the Moon. It did not reach a high enough speed to reach the Moon, but reached a height above Earth of more than 110,000 km. In December 1958 Pioneer 3 also failed to leave high Earth orbit. It did, however, discover a second Van Allen belt of radiation around Earth.

On January 2, 1959, after two earlier failed missions, the USSR launched Luna 1, which was expected to hit the Moon. Although it missed its target, Luna 1 did become the first artificial object to escape Earth orbit. On September 14, 1959, Luna 2 became the first artificial object to strike the Moon, impacting east of Moon's Mare Serentitatis. In October 1959, Luna 3 flew around the Moon and radioed the first pictures of the far side of the Moon, which is not visible from Earth.

In the United States, efforts to reach the Moon did not resume until 1962, with a series of probes called Ranger. The early Rangers were designed to eject an instrument capsule onto the Moon's surface just before the main spacecraft crashed into the Moon. These missions were plagued by failures—only Ranger 4 struck the Moon, and the spacecraft had already ceased functioning by that time. Rangers 6 through 9 were similar to the early Rangers, but did not have instrument packages. They carried television cameras designed to send back pictures of the Moon before the spacecraft crashed. On July 31, 1964, Ranger 7 succeeded in sending back the first quality images of the Moon before crashing, as planned, into the surface. Rangers 8 and 9 repeated the feat in 1965.

By then, the United States had embarked on the Apollo program to land humans on the Moon. With an Apollo landing in mind, the next series of U.S. lunar probes, named Surveyor, was designed to "soft-land" (that is, land without crashing) on the lunar surface and send back pictures and other data to aid Apollo planners. As it turned out, the Soviets made their own soft landing first, with Luna 9, on February 3, 1966. The first pictures of a dusty moonscape from the lunar surface were radioed by Luna 9. Surveyor 1 successfully reached the surface on June 2, 1966. Six more Surveyor missions followed, but only two were successful. Thousands of pictures of the lunar surface were sent back by the Surveyors. Two of the probes were equipped with a mechanical claw, remotely operated from Earth, that enabled scientists to investigate the consistency of the lunar soil.

At the same time, the Lunar Orbiter probes were launched by the United States, which began circling the Moon to map its surface in unprecedented detail. Lunar Orbiter 1 began taking pictures on August 18, 1966. Four more Lunar Orbiters continued the mapping program, which gave scientists thousands of quality photographs covering nearly all of the Moon.

Beginning in 1968, unpiloted Zond probes—actually a lunar version of their piloted Soyuz spacecraft—were sent around the Moon by the USSR. These flights, initially designed as preparation for planned piloted missions that would orbit the Moon, returned high-quality photographs of the Moon and Earth. Two of the Zonds carried biological payloads with turtles, plants, and other living things. Both the United States and the USSR were achieving successes with their unpiloted lunar missions. However, the Americans were pulling steadily ahead in their piloted program. As their piloted lunar program began to lag, the Soviets made plans for robotic landers that would gather a sample of lunar soil and carry it to Earth. Although this did not occur in time to upstage the Apollo landings as the Soviets had hoped, Luna 16 did carry out a sample return in September 1970, returning to Earth with 100 g (4 oz) of rock and soil from the Moon's Mare Fecunditatis. In November 1970 Luna 17 landed with a remote-controlled rover

called Lunakhod 1. The first wheeled vehicle on the Moon, Lunakhod 1 traveled 10.5 km (6.4 mi) across the Sinus Iridium during ten months of operations, sending back pictures and other data. Only three more lunar probes followed. Luna 20 returned samples in February 1972. Lunakhod 2, carried aboard the Luna 21 lander, reached the Moon in January 1973. Then, in August 1976 Luna 24 ended the first era of lunar exploration.

Exploration of the Moon resumed in February 1994 with the U.S. probe called Clementine, which circled the Moon for three months. In addition to surveying the Moon with quality cameras, Clementine gathered the first exact data on lunar topography using a laser altimeter. Clementine's laser altimeter bounced laser beams off of the Moon's surface, measuring the time they took to come back to determine the height of features on the Moon.

Questions 27-31

Complete the table below.

Write a date for each answer.

Write your answers in boxes 27-31 on your answer sheet.

EVENT	DATE
Example The US launched Pioneer 1 to orbit the Moon	**Answer** 1958
Discovery of the second Van Allen belt around Earth	(27) .....
Ranger7 crashed into the surface of the Moon	(28) .....
The resumption of US efforts to reach the Moon	(29) .....
Clementine went round the Moon for three months	(30) .....
Lunar Orbiter1 began taking photos of the surface of Moon	(31) .....

Questions 32-36

Do the following statements reflect the claims of the writer of the Reading Passage?

In boxes 32-36 on your answer sheet write

YES *if the statement is true*

NO *if the statement is false*

NOT GIVEN *if the information is not given in the passage*

EXAMPLE	Answer
Pioneer1 could not leave the highest orbit of the earth.	**YES**

32 The first Van Allen belt of radiation around the earth was discovered by Pioneer1.

33 The Russian were more successful in their attempts to reach the Moon than the American.

34 Ranger4 was destroyed by a meteor before it managed to land on the Moon.

35 Scientists had no other choice than sending astronauts to the Moon to investigate the consistency of the lunar soil.

36 Biological payloads succeeded in landing on the Moon before human beings did so.

Questions 37-40

*Complete each of the following statements (**Questions 37-40**) with a name from the Reading Passage.*

Write your answers in boxes 37-40 on your answer sheet.

Scientists use the name ... **(37)** ... to refer to the radiation belts around the Earth.

Luna2 was the Russian spacecraft that hit the Moon in the ... **(38)** ... region.

The USSR sent the unpiloted ... **(39)** ... to orbit the Moon in 1968.

Lunakhod1 was the first wheeled vehicle which traveled more than ten kilometers across Moon's ... **(40)** ....

TBRT-AM

SECTION 1 QUESTIONS 1-13

QUESTIONS 1-5

Look at the information below about CHAIN STORES.

In boxes 1-5 on your answer sheet write

TRUE *If the statement is true*

FALSE *If the statement is false*

NOT GIVEN *If the information is not given in the passage*

EXAMPLE	Answer
Chain stores have been expanding their activities in the past.	TRUE

1 In order to gain more money, chain stores should change their services.

2 A private–label should always be launched if a chain decides to evolve.

3 Chain stores cannot be found in rural areas of the United States.

4 Unlike franchised operations, chain stores involve individual ownership of many units carrying a single trade name.

5 Department stores are normally distinguished by central ownership.

CHAIN STORES

Despite acts to slow their activities, chain-store systems continued to evolve. Today the major chains are diversifying into related fields to broaden the profit base and to gain sales expansion which is both horizontal and vertical. General-merchandise chains have launched vital private-label programs for their merchandise.

Although there are chain stores all over the U.S., the major chains tend to be concentrated in large urban areas to take advantage of mass markets. Besides retail stores, chain operations also include theaters, banks, hotels, and some public utilities. Chain stores, distinguished by central ownership, should not be confused with franchised operations, which involve individual ownership of many units carrying a single trade name. Some widely known fast-food establishments, for instance, operate internationally under a franchise arrangement. Chain stores can now be found in many parts of the world, including Canada, Europe, and Latin America. Convenience stores, supermarkets, department stores, and housewares stores are among the most important.

QUESTIONS 6-13

*Look at **INTEREST** below.*
*Match each of the following sentences with **TWO** possible endings **A-M** from the box below.*
*Write the appropriate letters **A-M** in boxes 6-13 on your answer sheet.*

EXAMPLE	Answer
The rate of interest fluctuates in direct relation to ...	**A and M**

Questions 6 and 7

Economists usually regard interest as ...

Questions 8 and 9

Banks, businesses and companies usually pay simple interest on ...

Questions 10 and 11

The rate of interest depends on ...

Questions 12 and 13

Business executives borrow money and pay interest to increase ...

Possible Endings

A the supply of money.
B the reward for thrift.
C the amount of profit they make.
D the principal.
E the thrift.
F the return the money is expected to yield.
G the rate of interest.
H the sum of money loaned.
I the cumulative total of past interest payments.
J the amount of money borrowed.
K the payment made for capital.
L the motivation of individuals to become thrifty.
M the demands of the borrower.

INTEREST

Many companies, businesses and banks borrow money from people. They pay interest in return. Economists regard interest more specifically as a payment made for what they call capital. Economists also regard interest as the reward for thrift. Thrift is payment offered to people to encourage them to save and to make their savings available to others. Many companies, businesses and banks usually pay simple interest only on the principal. The principal is the sum of money loaned. Compounding the interest refers to the circumstances in which borrowers pay interest both on the principal and on the cumulative total of past interest payments. Compound interest is the term to describe the amount paid in this way.

The rate of interest is based on the percentage of the principal paid for its use for a given time. Thus, a loan of $100 at 10 percent per year earns interest of $10 a year. The relation of the supply of money to the demands of borrowers determines the current, or market, rate of interest. Interest rates tend to fall when the supply of money increases faster than the demands of borrowers. Sometimes the demand for investment funds grows much faster than the available supply of funds to meet that demand. In this case interest rates generally rise. Business executives will borrow money at an interest rate that does not exceed the return they expect the use of the money to yield.

SECTION 2 QUESTIONS 14-20

Questions 14-20

*Look at **ON THE FUNCTION OF CLEARINGHOUSES** below and at the following statements **(Questions 14-20)**.*

In boxes 14-20 on your answer sheet write

TRUE	*if the statement is true*
FALSE	*if the statement is false*
NOT GIVEN	*if the information is not given in the passage*

14 Members of a clearinghouse settle their accounts in virtually every city of the United States.

15 The services that clearinghouses provide include settling accounts, receiving checks, and enlisting the claims of banks.

16 Clearinghouses are legally authorized to debit the accounts of banks that are unable to balance their accounts.

17 In order for a bank to join the New York Clearinghouse, it should first become a member of the Federal Reserve System of the United States.

18 Any bank should be furnished with a memorandum to be able to debit or credit the accounts of other banks on its books.

19 Federal and state audits are in charge of investigating the books of the Federal Reserve Bank of New York.

20 Out-of-town banks cannot join the Federal Reserve System of the United States due to their distance from clearinghouses.

ON THE FUNCTION OF CLEARINGHOUSES

Banks or railroads usually form a type of association—a clearinghouse—as a central agency where members can settle accounts. In virtually every city of the United States a banking clearinghouse serves commercial banks in that area. Each member bank sends a representative to the clearinghouse every banking day to clear checks drawn on other member banks.

At the clearinghouse each representative presents checks received that represent claims against other banks, and in turn receives the checks that represent claims against his or her own bank. The clearinghouse then prepares a tally sheet enlisting the claims of each bank against every other bank. If the total of the claims of a bank exceeds the total of the claims against it, the clearinghouse pays the balance to the bank or credits the balance to the account of that bank, but when the total of the claims against a bank exceeds the total of its claims, the bank pays the clearinghouse or the clearinghouse debits its account.

Clearinghouses have made different arrangements for settling the daily accounts. Some pay and receive payment in cash; others receive drafts from debtor banks and pay creditor banks with special certificates negotiable only by member banks. The New York Clearinghouse—all of whose members belong to the Federal Reserve System of the U.S.—settles accounts by furnishing the Federal Reserve Bank of New York with a memorandum to debit or credit the accounts of the member banks on its books. The 12 district Federal Reserve Banks often provide check-clearing services for their members and are responsible for clearing almost all out-of-town checks received by banks in the United States.

The first clearinghouse in the U.S. became operational in New York City in 1853. Since then, the business transacted by that house has assumed enormous proportions. It is now the most important clearinghouse in the country.

Clearinghouses act for their members in dealing with out-of-town banks. A number of clearinghouses examine the books of their member banks, independently of the usual federal and state audits, and investigate the soundness of the loans and discounts of member banks. They keep the reports confidential and give them only to the investigated banks. They also direct these bank's attention to assets considered too great a risk by the examiner.

Questions 21-26

*Look at **ASSETS AND LIABILITIES** below. From the following list **(i-xi)** choose the most suitable summaries for the paragraphs **A, C,** and **E-H.***

*Write the appropriate numbers **(i-xi)** in boxes 21-26 on your answer sheet.*

***NB** There are more summaries than paragraphs, so you will not use them all.*

i	What comprises a firm's assets and liabilities
ii	The impact of individuals and foreigners on a nation's assets and liabilities
iii	Some delicate points about the assets and liabilities of firms
iv	The place of a nation's labor force in its inventory of assets and liabilities
v	An account of net liabilities of individuals
vi	The composition of an individual's assets and liabilities
vii	Raw material deposits and forests as natural assets
viii	What comprises the assets and liabilities of a nation
ix	How companies are liable for delivering commodities to their customers
x	What economists mean by assets and liabilities
xi	How balance sheets are used to show the wealth of individuals, companies, etc.

21 Paragraph **A**

EXAMPLE	Answer
Paragraph **B**	**Xi**

22 Paragraph **C**

EXAMPLE	Answer
Paragraph **D**	**I**

23 Paragraph **E**

24 Paragraph **F**

25 Paragraph **G**

26 Paragraph **H**

ASSETS AND LIABILITIES

A

Assets and Liabilities are terms used in economics and accounting. Assets represent property or rights to property and liabilities are debts owed to others. Assets and liabilities together show the wealth of a person, a firm, or a nation.

B

The usual practice is to use a balance sheet to show an entity's wealth as of a specified date—with assets on one side and liabilities and owner's equity on the other side. However, individuals, firms, and even nations have somewhat different assets and liabilities.

C

An individual's assets might include cash, bank deposits, stocks, rights to future pension payments, and a house and its contents. An individual's liabilities may consist of, for instance, a home mortgage, debt incurred on an automobile or other individual possessions, and other financial commitments, such as income tax liabilities.

D

The composition of assets and liabilities for a firm would be different. A firm's assets might include its plant and machinery, its inventories of raw materials or commodities in the process of production, or finished goods not yet delivered to consumers. A firm's assets should include receivables—perhaps for commodities which the firm has delivered but for which the customers have not paid yet.

E

Firms will also usually be more valuable than the sum of their assets because they expect to earn income—in a more technical sense ‘Goodwill’—as a result of the existence of the firm as a going concern, a unit producing goods or services for people. On the liability side of the balance sheet, the firm will have its financial obligations. If the firm has borrowed money from any other bank or has issued bonds to raise money, liabilities will include these obligations as well.

F

A nation has still another set of assets and liabilities. A national balance sheet will not simply be the sum of the balance sheets of individuals and firms. A nation's assets also include national capital, such as public buildings; publicly owned parts of the transportation infrastructure; or certain natural assets, such as raw material deposits, or national forests.

G

The balance sheet of any other entity may not include the above-mentioned items. It is also arguable that the most important asset of a nation is its labor force. The balance sheet should include a nation’s labor force in some way.

H

Obligations and liabilities between firms and individuals in the same country will cancel out. One person's liability to pay might be another person's asset. But a nation may own assets overseas, and foreigners may own capital within a nation. The accounting of a nation's wealth, therefore, should take account of net liabilities to the citizens, firms, and governments of other countries.

SECTION 3 Questions 27-40

Read the following passage and answer questions 27-40.

CORPORATE FINANCE

Corporate Finance is a branch of economics concerned with how businesses raise and spend their money. Corporate finance involves selecting projects that maximize profits and make the best use of a company's funds. Usually businesses can fund these projects on their own although they may often have to raise funds from outside the company. Corporate finance also involves finding the best way for businesses to fund their projects. A single individual who provides the funds may own small businesses, but major corporations are far too large to be owned in this way. Instead many people—shareholders—who possess shares of stock own these larger corporations. Investors purchase stock because it allows them to share in the company's profits, although there are no guarantees that the company will be successful. Each share of stock represents ownership of a portion of the firm and its possessions, or assets. For instance, in 2001 Exxon Corporation had about 600,000 shareholders, who together owned a total of about 1.2 billion shares of stock.

Shareholders usually purchase or sell shares of stock on a number of stock exchanges. For instance, Exxon's shares are regularly bought and sold on the New York Stock Exchange. At the end of 1995 Exxon's shares were priced at $80 each. At that price it would have cost about $100 billion to buy all of Exxon's stock.

Corporate investment decisions often involve substantial amounts of money. Many investment decisions are also difficult to reverse and can affect the company's business far into the future. For instance, in 1966 Boeing Company decided to invest about $1 billion to develop the 747 jumbo jet. This investment delivered long-term benefits as the company was still selling the jets 30 years later. Investment decisions are based on the two criteria of the expected rate of return and the risk or uncertainty of achieving the expected rate of return. Ordinarily, a financial manager estimates the return based on forecasts of potential sales, expenses, and profits that might occur from an investment. Thus, whether or not the company should go ahead with the project depends on what the investment could earn if invested differently. The company should accept any project that is expected to earn a higher return than shareholders can earn with another investment. If a firm fails to correctly predict returns, it may go bankrupt, and may have to merge into another company or to vanish entirely.

Investments require cash, which a corporation may be able to obtain by paying smaller dividends, by borrowing through issuing bonds or getting loans, or by selling more stock any of which has its pros and cons. A firm can finance projects by paying smaller dividends—paying out less of its profits in dividends—whereby the company can keep more of its profits as retained earnings to fund its investments. A company can also choose to borrow money to fund its projects either through getting loans from a bank or directly from investors by issuing bonds. Selling stock is a third way companies can raise funds which, unlike loans, they do not have to repay since these funds belong to the company itself. As a result of gaining funds through selling their stocks, firms do not have the expense of paying interest.

A financial manager must of necessity consider factors other than cost when deciding how to raise money. For instance, if a firm tries to raise new funds, the investors will speculate about the firm's plans, and if they think the plans are a bad idea the firm's stock price could fall.

International financial markets have become increasingly important sources of funds. United States firms often raise money in foreign financial centers such as London or Tokyo. Loans from abroad often have a lower interest cost to domestic U.S. firms because foreign banks are not subject to the restrictions of the U.S. Federal Reserve System. For example, instead of borrowing dollars from a bank in the United States, American firms may borrow dollars that have been deposited in London or Tokyo banks. These are known as Eurodollars. Eurodollars

are U.S. dollars held in banks outside of the United States. Similarly, instead of issuing bonds in the United States, U.S. firms may issue bonds in a foreign country to a group of international investors. These are called Eurobonds. Eurobonds are bonds sold outside the country whose currency is used to write the bond. For instance, a bond denominated in U.S. dollars issued by a Japanese bank is a Eurobond.

Events outside the control of a corporation can affect the firm and its financing decisions. For instance, a change in the interest rate can suddenly make borrowing money very inexpensive or very costly. From 1975 to 1995, interest rates in the United States were as high as 15 percent and as low as 3 percent. Many economic factors, such as changes in the price of oil or the price of foreign currency, can affect businesses as well.

Questions 27-31

Complete the table below.

Write a date for each answer.

Write your answers in boxes 27-31 on your answer sheet.

EVENT	DATE
Example Exxon had 600,000 shareholders	**Answer** 2001
Interest rate of 3% in the US	(27) ……
Boeing invested $1billion on 747	(28) ……
Shareholders owned $1.2 billion of Exxon	(29) ……
Interest rate of 15% in the US	(30) ……
Exxon's shares of stock cost $80 each	(31) ……

Questions 32-36

Do the following statements reflect the claims of the writer of the Reading Passage?

In boxes 32-36 on your answer sheet write

YES *if the statement is true*

NO *if the statement is false*

NOT GIVEN *if the information is not given in the passage*

EXAMPLE	Answer
Investment decisions are critical for any firm.	**YES**

32 Paying smaller dividends is the best way to gain larger revenue.

33 Events outside the control of a firm curtail its funds.

34 Clinton was successful in lowering the rate of interest during the 1990s.

35 Boeing gained a lot of benefit from its 747 project.

36 Bonds issued in the US are more expensive than those issued in Japan and Europe.

Questions 37-40

*Complete each of the following statements (**Questions 37-40**) with a name from the Reading Passage.*

Write your answers in boxes 37-40 on your answer sheet.

American firms prefer to borrow … **(37)** … deposited in European banks.

Recently US firms issue … **(38)** … in a foreign country to a group of international investors.

Selling jumbo jets has earned … **(39)** … a lot of revenue over the past thirty years.

Shareholders and investors usually refer to New York Stock Exchange to purchase or sell the shares of stock of … **(40)** ….