DEVELOPMENT AND VALIDATION OF A MULTIMEDIA COMPUTER PACKAGE FOR THE ASSESSMENT OF ORAL PROFICIENCY OF ADULT ESL LEARNERS: IMPLICATIONS FOR SCORE COMPARABILITY

Seyyed Abbas Mousavi

B.A., English Literature, Tabriz University, 1994
M.A., Applied Linguistics and TESOL, Shiraz University, 1997

A THESIS SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
In the Faculty of Arts

GRIFFITH UNIVERSITY
March 2007

© Seyyed Abbas Mousavi, 2007
ABSTRACT

This thesis is about the conceptualization, design, development, trial, and validation of a multimedia package for the computer-based administration of an interview in testing the general English language proficiency of adult ESL learners. This research is significant at both theoretical and practical levels. Theoretically, it fills a gap in the comparability studies of computer-based tests and conventional face-to-face interviews. It also sheds light on the usability and validity of a new mode of presentation which treats the computerized test driven language production as an aspect of the target language use in assessing performance and interpreting test scores. This study employs a quantitative and qualitative survey to investigate the interaction between test-takers and computer-delivered tests. It explores the effects of test-taker characteristics (such as age, gender, language background, and computer familiarity) on test performance and draws upon qualitative feedback provided by the examinees in interpreting test usefulness and substantiating validity generalizations.

At a practical level, this project contributes to the language testing industry by capturing the potential of the computer and digital media for developing tests and tasks and introducing a new set of innovative tasks for the assessment of speaking. It further formulates a practical process model for future test developers. The end product of this research is a working prototype of a multimedia language testing instrument using video and audio to present the tasks, which may be used as an entry/exit, gate-keeping, accreditation, certification, or placement mechanism. Along with the substantial findings about the comparability of computer-based tests with
face-to-face interviews, this study provides a set of practical guidelines for researchers who embark on the design and development of computer-based language tests.

Given the rate of innovation in the digital media, natural language processing and voice recognition technology, the present era must be considered a transitional one and the future is difficult to predict. This thesis, therefore, concludes with two principal suggestions regarding further research at conceptual and practical levels. First, due to the complexity of the nature of human-machine interaction, researchers in language testing (particularly speaking tests) are advised to exercise caution in validity generalizations, because modifications in the delivery mode can result in changes in the quality and nature of the task and, as a consequence, the quality of the speaking performance. Second, this study was a small-scale prototype and a working example of the use of digital video in oral testing. The results showed that, for large-scale test development projects, language testing professionals need to utilize the services of a team of information technology experts in developing tests of speaking proficiency with a view to increasing the number and variety of tasks as well as enhancing the security and usability of the test.
ACKNOWLEDGEMENTS

There are many people to whom I would like to express my very special appreciation in the completion of this thesis. First of all, particular thanks go to my supervisors, Professor Mike Levy, Dr. Claire Kennedy, and Associate Professor Shirley O’Neill for their invaluable advice, inspiration, encouragement and support throughout the years of my study. The completion of this thesis would have been impossible without the Griffith University Postgraduate Research Scholarship award offered to me in 2003. I would also like to thank Professor David Ingram, Elaine Wylie and Hilda Maclean for their generous support in providing me with material about the ISLPR ratings. My heartfelt thanks go to academics and friends who supported me during the slow times, hard times and hopeless times. I am truly indebted to Ms Anne Lin’s very special talent, curiosity, and wonderful new ideas. I wish to express my greatest thanks to Mrs Adriana Diaz for her truly admirable suggestions and advice. Thanks go to LAL staff and colleagues, Peter Smith and Ron Magarry, for their professionalism and patience. Very special thanks to Mr. Mohsen Mohajeri who helped me with the demanding process of programming the software package. Last but certainly not least, my family who were with me all the way through; I cannot thank you enough. I am honoured to dedicate this thesis to my parents whose role was pivotal in my intellectual and personal growth.
STATEMENT OF ORIGINALITY

I certify that this work has not previously been submitted for a degree or diploma in any university. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

Signed: ……………………………

Seyyed Abbas Mousavi

Date: ……./……../………………
TABLE OF CONTENTS

ABSTRACT ........................................................................................................................................ ii

ACKNOWLEDGEMENTS ........................................................................................................ iv

STATEMENT OF ORIGINALITY ....................................................................................... v

TABLE OF CONTENTS ........................................................................................................ vi

LIST OF TABLES ................................................................................................................ ix

LIST OF FIGURES ............................................................................................................. xi

LIST OF APPENDICES .................................................................................................... xiii

CHAPTER 1: INTRODUCTION ..................................................................................... 1

1.1 Background ..................................................................................................................... 1

1.2 Identification of the Problem .......................................................................................... 3

1.3 Significance of the Study ............................................................................................... 8

1.4 Purpose of the Study ....................................................................................................... 10

1.5 Thesis Structure .......................................................................................................... 12

CHAPTER 2: LITERATURE REVIEW ....................................................................... 13

2.1 Introduction ................................................................................................................... 13

2.2 Background: A Brief History of Language Testing .................................................... 13

2.2.1 Early Work .............................................................................................................. 13

2.2.2 Pre 1920s: The Essay Translation Approach ........................................................... 14

2.2.3 1930s-1960s: The Growth of Objective Testing and Structuralism ................... 16

2.2.4 The 1970s: Integrative Testing ................................................................................ 18

2.2.5 The 1980s to the Present: Modern Communicative Testing .................................. 19

2.3 Defining Language Proficiency ..................................................................................... 20

2.4 Definition of the Speaking Construct .......................................................................... 28

2.5 Test Usefulness and Its Components ........................................................................ 33

2.6 Testing Speaking Ability: Issues and Approaches ...................................................... 40

2.6.1 ACTFL OPI: An Overview ..................................................................................... 49
4.2.4 Descriptive Comments ..................................................................................................163
4.3 Analysis of Test-taker Language Samples ........................................................................164
  4.3.1 Pre-intermediate .......................................................................................................164
  4.3.2 Intermediate .........................................................................................................167
  4.3.3 Advanced ...................................................................................................................170

CHAPTER 5: DISCUSSION ...................................................................................................176
  5.1 Score Comparability and Self-assessment ........................................................................176
    5.1.1 Research Question 1 ..........................................................................................176
    5.1.2 Research Question 2 ..........................................................................................180
  5.2 Multimedia Effect ..........................................................................................................183
    5.2.1 Research Question 3 ..........................................................................................183
    5.2.2 Research Question 4 ..........................................................................................187
    5.2.3 Research Question 5 ..........................................................................................190
  5.3 Task Difficulty and Contributing Factors ........................................................................193
    5.3.1 Research Question 6 ..........................................................................................193
  5.4 Attitude Analysis and Anxiety .....................................................................................198
    5.4.1 Research Question 7 ..........................................................................................198
    5.4.2 Research Question 8 ..........................................................................................201

CHAPTER 6: CONCLUSIONS AND IMPLICATIONS ......................................................207
  6.1 Conclusions ................................................................................................................207
  6.2 Implications ................................................................................................................211
    6.2.1 Implications for Test Developers .........................................................................211
    6.2.2 Implications for Test Administrators ....................................................................212
    6.2.3 Implications for Test-takers ................................................................................213
  6.3 Limitations of the Study .............................................................................................214
  6.4 Suggestions for Further Research .............................................................................217
  6.5 Final Remarks .............................................................................................................219

REFERENCES ................................................................................................................220

APPENDICES .................................................................................................................236
LIST OF TABLES

TABLE 2.1 Overview of proficiency development according to the ISLPR [from Wylie
and Ingram (1979)] ................................................................. 54
TABLE 2.2 The ISLPR-IELTS conversion table ................................. 56
TABLE 2.3 Research questions and corresponding hypotheses ............... 95
TABLE 3.1 The minimum system requirements for the SAM Interview program .... 121
TABLE 3.2 Conventional symbols in drawing flowcharts ....................... 126
TABLE 4.1 Testees’ characteristics, self-assessment results, and performance on
IELTS and SAM Interview ....................................................... 144
TABLE 4.2 Results of the correlation coefficient between the SAM Interview and the
IELTS ...................................................................................... 147
TABLE 4.3 Ranked comparisons of testees’ SAM Interview scores with their IELTS
scores .................................................................................... 148
TABLE 4.4 Descriptive statistics for the participants and their scores on the SAM
Interview and the IELTS ............................................................ 149
TABLE 4.5 Descriptive statistics for anxiety level and gender ................. 153
TABLE 4.6 Chi-square results for multimedia effect and anxiety .......... 153
TABLE 4.7 Descriptive statistics for age and anxiety levels .................... 154
TABLE 4.8 Results of the correlation coefficient between age and anxiety levels .... 154
TABLE 4.9 Descriptive statistics for difficulty level and gender ............... 157
TABLE 4.10 Chi-square results for multimedia effect and difficulty level ......... 157
TABLE 4.11 Descriptive statistics for age and difficulty level .................. 158
TABLE 4.12 Results of the correlation coefficient between age and difficulty level .. 158
TABLE 4.13 Descriptive statistics for gender and user interface design ........ 161
TABLE 4.14 Chi-square results for age and user interface design..............................162

TABLE 4.15 Descriptive statistics for age and user interface design.........................162

TABLE 4.16 Results of the correlation coefficient between age and approval of user interface design........................................................................................................162
LIST OF FIGURES

FIGURE 3.1 Research design and methodology ............................................................ 99
FIGURE 3.2 Flowchart of the SAM Interview software ............................................ 128
FIGURE 4.1 Self-assessment results ............................................................................. 145
FIGURE 4.2 Scatter plot representing performance on the self-assessment and the SAM Interview ........................................................................................................ 145
FIGURE 4.3 A cross comparison of the IELTS, SAM Interview and self assessment results ............................................................................................................ 146
FIGURE 4.4 Scatter plot for the SAM Interview and the IELTS results ................. 147
FIGURE 4.5 A comparison of results from the IELTS and the SAM Interview ...... 149
FIGURE 4.6 Responses to Item 9.................................................................................. 151
FIGURE 4.7 Responses to Item 11............................................................................... 152
FIGURE 4.8 Responses to Item 14............................................................................... 152
FIGURE 4.9 Responses to Item 15............................................................................... 152
FIGURE 4.10 Responses to Item 6................................................................................ 155
FIGURE 4.11 Responses to Item 7............................................................................... 155
FIGURE 4.12 Responses to Item 8............................................................................... 156
FIGURE 4.13 Responses to Item 10............................................................................. 156
FIGURE 4.14 Responses to Item 12............................................................................. 156
FIGURE 4.15 Responses to Item 13............................................................................. 157
FIGURE 4.16 Responses to Item 1............................................................................... 159
FIGURE 4.17 Responses to Item 2............................................................................... 160
FIGURE 4.18 Responses to Item 3............................................................................... 160
FIGURE 4.19 Responses to Item 4............................................................................... 160
FIGURE 4.20 Responses to Item 5............................................................................... 161
FIGURE 5.1 The location of the recorded audio files........................................................................192
LIST OF APPENDICES

Appendix I: Informed Consent ................................................................. 236
Appendix II: Information Sheet ............................................................... 238
Appendix III: Pre-CBT Self-assessment .................................................... 241
Appendix IV: Instructions ...................................................................... 243
Appendix V: Post-CBT Questionnaire of the SAM Interview Program ......... 244
Appendix VI: The Questions for the Tasks ............................................... 246
Appendix VII: How to Install the SAM Interview Software Package ........... 249
CHAPTER 1: INTRODUCTION

1.1 Background

The testing of speaking has a long history (Alderson & Banerjee, 2002; Spolsky, 1990, 1995) but the use of computer technology in the assessment of speaking ability is relatively new (Alderson, 1986, 2000; Davidson, 2003; Fulcher, 2000; Larson, 2000; McDonald, 2002; Malabonga & Kenyon, 1998). This research project, which explores the development and validation of a computer-based test, is conducted with the aim of bringing together the two areas of testing speaking ability and technology. At each stage of conceptualization, design and implementation, this study seeks to address the novel, and critical, issues resulting from the merging of digital media with language testing, with an eye on test usefulness (Bachman & Palmer, 1996). It demonstrates validity arguments in support of the use of the new media in the assessment of language ability. Validity argument refers to any statement of generalization made about the validity assumptions made on the basis of test results (Bachman, 1995). This research also tries to explore what the scores from a computerised speaking test mean compared to those from conventional tests.

The testing of speaking is a sub-field of language testing in general which, in turn, is a sub-discipline of the broader field of applied linguistics. The history of testing second language speaking has been neatly delineated by Fulcher (2003) who looks at the testing of speaking before and after the Second World War. One major concern in much of the literature in the testing of speaking surrounds the reliability of such tests;
Introduction

others are the administration of the test and the subjectivity of the scoring system, which ultimately influence the reliability of the test.

Reliability and validity are the two integral parts and essential requirements of any measurement process. Reliability has to do with the consistency of scores across different administration times, test forms, raters and other characteristics of the measurement context (Mousavi, 2002). Unless test scores are reasonably consistent, they cannot provide the tester with any useful information about the ability (trait or construct) being measured.

However, due to the complex and dynamic nature of the speaking ability, described by Fulcher (2003, p. xv) as “fleeting, temporal and ephemeral”, this human attribute is most vulnerable to the conditions and the temporal aspects of the situations in which it is carried out, hence most difficult to assess. As a result, individuals taking a speaking test of any kind could get different results upon the delivery of such tests under different time and psychological conditions. Alternatively, the testers (raters) could assign them different scores based on their differential, subjective, interpretations or judgements of the testees’ performance. Subjectivity of scoring has been described by Bachman (1990) as one of the limitations on measures of language ability. Tests of language ability, specially speaking and writing as the two productive skills, are subjective at all levels of theorisation, conceptualisation, development, production, administration, and particularly scoring. The subjectivity present at all the above levels can influence the reliability of test results and the validity of the test use and interpretations.
Understandably, the issue of assessing speaking ability becomes logistically more complex when it comes to large-scale testing situations where factors related to the practicality of the test also play an important role in reducing, or at least affecting, the reliability of speaking tests. Reliability and practicality issues are causes for concern in conducting research in the area of testing second language speaking (Bachman & Palmer, 1996). The issues related to the testing process of speaking ability are so far-reaching that, interestingly, until 1926, language testers tried to avoid the complex problem of testing this ability as it lent itself to subjectivity in scoring and interpretation of results (Fulcher 2003).

1.2 Identification of the Problem

The aim of this research project is to establish validity arguments in support of computerisation of the delivery medium of a speaking test. The surrounding issues of practicality and usability which arise in the ongoing development and validation process are also dealt with. According to Davidson (2006), one of the key issues that has been hindering the development and use of computerized testing is the equivalence between computer-based and paper-based tests. In the current study, in order to establish the equivalence between computer-based tests and conventional tests, the degree of comparability of scores obtained from these two tests is investigated. This is carried out by the introduction of a new computer-based test to deliver oral tests. Research indicates that the speaking ability, compared to the other skills of listening, reading, and writing, is the most difficult to assess (Brown, 1987; Luoma, 2004). This difficulty is partly rooted in the way these tests are delivered but it is mostly due to the productive aspect of the skill under investigation. More importantly, factors that render such tests more difficult to design include: a) the
dynamic nature of the speaking ability (difficulty in defining the construct of speaking), b) the subjective nature of the assessment procedure and judgemental fluctuations in the interpretation of the scores assigned to such tests; c) the variation in the environmental factors related to the administration of such tests; and d) the temporal effects of the testing on the testees, i.e., the same testees may perform differentially on the same test on different occasions.

In a typical conventional speaking test, the examinees are interviewed by one or more professionally trained interviewers to determine how well they use the spoken language. In the language testing literature there is a range of definitions used for language assessment interviews (Lazaraton, 2002). However, the definition of language proficiency interview (LPI) proposed by Young and He (1998, p. 10) seems to be reasonably comprehensive:

…a face-to-face spoken interaction usually between two participants (although other combinations do occur), one of whom is an expert (usually a native or near native speaker of the language in which the interview is conducted) and the other a non-native speaker (NNS) or learner of the language as a second or foreign language. The purpose of the LPI is for the expert speaker – the interviewer – to assess the NNS’s ability to speak the language in which the interview is conducted.

Despite its comprehensiveness, the above definition, fails to include the important requirement of ‘professional training’ for the interviewers. Trained interviewers will conduct the interview procedure in a more standardized and uniform fashion and
thereby reduce threats to the reliability of this technique. As stated in the above definition, an interview is a face-to-face encounter, and therefore, is the most common mode of delivery of the oral proficiency test, because it is believed to be the closest replica of a real-life situation, hence the most authentic. The interview is also a measurement and research device whose value should not be underestimated (Lazaraton, 2002). In relation to the difficulty of designing and conducting interviews, Weir (2005, p. 191) admits that

there are enormous practical constraints on the large-scale testing of spoken language proficiency. These include the administrative costs and difficulties and the sheer logistics of testing a large number of candidates either individually or in very small groups. The extent to which the demands of validity can be met through direct tests of spoken language ability will depend on the particular situation test developers find themselves in.

It is largely due to the logistical and practical issues identified above that there has been considerable interest in developing speaking tests delivered through other modes. These include tape-mediated tests (Simulated Oral Proficiency Interviews, SOPI); in language laboratories (Osa-Melero & Bataller, 2001); via video conferencing (Clarke & Hooshmand, 1992); over the telephone, as in the PhonePass Test (Bernstein, 2000) of Ordinate Corporation, (Ordinate Corporation, 1998); and the FBI’s modified oral proficiency test (Cascallar, 1997). A more recent development is the introduction of computer-mediated tests (Kenyon, Malabonga & Carpenter, 2001; Larson, 2000; Stauffer & Kenyon, 2001; Strong-Krause, 2001).
Important for this research is the fact that the use of computer technology in the delivery of speaking tests is the most novel and therefore the least researched field. Thus, many language testers and researchers in the assessment of language ability are currently interested in exploring how computer technology and computer-based tests can be used to address the problems of oral proficiency testing. Although the use of technology is attractive for the flexibility it offers in the testing process, these innovations have been problematic since there are inevitably questions about the effect of the mode of test delivery on test-taker performance (Alderson & Banerjee, 2002) as well as over-estimated ability of computers in assessing language ability (Meunier, 1992). In addition to the effect of test delivery mode, researchers have sought to explore the comparability of scores achieved through conventional and computerized tests of speaking ability (Choi, Kim, & Boo, 2003; Sawaki, 2001). However, the findings are far from complete.

Before dealing with the specific research aims, it is worth discussing the purpose and function of a test of speaking. Such a test is used to motivate the test-takers to speak so the examiner can draw meaningful and valid inferences from the testees’ language sample. This brings us to the critical issue of validity. In the traditional definition, a test is said to be valid if it measures what it is supposed to measure (Henning, 1987). A more recent definition is that a test is valid to the extent that the inferences or decisions made on the basis of the findings (test scores) are meaningful, appropriate and useful (Bachman, 1990). Validity has been described as “the single most critical element in constructing foreign language tests” (Nakamura, 1995, p. 126). Validity has been a major concern, if not the major concern, in language testing. Highlighting this concern, Stevenson (1981, p. 40) states:
…the spirit of validation … not only tries to prevent the willful and careless misuse of tests; it also tries to protect the test constructor from his or her own self-confidence. Moreover, and perhaps most importantly, it tries to protect the test constructor and any future examinees from the acceptance of a measure by those test users who impatiently argue that their practical needs are of primary importance, and that test validation, while nice, is not.

In relation to construct validity, Beale (2003) states that the most fundamental kind of validity relates to the underlying theory of language on which the test is constructed. The process of construct validation is a central issue in language testing and any test development project. A construct is a key component in a theory. To establish the construct validity of a test, the theory underlying a test must be articulated and then the results compared with that theory. Construct validation involves assessing how well a test measures the construct it is based on (Alderson, Clapham & Dianne, 1995). A more detailed discussion of construct definition is given in Chapter 2.

The present research addresses the conceptualization, development, trial, and validation of a computerized test of spoken English. Given the potential effect of the new mode of testing on test-takers’ performance and the concerns raised as to the construct validity of computer-based tests (Green, 1998; Noijons, 1995), this project aims to shed light on:

\( a) \) the comparability of scores obtained from computerized speaking tests with those from conventional interviews.
b) the effects of the new medium of testing on test-takers’ performance.

This is sometimes called “test method facet” (Bachman, 1990, p. 111) or “method effect” (Shohamy, 1984, p. 147).

c) the test-takers’ attitudes towards computerized testing; and

d) the overall usefulness of such tests.

To investigate the usefulness of the new instrument, the framework of ‘Test usefulness’, proposed by Bachman and Palmer (1995) is used. Test usefulness is the most important consideration in designing and developing a new language test and includes the six test qualities of reliability, construct validity, authenticity, interactiveness, impact, and practicality. Each of the above qualities contributes in a unique but interrelated way to the overall value of the test. Depending on the purpose and function of the test, the test developer decides which quality is to be given priority in terms of consideration and attention. In this study, the two qualities of construct validity and practicality are regarded as of primary importance; however, the other qualities will also be given due consideration. This study is called for filling the gap in the current language testing of the usefulness and comparability of computer-based tests. A more detailed discussion of test usefulness and its components will be presented in Chapter 2.

1.3 Significance of the Study

This study is significant within the present context of applied linguistics and second language testing because it brings together conventional language testing practice and the use of new computer technology. It investigates the comparability of scores from face-to-face tests and those from computer-based tests. It is worth mentioning that the
available literature in the field of computers in the assessment of second language ability is very much in favour of the use of technology (Brown, 1997; Chapelle, 2001). However, not much has been done in the use of computers in testing speaking. Computer-Based Tests (CBTs) in general offer many advantages over their paper-and-pencil counterparts. According to Alderson (2000), for example, there are technical, administrative, and pedagogical advantages associated with CBTs (for more on the advantages of computer-based tests see section 2.6). Also, this research reinforces the importance of considerations given to the potential effects of computer-delivered tests on test-taker performance.

Drawing upon the digital media, this project uses a sophisticated computer package to deliver an English oral proficiency interview. Theoretically, it tries to meet the test usefulness criteria proposed by Bachman and Palmer (1996) and practically it provides a greater standardization of test administration conditions. These steps result in increased test security and the use of innovative item types and performance tasks which are not feasible in a paper-and-pencil counterpart (Carr, 2006). This project provides a detailed blueprint for other researchers who intend to design and develop a CBT. Potentially, the software package accompanying this research can accommodate a large number of test tasks, which in turn increases the reliability of the assessment. Therefore, a wide range of professionals in the following areas will benefit from the findings of this study including researchers in test development and task design; researchers in validation and reliability studies in language testing; measurement specialists; test administrators and proctors; language centres and language learning institutions; English language teachers and learners; EFL/ESL interviewers;
researchers interested in the impact of multimedia in language production; researchers in human-machine communication; and computer programmers.

Practically, the findings of the current study have important implications for areas such as:

a) the user interface design in computer-based test development projects;

b) validation studies of new tests of speaking;

c) item writing procedures;

d) task development; and

e) screening and placement procedures. In addition, upon the use by language institutes, the accompanying CD-ROM of this project offers new opportunities to test administrators in terms of convenience regarding the place and time of testing. The test location can be chosen by the testees or the organisation administering the test. This decision can have a potential effect in reducing the test anxiety which is traditionally associated with group administration of conventional large-scale tests. Finally, an obvious advantage of this project, upon practical application of the package, is the possibility of providing quick feedback on the testees’ performance and the ease of adding or removing tasks from the program.

1.4 Purpose of the Study

Given the increasing number of non-English-speaking background students aiming at studying in English-speaking countries, the job of determining their language proficiency levels reliably, validly and practically, for placement purposes prior to the commencement of their study, demands extensive time and budget-related pressures upon individuals and institutions alike. In order to increase the efficiency of test delivery and administration, consistency of scoring procedures, reliability of test results, and quality of performance, it is imperative that we deliver tests and tasks in a unified and standardized format. In the current project, the researcher embarked upon
the conceptualisation, design, development, trial, implementation and validation of a new computer-based test of speaking proficiency, namely, *The SAM Interview*. The SAM Interview is a software package whose purpose is to make best use of computer capabilities to facilitate the delivery of the interview procedure. The results obtained from this test are compared with those of the live interview, particularly from the viewpoint of the test-takers. This will give the researcher(s) a great deal of information and insight in trying to improve on various aspects of task design and test characteristics and input of time and resources. The purpose of this study, therefore, is to demonstrate validity and usability arguments for the SAM Interview package. In developing a new computer-based test, especially for the productive skill of *speaking*, many technical and controversial issues arise. These issues predominantly relate to the technology (logistics and practicality) side of the development of the program and to the psychological aspect of the performance of students on such tests. These issues will be dealt with in more detail in Chapter 2.

Theoretically, this research project sheds light on the validity and usefulness of computer-based tests for speaking. It investigates the comparability of the scores and what they really mean in the context of technology based assessment. Furthermore, this research highlights the significance and nature of the interaction between test-takers and the computer and investigates the impact of and the attitudes towards the CBT, examinee perceptions of CBT task difficulty, and anxiety levels experienced in taking a CBT.
1.5 Thesis Structure

This thesis comprises six chapters. Chapter 1 introduces the background to the research and highlights the significance of the problem, situating it within the broader context of language testing. Chapter 2 presents a detailed review of relevant literature in the field of language testing approaches and the use of information and communication technology, emphasising the need for this research. In addition, this chapter presents the research questions and their corresponding research hypotheses. Chapter 3 deals with the methodology, design and procedure of the study along with a detailed description of the test specifications. It also covers the step-by-step developmental procedures for the SAM Interview software package. The results of the collected data are presented in Chapter 4 and discussed in Chapter 5 in light of the research hypotheses. In Chapter 6, conclusions are drawn based on the findings; and implications of the study are discussed. This Chapter also discusses the limitations of the study and suggests directions for further research.
CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

The aim of this chapter is to review the literature in language testing and technology from which the rationale for the current study is drawn. First, a short historical background to the development of major approaches to language testing is presented. This discussion covers developments from the early work in language testing up to the present time. These trends follow in order but overlap in timing and approach. In sections 2.3 and 2.4, language proficiency and the construct of speaking ability are defined. These sections are followed by a detailed discussion of test usefulness criteria and their relationship to the current thesis. In section 2.6 the ACTFL (American Council on the Teaching of Foreign Languages) guidelines and the ISLPR (International Second Language Proficiency Ratings) scale are discussed. An analysis of the use of computers in language testing in general and in the testing of the speaking ability in particular is presented in sections 2.7 and 2.8. Towards the end of this chapter, in sections 2.9 to 2.11, the literature related to comparability studies, task difficulty in CBTs and attitudes towards computer based tests are reviewed.

2.2 Background: A Brief History of Language Testing

2.2.1 Early Work

Although many language testers consider the professional field of language testing to have formally begun in the early 1960s (Lado, 1961), language tests were in use long before. The first political language test was the Shibboleth test as recorded in The Book of Judges in The Bible. It was an extremely high-stakes assessment which
included a single-item, objective, oral, phonological test, individually administered, and the 42,000 who failed it were slaughtered on the spot (Spolsky, 1995). In the eighteenth and nineteenth centuries, formal examinations in Prussia, France, and Germany played a role in controlling schools and selecting civil servants. And in Britain, Oxford and Cambridge Universities started examinations with public oral disputations and then added written examinations. In the United States, written examinations were introduced at Harvard University in the mid-nineteenth century. In the first half of the twentieth century, language tests gained in stature in the US as the Arm Alpha tests were instituted after the US entered World War I and Cambridge University in Britain began to administer the Certificate of Proficiency and the First Certificate tests in English to assess the language proficiency of their overseas students.

2.2.2 Pre 1920s: The Essay Translation Approach

Until the 1920s, most public examinations were confined to extended written examinations. This was well before the introduction of objective psychological testing (Spolsky, 1995). This era of language testing is commonly referred to as the pre-scientific stage in which no special skill or expertise in testing was required. The pre-scientific movement is generally associated with the grammar-translation approach to language teaching. Since such approaches have existed for many years, the end of this movement is usually clearer than its beginning. The pre-scientific movement “ended with the onset of the psychometric-structuralist movement but clearly such movements have no end in language teaching because teaching and testing practices are going on in many parts of the world” (Mousavi, 2002, p. 675). The pre-scientific era of language testing has three distinct characteristics, as follows:
a) **Abundant translation and composition:** This trend is especially characterized by translation and free composition tests developed exclusively by classroom teachers who are on their own when it comes to developing tests. In this era, oral examinations of any kind are the exception; language testing is assumed to be a matter of open-ended written examinations. Depending on the language teaching philosophy, such examinations would typically consist of passages for translation into or from the foreign language (including free composition) and selected items of grammatical, textual or cultural interest (often in the form of comments about the language being learnt). During this period, language tests were clearly the business of language teachers, or, in more formal situations, of language teachers promoted or specially appointed as examiners. No especial expertise in language testing was required; if a person knew how to teach, it was assumed that he could judge the proficiency of his students.

b) **Lack of objectivity and reliability:** A major issue during this era was that it was very difficult to score examinations objectively. Thus subjectivity was an important factor in scoring such tests. In its simplest form, it assumed that one could and should rely completely on the subjective judgement of an experienced teacher, who could tell after a few minutes’ conversation, or after reading a student’s essay, what mark to give.

c) **Lack of statistical support:** There were no language testing specialists involved in the pre-scientific movement. It follows that there was little concern for statistical matters or for such notions as objectivity, reliability, validity studies, correlation coefficient and so forth.
2.2.3 1930s-1960s: The Growth of Objective Testing and Structuralism

During the 1930s, the growth of objective testing driven by the psychometric approaches to mental tests encouraged the development of objectively scored language tests (Spolsky, 1995). This period was informed essentially by a theoretical view of language ability as consisting of skills (listening, speaking, reading, and writing) and components (e.g., grammar, vocabulary, pronunciation) and an approach to test design that focused on testing isolated, discrete points of language, while the primary concern was with the psychometric reliability (Harris, 1969; Bachman, 2000). This era came to be known as the structuralist period of language testing. This approach is characterized by the interaction of two sets of experts, agreeing with each other mainly in the belief that testing can be made precise, objective, reliable, and scientific. The two groups of experts were those in educational measurement and those specializing in structural linguistics.

a) Educational measurement experts: The first of these groups of experts were the testers and psychologists responsible for the development of modern theories and techniques of educational measurement. Their key concern was to provide ‘objective’ measures using various statistical techniques to ensure reliability and certain kinds of validity. Their first thrust was to demonstrate the unreliability of traditional examinations. This done, they moved to develop more reliable measures, working to find either techniques for making judgements more reliable, or new kinds of test items more amenable to control. The better known work of the testers was the development of short, objective tests including multiple-choice items.
b) Structural linguists: The second major impetus of the scientific period, or approach then, was when a new set of experts added principles from the science of language to those from the science of educational measurement. These were the structural linguists who held the view that language elements could be separated, and that these elements (phonology, vocabulary, and grammar) could be taught and therefore tested separately. They believed that students’ mastery could be tested using words and sentences completely divorced from any context on the grounds that larger samples of language forms can be covered in the test in a comparatively short time. In this approach, the skills of listening, speaking, reading, and writing are also separated from one another because it is considered essential to test one thing at a time. This era witnessed the birth of discrete-point tests whose purposes were to diagnose learner strengths and weaknesses; prescribe curricula at particular skills; and develop specific teaching strategies to help learners overcome particular weaknesses. The arguments of objective testing were particularly effective in setting the model for the development of the TOEFL (Spolsky, 1995).

Despite the above desirable goals, however, there have been strong attacks on the theoretical basis of the structuralist approach. This approach has been especially challenged by two trends in contemporary linguistics. The first, called the ‘language competence trend,’ is connected to various views of psycholinguists. It is based on a belief in a construct called the overall language proficiency, and a feeling that knowledge of a language is more than just the sum of a set of discrete parts (Oller, 1979). They believed that discrete point analysis necessarily breaks the elements of language apart and tries to teach them (or test them) with little attention being paid to
the way those elements interact in the larger context of communication. What makes it ineffective as a basis of teaching or testing language is that crucial properties of language are lost when its elements are separated. For example, the core concept of the ‘meaning’ of utterances is not given the required attention and the authenticity of tests and their washback are not considered at all. The second challenge, coming from what is called the ‘communicative competence trend,’ is connected with views of modern sociolinguists; it accepts the belief in integrative testing, but insists on the need to add a strong functional dimension to language testing (Oller, 1979).

2.2.4 The 1970s: Integrative Testing

The criticisms against discrete-point testing resulted in the emergence of what came to be known as the integrative approach to testing in the late 1970s (Oller, 1979). In this era of testing, tests involved the assessments of language in context and were thus primarily concerned with meaning and the total communicative effect of discourse. The rationale of this approach to testing is derived from the argument that language is creative and dynamic. More precisely, in this approach, language professionals believe that language is more than the sum of the discrete parts being treated in the structuralist approach. In this approach, tests do not seek to separate language skills into neat divisions in order to improve test reliability; instead, they are often designed to assess the learners’ ability to use two or more skills simultaneously. Thus, tests based on the integrative rationale are concerned with a global view of proficiency. At the onset of this approach, it was felt that the development of communicative competence depended on more than simple grammatical control of the language; communicative competence also hinged on knowledge of the language appropriate for different situations. Tests in this era included the cloze, dictation, oral interview,
translation and essay writing, all of which try to assess the test-takers’ ability to manipulate language within a context of extended text rather than in a collection of discrete-point items.

2.2.5 The 1980s to the Present: Modern Communicative Testing

The past 20 years have witnessed the refinement of a rich variety of approaches to the development of language tests (Bachman, 2000). This era is known as the communicative era, during which language tests are concerned primarily, if not totally, with how language is used in communication. Therefore, most tests, in this approach, aim to incorporate tasks which approximate as closely as possible those facing the test-takers in real life. In the communicative approach, tests should meet certain criteria. For example: a) they should be interactive; b) they should be direct in nature, with tasks reflecting realistic discourse-processing activities; c) texts and tasks should be relevant to the intended situation of the target population; d) these tests should be based on an explicit set of *a priori* (test) specifications; and e) ability should be sampled within meaningful and developing contexts (Mousavi, 2002).

In this approach, success is judged in terms of the ability to communicate effectively through language in particular settings and contexts rather than formal linguistic accuracy or performance on discrete-point tests. This involves the notion that linguistic activity in the test should be of the kinds and under the conditions which approximate those of real life (Bachman, 2000; Bachman & Palmer, 1996; Canale & Swain, 1981). The 1980s also witnessed the influence of second language acquisition theories on our understanding of the relationship between notions such as field independence/dependence and language testing (Stansfield & Hansen, 1983), and the
influence of background knowledge on test performance (Alderson & Urquhart, 1985) as well as the use of computers in language testing (Brown, 1997; Chalhoub-Deville, 2001, 2002; Chapelle, 2001).

According to Bachman (2000), during the 1990s language testing issues were broadened and covered other areas of applied linguistics such as research methodology, assessing English for Specific Purposes (ESP) and cross-cultural pragmatics, factors that affect test performance on test-takers, issues related to authenticity and the professionalization of the field of language testing. Furthermore, in the past few years, “advances in the technology of test design and development, along with the availability and use of ever more sophisticated computer- and web-based applications for test administration, scoring and analysis, have resulted in a greater range of test formats and assessment procedures than has ever been available” (Bachman, 2000, p. 2). Despite these advances, however, Bachman (2000, p. 9) continues to stress that “there is clearly a need for research” in the area of the use of computers for design and development of tests. This has prompted the researcher to undertake the design and development of a new computer-based test in the present study.

2.3 Defining Language Proficiency

In this section, a review of fundamental concepts in the language testing literature is presented, with particular focus on the impact and role of technology in assessing second language proficiency. Drawing on the test usefulness model proposed by Bachman and Palmer (1996), this section also presents, in an itemized format, a detailed discussion of a) what we mean by general language proficiency; b) speaking
Another equally practical definition of language proficiency is predicated on a socio-theoretical foundation. What this means is that language is more than just the sum of discrete parts (e.g., pronunciation, vocabulary, grammar). It develops within a culture for the purpose of conveying the beliefs and customs of that culture. Anyone who has ever tried to translate an idiom from one language to another understands this premise. A ‘bump on a log’ in English means someone who is lazy or a do-nothing, but the non-English speaker has to assimilate the idiom rather than the meaning of each individual word in order to make sense of the phrase. This is because language usage is a) dynamic and contextually-based (varies depending upon the situation, status of the speakers, and the topic); b) discursive (requires connected speech); and c) requires the use of integrative skills to achieve communicative competence. In other words, language proficiency is a coherent orchestration of discrete elements, such as vocabulary, discourse structure and gestures, to communicate meaning in a specific context (e.g., the school).

To be able to justify the validity of the use of a test and the interpretation of its results, any language test development project must clarify the theory or model of language proficiency on which it is based. Test development refers to the entire process of creating and using a test, beginning with its initial conceptualization by asking, among other things, the question: “What is this test supposed to measure?” The development
then proceeds to the design, implementation, and archiving stages and the iterative process of validation. Whether it is a small scale, low-stakes (classroom) test, or a large scale, high-stakes test, the question of what the test is trying to measure remains pivotal. To design a test of general English proficiency, once an operational definition of “proficiency” has been presented as an initial step, the later stages of the development process are carried out. In the broadest sense, the term ‘general language proficiency’ refers to what an individual can do with the language at the present moment, no matter how she/he has learnt the language. A more detailed discussion of language proficiency is presented below.

Contrary to common misconceptions, language testing is not a discipline in isolation. Upon appropriate use and interpretation,

“language tests can be a valuable tool for providing information that is relevant to several concerns in language teaching. They can provide evidence of the results of learning and instruction, and hence feedback on the effectiveness of the teaching program itself” (Bachman & Palmer, 1996, p. 8).

Moreover, in teaching and learning contexts, testing has important pedagogical functions too. For example, Mousavi (2002) believes that language tests allow teachers to:

a) find out about a student’s suitability to follow a course of study;

b) find out how a student is progressing during a course of study and possibly identify problem areas before a course ends;
c) compare a student’s performance with that of other students;

d) find out how much a student has learned during the course or academic year, 
   i.e., compare what students can do at the end of the course compared with the 
   beginning of the course;

e) provide evidence of achievements in certain areas;

f) give the students a sense of achievement;

g) facilitate learning through diagnosing the problematic areas;

h) give promotion to students;

i) predict the possible success or failure in language acquisition of certain 
   individuals;

j) show evidence of the effectiveness and efficiency of instruction;

k) see if the materials for instruction were at the right level of difficulty;

l) see if all the language skills are being equally emphasized; and

m) learn more about the language acquisition process at various age levels in 
   order to recommend either an earlier or later starting point, or more effective 
   instructional measures for a target age group.

In fact, language teachers and language testers have a lot to say to each other about 
foreign language proficiency: how it is gained, how it is lost, how it can be recovered, 
how it develops, how it can be measured, and most importantly what it is. Davies 
(1995, p. 1) believes that language “proficiency (unlike [for] the native speaker) does 
not occur in nature, we have to invent it, define it, find something that stands for it.”

The term ‘language proficiency’ has been traditionally used in the context of language 
testing to refer, in general, to knowledge, competence, or ability in the use of 
language, irrespective of how, where, or under what conditions it has been acquired
Literature Review

(Bachman, 1990). One interesting way of defining proficiency is to examine what it is not. To this end, the tradition of distinguishing clearly between achievement (or attainment) and proficiency is a convenient one. Proficiency is general; achievement is local and specific. Proficiency is theoretical or theory-based; achievement is syllabus or materials or curriculum-based. Achievement is parasitic, in the sense that information from achievement tests describes the learning of a single program. However, general proficiency is free standing and describes learning in some absolute sense (Davies, 1995). To obtain a better understanding of proficiency as opposed to achievement, it is perhaps more convenient to contrast proficiency tests with achievements tests.

An achievement test is a test which measures how much of a language someone has learnt with reference to a particular course of study or programme of instruction; hence the term parasitic. The notion of achievement or attainment refers to the mastery of what has been learnt, what has been taught or what is in the syllabus, textbook, or, materials. For this reason, achievement tests are directly related to language courses, their purpose being to measure the extent of learning in a prescribed content domain, often in accordance with explicitly stated objectives of a learning programme within or up to a given time. Depending on the purpose of the instruction and testing, achievement tests may be given as unit, chapter, phase, midterm, end-of-volume, or final tests or any combination of these. What is clear, however, is that the content of these tests must be related to the courses with which they are concerned.

A proficiency test, on the other hand, is a test which aims to measure how much of a language someone has learnt. It seeks to answer the question: Having learned this
much, what can the student do with it? Proficiency tests are designed to measure students’ ability in a language regardless of any training they may have had in that language. The content of a proficiency test, therefore, is not based on the content or objectives of the language courses which test-takers may have followed. Rather, it is based on a specification of what candidates have to be able to do in the language in order to be considered proficient. Despite differences between them of content and level of difficulty, all proficiency tests have in common the fact that they are not based on courses that candidates may have previously taken, because they are concerned with the students’ current standing in relation to their future needs. Although this may be a result of previous instruction and learning, in proficiency testing these are not the focus of attention.

A review of the published literature in second language acquisition indicates that language proficiency models afford language testers numerous but diverse frameworks and components, and at the same time reveal a lack of consensus among models in their definition of proficiency (Bachman, 1990; Davies, 1986; Egan, 1999; Oller, 1979; Spolsky, 1986). Cummins (1984), for example, states that the nature of language proficiency has been understood by some researchers as consisting of 64 separate language components while others, e.g., Oller (1979), see it as consisting of only one global factor.

Proficiency in a language is, therefore, a complex concept. Second language scholars are still trying to define it and identify what its components are. For instance, practitioners in large-scale test development and implementation such as the ACTFL (American Council on the Teaching of Foreign Languages) and the ILR (Interagency
Language Roundtable) continue to work towards a framework for describing language proficiency at different levels that incorporates the elusive components of communicative ability (American Council on the Teaching of Foreign Languages, 1986; Canale & Swain, 1981). For this lack of consensus about the definition of language proficiency, debates about the nature of language proficiency have influenced the theory, design, and implementation of language tests and more importantly the interpretation of test results and justifying the validity generalizations (Bachman, 1990). Validity generalization refers to the extent to which the validity evidence obtained from a particular situation can be applied to other similar situations.

Looking at language proficiency from a historical point of view, Davies (1995, p. 3) states that language “proficiency was widely used in the 1970s and early 1980s under the label general language proficiency synonymous with the unitary competence hypothesis.” More recently, however, appreciation of the complexity of fundamental notions like ‘language proficiency’ and acceptance of the resulting impossibility of finding an interpretable single measure has led to a realization that assessment of language knowledge is both multipart and multifaceted as well as intricate, thus more likely to be served by profiles than by simple and/or single scores. The term ‘proficiency’ has come to be associated in foreign language teaching circles almost exclusively with a specific language testing procedure. But the most general phrase which encompasses all the meanings and connotations of language proficiency is ‘language ability’ (Bachman, 1990). Language ability is the capacity for using the knowledge of language in conjunction with the features of the language use context to create and interpret meaning, while language proficiency is a person’s skill for using a
language for a specific purpose. The phrase ‘communicative competence’ is sometimes synonymously used with language proficiency (Bachman, 1990).

The successful development, implementation and validation of any new testing instrument must, therefore, be based upon a sound theory of language proficiency which clearly defines the construct to be measured as well as the ways in which the language sample is elicited for measurement. The Communicative Language Ability (CLA) framework formulated by Bachman (1990) is used as a guide in this research.

In this framework, Bachman describes communicative language ability as consisting of both knowledge, or competence, and the capacity for implementing or executing that competence in the appropriate, contextualised communicative language use. He views language as multicomponential and proposes an interactional model of language test performance in which language ability is seen as interacting with test method (characteristics of the environment, rubric, input, expected response, and the relationship between input and expected response) to produce a performance that can be described and reported. This framework essentially includes three components: language competence, a set of specific knowledge components that are utilised in communication via language; strategic competence, the mental capacity for implementing the components of language competence in contextualised communicative language use; psychophysiological mechanisms, the neurological and psychological processes involved in the actual execution of language as a physical phenomenon. Bachman’s (1990) model is perhaps the most widely adopted model of language ability in language testing. It has provided a principled, systematic basis for
the development of language tests, such as the English proficiency test for non-native English language teachers in Australia (McDowell, 1995).

Therefore, in this study, the performance of the test-takers on a computer-based test will be analysed in light of Bachman’s communicative language ability framework. Furthermore, the test usefulness framework proposed by Bachman and Palmer (1996) will be used as an approach to test design and development.

### 2.4 Definition of the Speaking Construct

Of the four skills of listening, speaking, reading, and writing, this research project focuses on the assessment of speaking ability only. As noted earlier, what speaking ability is and what constitutes ability to speak requires what is called ‘definition of the construct’. In order to be able to measure a particular construct, we need to know what that construct is and try to operationally define it. An operational definition of a construct enables a test designer to relate the theoretical construct to the actual observations of behaviour, the behaviour being the individual’s spoken language samples. Speaking ability, among other abilities, is a mental construct (Fulcher, 2003). A mental construct, also called a trait, refers to a set of mental tasks that an individual is required to perform on a given test. In general, mental and internal abilities are psychological qualities that can only be observed indirectly (Anastasi & Urbina, 1997).

Unlike physical constructs such as age, height, eye colour and hair whose characteristics and magnitude can be observed and measured directly, we cannot directly observe psychological constructs, but we can observe the behaviours resulting
from the existence of those underlying constructs. In other words, “constructs are unobservable and abstract qualities of which only the representations or manifestations can be observed and therefore assessed and measured” (Mousavi, 2002, p. 775).

According to Kerlinger and Lee (2000), a construct differs from a concept in two important ways. Firstly, it is defined in a way that can be observed and measured. Secondly, the relationship between different constructs constitutes a theory. If we consider speaking as a mental construct, then we need to look for observable manifestations of this construct to be able to measure them. The observable manifestation of this construct occurs when an individual actually speaks or produces spoken language samples.

On the face of it, this may appear to be quite simple. However, speaking is perhaps one of the most sophisticated human behaviours, the quality and nature of which is dictated by numerous external factors. Speaking does not normally occur in isolation. In other words, in an actual language use situation, speaking takes place within a surrounding context which is influenced by a host of contextual factors including other individuals and their social role, status, power and inter-personal relationships, as well as the interaction of the above factors.

As a result, the construct definition of speaking will include extralinguistic and contextual factors (Chapelle, 1999) and aspects of the target language use domain (Bachman & Palmer, 1996) as well as a description of language outcomes (Department of Education Queensland, 1994). Those language outcomes are the
observable features of language use, related attitudes, processes, skills and knowledge that are essential and distinctive to the (English) language. These outcomes identify the kind of performance typically demonstrated as individuals become more proficient in learning about and using English.

In defining the speaking ability, it is advisable to draw a distinction between spoken and written language. A careful analysis and description of the spoken genre will give us a better understanding of the nature of speaking. As noted by the Department of Education Queensland (1994), spoken and written texts differ in their social context and textual features (cohesion, vocabulary, grammar) and therefore serve different purposes.

Other equally crucial factors to be included in any construct definition of speaking are the tasks (e.g., formats and difficulty) and mode of test delivery (e.g., face-to-face, tape, video, telephone, computer) as well as methods of scoring the language sample (e.g., analytic or holistic). Due to the richness of the nature of human communication, especially at the speaking level, Fulcher (2003, p. 19) admits that the list of requirements to be included in the definition of a construct cannot be “exhaustive in any way.” This implies that, for different test purposes, we are likely to have different sets of factors comprising the construct definition. Luoma (2004, p. 27), for example, brings a social aspect into construct definition and defines speaking as a “meaningful interaction between people” who intend to communicate issues of mutual interest, exchange information and experience, or even pass the time or amuse each other, share opinions and get something done. The common denominator to the above
activities is that they can take place simultaneously, which adds to the complexity of the speaking phenomenon both as a social interaction and as a subject of assessment.

Fulcher (2003, p. 20) states that “from what we know about speaking and testing second language speaking from research, we can therefore ‘pick and mix’ to make a construct.” To do this, for every unique target language use domain, a set of empirically convincing and theoretically appealing arguments needs to be provided to make the test useful for that particular domain of language use. The target language use for the current project is considered to be situations where individuals are required to perform in tasks of the following kinds:

- Describe a phenomenon
- Explain a procedure or set of events
- Define a situation
- Report an accident
- Narrate a story
- Exchange ideas and information

Obviously, “mastering the grammar and lexicon of a language is not sufficient for using a language to achieve ends in social events” such as the ones in the above activities (Mislevy, 2002, p. 1). In real-life situations, language use is undertaken in contexts that are more realistic and complex than in, say, discrete-point tests or isolated items of assessment. Alternatively stated, for a successful performance in social contexts, speakers need knowledge beyond that of, or about, the language itself. This knowledge entails the integration of topical, social, and or pragmatic knowledge.
along with knowledge of the formal linguistic elements (Mislevy, 2002). An approach in language testing which tries to emulate the activities in real-life situations in testing contexts is task-based assessment. Tasks, by virtue of their nature, format, difficulty, and complexity, are crucially important in guiding testees to perform at their best in a given assessment situation. Tasks are defined by Luoma (2004, p. 30) as “activities that people do.” In general terms, tasks refer to any activity in which students are engaged with comprehending, manipulating, producing or interacting in a target language with a focus on meaning (Nunan, 1993). But in more precise terms, the definition of a speaking task is that given by Luoma (2004, p. 31): “activities that involve speakers in using language for the purpose of achieving a particular goal or objective in a particular speaking situation.” Luoma (2004) categorizes speaking tasks into two major groups: open-ended and structured tasks. Each has a set of different task types as summarised below:

I: Open-ended Speaking Tasks

A: Discourse type tasks

1. Descriptive
2. Narrative
3. Instruction
4. Comparison
5. Explanation
6. Justification
7. Prediction
8. Decision

B: Role play

1. Social simulation
2. Professional simulation

II: Structured Speaking Tasks

A. Reading aloud
B. Sentence repetition
C. Sentence completion
D. Factual short-answer questions
E. Reacting to phrases

The open-ended tasks are the ones which require examinees to do something with the language as an indication of their ability. Activities such as a presentation and making a request are among these tasks. On the other hand, structured speaking tasks “are the speaking equivalent of multiple choice tasks and therefore the expected answers are usually short, and the items tend to focus on one narrow aspect of speaking at a time” (Luoma, 2004, p. 50). The above classification is analogous to the traditional categorization of item types into structured-response items and selected response (true/false, multiple-choice, or matching) items. Tasks performed in a speaking test (including the face-to-face interview and a computer-delivered test of speaking) in which the test-taker is expected to generate connected discourse are considered to be instances of discourse type tasks and may allow for some creativity.

2.5 Test Usefulness and Its Components

In this section, the framework against which the quality of the current project is assessed will be described. It is self-evident that the design, development, and use of a new test should have a metric by which its value can be evaluated. Bachman and Palmer (1996) noticed a gap in the language testing literature regarding the definition
of what makes a test useful. As a result they came up with a model of ‘test usefulness’ which they think can be used as a yardstick to evaluate the quality of an existing test or a test under development. Test usefulness comprises the six qualities of:

   a) Reliability
   b) Construct Validity
   c) Authenticity
   d) Interactiveness
   e) Impact
   f) Practicality

Each of the above qualities is a stringent requirement in its own right and they have traditionally been studied and discussed independently of each other. However, Bachman and Palmer (1996, p. 18) suggest that, rather than emphasizing a certain quality separately (while ignoring others), test developers need to recognize the complementarity of the qualities and try to establish a balance among them, because “[these] qualities contribute in unique but interrelated ways to the overall usefulness of a given test.” To operationalize the usefulness in the design and development of a new test, Bachman and Palmer (1996, p. 18) believe that the following three principles must be used:

   Principle 1: It is the overall usefulness of the test that is to be maximized, rather than the individual qualities that affect usefulness.
Principle 2: The individual test qualities cannot be evaluated independently, but must be evaluated in terms of their combined effect on the overall usefulness of the test.

Principle 3: Test usefulness and the appropriate balance among the different qualities cannot be prescribed in general, but must be determined for each specific testing situation.

In order to be useful, any given test must be developed with a specific purpose and for a particular group of test-takers and a specific language use domain. Depending on the importance and purpose of a given test, evaluating the overall usefulness of a test is essentially subjective, since this involves value judgements on the part of the test developer. The following are brief descriptions of the six test usefulness qualities.

Reliability: Reliability is a quality of test scores which refers to the consistency of measures across different times, test forms, raters, and other characteristics of the measurement context. Concern for reliability comes from the necessity for dependability in measurement. In general, synonyms for reliability are: dependability, stability, consistency, predictability, and accuracy. A reliable man, for instance, is a man whose behaviour is consistent, dependable, and predictable; that is, what he will do tomorrow and next week will be consistent with what he does today and what he did last week. This characteristic of reliability is sometimes termed consistency. For example, we can readily see how measurement with a steel tape measure would give more reliable or consistent results than measurement with an elastic tape measure. Thus, we infer that the steel tape measure is a more reliable instrument. By the same token, a reliable language test is one which yields consistent results across different
administrations, markings, forms, and raters. So reliability is essentially the “quality of test scores” but not the test itself (Bachman & Palmer, 1996). In order to maximize reliability of test scores we should try to minimize measurement error. Any construct irrelevant factor — such as poor health, fatigue, lack of interest or motivation, differential familiarity with computers (in taking computer-based tests), computer anxiety, or attitudes towards computers — that can affect individuals’ test performances, but which are not generally associated with the construct of language ability, and thus not characteristics to be measured by language tests, are considered as threats to reliability (Bachman, 1990). When the effects of these various factors are minimized, the measurement error is also minimized, which contributes to enhanced reliability. In this study, efforts are made to standardize the delivery of the interview with the aim of reducing the effect of construct-irrelevant factors, or unintentional testing of irrelevant constructs such as the ability to use a mouse or use the scroll bar (Kobrin, 2000).

**Construct Validity:** Bachman and Palmer (1996, p. 2) define construct validity as “the meaningfulness and appropriateness of the interpretations” that are made on the basis of test scores. In other words, for test scores to be meaningful researchers need to be able to provide evidence that the scores reflect those areas of language ability which are supposed to be measured. Therefore, an operational definition of the abilities under investigation becomes necessary. It follows that once we have established a theory defining the constructs of language or language learning, the test needs to be able to measure only that construct. For example, if the assumption is held that systematic language habits are best acquired at the elementary level by means of the structural approach, then a test which emphasizes the communicative aspects of the
language will have low construct validity. Conversely, if a communicative approach to language teaching and learning has been adopted throughout the course of instruction, a test comprising chiefly multiple-choice items will lack construct validity. By the same token, if we assume that communicating with the new technology and computer screens is essentially part of the daily lives of the new generation of students, then the use of the computer delivered tests can be justified in that they are not likely to threaten the construct validity.

**Authenticity:** Authenticity is an important test quality and has been the topic of discussion among language testing scholars for more than a decade (Hoehje & Linnel, 1994; Lewkowicz, 1997; Shohamy & Reves, 1985). It refers to the degree of correspondence between performance on the test and the specific domains other than the test itself. It is the extent to which the tasks required on a given test are similar to normal real-life language use. Bachman and Palmer (1996, p. 23) refer to normal real-life language uses as “target language use (TLU) domains.” In designing and developing a new test, the clearer the correspondence between target language use domains and the tasks on the test, the more authentic the test will be (Wu & Stansfield, 2001).

In this project, performance on a computer-based test is believed to be a reflection of performance on a larger domain of language use, i.e., using the technology to communicate orally in the real world. In other words, interpretations drawn from the quality and nature of performance on computer-based tests can be generalized beyond test context to language use in the TLU domain or to other similar non-test language use domains. Once a correspondence between test performance and performance on
TLU domain has been established, it allows the researcher to generalize the score interpretations in the interest of the construct validity of the test. In this study, communicating with the computer is considered to be an aspect of real-life and authentic language use at the university context, and therefore it is reflected in the design of the test tasks.

**Interactiveness:** As the fourth component of test usefulness, interactiveness is defined by Bachman and Palmer (1996, p. 25) as “the extent and type of involvement of the test-taker’s individual characteristics in accomplishing a test task.” Individual characteristics such as language ability, topical knowledge, metacognitive strategies, and affective schemata are considered to be factors influencing the nature and the extent of interactiveness between the test and the individual taking test. In this study, the nature of the interaction (McNamara, 1997) between the test-takers and the technology-based test is investigated from the viewpoint of their reactions, attitudes and perceptions towards the tasks and their characteristics. In taking a computerised oral proficiency test, as opposed to a selected response test, individuals are actively attending to the tasks by listening to the questions and responding in extended discourse.

**Impact:** The impact of language tests is a relatively new topic in language testing. A test is believed to have an impact on not only individuals, but the education system and society at large. Test impact, as another test usefulness quality, refers to the prospective and potential behavioural changes that may occur in education and society as a result of the introduction of a new test or testing approach. The introduction of a new testing instrument (most notably a computer-based test) may influence different
aspects of the language education industry including curriculum, teaching methods 
(Saif, 2006), teaching and learning strategies (Watanabe, 1996), material and 
courseware, assessment practices, logistical demands, gate keeping policies, and even 
the content of instruction. Therefore, efforts need to be made to maximize the positive 
impact of tests on the above areas. Washback or backwash is considered by Bachman 
and Palmer (1996, p. 30) as “an aspect of impact that has been of particular interest to 
both language testing researchers and practitioners” (Alderson & Wall, 1993; Bailey, 
1996; Prodromou, 1995; Shohamy, Donista-Schmidt & Ferman, 1996; Wall, 1997; 
Wall & Alderson, 1993; Wesdorp, 1982). In this study, the impact of the introduction 
of a new computer-based test is investigated from the view point of test-takers’ 
reactions and attitudes towards the test.

**Practicality:** The practicality of a test, as the last quality of test usefulness, is 
sometimes referred to as efficiency, usability or financial validity (Alderson, et al., 
1995; Brown, 1996). It refers to a whole host of practical characteristics of a test such 
as costs, the amount of time it takes to construct and to administer, ease of scoring, 
and ease of interpreting and reporting the results. Bachman and Palmer (1996) believe 
that, for any given situation, if the resources required for implementing the test exceed 
the resources available, the test will be impractical, and will not be used unless 
resources can be allocated more efficiently, or additional resources can be allocated.

In designing and developing a new test, efforts must be made to achieve the optimum 
balance among all the above test usefulness qualities for a particular testing situation. 
It is self-evident that the use of digital media in the design and delivery of a test is a 
necessary undertaking which has already started to revolutionize the assessment
industry. In the current study, the practicality of the new test is investigated with respect to financial considerations, time (for developing and using the test), space (for implementing the test and storing test data), administration procedures, scoring criteria, and special material and equipment required (including computers, microphones and headsets).

2.6 Testing Speaking Ability: Issues and Approaches

In this section, the rationale and importance of tests of speaking ability are highlighted, with the focus on issues such as their initial conceptualization, administration, scoring, reliability, validity and practicality. Several widely used tests of speaking ability are also described in terms of their practicality and test usefulness.

In a historical overview of speaking tests, Fulcher (2003) reports that, during the Cold War, the initial impetus for the development of speaking tests in the US undoubtedly came from a military need. He further states that it was not until 1952 that it became necessary to measure the speaking skills taught by the Foreign Service Institute (FSI) instructors, when the Civil Service Commission decided to create a register of personnel documenting their familiarity with foreign languages and cultures.

Speaking is at the heart of second language learning and arguably the most important skill for business and government personnel, but it is also liable to deteriorate quickly over time if not practiced regularly. Despite its importance and fragility, however, speaking has been largely ignored in schools and universities, primarily for logistical and administrative reasons which resulted in the excessive emphasis on grammar and culture as well as unfavourable teacher-student ratios (Egan, 1999). “Testing the
ability to speak a modern language set a daunting challenge” (Spolsky, 1995, p. 77) and therefore was absent from the curriculum because of the difficulty in evaluating it objectively and the time it takes to conduct such tests reliably (Clifford, 1987). As noted by Spolsky (1995, p. 318, “even when the specifications for the TOEFL were being developed in the early 1960s, there seemed to be no pressure to give high priority to the testing of foreign students’ speaking ability.” This lack of attention to speaking ability is in sharp contrast to the popular, but reasonable, assumption that knowing a language means being able to “speak” it. The testing of oral proficiency became more commonplace in the 1980s and was no longer driven by military needs. This coincided with the communicative language teaching era (Brumfit, 1984; Savignon, 2002).

The situation is quite different now, in the sense that an increased emphasis on the acquisition of communicative competence (Canale & Swain, 1981) requires language learners to develop a reasonable oral proficiency in English. This has become an essential prerequisite for study in universities and academic institutions in English speaking countries. International students coming from non-English speaking countries are required to sit for speaking tests as part of their general proficiency test, to prove their oral ability in order to be able to commence academic courses. A key principle related to assessment is that oral testing must be done orally, not through written tests (Brown, 1987). This places a heavy burden on teachers who can have hundreds of students to evaluate. At times, individual students can be evaluated in a whole class or group setting, but, for the best and fairest result, it should be done on a one-to-one basis. However, given the large number of students that foreign language teachers must evaluate, usually over a short span of time, this would go against the
testing principle of ‘practicality.’ Defining practicality as “the relationship between the resources that will be required in the design, development and the use of test and the resources that will be available for these activities,” Bachman and Palmer (1996, p. 36) also believe that, although a consideration of practicality logically follows the consideration of other qualities, this does not imply that practicality is any less important than the other qualities of test usefulness.

Another major consideration in testing oral ability is setting tasks that form a representative sample of the population of oral tasks that a tester expects examinees to be able to perform. According to Hughes (1989, p. 101), “the tasks should elicit behaviour which truly represents candidates’ ability and which can be scored validly and reliably.” Weir (2005, p. 103) emphasises that, for purposes of validity:

> there is a strong case for testing spoken language performance directly, in realistic situations, rather than testing hypothetical knowledge of what might be said. If we wish to make statements about capacity for spoken interaction, we are no longer interested in multiple-choice, pencil-and-paper tests, that is indirect tests of speaking where spoken language is conspicuously absent.

It can be concluded from the above statements that it seems best to test speaking by asking individuals to speak. The complexity and nature of the elicitation techniques employed to encourage individuals to speak may range from simple short-answer questions, in which test-takers produce a very limited response to a stimulus, to more complex open-ended tasks like interviews on which individuals must produce extended cohesive and meaningful discourse. Whatever the task or procedure, the
technique must be able to elicit representative amounts of naturally produced language from the testees, to provide the tester with opportunities to make meaningful inferences and interpretations of the test results. This is what is technically called the validity of a test, one of the six qualities of test usefulness discussed above. Not only should the interpretations be meaningful, but they should be meaningful consistently.

Another of the six qualities of test usefulness is a notion which emphasizes the consistency of accurate interpretations. This is called reliability. As two fundamental concepts in language testing, reliability and validity are also two of the stringent requirements of test usefulness (Bachman & Palmer, 1996). The relationship between reliability and validity is in principle a simple one, but in practice rather complex and not well understood. Traditionally, it is believed that, for most kinds of validity, reliability is a necessary but not sufficient condition. Alternatively stated, it is possible for a test to be reliable (produce consistent results) without being valid for a specified purpose, but it is not possible for a test to be valid (measure what it is supposed to measure) without first being reliable. That is, if a test does not measure something consistently, in this case speaking ability, it follows that it cannot always be measuring it accurately. Reliability, therefore, is a requirement for validity. On the other hand, it is quite possible for a test to be reliable but invalid, which means that a test can, for example, consistently give the same results, although it is not measuring what it is supposed to measure. Therefore, although reliability is needed for validity, it alone is not sufficient.

Validity, on the other hand, is a more conceptual issue and has been around for a long time. In reflecting the traditional views about validity, Kelly (1927, p. 14) noted that
“the problem of validity is that of whether a test really measures what it purports to measure.” A test is said to be valid to the extent that it measures what it is supposed to measure or can be used for the purposes for which it is intended (Lado, 1961).

The modern definitions of validity, however, ask whether the test results or the interpretations drawn from them are meaningful for the specified purpose of the test (Bachman, 1990). It follows that the term valid, when used to describe a test, should usually be accompanied by the preposition for. Any given test then may be valid for some purposes, but not for others, depending on how we interpret the results. The matter of concern in testing is to ensure that any test employed is valid for the purpose for which it is administered. Validity tells us what can be inferred from test scores. As stated by Weir (2005, p. 12):

validity is perhaps better defined as the extent to which a test can be shown to produce data, i.e., test scores, which are an accurate representation of a candidate’s level of language knowledge or skills. In this revision, validity resides in the scores on a particular administration of a test rather than in the test per se.

After all, validity is increasingly seen, according to Bachman (1990), as a unitary entity, evidence for which includes content relevance, criterion relatedness, and meaningfulness of the construct(s). However, Weir (2005, p. 13) believes that “validity is multifaceted and different types of evidence are needed to support any claims for validity of scores on a test. These are not alternative but complementary aspects of an evidential basis for test interpretation.” Finally, validity is also now
recognized as a matter of degree rather than an all-or-none issue (Messick, 1989). This means that a test may not produce a perfect fit in one particular aspect of validity; however, it may demonstrate a stronger match with certain test specifications. Therefore, validity should be viewed as a relative concept. This study is a step forward in generating validity evidence for the new generation of computerized oral proficiency tests. It does so by employing the multimedia facilities of computers in delivering the tests.

Apart from the discussion of validity and reliability, an important consideration in using oral tests is the range of rubrics and elicitation techniques that may be used to get the testees to produce language for assessment. Some of the commonly used techniques include impromptu speeches, interviews, role play, discussion, and picture stories (Underhill, 1987). Among all these techniques, however, the most direct, obvious, and supposedly valid measure of speaking ability is the interview. This is because an interview consists of a direct face-to-face encounter between the interviewee and the interviewer(s) (Underhill, 1987). It may take several configurations, including one-on-one interview, paired tasks between examinees, and group testing. More recently, live interaction is being tested through the telephone or teleconferencing. This last option is normally taken into consideration when it is logistically difficult to bring the tester and the examinee face-to-face for geographical reasons. Luoma (2004) believes that the main characteristic of the live interview is that it is interactive and two-directional, so that speakers can take turns and clarify each other’s responses to facilitate the smooth and natural flow of the conversation. The interview technique is thought to be a direct test in the sense that it duplicates the setting and operation of the real-life situations in which proficiency is normally
demonstrated. Direct tests are normally contrasted with indirect tests. Indirect tests attempt to measure the abilities which underlie the skills under consideration. Such tests do not require the test-taker to perform tasks that directly reflect the kind of language use that is the target of assessment; rather, an inference is made from performance on more artificial tasks (Mousavi, 2002).

On the negative side, however, Hughes (1989, p. 104) believes that the interview, “in its traditional form, has at least one potentially serious drawback. The relationship between the tester and the candidate is usually such that the candidate speaks to a superior and is unwilling to take the initiative.” This is why an interview in itself does not necessarily guarantee a valid or reliable assessment of speaking ability, unless the test is very carefully arranged, properly administered, and fairly scored. The validity of the use of the interview procedure as an elicitation technique and the reliability of scoring it depend on a plethora of factors including the structure and procedure of the interview, the preparedness and expertise of the scorers, the subject matter or contents of the questions to be asked, and the time available.

A further issue in establishing the validity of speaking tests is the way the speaking samples are scored. This is because, in university contexts, there are enormous practical constraints on large-scale testing of spoken proficiency (Weir, 2005). These constraints include the time and personnel required to score the language samples of large groups of test-takers. Furthermore, Weir (2005, p. 192) admits that “the assessment of spoken language is potentially more problematic than the rating of spoken scripts, given that no recording of the performance is usually made.”
The literature on testing oral ability offers a series of techniques to get around these problems. Hughes (1989, p. 105), for example, suggests testers make the oral test last as long as is feasible in order to be able to elicit as much information as possible. The reliability of a test is affected by the length of the test; in other words, the longer the test, the more reliable it will be. However, adding more items to a test improves the reliability to a certain point of asymptote where little further contribution is made to reliability regardless of how many more items are added (Farhady, Jafarpur & Birjandi, 1994). Hughes (1989) also argues that interviewers must be selected and trained carefully, and that successful interviewing is by no means easy and not everyone has an aptitude for it.

In the literature on conducting oral tests, spontaneous interviews are not recommended (Underhill, 1987). It is essential that the interviewers decide on the general content of the encounter in advance. It is also advisable that first-time interviewers give themselves some practice with a number of specific questions. To obtain dependable results in rating an interview, it is necessary to utilize the services of at least two trained raters to ensure inter-rater reliability of the results. The average of two ratings represents a better appraisal of the examinees’ performance than any single rater’s judgement.

Along with face-to-face interviews, there are also tape-based tests, which are basically one-directional in the sense that the examinee is expected to accommodate to the tape while the tape cannot possibly accommodate to the examinee (Luoma, 2004). Tape-based tests are usually used when there are large numbers of examinees and it would be difficult to get enough live testers to interact with each and every examinee at a
given time. From a practical point of view, developing a master for a tape based test requires a large amount of work, but once it has been developed, it can be consistently used in many locations and occasions. In addition, tape-based testing offers efficiency in administration as well as opportunities for score comparability studies. An important finding about the use of tape-based testing is that it seems to be measuring the same skill as a face-to-face test (Stansfield & Kenyon, 1991) which confirms the validity of its use. But it seems that “when the examinees talk to a tape recorder, their language is a little more literate and less oral-like, and many of them feel more anxious about the test because everything they say is recorded and the only channel they have for communicating is speaking” (Luoma, 2004, p. 45). The above quote from Luoma has important implications about the nature of the language produced and the anxiety induced by the use of tapes. In spite of this, however, many testees also report positively, admitting that the tape-based test can be a good test of their speaking skill even if they prefer live testing (Stansfield & Kenyon, 1991).

Following the demise of the cassette tape in the 1990s in most parts of the world, the new computer technology took over and revolutionised the audiovisual world. The first computer-mediated test of speaking ability was perhaps the Computerized Oral Proficiency Instrument (COPI) developed by Malabonga and Kenyon in 1998 at the Centre for Applied Linguistics in the United States. (For more detailed discussion of the use of computers in oral testing see section 2.8 below). As a result of merging new computer technology with conventional testing procedures using face-to-face interviews, several interesting questions have been raised by researchers as to the nature of the language sample produced in the interaction, the effect of the delivery medium on the test-takers’ performance (Goldberg & Pedulla, 2002), and test-taker
reactions to the new mode of test delivery (Horton, 1987). Also, the new technology has triggered research into the comparability of scores obtained from computerized testing and those from conventional assessment (Chan & Leung, 2003, Choi, et al., 2003; Kenyon & Malabonga, 2001; Rico & Vinagre, 2000; Neuman & Baydoun, 1998; Sawaki, 2001).

Before discussing computer-based testing, and its potential in testing spoken ability in particular, it is necessary to close the current section with an overview of two important sets of general proficiency scales used in scoring speaking tests: the American Council on the Teaching of Foreign Languages (ACTFL) scale and the International Second Language Proficiency (ISLPR) scale.

2.6.1 ACTFL OPI: An Overview

The initial impetus for the ACTFL\(^1\) came in 1979 when US President Carter’s Commission on Foreign Language and International Studies published its report entitled *Strength through Wisdom (1979)*. One of the recommendations of that commission was the development of a standard means of rating the language proficiency of US government employees. According to the recommendation, the criteria were to be able to differentiate testable levels between ‘no knowledge’ of the foreign language and ‘total mastery.’ The result of that recommendation was the publication in 1982 by the American Council on the Teaching of Foreign Languages (ACTFL) of ACTFL Provisional Proficiency Guidelines.

---

\(^1\) For further information on this see Spolsky (1995).
The Guidelines describe in detail four proficiency levels — *novice, intermediate, advanced* and *superior* — for each of the four language skills of listening, speaking, reading, and writing. The Guidelines are based on the claim that all second language users must be measured ultimately in relation to the educated native speaker, who is thus taken as the norm against which all second language speaker performance is judged.

Following their publication, the Guidelines were widely distributed for comment throughout the foreign language teaching profession. Several hundred individuals were later trained to administer a face-to-face speaking test to assign one of the proficiency levels defined in the Guidelines to each person tested. As a result of their field-testing, the Guidelines were determined to be an appropriate scale for assessing language proficiency among secondary and college level students of foreign languages. Following minor revisions, the scale was republished in 1986 as the ACTFL Proficiency Guidelines. The 1986 version represents a hierarchy of global characterizations of integrated performance in listening, speaking, reading and writing. Each description is a representative, not exhaustive, sample of a particular range of ability, and each level subsumes all previous levels, moving from simple to complex in an *all-before-and-more* fashion. The ACTFL Proficiency Guidelines are available from a number of sources (e.g., Breiner-Sanders, Lowe, Miles & Swender, 2000).

The ACTFL Guidelines are employed in assessing the language proficiency level of test-takers in the Oral Proficiency Interview (OPI). The OPI is a structured interview whose purpose is to elicit pronunciation, fluency, integrative ability, sociolinguistic
Literature Review

and cultural knowledge, grammar, and vocabulary (Henning, 1992). The OPI has long
been regarded as a direct and therefore valid test of speaking ability of foreign
language learners. A typical OPI is conducted in four notional phases as follows:

1: Warm-up: the tester puts the examinee at ease, leads him/her into
communication in the language, and makes a preliminary assessment of
his/her skills.

2: Level-check: the tester finds the level at which the examinee can function
comfortably.

3: Probes: the tester uses questions and language tasks at a higher proficiency
level, and finds the level at which he/she can no longer function adequately.

4: Wind-up: the testee terminates the interview session by returning to the level
of language use at which the examinee can communicate comfortably.

The speaking performance of individuals on a given speaking test is usually rated
using a rating scale which, according to Upshur and Turner (1995) and Turner and
Upshur (1996), has been in use for many decades. In this study a holistic rating
approach to scoring the tests is used. A holistic approach is one in which one or more
raters listen to the language sample produced by an examinee and mark it as a whole,
assigning a single score based on the total impression of a performance (Carr, 2000;
Huot, 1990). The point of the holistic approach to scoring a given performance is that
it employs the raters’ full impression of a speech sample without trying to reduce their
judgement to a set of generalizable skills. In other words, the holistic scoring method
follows the principles of Gestalt psychology (Oller, 1979) and is based on the theory
that a whole piece of performance is greater than the sum of its parts. An obvious
advantage of the holistic approach is that it asks the rater to concentrate on what the
speaker has done, given the conditions under which the test is administered. Another
advantage of the holistic approach is that, since it requires a response to the speaking
task as a whole, the test-takers do not run the risk of being assessed simply on the
basis of a single aspect of their speaking (e.g., pronunciation only). In other words, the
test-takers are not unfairly penalized for a poorer-than-expected performance of some
aspect of their speaking.

On the negative side, however, the holistic approach has a potential drawback. When
the assessors use a set of guidelines as a point of reference for evaluating a speech
sample, they may have personal or even idiosyncratic interpretations of the same set
of guidelines and therefore such ratings are likely to be subjective due to bias, raters’
fatigue, internal lack of consistency, previous knowledge of the student, and/or
shifting standards from one speech sample to the next. Some safeguards such as
training raters can increase objectivity of the ratings.

2.6.2. The ISLPR

The criteria for the International Second Language Proficiency Ratings (ISLPR) are
used in the current research for rating the speaking performance of the examinees.
The ISLPR is a proficiency scale with twelve levels, ranging from zero to native-like
ability, in four classes, for the macro-skills of listening, speaking, reading, and
writing. The scale is used to determine the general language proficiency of test-takers.
According to its developers, Wylie and Ingram (1999), this scale is fully adaptive to
the candidates in terms of their proficiency levels — as hypothesised and continually
re-hypothesised by the tester as the test progresses — and communicative needs and
interests. In the ISLPR, the abilities of speaking, listening, and reading are tested in a face-to-face interview. The interview is usually one-to-one but in very high-stakes situations a second person may be involved, acting as an observer, and can involve him/herself in appropriate parts of the interview with higher level candidates.

The first version of this scale, called ASLPR (Australian Second Language Proficiency Ratings), was developed in 1979 at the Centre for Applied Linguistics and Languages, Griffith University, Brisbane, Australia. The authors decided to change the name from the Australian Second Language Proficiency Ratings (ASLPR) to the International Second Language Proficiency Ratings (ISLPR) in 1997. The main reason for the name change was the rapid growth in the international use of the scale and associated assessment procedures. The original English version of the ISLPR drew upon the Absolute Language Proficiency Ratings developed by the US Foreign Service Institute (FSI) School of Language Studies in 1979. The authors of the ISLPR (Ingram & Wylie, 1979) note that the most recent (1995/99) versions of the test are the result of a series of extensive research and feedback from international test-takers and professional test administrators.

In the ISLPR, the subscales trace the development of learners of a second or foreign language from 0 (no ability to communicate in the target language) to 5 (indistinguishable from a native speaker of the same sociocultural background). There are 12 levels including intermediate ‘plus’ or ‘minus’ levels. The description at each level includes a statement of the kinds of tasks that people at that level can perform, with the contexts they can perform them in, and the kinds of language forms they use when performing those tasks, with details about accuracy, fluency, and
appropriateness. The descriptions assume real-life communicative language use and are presented in the following table.

**Table 2.1 Overview of proficiency development according to the ISLPR [from Wylie and Ingram (1979)]**

<table>
<thead>
<tr>
<th>Level</th>
<th>Name of Level</th>
<th>Major Parameters of Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Native like Proficiency</td>
<td>Proficiency equivalent to that of a native speaker of the same sociocultural variety.</td>
</tr>
<tr>
<td>4+</td>
<td>Advanced Vocational Proficiency</td>
<td>Subtlety in language use, with beneath-the-surface meaning dependent on wide knowledge and deep understanding of the target culture.</td>
</tr>
<tr>
<td>4</td>
<td>Vocational Proficiency</td>
<td>Able to perform very effectively in almost all situations pertinent to social and community life and everyday commerce and recreation, and generally in almost all situations pertinent to own ‘vocational’ fields.</td>
</tr>
<tr>
<td>3+</td>
<td>Basic Vocational Proficiency Plus</td>
<td>Use of complex, appropriate language.</td>
</tr>
<tr>
<td>3</td>
<td>Basic Vocational Proficiency</td>
<td>Able to perform effectively in most informal and formal situations pertinent to social and community life and everyday commerce and recreation, and in situations which are not linguistically demanding in own ‘vocational’ fields.</td>
</tr>
<tr>
<td>2+</td>
<td>Social Proficiency</td>
<td>Use of precise, personalized language. Able to satisfy basic social needs, and routine needs pertinent to everyday commerce and to linguistically undemanding ‘vocational’ fields.</td>
</tr>
<tr>
<td>2</td>
<td>Basic Social Proficiency</td>
<td>Able to satisfy everyday transactional needs and limited social needs.</td>
</tr>
<tr>
<td>1+</td>
<td>Transactional Proficiency</td>
<td>Creativity in language use. Able to satisfy basic everyday transactional needs.</td>
</tr>
<tr>
<td>1</td>
<td>Basic Transactional Proficiency</td>
<td>Able to satisfy immediate, predictable needs, using predominantly formulaic language.</td>
</tr>
<tr>
<td>1–</td>
<td>Minimum Creative Proficiency</td>
<td>Use of formulaic language. Able to perform in a very limited capacity within the most immediate, predictable areas of need, using essentially formulaic language.</td>
</tr>
<tr>
<td>0+</td>
<td>Formulaic Proficiency</td>
<td>Unable to communicate in the language.</td>
</tr>
<tr>
<td>0</td>
<td>Zero Proficiency</td>
<td>Unable to communicate in the language.</td>
</tr>
</tbody>
</table>

2.6.2.1 The ISLPR: Applications

The ISLPR scale has important pedagogical and assessment applications. For instance, in the Australian context of second language assessment, the need for the use of the ISLPR is reflected in the increasing number of international students coming to study at Australian tertiary institutions. According to Wylie (1998) the ISLPR has three broad applications which include assessing the proficiency of individual
language learners, research and policy making in language and language education, and providing a framework for language curriculum development. Along with the IELTS as a language proficiency assessment instrument, the ISLPR is also commonly used in different states and territories in Australia to assess the general English language ability of international students before they enter the mainstream Australian tertiary institutions. It is used by the Queensland Nursing Council and accepted by other nursing registration bodies in Australia. The ISLPR scores are accepted by DIMIA (Department of Immigration and Multicultural and Indigenous Affairs) for the purpose of granting visas for nurses for entry to Australia. Griffith University uses the ISLPR scale for entry to most undergraduate programs. It is also used by IBT Education, a national VET-sector institution for entry to its Certificate IV or Diploma programs. In addition, the ISLPR is used to specify proficiency in government legislation and in a wide range of educational contexts. For example, it is used as a framework for curriculum development (Ingram & Wylie, 1991), as a means of determining entry requirements for particular courses and other contexts such as determining eligibility for professional registration or need for interpreters in legal situations (Wylie, 1998).

In assigning proficiency ratings to test-takers in the current research, the ISLPR level description is used. The scores are then converted to the 0-9 IELTS scale to make the comparison easier. For instance, an individual with a band score of 9 on the IELTS is described as having fully operational command of the language, appropriate, accurate, and fluent with complete understanding. The equivalent level in the ISLPR is the score of 5. That is, an individual with a score of 5 on the ISLPR is interpreted as having a proficiency level equivalent to that of a native speaker of the same sociocultural variety. The following table shows the ISLPR Levels and their
corresponding IELTS scores (Wylie, 1999, 2006) and the approximate SAM Interview levels.

**TABLE 2.2 The ISLPR-IELTS conversion table**

<table>
<thead>
<tr>
<th>SAM INTERVIEW</th>
<th>ISLPR</th>
<th>IELTS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Advanced</strong></td>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>4+</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>8 to 9</td>
</tr>
<tr>
<td></td>
<td>3+</td>
<td>7 to 8</td>
</tr>
<tr>
<td><strong>Intermediate</strong></td>
<td>3</td>
<td>6 to 7</td>
</tr>
<tr>
<td></td>
<td>2+</td>
<td>5 to 6</td>
</tr>
<tr>
<td><strong>Pre-intermediate</strong></td>
<td>2</td>
<td>4 to 5</td>
</tr>
<tr>
<td></td>
<td>1+</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>3 to 4</td>
</tr>
<tr>
<td><strong>Not Applicable</strong></td>
<td>1-</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>0+</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

### 2.6.2.2 The ISLPR: Validity and Reliability

As a general English proficiency rating scale, the ASLPR (Australian Second Language Proficiency Ratings) correlates highly with the Comprehensive English Language Test (0.90), the cloze test (0.89), and a dictation test (0.90) (Ingram, 1984). Several other studies have investigated the validity of the ASLPR. In a study involving 76 Indonesian post-graduate students, Phillips, et al. (1985) confirmed that ASLPR Level 3+ was an ideal pre-requisite for students entering postgraduate courses in Australia. The construct validity of the ISLPR has also been investigated (Gariano, 1997; Lee, 1999). Using a multiple regression model, Gariano (1997) analysed records of 32,735 Australian Adult Migration Education Program (AMEP) clients and found that the ISLPR adequately discriminated a person’s English language proficiency. Lee (1999) used a many-faceted Rasch program to analyse the ratings of 329 test-takers. He found that the ASLPR and its subclasses were able to uncover a proficiency development path of learners of English as a second language from diverse backgrounds and age groups. Kellett and Cumming (1995) rated the language
proficiency of 35 non-English speaking background students upon entry to university and found that language proficiency level was strongly related to academic success and could not be compensated for by other positive attributes of the students. They concluded that the students needed at least ASLPR 2+ in all macro-skills to be able succeed at the university; and those with at least Level 3 were much more likely to succeed.

Studies of the inter-rater, intra-rater, and test-retest reliability of the ISLPR have also reported high correlation coefficients. Ingram and Wylie (1991) found an inter-rater reliability index of 0.96 for the speaking section of the ISLPR. Ingram (1984) repeated the study in China with a group of non-native English speakers of ISLPR trainees and found an inter-rater reliability of 0.95. As to the intra-rater reliability, Ingram and Wylie (1991) found a correlation of 0.97 for the speaking component. In addition to the above findings, test-retest reliability studies of the ISLPR report correlation coefficients of as high as 0.90.

The above psychometric qualities achieved over a period of 30 years of research with the ISLPR have made it a rigorous rating scale in Australian context (Ingram & Wylie, 1991; O’Neill & Hatoss, 2003). As a result of this and the lack of such comprehensive tests and approaches that facilitate both curriculum content and a more authentic form of testing, this scale is used in the current research to assess the general proficiency of the participants.
2.6.3 Common European Framework of Reference of Reference (CEFR)

The Common European Framework of Reference for Languages: Learning, Teaching, Assessment is another widely used guidelines to describe achievement of languages among learners across Europe (Little, 2002, 2005; Council of Europe, 2001). It was created in 2001 by the council of Europe which has 47 member states and five observer states including the United States, Canada, Mexico, the Holy Sea, and Japan. The principal purpose of the CEFR is to provide a method of assessing and teaching languages which applies to all European languages. It is also aimed at improving the recognition of language qualifications and help language teachers cooperate, eventually leading to improved communication and cooperation generally in Europe.

2.7 Computers in Language Testing

The application of computer technology is increasing in almost every field of endeavour including assessment. In an era of steadily rising costs, computing costs have been decreasing dramatically and, especially in recent years, we have been witnessing the mushrooming of more powerful and cheaper technology, both software and hardware (Deitel, Deitel & Nieto, 1999). Computers that might have filled large rooms and cost millions of dollars 20 years ago can now be inscribed on the surfaces of silicon chips smaller than a fingernail that cost perhaps only a few dollars each.

Technology has been used in education for decades. Second language (L2) testing in general and measurement of second language proficiency in particular are no exceptions. The fascination with machines in testing and assessment is far from new (Fulcher, 2000). Although the technological innovation was accepted reluctantly at first by only a few educators, it rapidly became the dominant force in American
institutional testing (Spolsky, 1995). As for the history of the use of computers in language testing, Fulcher (2000, p. 93) mentions that “computers have played a key role in language testing since 1935” when “the IBM model 805 became commercially available.”

Since the early 1990s, the research and studies in the application of computers in language testing have shown an increasing interest in the use of the new technology in the field of second language assessment (Alderson, 2000; Brown, 1997; Chalhoub-Deville, 2001, 2002; Chan & Leung, 2003; Chapelle, 2001; Choi, et al. 2003; Dunkel, 1991; Fulcher, 2000). It is not surprising, then, that the second language testing profession adopted "language testing and technology" as the theme for their 2001 annual conference, the Language Testing Research Colloquium, LTRC, (Chalhoub-Deville, 2001). Earlier, the Seventh Annual Language Testing Research Colloquium had focused on computers in language testing in 1985. The conference proceedings were then published in a book entitled: Technology and Language Testing (Stansfield, 1986). The first section of this book deals with the new developments in measurement theory, computerized adaptive testing (CAT) and the applications of latent trait models to test and item analysis. The second section is a compilation of a few articles of a performance-assessment nature: using technology to develop new measures of speaking, reading and writing. For the very reason that performance-based assessment does not readily lend itself to computerized administration, most computer-based assessment procedures include more traditional and selected-response item formats (Chalhoub-Deville, 2001).
Another major work on the use of technology in language testing is an edited book by Dunkel (1991). This work delineates the developmental procedures and validation studies of Computer Adaptive Test (CAT) programs with special focus on CAT applications for assessing second language proficiency in school and university settings, which is yet another indication of the use of technology in language assessment. Also in the same field, Chapelle’s (2001) book offers a comprehensive survey of the use of computers in second language acquisition with implications for teaching, testing and research. As mentioned above, the computer offers a great potential in manipulating the delivery format of the assessment tools. The majority of the available computer-based tests use the computer to simply display items in text format and record responses.

However, given the advances in the digital media, including audio visual provisions, the true potential of the computer’s capability goes well beyond the presentation of test material in text format. The present project uses the capacity of the computer to present audio and video and capture sound as alternatives to presenting the test items in a simple text format. The researcher is aware that the development of audio and video format test items is much more challenging than the conventional computer-based tests. However the outcome of the research adds a great deal to the scarce literature in the use of alternative item formats in computerized testing.

More importantly, the benefits of this research are realized when we consider the audio and video as central to human communication. According to Parshall and Balizet (2001, p. 5) “audio is a critical element in human cognition. The quality of measurement of language abilities could be improved by appropriately using speech
and non-speech sound.” Parshall and Balizet propose an initial framework for the use of sound in computerized tests, but they stop short of implementing video in their tests. However, they believe that listening is fundamental to human communication: 45% of a person’s total communication is spent in listening, while speaking employs 30% of communication time, reading uses 16% and writing uses only 9% (Feyten, 1991; Parshall & Balizet, 2001). Listening is also important for learning in children. It is equally important for adults in college education where lectures are primarily delivered in spoken form. However, perhaps for technical and logistical reasons, “audio may be the only underutilized medium for instructional material today” (Hartley, 1999, p. 149).

In university teaching and assessment contexts, students will be exposed to audio rich environments where they will have to listen to lectures, listen for specific purposes including taking notes, watch videos, attend audio-visual presentations, report in oral or written form and a whole host of similar activities. In the above range of situations, which Bachman (1990) calls the Target language Use (TLU) situations, the audio-visual element is present. Therefore, it seems logical and relevant to employ techniques to evaluate students’ performance in such audio-video rich contexts. To predict the quality of their performance, we can design a delivery instrument which makes an effort to simulate the TLU situations in which the students will be using the second language.

A short note is worth mentioning about the benefits of computer-based listening tests too. Traditionally, second language listening tests have been delivered via conventional audiotapes to groups of examinees. This delivery mode imposes severe
operational constraints, both in terms of implementation and in terms of the quality, storing, reusing, scoring, searching and transporting of the spoken samples of the testees. Above all, it divorces the listening context from all the visual clues considered to be part of natural language use. However, in computer-based tests, we have the option of adding real-life videos to listening tests and making the tasks more meaningful and contextualized. The computer technology not only offers the test-takers additional visual and contextual clues in the test, but it also makes it substantially easier to save the language samples of the test-takers for immediate transmission over the Internet for rating purposes by proficiency professionals who do not even have to be in the same country.

Another obvious benefit of computer generated audio files is the high sound quality of such speech samples which, unlike conventional audiotapes, never deteriorate in quality. However, given all the above advantages, Parshall and Balizet (2001, p. 8) surprisingly admit that “there are relatively few second language listening CBTs.” Among such tests is the ESL computerized adaptive test of listening comprehension developed by Dunkel (1991) in which she provided the test-takers with the opportunity to use visual support such as images and graphics. In a computerized task-based test development project, van den Branden, Depauw, and Gysen (2002) used sounds, pictures and images to assess the language proficiency of non-native adult speakers of Dutch. However, to the best of this researcher’s knowledge, the use of digital video in language proficiency testing has not been reported in the literature.

A review of the published literature on the use of computers and computer-related technology in L2 testing reveals that this topic falls into three major categories:

2.7.1 Computer-Based Tests (CBTs)

Computer-Based Tests (CBTs) are tests that are administered at computer terminals or on personal computers. Many of the receptive item formats, such as multiple-choice, True/False, and Yes/No, and some of the productive item types, like short-answer items and limited-response items can be included in the testing (Phinney, 1991).

Computer-based tests have become “an appealing and viable medium for the administration of standardized second language tests in academic and nonacademic institutions” (Chalhoub-Deville, 2002, p. 471), and therefore offer a number of advantages over their paper-and-pencil counterparts. According to Alderson (2000), for instance, there are technical, administrative, and pedagogical advantages associated with CBTs. First of all, if properly designed and administered, a CBT obviates the requirement for a fixed date and location normally observed in conventional paper-and-pencil tests. The CBTs can be conveniently administered upon the availability of personal computers or computer labs. The convenience for the testee of choosing the place of testing has a potential effect in reducing test anxiety, which is traditionally associated with group administration of large-scale tests. The reduced anxiety as a result of this extraordinary convenience can lead to a positive washback effect in the test-takers too.

One obvious advantage of using a CBT to assess the receptive skills is the possibility of providing immediate feedback on the testees’ performance on any given test task.
Immediate feedback in general is believed to be more meaningful than delayed feedback and has a more direct impact on preventing fossilization of errors and thus has a positive pedagogical effect on the learning process (Patri, 2002). In computer-based tests, “diagnostic feedback can be provided very quickly to each student on those items answered incorrectly” (Brown, 1997, p. 48).

As further confirmation of the usefulness of CBTs, Alderson (2000) also maintains that “CBTs, and especially Internet-delivered (online) tests, can access large databases of items.” He then concludes that this will result in enhanced test security “since tests can be created by randomly accessing items in the database and producing different combinations of items” (2000, p. 596).

Yet another advantage of the use of CBTs is their user friendliness. This aspect is considered to be their most outstanding pedagogical advantage, in the sense that they can encourage and motivate the test-takers on one hand, and facilitate the assessment process on the other hand (Stevenson & Gross, 1991).

In a broad overview of computerized testing, Chalhoub-Deville & Deville (1999, p. 274) list a series of potential benefits for CBTs, particularly the “greater standardization of test administration conditions.” Other benefits include increased test security and the use of innovative item types and performance tasks which are not feasible in paper-and-pencil tests. Also, the provision of special accommodation for test-takers with disabilities can be mentioned as another major benefit of CBTs.
A particular area in which CBTs can be beneficial is self-assessment. Computer-based tests allow researchers and testers to avoid the labour-intensive process of self-assessment. A prominent study investigating the use of computer-administered self-assessment in determining test-takers’ approximate proficiency level is that of Malabonga, Kenyon, and Carpenter (2005). In their study, the authors found that 92% of the examinees who took a computer-administered self-assessment were able to select test tasks at appropriate difficulty levels. Self-assessment has been found to be a reliable indicator of examinees’ actual ability (Adair-Hauck & Pierce, 1998; Williams, 1992). Kenyon (1996) investigated the relationship between a computer-administered self-assessment and performance on an oral interview and found a moderate correlation of .78. As computer-delivered self-assessment is a new instrument, more research is needed into the validity and usefulness of this approach.

On the negative side, however, there are limitations and disadvantages associated with innovations in IT-based testing. Most importantly, “direct measurement of examinees’ speaking proficiency is impossible by means of a computer” (van den Branden, et al. 2002, p. 444). As yet, CBTs do not easily lend themselves to the assessment of the highly valued productive skills of speaking and writing, as they do to that of the receptive skills through the traditional selected-response item types such as multiple-choice, true-false, or matching items (Alderson, 1986, 1996, 2000). After all, computers are not able to act as a fully-fledged conversation partner. Furthermore, since computers easily provide the possibility of including multiple-choice items in the testing process, there seems to be a widespread re-emergence of this item format despite its being seriously challenged on theoretical grounds (Oller, 1979).
One of the most logistically difficult tasks researchers are faced with is the automated scoring of constructed response item types, particularly the productive skills of speaking and writing. In fact, the more productive skills are included in a test, the more difficult it is for the machine to score it. Given the present capabilities of computer technology, Brown (1997, p. 45) believes that “the more interesting types of language tasks (e.g., role plays, interviews, compositions, oral presentations) prove much more difficult to develop for computer-assisted testing.” This shortcoming is also highlighted by Chalhoub-Deville (2002) in her description of the new Test of English as a Foreign Language (TOEFL) which, in 1988, began to be administered in computerized format in the United States and in other countries around the world. The CBT version of the TOEFL test “includes traditional multiple-choice items, as well as other selected response items, including selection of a visual, matching, and ordering objects and texts” (Chalhoub-Deville, 2002, p. 473). The Educational Testing Service has only recently added the Test of Spoken English (TSE) to its assessments. The TSE is a measure of the oral language proficiency of non-native speakers of English. It includes a set of recorded questions on an audio instrument to which the test-takers listen and respond.

In addition to the above shortcomings, the initial development of CBTs is relatively costly as it involves programming expertise. The cost of developing a test is related to the practicality component of test usefulness highlighted by Bachman and Palmer (1996). The implementation of a technology-based test requires hardware and software and familiarity with their use. However, if test designers are trained in computer programming, once a package of a given CBT is developed, the cost of maintaining, upgrading, and modifying the package, including the test delivery
format, will be reduced. Amiri (2000, p. 83) argues that the language teaching profession “increasingly needs qualified language teachers who at the same time are highly skilled in the relevant aspects of IT including design, development, and programming, of computer-based materials.”

A potentially promising field in the use of computers in testing second language speaking is the application of continuous speech recognition technology. According to Coniam, “the new speech recognition technology claims to be able to process continuous speech and no longer requires pauses between words. Continuous speech recognition certainly holds greater promise for man-machine interaction than discrete word speech recognition” (Coniam, 1998, p. 8). However, in a study investigating the feasibility of the speech recognition technology in testing speaking ability, he found that the current implementation of such technology “is still far from being readily useable primarily because the technology is speaker dependent and has to be trained to recognize each person’s voice” (Coniam, 1998, p. 7).

Another important issue in the use of computers in assessment is the psychology of reading a text from a screen, which is believed to be different from reading a text on paper. Flipping pages back and forth involves a different feeling from scrolling a screen up and down. Ten Have (1999) argues that the basic requirements for reading from a screen are similar to those for reading a printed page, or a hand-written one for that matter: the eyes move from left to right, then search for the start of the next line, and when the eyes meet the bottom frame of the screen, the finger touches a key, such as ‘PgDn’, instead of turning the page. However, the differences can appear when it comes to anything beyond users reading page by page. To locate a certain page in a printed document, the reader has various means. For instance, there will be page
numbers and possibly headers or footers repeating a chapter title. Furthermore, one has the physical object in one’s hand: especially with books, feeling the thickness of the pages already read and those still ahead gives an idea of how one is progressing. However, in the case of reading a text from a screen, getting a similar overview can be much harder or at least of a different nature. Ten Have (1999, p. 275) further argues that “this is partly dependent on the reading ‘environment’, i.e., the program used to project the text. Word processing programs will have various devices to provide some orientation and to ‘scroll’ the text, or, to put it in terms of a reader’s perspective, to ‘travel’ (or navigate) through the text.” Although the current page number and the line the cursor is on can be projected at the bottom of the screen, not many readers will be as aware of this as they are of the page number of a book.

A significant proportion of computer experience in test takers results from factors such as the degree of differential familiarity with the computer, individual personality differences, background, and attitude toward new forms of language tests among testees (McDonald, 2002). These factors are some of the concerns related to the construct validity of CBTs. Investigation of these issues is important because a test score obtained from a computerized test is supposed to reflect only the construct of interest only (Sawaki, 2001). That is, if the test score represents both language ability (or the trait under investigation) and computer familiarity, for example, then valid generalization of test scores across testing modes is no longer possible. Many studies have been conducted to determine the influence of computer familiarity on test-takers’ performance on computerized tests, some of which are highlighted below.
There seem to be big differences in the computer familiarity levels of different language learners in different parts of the world. Up to the early 2000s, studies show a wide variation in familiarity with computers from one country to another and from one point in time to another. For example, Pelgrun, et al. (1993, cited in McDonald, 2002, p. 303) “surveyed 10 countries and found that between 3% and 89% of elementary and secondary school children did not use computers.” In a similar study, Weil and Rosen (1995) assessed computer experience in university students from 23 countries between 1992 and 1994, and found this to vary from 25% to 98% between countries. To determine the effect of IT-familiarity of test-takers on their performance in the TOEFL test, for example, the Educational Testing Service conducted research which showed that a significant 16% of the TOEFL test-takers had negligible computer familiarity (Kirsch, Jamieson, Taylor & Egnor, 1998; Taylor, Jamieson, Egnor & Kirsch, 1998).

However, given the widespread availability of computers at the workplace and home, the effect of computer familiarity may be diminished by the passage of time; “ten years from now lack of familiarity with computers, for example, may be a less significant issue than it is currently” (McDonald, 2002, p. 303). This prediction is supported by the fact that computers are becoming more ‘intelligent’ and interactive, hence easier to use.

In any case, the effect of familiarity with computers may actually be only a minor issue in the use and administration of computerized tests. Sawaki (2001) cites two studies focusing on examinees’ reactions to new forms of testing. One of these, Madsen’s (1991) study, is one of the few studies that provides details based on a self-
report questionnaire. Madsen administered an attitude questionnaire on the CAT (computer adaptive test) version of an ESL placement test of reading, structure and listening at Brigham Young University. He found that, although students' reactions to the new test were generally positive, differences in attitudes were observed across language groups. Spanish speakers in his study reported that it was easier to read on the computer screen than in print and that they were interested in and willing to take the CAT in the future. On the other hand, Japanese students' reactions were rather negative. They claimed that it was more difficult to read on the screen and reported anxiety about taking the CAT, even though the Japanese subjects were "more experienced" users of computers than the Spanish-speaking students. Overall, Madsen concluded that experience with computers does not reduce test anxiety, and effects of examinee language background on affect must be investigated more closely. The current research will shed light onto the issues of anxiety and attitude in taking a computer-based test of oral proficiency.

2.7.2 Computer Adaptive Tests (CATs)

A particular advantage of Computer Adaptive Tests (CAT) is, of course, their adaptive nature. In these tests, the selection and sequence of items depend on the emerging pattern of success or failure of the test-taker on the preceding items. That is, if a test-taker gets a particular item right, the next item presented will be of slightly higher difficulty level, and vice versa. Computer-adaptive testing involves technologically advanced assessment processes based on Item Response Theory (IRT) and probabilistic theory (Henning 1991).
The first step in developing CATs is to ensure that item difficulties are set for the population for which the test is intended, and this requires that items be pre-tested on large numbers of students (Fulcher, 2001). The second step is to place all items on a ‘difficulty scale’, so that when learners take the test their ability level can be calculated. This means that a large number of items are needed in the ‘item bank’. The entire difficulty/ability scale must be represented, and there must be enough items at each point on the scale to ensure that the item bank is not revealed too quickly to the test-takers.

Most of the advantages of computer-adaptive tests result from the capability of computers and the adaptive nature of such tests. In other words, the increasingly huge memory, accuracy and high speed of computers allow the storage and subsequent retrieval of a large amount of data for analysis. The adaptive nature of the test makes it possible to immediately focus on the testees’ ability level. Unlike paper-and-pencil tests, in which we may need a large number of items to elicit a representative sample of test-takers’ language behaviour, in CATs only a few items are required to estimate test-takers’ abilities with the same degree of precision.

As the items in a CAT are algorithmically selected from a large pool of items, test security is significantly enhanced. In other words, two individuals sitting next to each other are administered two different sets of items, so the possibility of cheating is removed (Chalhoub-Deville, 2001; Chalhoub-Deville & Deville, 1999).

In spite of all the above merits, a review of the related literature reveals that there are also limitations associated with the development and use of CATs. Fulcher (2000, p.
98) believes that “CATs can only work if there is a large number of items in an item bank, which are calibrated to a measurement scale using Item Response Theory. Building up a sufficiently large item bank can be time consuming and costly.” Another problem is the difficulty of “establishing criterion-referenced meaning at various points on a scale, especially for cut scores” (Fulcher, 2000, p. 99). The third problem is related to content/domain sampling in such tests. CATs make use of only a few items to arrive at the ability level of the test-taker, so the representativeness of the items is a critical issue. Unlike paper-and-pencils tests, a test-taker cannot skip any of the items on a CAT. This is because, in skipping items the test-takers would probably try to answer only those they think they are going to get right; therefore, the ability estimate would be unnaturally high.

2.7.3 Web-Based Tests (WBTs)

Web-Based Tests (WBTs) are assessment instruments written for web delivery, in the language of the web: HTML (Hypertext Markup Language) or (XML) eXtensible Markup Language and many more. The test consists of one or several HTML or XML file(s) located on the tester’s computer, the server, which are downloaded to the test-taker’s computer, the client. Downloading can occur for the entire test at once, or item by item (Roever, 2001). Web delivery of large-scale standardized tests has become appealing to the extent that the computer version of the TOEFL was launched in 1998 by the Educational Testing Service in the US and many countries around the world (Educational Testing Service, 1998).

In a similar vein, the Council of Europe has sponsored the DIALANG project, which provides prognostic assessment in 14 European languages (http://www.dialang.org).
According to the Language Acquisition Resource Centre (LARC, 2006), DIALANG is an ambitious project for the development of a computerized test of listening, reading, writing, grammar, and vocabulary skills, in fourteen European languages, eleven of which are European Union languages. The scale used is the Council of Europe scale, developed by the Association of Language Testers in Europe (ALTE). The scale has steps for high and low levels of skill for the ‘Proficient User’ (advanced), ‘Independent User’ (intermediate), and ‘Basic User’ (elementary); these are, respectively, C2, C1, B2, B1, A2, A1. The scale’s most general form is called the ‘Common Reference Levels: global scale,’ and there are specific forms for reading, writing, speaking, and listening. The test-taker may choose the language of instructions and feedback, then the language and skill in which he wishes to be tested. After a ‘placement test’ which tests vocabulary to approximately determine level of skill, the test-taker is then tested in the particular area he has selected. If the test-taker has completed the introductory placement test, the test will be adaptive. On completion, the test-taker is given a score and an appraisal of what the score means. He may also receive a more detailed ‘scorekeeping’ of each of his answers, showing which were correct and which incorrect. Administered and scored by computer, the test has no speaking component. It is intended that the finished DIALANG tests will be available on the Web at no charge, at present writing, it seems that they are available as downloaded files.\footnote{For a review of DIALNG, see Chapelle (2006), and for a comprehensive description of the test, see Alderson and Huhta (2005), or visit DIALANG homepage: http://www.dialang.org/intro.htm.}

Another well-known test which attempts to use a sophisticated database to machine-score fluency is the \textit{Phone Pass Set-10 test of English}, developed by Ordinate.
Corporation in 1999 (Ordinate Corporation, 1998; LARC, 2006). This test, administered by telephone, is a test of basic conversational skills, and bases scores on particular lexical items, their pronunciation, and the pace and fluidity of the examinees’ speech as measured by the computer. While some responses to the test are computer-scored, others are stored for human scoring. Computer scoring is done by special-purpose speech processing and recognition software, which uses pronunciation models, pronunciation dictionaries, and expected-response catalogues. Computer scoring of fluency does not appear to consider parameters of particular languages.

Given the abundant research and findings on the topic, we can say that WBTs are having some success in making “assessment more efficient and serviceable” (Chalhoub-Deville, 2001, p. 96). WBT assessments also provide the opportunity for a more customized and learner-specific administration of tests, tracking of students’ progress through a course of study, immediate test feedback, and, potentially, a vast number and variety of item types. An individual who has access to the Internet can take such a test at his own convenience. Tests on the Internet may include a large database of item types and formats which can be constantly revised, updated, and modified. Using the Internet allows test developers to have access to a large group of item writers who would like to make a contribution to the development of a particular test battery. On the negative side, however, maintaining an accumulating question bank with valid items requires a lot of time and expertise. Roever (2000) mentions four potential problems with WBTs, namely: identification, item confidentiality, secure data storage and secure data transfer. Given the security concerns, he then decides that WBTs should be restricted to low-stakes tests or tests where it is in the test-takers best interest not to cheat, such as self-assessment instruments. However, if
testing is conducted at testing centres under supervision, security problems can be
minimized, although that would defeat the ‘anyplace, anytime’ principle of WBT.
Lastly, on the technological side, for clients who have a slow download rate on the
Internet, tests which include large graphic, audio or video files will not be efficiently
administered.

2.8 Computers and Oral Proficiency Testing

Computerization of second language oral ability tests has been of interest among
language testers for the past decade (Malabonga & Kenyon, 1998), but few empirical
studies have evaluated the equivalence of the construct being measured in
computerized and conventional face-to-face speaking tests. Sawaki (2001), who
investigated the comparability of conventional and computerized tests of reading in
second language points out that, despite the rapid growth of demand in the area of
computerized language testing, development and implementation of this new mode of
testing is currently in its initial stages. Therefore, sufficient empirical data, which
would allow researchers to look into the soundness or usefulness (Bachman, 1990) of
computerized language tests with regard to construct validity and fairness, are yet to
become available. One reason for the slow pace of development in this area may be
the need for sophisticated authoring software and the ability to integrate sound and
video into the tests.

In this section, the most significant of the available computerized speaking tests will
be analysed, with special attention to their specifications and usefulness. These have
been selected on the basis of their relevance to this research and quality in terms of
modelling good practice. Oral proficiency of ESL learners can be measured through
the use of a variety of testing instruments and the Oral Proficiency Interview (OPI) is selected as the first example. The OPI is used by various US government agencies involved with language training, including the Foreign Service Institute, where it was originally developed in the 1950s to assess the readiness of US personnel for overseas diplomatic missions.

In the mid 1990s, the OPI was developed into the Simulated Oral Proficiency Interview (SOPI) by the Centre for Applied Linguistics (www.cal.org). The SOPI is a type of tape-mediated test of speaking proficiency based on the OPI. All SOPI items are based on the speaking proficiency guidelines of the American Council on the Teaching of Foreign Languages (ACTFL). The items included in the SOPI are picture-based, topic-based and situation based. Once the test-takers’ language samples have been recorded on tape, the recorded performance can then be sent to trained OPI raters for evaluation. Developers of the SOPI report correlations as high as 0.92 between the results of the OPI and those of the SOPI (Stansfield & Kenyon, 1991). The format of the SOPI obviously brought about promising advantages in standardised administration of large scale tests. Further adaptations of the SOPI were developed, including the Video Oral Communication Instrument (VOCI) by the Language Resources Centre at San Diego State University, in which a video rather than a tape was used along with the test booklet to elicit examinee speech performances. Malabonga and Kenyon (1998) believe that one of the pitfalls of the SOPI is the lack of control by the testee over the testing procedure or the choice of appropriate tasks. It is the interviewer who actually adapts the level of difficulty of the questions to the proficiency level of the examinee. Besides, all the questions on the tape must be answered.
This lack of control on the part of the test-taker prompted the test developers at CAL to come up with an enhanced version of the SOPI called the Computerized Oral Proficiency Instrument (COPI). The items of the COPI are also based on the ACTFL Speaking Proficiency Guidelines (Kenyon & Malabonga, 2001). The developers of the COPI claim that the purpose of COPI is to use the advantages of multimedia computer technology to improve the SOPI by giving examinees more control over various aspects of the testing situation, including the time they need to prepare for and respond to a task, and by increasing raters’ efficiency in scoring the test. The COPI, like its predecessor SOPI, uses simulated real-life tasks to elicit speech rateable by the ACTFL Guidelines’ criteria. The task pool of the COPI includes 100 items codified according to the four levels (Novice, Intermediate, Advanced, or Superior) of the ACTFL Guidelines. The tasks include a picture with a single-sentence written or spoken description of the picture. Although the test is claimed to be a measure of the speaking proficiency, there is a potential effect of other abilities on the performance of the testees too. For instance, the testee looks at the pictures, reads the descriptions, listens to prompts before answering the question in spoken language. Depending on the choice that the examinee makes, the test takes anywhere from 30 to 50 minutes to complete. In taking the COPI, the examinees go through the following 8 steps:

1. Welcome and information on the purpose and structure of the COPI.
2. Input and correction of personal information.
3. Self-assessment of proficiency level (by answering an 18-item questionnaire about the ability to, for example, give directions, ask questions, hypothesize, and so on).
4. Listening to an adequate response to a sample task.

5. Practice with the same sample task.

6. Responding to the actual performance tasks. At this stage, the examinee responds to a minimum of four tasks at the level of the first task selected and three at the next higher level.

7. Feedback on the levels of the tasks that the examinee took.

8. Closing.

One of the potential concerns regarding the design of the COPI is the way it presents the tasks to the examinees. As described above, the tasks are in graphic form accompanied by a written description and followed by another written or spoken prompt which is the actual question. The fact that the test invites the examinee to ‘read’ a description of the pictorial task brings in a different ability (i.e., reading) from the one the test purports to measure (i.e., speaking) is a potential threat to the validity of the test. The validity of a test is the extent to which it measures what it is supposed to measure, or the extent to which the inferences and decisions made on the basis of test scores are meaningful and useful. As an example, a tester may overload a vocabulary test with items full of proper nouns and historical facts, rendering it less valid as a measure of general vocabulary knowledge. Or a tester may select appropriate test items for his third-graders but ‘write instructions’ that only the better readers are able to understand, rendering the results of this test invalid for their intended use.

In the case of the COPI, there is no way to make sure that the quality and level of performance demonstrated by the test-takers is or is not affected by their ability to
‘read’ the descriptions. In the COPI, the ability under investigation is therefore likely to be affected by other abilities which are not supposed to influence the test taskers’ performance. Therefore, there is the potential problem that interpretations, inferences or decisions made on the basis of test scores may be contaminated by the undue presence of a construct irrelevant factor — reliance on ‘reading’ in a ‘speaking’ test. Further criticisms can be of the COPI in relation to timing. The COPI provides the test-takers with a certain amount of time to think about their responses and time to give their responses. The amount of time allotted for the responses is indicated by the use of a series of balls appearing in a horizontal line, each ball representing 15 seconds. Using balls appearing at an interval of 15 seconds could be distracting and even stressful for the test-takers. It might be more appropriate for the balls to be replaced by a more visually appealing graphic such as progress bar or percentage indicator. Also, especially for large scale testing, the period of 30-50 minutes required for the completion of a test seems too long and may even be boring for some of the test-takers. This might warrant change for practicality reasons. In large scale standardized tests such as the IELTS, a rateable and representative sample of a test-taker’s speech is elicited within only 12-15 minutes.

In 2003, CAL introduced a new computerized test called BEST Plus. The Best Plus is presented as an advanced adaptation of the Basic English Skills Test, whose purpose is to assess the ability in interpersonal communication using everyday language. The BEST is used to evaluate elementary listening, speaking, reading, and writing of limited-English adults for the purpose of class placement, progress checking in survival and pre-vocational classes, and diagnosis. It is made up of two parts:
1. Core section: a one-on-one interview on social survival topics, plus a sight-word reading task and a simple bio-data writing task.

2. Literacy skills section: real-life tasks such as reading ‘wanted’ ads and completing job applications (Madsen, 1983).

The BEST Plus comes in two versions: a computer-adaptive version and a semi-adaptive print-based version. The items on BEST Plus reflect “language used in everyday American life – at home, at work, and in the community” (http://www.cal.org/bestplus/intro.html). In the computer version of the BEST Plus, the test administrator asks the examinee a question, listens to the examinee’s response, uses a rubric to score the response, and enters the score into the computer. The computer then selects the next test item, choosing questions most appropriate for the examinee’s demonstrated ability level. Its items are claimed to reflect language used in everyday American life and cover communicative language functions ranging from providing personal information to giving and supporting opinions. The test takes 5-20 minutes to administer, which is a significant reduction in time compared to its predecessor, COPI. The performance of the testees on the BEST Plus is rated on a 0-10 scale.

2.9 Comparability Studies

While computer-based tests and computer adaptive tests are attracting the attention of researchers, language learners and test users for developing, administering, and scoring tests (Sawaki, 2001), this area is still in its infancy for two principal reasons. One of these, as Choi, et al. (2003, p. 296) state, is that “the application of Item Response Theory (IRT) to language testing had been delayed by persistent
controversies over the dimensionality of language ability.” The other, perhaps more important, reason has to do with the potential effects of multimedia on interactivity and thus test-taker performance on language tests. This is closely related to the impact of the new technology on test taker performance. Impact and or backwash are components of test usefulness (Bachman & Palmer, 1996). The new method facets in computer-based tests, including the delivery mode and familiarity with computers, raise questions pertaining to the construct validity of such tests. For example, questions may include: To what extent are the results of a paper-based test comparable to those of the computerized version of the same test? Or, if a paper-based test has satisfactory reliability and validity indices, will the computerized version enjoy the same psychometric properties? Which test version results in a more efficient, cost-effective, standardized, and accurate measurement of the intended language abilities?

Given the potential effect of multimedia on the delivery of language tests, on one hand, and the positive or negative washback effects of computer delivered tests on the performance and attitude of the test-takers, on the other hand, we realize that the use of computers in assessment will become a fact of life. Before dealing with a few research projects particularly focusing on score comparability studies, it is important to define a test score. Simply put, scores are numbers or any other quantitative values used to represent:

\[ a) \text{ an individual response to language stimuli, where the term response is meant to encompass any measurable act of an individual; } \]
In educational measurement and language testing, it is the second meaning which is usually under investigation. A thorough description of test scores and what they mean precisely in an educational context was given by Lyman (1986). The description of the numerous types of scores (and their conversions) is beyond the scope of the current study. However, for the purposes of this research the second meaning of the test score above will be used.

Comparability studies of computer-based tests and paper-and-pencil tests are sometimes referred to as score comparability studies. This is mainly because test scores are the concrete observable data most immediately and readily available to researchers to analyse and study. Performance on language tests is affected by a wide range of factors such as cognitive and contextual ones and these influences on performance are represented in test scores.

Chalhoub-Deville and Deville (1999) point out the scarcity of score comparability research in second language tests and the importance of conducting comparability studies in local settings to detect any potential test-delivery-medium effect when a conventional test is converted to a computerized test. The effect of the format of a test on test performance is thought to be a fruitful avenue of research in test validation (Lazaraton, 2002). Shohamy (1994), for example, compared the candidate output on two tests of spoken Hebrew, one tape-mediated Semi-Oral Proficiency Interview (SOPI) and the other a face-to-face Oral Proficiency Interview (OPI). Her results
indicated that the SOPI format samples more widely for low-level candidates while the OPI seems to be better suited to high-level candidates. She also found that candidates self-corrected and paraphrased significantly more on SOPI and switched to their first language significantly more on the OPI. Finally she concluded that the long process of test validation can be enhanced by analysing test data from multiple perspectives.

In a research study Choi, et al. (2003) utilized an EFL test battery (Test of English Proficiency developed by Seoul National University) with the aim of addressing the issue of the comparability between paper-based and computer-based language tests based on content and construct validation. Their tests included listening comprehension, grammar, vocabulary, and reading comprehension, but not speaking. They report that the findings support the comparability between the two versions of the test.

Along the same line of research, as a result of comparing transcripts from 10 Spanish SOPI and OPI performances on accuracy, various management strategies, and several structural components, Koike (1998) found that the SOPI generated significantly more fillers, and fewer turns, quotes, speech acts, and switches to the first language than the OPI did. However, she cautions that task type and specific topic influenced language production more than test format did.

Factors that are of prime importance in most comparability studies include prior exposure to computers, the ‘differential familiarity’ component, computer anxiety, gender, computer attitudes, task types of selected response items (predominantly
multiple-choice items as they most easily lend themselves to computer testing), the effect of graphics or animations, the length of reading passages on the screen, speededness and time constraints, and computer interface design. However, there are very few studies of speaking tests and no study includes video in the design or development of computer-based tests.

Choi, et al. (2003, p. 297) define comparability studies as “an investigation into the comparability of test methods or test tasks represented in different testing modes.” In relation to mode effect, Sawaki (2001, p. 40) maintains that “in order to support construct validity of computerized tests such that the construct being measured is not being affected by the mode of presentation, the equivalence of conventional and computerized test forms must be established from various directions.” In other words, efforts must be made to make the testing conditions as similar as possible to real-life conditions in order to diminish the effect of the new medium of presentation on testees’ performance. For example, the content, format, length, and difficulty level of the two tests should be comparable.

Another closely related and major consideration in converting conventional paper-and-pencil tests to computerized form is the issue of ‘parallelism.’ Two tests are said to be parallel if they contain the same kinds of items of equal difficulty, are constructed to the same specifications, are identical in the nature of their sampling, are of the same test rubrics, are highly correlated in the scores they produce (Bachman, 1990). Bachman (1990, p. 168) defines two parallel tests as tests “of the same ability that have the same means and variances and are equally correlated with other tests of that ability.”
Research in this field indicates that, upon the presentation of a test in a different mode, there is the potential for ‘task change’ (Green, 1988), a condition that the nature of the task being altered, which in turn may induce unexpected changes in item difficulty. Green (1988, p. 78) states that “if computer presentation changes tasks, so that the correlation between scores on the computer and conventional versions is low, then validity is threatened.”

In a meta analysis, Mead and Drasgow (1993) analysed the potential presentation-mode effects on cognitive ability in speeded test performance. They found that, after correcting for measurement error, the estimated cross-mode correlations were .97 and .72 for timed power tests and speeded tests respectively. Based on these results, the authors concluded that the mode of presentation may affect speeded tests but not timed power tests. In a similar study of mode effects on a speeded clerical test, Neuman and Baydoun (1998) found consistent high cross-mode correlations between paper-and-pencil and computer modes for the instrument’s subtests, and a structure equation modelling suggested that the constructs being measured in the paper-and-pencil and CBT versions of the tests were equivalent. In a study investigating the use of a partly auditory and partly visual mode of presentation, Mousavi, et al. (1995, p. 319) found that “effective working memory may be increased by presenting material in a mixed rather than a unitary mode.”

Another source of task change may be differences in what Bachman (1990, p. 118) calls “testing environment,” or test administration conditions, across modes of presentation. The testing environment includes facets like: 1) familiarity of the place
and equipment used in administering the test; 2) the personnel involved in the test; 3) the time of testing; and 4) physical conditions.

In a comparative study of paper-and-pencil and CBT versions of three end-of-unit tests for the Ground Radio Repair Course at a Marine Corps Communication-Electronics School, Spray, Ackerman, Reckase, & Carlson (1989) argue that presentation mode effects on test performance found in previous research may be partly due to differences in the flexibility of test administration conditions, control of the testees over the testing situation being one element. In this study, the authors allowed the test-takers to skip items and to review and change answers after completing the test, an option not normally provided in many computer administered tests. As a result, mean scores and cumulative score distributions for the raw scores across modes on this test were not significantly different between the paper-and-pencil and computerized testing groups. The authors concluded that paper-and-pencil and computerized test results would be equivalent when the same test-taking condition flexibility is maintained across modes.

Finally, to ensure the achievement of psychometric equivalence or parallelism between paper-and-pencil tests and computerized tests, the American Psychological Association published Guidelines for Computer-based Tests and Interpretations (American Psychological Association, 1986). The Guidelines maintain that when interpreting scores from computerized versions of conventional tests, the equivalence of scores from computerized versions should be established and documented before using norms or cutting scores obtained from conventional tests. According to the Guidelines, if the scores of individuals tested in alternative modes closely
approximate each other, this will be considered a prerequisite for achieving psychometric equivalence of paper-and-pencil and computerized tests. As a second requirement, the Guidelines also maintain that scores from conventional and computer administrations may be considered equivalent if the means, dispersions, and shapes of the score distributions are approximately the same or have been made approximately the same by rescaling the scores from the computer mode.

Considering the above comparability studies and their findings, there seems to be a stronger tendency towards computerizing selected response items than productive tasks like speaking. The current project investigates the comparability of the face-to-face interview with that of a computer-based oral test.

2.10 Task Difficulty
What constitutes task difficulty is a critical and appealing issue in language testing. According to Luoma (2004), difficulty is not a direct characteristic of a given task but rather a sum of task characteristics as well as the conditions under which someone performs the task. Task difficulty of speaking tests has only recently become a topic of investigation in language testing (Elder, Iwashita & McNamara, 2002; Fulcher & Marquez Reiter, 2003; Lumley & O’Sullivan, 2005; Norris, Brown, Hudson & Bonk, 2002). In second language acquisition and language testing, several studies have investigated factors influencing the difficulty of tasks (Elder, et al., 2002; Robinson, 2001; Skehan, 1998). The findings from different studies sometimes contradict each other. This could possibly be the result of the difficulty in predicting task difficulty as well as the nature of speaking tests and scores. Luoma (2004, p. 46) has summarized the factors in two groups: “[the]… first between the examination participant(s), the
interlocutor(s) and the task, and the second between the performance, the rater and the scoring criteria.” All of these factors and their interactions have an influence on the scores. The studies on speaking task difficulty have yielded some useful advice for the planning of speaking tasks, and possibly some general guidelines for manipulating task difficulty, as long as comparisons are made between tasks of the same type. Among the factors influencing task difficulty is the number of objects or individuals to be discussed during the test performance. Brown and Yule (1983), cited in Luoma (2004, p. 46), found that “it was easier to tell (and understand) a story about an accident that involved two or three cars than one involving four or five.” The distinguishability of the objects and individuals (e.g., number of people or pronouns) as well as the number and complexity of events was also found to be a significant factor in contributing to the difficulty of a given task. Robinson (2001, p. 30) identifies two different sets of factors contributing to task complexity: “resource-directing factors” which include the number of task elements, reasoning demands of the task, immediacy of information provided; and “resource depleting” factors which include planning time, number of tasks, and prior knowledge. For task difficulty, Robinson (2001, p. 30) presents “affective” and “ability” variables. Affective variables include motivation, anxiety and confidence, while ability variables include aptitude, proficiency and intelligence. He also believes that by modifying these factors the cognitive demand (e.g., amount of attention, memory, reasoning, and other information processing) required for task performance will vary, leading to variation in the quality of language produced.

On the other hand, Skehan (1998) proposes a model in which he defines task difficulty in terms of three different factors, namely code complexity (linguistic
complexity and vocabulary load) *cognitive complexity* (cognitive processing factors such as information type and organizational structure, as well as the familiarity of task topic discourse and genre) and *communicative stress* (time pressure, nature of the prompt and number of participants). Both Robinson’s and Skehan’s models provide substantial opportunity in language testing for the manipulation and presentation of tasks to test-takers in order to understand the true nature of the interaction between the testee and the task in question. The most recent model for defining task difficulty in language testing was proposed by Fulcher and Marquez Reiter (2003) in which they incorporate pragmatic ability as an indicator of success in performing on culturally loaded tasks. In the current study, the resource depleting factors (planning time, number of tasks, and prior knowledge) proposed by Robinson are considered as elements contributing to the difficulty of the tasks.

Given the circumstances of a computer-administered test, an inherently influential factor contributing to the difficulty and complexity of tasks is the user interface. Fulcher (2003) was probably the first researcher in language testing who highlighted the significance and potential impact of the user interface design on test-takers’ performance by establishing a process model. Given the scarcity of adequate models of user interface design, he encourages computer-based test producers to “publish the details of the process of” such designs (Fulcher, 2003, p. 405). He believes that such publications would form part of an important mix of validity evidence available to support test use and interpretation. The current study aims to contribute to the resources available on this question.

An issue closely related to the user interface design is the inclusion of videos in tests of speaking and listening. Coniam (2001) examined the use of video as a mode of
assessments for a group of teachers taking an English language certification test. He justified the use of video in language testing by emphasizing the point that “video lends a degree of authenticity in terms of context, discourse, paralinguistic features and culture which are somewhat lacking in purely audio medium” (Coniam, 2001, p. 10). He also highlights the lack of sufficient studies incorporating the use of videos in language tests. Most of the studies investigating the effect of incorporating video in tests have focused on listening comprehension tests (Baltova, 1994; Brett, 1997; Gruba, 1993, 1997; Secules, Herron & Tomassello, 1992), while the current study approaches the effect of the video on tests of speaking ability. However, the use of video in tests, in general, is supported by the fact that many ESL learners have more contact with visual media than purely audio media (Progosh, 1996). In fact, user interfaces which include more life-like qualities such as speech or video enable researchers to better understand and assess the users’ interaction with the machine (Parise, Keisler, Sproull & Waters, 1999).

2.11 Attitude Analysis

With the expansion of the use of computers in almost every aspect of our daily lives, people have not only been affected by the new media but have also changed their attitudes towards computers. For some people, computers “still suffer from the image problem” (Amiri, 2000, p. 80) which is mainly a legacy of their origin as number-crunching tools for scientists and mathematicians. Attitudes towards computers are closely associated with issues such as computer familiarity, computer anxiety, previous experience and popular fear of mathematics and numbers. There have been several studies investigating the attitude of individuals towards technologically enhanced education (Chan & Leung, 2003; Huang, 2001; Yuen & Ma, 2001) and
students’ motivation in learning (Rizo & Vinagre, 2000) as well as sex differences in attitudes towards computers (Horton, 1987). The most commonly used computer attitude scale is the one developed by Gressard and Loyd (1986). This is a 30-item instrument designed to assess three areas by means of three 10-item sub-scales: computer anxiety, computer liking, and computer confidence. In their study Gressard and Loyd (1986) found that the reliability coefficients of the three sub-scales and the findings of the factor analysis supported their hypothesis concerning the effect of computer attitudes on performance in computer-related teaching. They also found that computer experience — that is, familiarity with computers — was significantly related to more positive attitudes on all three scales.

At the same time, research suggests that negative attitudes and unfavourable perceptions of computers adversely affect computer literacy (Volman, 1997). Increased computer competency is related to more positive perceptions and attitudes towards computer use. Generally, these studies show that a positive computer attitude is a necessary prerequisite and an integral part of computer literacy (Kay, 1993). For example, in a comparative study of computerized and paper-based tests, Rico and Vinagre (2000, p. 464) found that computers enhance motivation and effective learning and concluded that “computer-oriented tasks motivate students to make more extended use of the references provided by allowing easier and quicker access to the contents, which leads to self-learning and individualization of the learning process.” Having developed a Chinese version of a computer attitude scale, Yuen and Ma (2001) concluded that affective and general usefulness aspects of computers were significant factors in influencing the usage among students. In a computer attitude scale validation study, Troutman (1991) administered a computer attitude scale to a
group of 292 pre-service teachers and found that those who feel secure in their own personal use of the computers also feel positive toward the use of computers in schools and educational contexts. In a similar study investigating examinee attitudes to a computer-based Test of English as a Foreign Language, Stricker, et al. (2004) found that attitudes to the computer-based TOEFL were positively correlated with test performance.

In a comprehensive study of the relationship between computer attitude and achievement among elementary students, Huang (2001) studied 153 sixth-grade students and found that students who had positive general computer attitude demonstrated more favourable computer attitudes towards web-based instruction than those who had low general computer attitude. He also found that female students demonstrated more favourable computer attitudes to web-based instruction than the male students.

In the field of second language assessment, however, only a few studies have involved the development and use of computer attitude scales for language testees. Kenyon and Malabonga (2001) compared examinee attitudinal reactions towards computer-based and other oral proficiency assessments across three languages: Spanish, Arabic, and Chinese. Participants in their study reported that control over choice of tasks, control over difficulty level, language of directions and thinking and response time were the best features of the computerized testing instrument. However, the participants also reported that some aspects of language related to the interactional nature of the conversation could not be captured in the computerized version of the test. Finally, it has been argued that upon the emergence and proliferation of technology-based and
online teaching/learning settings, many students are now well adapted to the new technology and therefore the issue of differential computer familiarity seems to be fading away (Kirsch, et al. 1998; Wang & Chuang, 2002). For example, in a comprehensive study investigating the relationship between computer familiarity and performance on computer-based language tasks, Taylor, Kirsch, Eignor and Jamieson (1999) administered a questionnaire focusing on examinee computer familiarity to 90,000 TOEFL test-takers. They found no meaningful relationship between level of computer familiarity and level of performance on the computerized language tasks, after controlling for English language ability. They also concluded that no evidence existed of an adverse relationship between computer familiarity and computer-based TOEFL test performance due to lack of prior computer experience.

In addition to familiarity with computers and attitudes towards the new technology, test-taker gender can also be considered a construct-irrelevant factor with the potential to influence the performance on computer-based oral tests (Lumley & O’Sullivan, 2005; O’Loughlin, 2002). In formulating validity generalizations and investigating task difficulty it is important to discover which test-taker characteristics actually constitute sources of true variance in their performance. The purpose of detecting construct irrelevant factors is to prevent such intervening variables from unduly influencing performance on tests. The effect of gender as a construct irrelevant factor has been investigated in studies by Horton (1987) in relation to attitudes towards computer-based tests, and by Lumley and O’Sullivan (2005) in tape-mediated tests, as well as by O’Loughlin (2002) in face-to-face interviews. In a similar study, Gallagher, et al. (2002) investigated whether change from paper-based administration to computer-based tests influenced group and gender differences in performance.
Gallagher, et al. (2002) found that female examinees’ performance was lower on the computer-based test than its paper-based counterpart. Lumley and O’Sullivan (2005) found that the gender of the test-taker had small but not negligible effects on their performance in tape-mediated tests of speaking. They concluded that task difficulty may be influenced by individual differences rather than group characteristics. O’Loughlin (2002, p. 188), however, found that gender did not have a significant impact on the IELTS oral interview. In the current research, the relationship between test-taker gender and reaction to computer-based tests is investigated.

2.12 The Research Questions and Corresponding Hypotheses

This study addresses four areas of concern in the use of computer-based tests: comparability of scores, multimedia effect, task difficulty, and attitudes towards computer-based tests. The central issue is the comparability of scores obtained from conventional and computerized tests. However, as discussed in chapter 2, dealing with a new medium of test delivery, also means addressing other issues such as the efficacy and accuracy of computerised self-assessment, multimedia effect on test performance, anxiety level in taking a CBT, attitudes towards computer-based tests, and the like.

The questions investigated in this study are categorised into the following groups:

a) score comparability and self-assessment;

b) multimedia effect on performance and ratings and its relationship with the practicality of the test (including interface design as a facet);

c) task difficulty and factors contributing to this; and

d) attitude analysis and test anxiety.

The research questions and the corresponding research hypotheses are outlined below.
<table>
<thead>
<tr>
<th>Categories</th>
<th>No.</th>
<th>Research Questions</th>
<th>Corresponding Hypotheses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score comparability and self-assessment</td>
<td>1</td>
<td>Are CBT and face-to-face speaking test scores comparable? Are the tests measuring the same attribute?</td>
<td>The CBT and face-to-face speaking scores reflect assessment of the same attribute. The new CBT will tap the same construct as the conventional speaking test.</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>What additional information can we learn from the provision of self-assessment alongside “normal” assessment, and how consistent are the results of self-assessment with the actual ratings of the language performance?</td>
<td>The provision of self-assessment to the testees will enable the researcher to obtain a better picture of the overall ability level of the test-taker. There will be a high positive correlation between testees’ self-assessment results and their IELTS test or equivalent test results.</td>
</tr>
<tr>
<td>Multimedia effect on test performance and ratings</td>
<td>3</td>
<td>Does the multimedia delivery satisfy the interactiveness criterion of test usefulness? Are there significant differences between the perceptions of male and female test-takers in relation to the interactiveness criterion?</td>
<td>The multimedia test mode will facilitate the delivery of the interview procedure enabling the examinees to take control of the testing situation hence enhancing the interactiveness of the test. There will be no significant differences between the perception of the male and female test-takers of the CBT in relation to the interactiveness of the test.</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Do both male and female students find the user interface design useful and effective for the required tasks in the CBT?</td>
<td>The design of the user interface will be perceived as conducive to the quality of performance by both male and female test-takers and there will be no significant differences between the performance of male and female test-takers.</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>Does the change in the testing format enhance the practicability of the test and accuracy of proficiency ratings?</td>
<td>The change in the testing format will enhance the practicability of the test and accuracy of the rating procedure.</td>
</tr>
<tr>
<td>Task difficulty</td>
<td>6</td>
<td>What aspects of the computer-based tests do students find difficult to handle, and what are the factors contributing to task difficulty in CBT?</td>
<td>Students will perceive the CBT tasks to be equally difficult to those of conventional tests.</td>
</tr>
<tr>
<td>Attitude analysis and test anxiety</td>
<td>7</td>
<td>Are differences in test-taker attitudes and perceptions of difficulty associated with actual differences in task difficulty as reflected in scores assigned to learner performance?</td>
<td>Students with more positive attitudes towards computers and the CBT will display better performance on computer-based tests.</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>Do male and female students experience test anxiety in taking a CBT?</td>
<td>Students of both genders will experience less anxiety on the CBT than on conventional tests.</td>
</tr>
</tbody>
</table>
2.13 Chapter Summary

This chapter has reviewed the available literature on the use of technology in language testing and its implications in the quality of test taker performance in taking a computerized test of speaking ability. This analysis of the state of language testing provides three key elements of the rationale for the current study. First, to the best of the researcher’s knowledge, no language testing project has yet utilised the advantages of digital videos in task design and development or in the delivery of speaking tests. In this research, the researcher has endeavoured to develop an Australian computer-based test, namely the SAM Interview, to facilitate the assessment procedure of oral proficiency of incoming international students at Australian universities. This project is a first and important step, which involves conceptualizing, designing, trialling and validating a CBT. The content of the tasks is tailored to the interests of the potential target group of test-takers, and therefore covers different aspects of the Australian life-style as well as university life.

Compared to the COPI and the BEST Plus, the current project uses more advanced applications of the computer, in delivering the speaking test via video. The amount of reading has been reduced to zero in taking the SAM Interview. (For more information on the developmental processes and test specifications, see Chapter 3).

Second, there is a significant practicality dimension to the current study. Presently, international students are required to sit for the IELTS test or the ISLPR as part of the language requirement for their admission to Australian universities. The IELTS and the ISLPR are therefore the two prominent gate-keeping measurement instruments within the Australian higher education context. The current project offers an alternative placement instrument for testing speaking and provides education
institutions with an opportunity to screen a large number of candidates within a shorter period of time. Third, the current study makes a significant contribution to the investigation and the establishment of comparability of scores obtained from computer-based tests and conventional tests. This chapter also introduced the research questions and their corresponding hypotheses.

The following chapter discusses the methodology employed in this study and the detailed procedures in designing the SAM Interview test.
CHAPTER 3: RESEARCH DESIGN AND METHODOLOGY

This chapter delineates the research design and methodology employed in the study to create and evaluate the SAM Interview software package for testing oral proficiency. First, the design of the study is outlined. Then, the research questions and their corresponding hypotheses will be presented. Later in the chapter, the terminology used in the thesis will be defined briefly. This is followed by a detailed account of the SAM Interview test specifications and procedures for the trial, implementation, and validation of the test, as well as the rating of the testees’ language performance.

3.1 Design of the Study

This study employs an *ex post facto* design investigating the comparability of scores obtained from the administration of a newly developed computer-based test of speaking with those obtained from the speaking component of the IELTS or its equivalent, the ISLPR. It also investigates, both quantitatively and qualitatively, the attitudes, reactions, preferences, and perceptions of the examinees in relation to the usefulness, appropriateness and usability of the software package as a means of delivering the speaking test. The first part of the study investigates the relationship between performance on the CBT and face-to-face tests of speaking. The second part takes into account the reactions, attitudes, and perceptions of the examinees reflected in a post-CBT questionnaire. Figure 3.1 represents the methodology in this research.
FIGURE 3.1 Research design and methodology
The coefficient of correlation between the results of the two testing modes will be used to measure the comparability of the two test forms, an indication of the validity of the new computerized test. The test-takers’ attitudes towards computer-based tests are investigated quantitatively and qualitatively. The findings of this section of the study have implications for the use of technology in language testing and task design as well as for improving understanding of test-takers’ attitudes and reactions to technology mediated assessment.

3.2 Definition of Terms

The following section defines the key terms used in this thesis.

**Algorithm**: the step-by-step procedure devised to process the given data and produce the required results in a computer program designed to solve a problem.

**Authenticity**: the degree of correspondence between tests, tasks, and activities of the actual use of the language in real life. The closer the correspondence between real-life language use and language activity used in the testing situation, the more authentic the test (Bachman, 1990; Spolsky, 1985).

**CBT (Computer-based language test/testing)**: any test that has been prepared or adapted so that it can be administered and/or scored by a computer (Brown, 1997).

**Coding**: the writing of the syntax required for computer operations in the programming language. Codes (the program syntax) are a series of statements written in computer-readable language which transform some operations into others.
**Debugging**: the process of tracking down and removing any errors in the computer program. Debugging is inherently a frustrating process and requires a meticulous detective work. Bugs (errors in a computer program) could be the result of typing mistakes, flaws in the algorithms, or incorrect use of the computer language rules such as using addition instead of multiplication.

**Flowchart**: the schematic (graphic) representation of a logical process in a computer program. Flow-charts (also called flow-chart, or flow chart) are considered to be the very first step in conceptualising the overall working plot of a software package.

**Interview**: a direct and face-to-face encounter between the testee and the test-taker. Face-to-face interview is the most valid measure of speaking ability (Underhill, 1987).

**ISLPR (International Second Language Proficiency Ratings)**: a proficiency scale (with subclasses for macro-skills of speaking, listening, reading and writing) with 12 levels, ranging from zero to native-like ability. The scale is used to determine the general language proficiency of test takers. (Wylie, 1998; Wylie and Ingram, 1999)

**Item bank**: a collection of test items administered and analysed for use with an intended population. In computer-based tests, item banks are grouped according to difficulty level (Brown, 1997).

**Language ability**: language ability is defined as the capacity for using the knowledge of language in conjunction with features of the language use context to create and interpret meaning. The communicative language ability framework proposed by
Bachman (1990), which includes areas of language knowledge and meta-cognitive strategies, best describes the language ability.

**Language proficiency:** a person’s skill in using a language for a specific purpose.

There are two views of the nature of language proficiency. Some researchers view the construct of language proficiency as like a sticky substance consisting of one unique factor; but for others it is more like a machine which can readily be broken down into component parts or discrete points (Bachman, 1990; Davies, 1995; Oller, 1979). As a result, to date, there is no unanimously acceptable definition of language proficiency.

**Oral proficiency:** a person’s ability to express their thoughts, feelings, ideas etc. in spoken language.

**Performance:** the way the test-takers actually perform on a particular task rather than with a paper-and pencil instrument (McNamara, 1996). Performance on any given task requires the integration and use of a number of skills.

**Performance assessment:** any form of assessment that requires test-takers to show what they know by performing on specified tasks. Performance requires the candidates to accomplish approximations of real-life authentic tasks, usually using the productive skills of speaking or writing, but also reading or writing or combining skills (Bachman & Palmer, 1996; McNamara, 1996).

**Practicality:** the practical consideration of a test such as the costs, the amount of time it takes to construct and to administer, availability of equipment and resources, ease of
scoring, and ease of interpreting/reporting the results (Alderson, et al., 1995; Bachman, 1990). Practicality consideration is a paramount issue in the current research.

**Proficiency test**: a test which measures how much of a language a person has learnt irrespective how they have learnt it. It seeks to answer the question: Having learned this much, what can the student do with it? Proficiency tests are designed to measure test-takers ability in a language regardless of any training they may have had in that language (Davies, 1995).

**Programming language**: a special computer-readable syntax which allows the programmer to write the codes in the program in order to control the behaviour of the machine. Like human language, programming languages have their own set of syntactic and semantic features (Pierce, 2002).

**Programming**: the actual writing of the codes in the programming language. Programming has elements of language, mathematics, science and engineering.

**Rating scale**: a systematic procedure used to measure language proficiency in which aspects of a person’s language use are judged using scales that go from worst to best performance in a number of steps (Davies, 1995).

**Reliability**: the quality of test scores which refers to the consistency of measures across different times, test forms, raters, and other characteristics of the measurement context. Reliability is considered to be the pivotal quality of any measurement
process. For unless test scores are relatively consistent they cannot provide us with any information at all about the ability we want to measure (Henning, 1987). The ability we are interested in measuring is called the construct and whether we are truly measuring it is another pivotal concept called validity.

**Software Testing**: the process of analysing a software item to detect the differences between existing and required conditions (i.e., bugs) and to evaluate the features of the software item (Black, 1999; Hutcheson, 2003).

**User Interface**: a collection of objects by which the end users interact with the computer package or program. The user interface receives input from users to control the system and allows output from the system to inform the users (Fulcher, 2003; Shneiderman, 1998).

**Test security**: any measure taken to guarantee the confidentiality of test content or material prior to the administration of the test. Test security is a major concern in high-stakes testing situations.

**Testee**: an individual who takes a test. In this thesis, the terms testees, test-takers, examinees, subjects, candidates, students and participants are used synonymously.

**Validity**: the extent to which the inferences or decisions we make on the basis of test scores are meaningful, appropriate, and useful (Bachman, 1990). The traditional definition of validity is the extent to which a test measures what it is supposed to measure (Farhady, Jafarpur & Birjandi, 1994).
Washback or backwash effect: the effect of the nature of a test on teaching and learning. If the test has some sort of pedagogical effect on students’ learning of the language material, the test is said to have positive washback. Or if a test and the testing techniques are at variance with the objectives of the teaching course, then there is likely to be negative washback (Alderson & Wall, 1993).

3.3 Instrumentation

In this study, for data collection purposes, two major instruments were employed. The main tool was the SAM Interview computer-based test and the other was the post CBT attitude questionnaire. The following section provides a detailed description of the two data collection tools.

3.3.1 The SAM Interview Test Specifications

Test specifications are, by definition, a clearly defined set of characteristics of a given measurement instrument. The purpose of the following specifications is to assist test constructors and test users to gain a thorough understanding of the test development process, test purpose, rationale, content, administration procedure, and scoring criteria. The aim of the current project was to develop a computer-based test to facilitate the delivery of the general English proficiency interview. In drawing up the test specifications for this test, a set of criteria proposed by Alderson, et al. (1995, p. 11-20) was followed with slight modifications to more suitably fit the context of the research. Typically, one would expect that for every newly-developed test, there would be a need to develop a slightly different set of criteria and characteristics. Thus, in keeping with the specific purpose of the test there was a need to develop a set of
specifications that would ensure the fulfillment of the intended criteria. The specifications of the current project are:

1) Test Type and Purpose
2) Characteristics of the Test-Takers
3) Test Sections and Duration
4) Target Language Use Situation
5) Test Level
6) Test Delivery Mode
7) Language Skills to Be Tested
8) Tasks
9) Number of Tasks
10) Instructions or Rubrics
11) Assessment/Marking/Scoring Criteria
12) Description of Typical Performance at Each Level (Pre-intermediate, Intermediate, Advanced)
13) Inventory of the Required Resources

A detailed description is provided for each specification below.

1) Test Type and Purpose

This is a computer-based test of English proficiency in the sense that it is designed to measure the test-takers’ ability in spoken English regardless of any training they may have had. The content of this test is not based on any particular language program or
Research Design and Methodology

course objectives. Unlike an achievement test, the tasks in this test are not based on any pre-specified course content or material. This proficiency test is also diagnostic in nature, in that the sample performance of the test-takers can be analyzed for the purposes of discovering weaknesses and strengths. However, the diagnostic function of this test is not investigated in this study.

2) Characteristics of the Test-Takers

The potential candidates to sit for this test are newly-arrived international students who intend to undertake tertiary studies in Australia. The candidates come from non-English speaking countries, mostly in Asia and the Middle East, and usually have an English proficiency equivalent to 4.5 on the IELTS or +2 on the ISLPR. The participants in the evaluation of the test for this study are 30 male and female students aged between 20-33 who have not achieved the minimum requirement of English ability to start their tertiary studies in an Australian university and so are currently taking the ELICOS (English Language Intensive Courses for Overseas Students) courses at CALL (the Centre for Applied Linguistics and Languages) at Griffith University. The minimum age is 18, the age of those who have just finished their high school and are intending to take tertiary education. Differential abilities in using the computer are considered to be a construct-irrelevant factor and may potentially threaten the validity of the computer-based test. For these reasons, the design of the program does not require the candidates to have a thorough knowledge of the computer. To achieve this goal, every effort has been made to minimize the requirement for this computer skill. To be able to take this test, a basic knowledge of using the mouse and the keyboard will suffice.
3) Test Sections and Duration

Tests usually include separate sections, depending on the purposes for testing. This test is to be used for two purposes: assessment of English language proficiency and collection of data for research. The assessment component has two major sections: a pre-CBT self-assessment of speaking proficiency (see Appendix III for the items of the self-assessment) followed by five speaking tasks (see Appendix VI). For an average reader, the pre-CBT self-assessment takes about 5 minutes to complete. The average length of the five tasks in the CBT is 30 seconds and each should be answered in 3 minutes (15 minutes in total). The research component requires each testee to answer a post-CBT attitude questionnaire as well. This questionnaire includes 15 Likert type statements and a single open ended question, and takes about 5-7 minutes to complete (see Appendix VI).

4) Target Language Use Situation

This is a test of speaking ability whose purpose is to deliver assessment tasks through digital audio/video format with the responses recorded in digital format for subsequent rating by human raters. The authentic and actual situations envisaged for the test include telephone conversations, audio/video conferencing, online discussions using web-cams, authentic descriptions of scenes, as well as retelling stories and analysis. Situations like online conversations with the use of web-cams and audio/video conferencing are simulated in the test content and mode of delivery.

5) Test Level

The tasks in this assessment are grouped into the three levels of Pre-intermediate, Intermediate and Advanced. The reason why a beginner level is not included in this
test is that the potential participants have a minimum language proficiency comparable to the description of the pre-intermediate level. In the task bank, each level includes 10 tasks. At any given testing session, every test-taker is presented with three tasks at their own ability level and two more tasks at a higher ability level to challenge their potential skill. The tasks have been drawn up and trialed by four experienced ISLPR testers and cover a wide range of topics.

6) Test Delivery Mode

The tasks in this test are delivered through the use of a multimedia package installed on a computer. The tasks are presented in the form of video clips accompanied with relevant audio questions. The question related to each task is read aloud by a native speaker and the testees hear the question as they watch the video. The accompanying video is used as a contextualizing tool to reinforce the meaningfulness and authenticity of the tasks and to make sure that the test-takers have a clear idea of the content area requirements for successful performance on the tasks.

7) Language Skills to Be Tested

In this test, the assessment of the speaking ability is the major consideration and the tasks have been drawn up in line with the principles discussed in section 2.4 with reference to the definition of the speaking construct. The testees are required to watch the videos and listen to the corresponding questions and produce extended language samples in response. To make sure that other abilities are not involved in this assessment, no reading or writing is required for this test. However, as any speaking activity will involve some degree of listening, in this test the test-takers are required to keep a limited amount of information in their working memory in order to use it for
their speaking performance. The tasks which require a higher ability level involve more complex syntactic structures and also contain more abstract vocabulary items with a higher number of syllables. The length of each audio question is about 20 seconds. The visual stimuli assist the test-takers in understanding the task and its requirements hence obviating the need to keep an unmanageable amount of information in their working memory.

8) Tasks

In general, tasks in a speaking test may include any “written or picture-based materials such as role-play cards, menus, schedules, suggested topics or sub-topics for discussion, short written texts, pictures and picture sequences or whatever materials that the examiners provide to the examinees” to establish contents, outlines or starting points for the test discourse with the purpose of generating talk (Luoma, 2004, p. 53).

The tasks in this test are recorded videos with no audio. The questions related to the content of the images are recorded separately and then added to the videos. The scenes in the videos have been carefully selected not to contain any sophisticated abstract issues for tasks at lower levels. In order for the questions of the tasks and those of the questionnaire to be neutral, unbiased and non-discriminatory, every attempt was made to develop tasks on general content. Therefore, questions which were thought to be pertaining to national pride, strong religious beliefs, culturally sensitive issues, and other potentially controversial topics were avoided.

As a starting point in taking a speaking test, tasks are perceived as highly important because they allow testers to give direction to the way the test is administered. An
important consideration in developing tasks for any computer-delivered test of speaking is the time and effort to be dedicated to conceptualising and generating such materials. As elicitation instruments, these tasks need to be inspiring, motivating, clear and unambiguous in order to allow potential test-takers to generate enough talk to be representative of their general language proficiency for assessment purposes.

The nature/content of the tasks is closely associated with their future university life or social events that the students may encounter in their tertiary studies in Australia. The topics include travel, the environment, sport, general science, leisure activities, and libraries. To avoid concerns about copyright issues, the videos have been shot specifically for the test showing a variety of authentic scenery from the university, Australian life-styles and the natural environment. Again, as a safeguard to protect the copyrighted materials, no recognizable human face has been recorded. Instead of using a *talking head*, the researcher decided to record authentic scenery for the tasks as a way of contextualizing the assessment and at the same time reinforcing the meaning of the tasks. The videos were first recorded from a wide variety of contexts and were then edited using *Adobe Premiere 6.0* and the free version of *Video Cut Pro* software packages. The video file was parsed into smaller 30-second long clips each containing an idea unit. These idea units easily lent themselves to specific questions. The written questions for the tasks were read out by a native speaker in a recording studio. Each question was then merged with the corresponding video clips. It was decided that the question should be played towards the end of the video clip so the students would have had sufficient contextual and visual support to understand the actual question.
9) Number of Tasks

As noted earlier, the actual test consists of five tasks. Three of them are at the level of the testee’s language proficiency according to the self-assessment and two of them are one level higher. This allows the rater to evaluate whether or not the test-takers can fulfill the criteria for performance at the next higher level. It is generally believed that five tasks will give test-takers ample opportunity to produce a representative amount of language behaviour for the assessor to build upon their judgments on the testees’ speaking proficiency (Farhady, et al., 1994). It is an accepted fact that the longer the test, the more reliable the results will be; however, when the test length exceeds a certain limit, the gains in reliability become so minimal that it is not worth the practical considerations imposed on the test. A speaking sample of 12-15 minutes (as in the IELTS test) is thought to give the tester a representative amount of language for assessment.

10) Instructions or Rubrics

The exact procedures for taking this test have been specified in the instructional video clip. This video is played at the beginning of the test. The purpose of the instructions is to let the test-takers know what exactly they are supposed to do. As stated by Genesee and Upshur (1996, p. 201), the test-takers “should be informed clearly what to do for each kind of task included in the test, and they should get enough information about the test as a whole to decide how best to explain their efforts in the time available.” Luoma (2004, p. 52) believes that “when writing instructions, the assessment designers need to consider how much about each of the rubric features has to be explained to the participants.” The importance of instructions and their crucial role in standardizing the administration process of a test cannot be emphasized
enough. They need to be brief and clear because they set the scene for how the test-takers will perceive the task and their own performance on it. To ensure validity of the test and uniformity in understanding the procedure by all the test-takers, the wording of the instructions has been carefully chosen by four professional test developers and ISLPR trained raters. The instructions are read out clearly by an experienced native speaker. The information given in the instructions includes the purpose and procedure of the test, each section of the test, the length of the test, as well as performance requirements on the tasks. The language of the instructions is straightforward and easy to understand (see Appendix IV). To make sure that the testees fully understand the instructions and to be on the safe side for validity, the examinees are given the option of listening to and watching the instructions as many times as they wish.

11) Assessment/Marking/Scoring Criteria

The assessment criteria in this test are the ones used in the speaking section of the ISLPR. The language samples produced by the examinees in the SAM Interview are rated by an experienced and trained ISLPR rater. More precisely, the criteria in assessing the test-takers’ performance include: pronunciation (intonation and accent), accuracy, syntactic structures, coherence, relation, creativity, stylistic appropriateness, the depth of vocabulary knowledge and fluency.

12) Description of Typical Performance at Each Level (Pre-intermediate, Intermediate, Advanced)

Obviously, as we move to higher levels, the difficulty of the tasks also increases and a higher speaking ability is required for successful performance on each task. In this assessment, tasks which require lower cognitive load and information processing are
categorized as lower level tasks and hence directed to lower ability test-takers), while
tasks which require complex information processing and more abstract thinking are
categorized as more difficult tasks which are directed to higher ability test-takers.

Examples of tasks at the lower ability level are those related to concrete topics such as 
library activities, first day experience in a foreign country, traffic conditions, food,
and similar issues. However, at higher ability levels, the tasks and topics are more
abstract; and therefore successful performance on them will require more analytical
and critical thinking and deeper information processing ability. Examples of more
difficult items include topics and tasks related to social issues, predictions of modern
life, or aspects of environmental protection. The following is a more detailed
description of tasks at each level:

a) Pre-intermediate (equivalent to 1-2 on the ISLPR):

The tasks in this level are designed for test-takers with lower order abilities in spoken
English and involve the following:

- Recounting
- Description
- Retelling
- Talking about personal experience

The questions at this level have a simple grammatical structure using familiar and
concrete vocabulary items as well as short sentences. Tasks may include simple
talking about seasons, food, and the like. The responses from the test-takers are also expected to include the type of language sample described in ISLPR 2. Not much extra-linguistic or world knowledge is required to respond. The topics are daily events, such as describing a library, telling childhood stories and first day in Australia, which are thought to have been experienced by a typical international student intending to study in Australia. Questions are simple and short, but to ensure they can be understood by all test-takers they are very much contextualized and reinforced by the accompanying videos of related scenes. According to the authors of the ISLPR, a person at this level of ability will be “able to use the second language to satisfy requirements of simple, controllable activities directly related to teaching in ‘regular’ second language programs aimed at developing general proficiency, and of routine non-teaching situations pertinent to the profession” (Ingram & Wylie, 1979).

According to Wylie & Ingram (1999, p. 18) a candidate at this level is:

“able to satisfy basic social needs, and the requirements of routine situations pertinent to own everyday commerce and recreation and to linguistically undemanding ‘vocational’ fields. The learner sustains basic social conversations on everyday topics, routine transactions pertinent to own consumption of goods and services, and routine tasks in a ‘vocational’ area which does not require high-level or specialised language skills. Can communicate some information in less routine but concrete situations, provided the register is familiar and there is strong support from the context. Can not express any significant degree of complexity or abstraction. Expresses own emotional attitudes and (tentatively) own intellectual attitudes about familiar topics with some, though not great, precision. There may be a significant difference between what the speaker
wants or intends to convey and the total message (including purposive and attitudinal elements) that is actually conveyed. Uses a variety of high-frequency connectives. Uses a variety of subordinate clauses, though not always securely. Usually handles simple, high-frequency structures fairly accurately, but lack of grammatical control shows in more complex or less familiar structures. The influence of L1 is likely to result in non-standard forms. Lack of mastery of cohesive devices, particularly in longer utterances, may cause temporary confusion on the part of interlocutors. Structures needed for the task types indicated are often lacking; to overcome this and gaps in vocabulary, circumlocutions are frequent. These circumlocutions and other strategies or hesitations necessitated by grammatical or lexical limitations affect the learner’s fluency, and the overall rate of imparting information is significantly less than that of a native speaker. Idioms are noticeably lacking or misused. Pronunciation is generally intelligible. Undue exertion on the part of the listener is not usually necessary in familiar situation types, although some repetition may be needed when there is relatively little support from the context. Register flexibility is limited. In very familiar situation types, language used may be appropriate to the situation in terms of, for example, level of formality, intimacy and technicality, but this appropriateness is unlikely to be sustained in a wide range of situations involving spontaneous language use.”

b) Intermediate (equivalent to 2+ and 3 on the ISLPR):

The tasks at the Intermediate level are more complex and demanding in nature. They require abilities such as:
• Recounting
• Analysis
• Basic argumentation
• Inference
• Extra-linguistic (world) knowledge
• Comparison and contrast

At this level the topics are more complicated and require test-takers to infer from available information and arrive at conclusions. In some tasks, an example may be given and the candidates are asked to elaborate on them or give examples and explanations of their own. The candidates may also be asked to talk about celebrations and customs of their country or compare and contrast aspects of Australian daily life (e.g., transportation) with those in their home country. According to the authors of the ISLPR, a person at this level of ability will be “able to use the second language effectively in most primary and secondary school activities directly related to teaching in ‘regular’ second language programs aimed at developing general proficiency; able to use the program effectively in non-teaching situations pertinent to the profession which do not require very complex, precise or specialized language” (Ingram & Wylie, 1979). According to Wylie & Ingram (1999, p. 24) a candidate at this level is:

“able to perform effectively in a wide range of informal and formal situations pertinent to social and community life and everyday commerce and recreation, and in situations which are not linguistically demanding in own ‘vocational’ fields. In most conversations in such situations, the learner conveys fairly precise meanings, and has sufficient control of discourse to be able to sustain to
some extent the juxtaposition of different ‘planes of meaning’. In interviews and discussions, however, can not pursue the argument to the complexity, depth, level of precision and/or abstraction that are often required in specialised fields. Gives relatively long narrative or descriptive monologues fairly effectively. In any unprepared expository monologue of any length, however, rhetorical structure is likely to be non-standard, and limitations, particularly in control of discourse and precision of vocabulary, often mean that the argument is not developed and sustained effectively. In complicated and/or less familiar situation types, there are likely to be differences between what the speaker intends or would like to convey and the total message (including purposive and attitudinal elements) that is actually conveyed. Uses a variety of conversational linkers and other connectives, and a wide variety of subordinate clauses, and of verb forms and other grammatical structures. Errors are made (sometimes reflecting the influence of Li), but they rarely interfere with understanding, and do not generally irritate or amuse native speaking interlocutors of a similar sociocultural background. Vocabulary range is such that, in conversations of the types indicated, the learner can readily overcome most gaps by circumlocution. This circumlocution and the pauses that are sometimes necessary as the learner searches for words or structures mean that the overall rate of imparting information is generally less than that of a native speaker. Makes secure use of only high-frequency idioms and colloquial forms. There may be a fairly strong accent, but other elements in utterances are sufficiently standard to compensate for non-standard phonological features, and problems with intelligibility rarely disturb the native speaker. Any failure to observe basic social conventions in informal conversation is unlikely to be attributable to L2 developmental factors,
although the influence of L, discourse norms may result in delayed ‘coming to the point’. Modifies language to a significant degree to meet the differing register requirements of familiar situation types.”

c) Advanced (equivalent to 3+ or higher on the ISLPR):

Tasks at this level are more advanced in nature and require abilities such as:

- Analytical thinking
- Argumentation
- Discussion of abstract concepts
- Prediction
- Comparison and contrast in relation to cultural aspects of home country and Australia

At this level, the test-takers are expected to respond confidently and critically and be proficient in talking about both familiar and unfamiliar topics. Successful test-takers at this level are those who produce sophisticated and convincing arguments and judgments about social, environmental, and political issues in relation to the tasks presented. According to the authors of the ISLPR, an individual at this level of ability will be “able to use the second language effectively in all primary and secondary and most tertiary-level activities directly related to teaching in ‘regular’ second language programs aimed at developing general proficiency, and in specified (vocational) purpose and immersion and bilingual programs in own fields of specialization; able to use the language very effectively in almost all non-teaching situations pertinent to the
profession” (Ingram & Wylie, 1979). According to Wylie & Ingram (1999, p. 29) a candidate at this level is:

“able to perform very effectively in almost all situations pertinent to social and community life and everyday commerce and recreation, and generally in almost all situations pertinent to own ‘vocational’ fields. The learner conveys his/her desired meaning in straightforward conversations, interviews, discussions and monologues with virtually the same fluency, precision and complexity, and to virtually the same depth as do native speakers of the same sociocultural variety. Usually needs no more support from the context to communicate than a native speaker does. The learner may for a short time in some situations produce language which is indistinguishable from that of native-speaking peers. In very complex texts, however, has less control of the argument than such peers do. Rhetorical structure in such texts may at times be non-standard, particularly in less familiar situation types. No grammatical structures are missing from the learner’s repertoire; errors of grammar are fairly rare, and are often picked up in a monitoring process and corrected immediately. Errors never interfere with understanding, although there may be occasional lapses in the use of cohesive devices (typically when the referent is well separated) which may momentarily distract listeners. Vocabulary range is close to that of a similarly educated native speaker and allows for some stylistic variation for aesthetic purposes (e.g., for euphony). High- and medium-frequency colloquial and idiomatic forms are secure but some non- or misuse of other items occurs. Is secure in the use of borrowings (from other languages or other varieties) that are in high- and medium-frequency use in the speech of native-speaking peers. There may be an
obvious ‘foreign’ accent, but this in no way impedes comprehension by a native speaker of the same or a similar variety. Has considerable sensitivity to register requirements. There are, however, occasional minor lapses in terms of appropriateness of expression (e.g., inappropriate influences of written text) and, particularly in less familiar situation types, in terms of what meanings may be (directly) expressed. Such lapses do not confuse interlocutors, and do not generally per se offend native-speaking peers.”

13) Inventory of the Required Resources

According to Luoma (2004, p. 56) “the way a speaking assessment is administered influences the examinee’s experience of the test, and therefore it also has a bearing on what is tested.” As a result, in the interest of validity, fairness and comparability, every effort should be made to provide the testees with the required equipment and inform them of the detailed procedures of the test. This assessment tool is recorded on a CD-ROM and under optimal circumstances it can be successfully administered in a quiet location using any platform. The following table shows the minimum system requirements:

<table>
<thead>
<tr>
<th>TABLE 3.1 The minimum system requirements for the SAM Interview program</th>
</tr>
</thead>
<tbody>
<tr>
<td>Processor</td>
</tr>
<tr>
<td>Memory</td>
</tr>
<tr>
<td>Hard Disk</td>
</tr>
<tr>
<td>Drive</td>
</tr>
</tbody>
</table>

3 See Appendix VII for instructions on installing the software package.
Display  Super VGA (1024 x 768) or higher resolution display with 256 colors
Mouse   Microsoft Mouse or compatible pointing device
Headset  Headphones and a microphone

To maintain audio quality, it is strongly recommended that the test be administered in a quiet room.

3.3.2 Attitude Questionnaire

In order to investigate the attitudes, reactions and experiences of the CBT test-takers, a questionnaire was designed and given to the students immediately after the administration of the computer-based test. This post-CBT questionnaire comprised two parts:

**Part I:** Part I of the questionnaire contained fifteen Likert-type items which were designed to test three major hypotheses concerning the following issues:

- **a)** User Interface
- **b)** Anxiety Level
- **c)** Task Difficulty

**Part II:** The second part was an open-ended item which required students to write down any other comments not covered by the preceding fifteen items. The testees’ responses to this open-ended question are investigated qualitatively.
3.3.3 Developmental Process: The SAM Interview Package

In this section a detailed account of the developmental process of the test delivery package for this study is presented. This process required a significant amount of time and the acquisition of significant new computer programming skills on the part of the researcher.

The special software package designed for this study was the primary data collection instrument in this research. This computer package was designed and developed using the programming application *Visual Basic 6 (VB6)*. At the time of conducting the current study, there was only one published article in the language testing literature on user interface design for computer-based language tests (Fulcher, 2003). The scarcity of relevant literature or adequate models for this operation meant that additional skill development was required for the researcher. This was essential to conceptualize, plan, design and write a computer program, debug, as well as test and trial the package before installing it.

VB6 provides an Integrated Development Environment in which the programmer can choose what type of software program to develop. For the purposes of this study, an Event-Driven Programming approach was used (Shneiderman, 1999). The EDP approach is one in which it is the user who dictates the order of program execution. In other words, instead of the program “driving” the user (in this case ultimately the testee), it is the user who “drives” the program (Deitel, Deitel & Nieto, 1999, p. 51).

At a deeper level of programming, it also means that the programmer designs the program to be handled or controlled by the end user, i.e., the test-taker will have control over the test-taking process.
VB6 is a high level language which allows the programmer to use instructions that more closely resemble the English language. High level languages require either an interpreter or a compiler to convert the English-like instructions (commands) into the 0s and 1s the computer can understand. In procedure-oriented high-level languages, the emphasis of a program is on *how* to accomplish a certain task. It is the programmer’s job to instruct the computer every step of the way, from the start of the task to its completion. With all this in mind, the programmer can visualise the ultimate test-taking process of a certain group of students and implement this process into the program.

In taking a computer-based test, the testees are actually dealing with a number of buttons, scroll bars, dialog boxes, etc. (technically called “objects”) on the screen to take control of a series of events — in this case clicking on a right answer or the next question. To create objects and events in a program we need a high-level language, namely, an object-oriented and even-driven programming language. VB6 is an object-oriented and even-driven programming language which is easy enough for a nonprogrammer to use, yet sophisticated enough to be used by professional programmers. In object-oriented even-driven languages, the emphasis of a program is on the objects included in the user interface (such as scroll bars or buttons) and the events or actions (such as scrolling and clicking) that occur when these objects are used.

The following are the key requirements of the standard steps taken to develop the testing software (Shneiderman, 2004; Zak, 2001).
STEP 1: Analysing (defining the problem): This was the first step in programming which is analogous to the definition of a problem in any research project. At this stage, a statement of the problem was articulated in order to give a clear direction to the rest of the programming steps. In other words, what exactly the programmer wants the program to do was defined at this point. The core problem at this stage is the low degree of practicality of administering the spoken test to a large group of EFL/ESL students through the use of a live face-to-face interview under time and budget constraints in tertiary education institutions in Australia.

The secondary problem from a validation perspective is whether the test scores, obtained from computed-based tests, are comparable to those obtained from face-to-face conventional interview tests. From the programming point of view, the statement of the problem was: “What characteristics should a computer package have in order to reliably deliver, or at least facilitate the delivery of, the speaking proficiency test through the computer?” This problem prompts the need for an event-driven program to deliver the interview test.

STEP 2: Planning the solution of the problem: The planning stage is to set out a logical sequence of steps (called an algorithm) to solve the problem. The procedure-oriented approach to problem solving required the researcher to think in a step-by-step, top-to-bottom fashion. This included planning tools such as pseudocodes and flow-charts. The Pseudocodes make use of English-like phrases to represent the required steps. However, in order to visually display the break-down of the steps, a standard flow-chart of the program is usually preferred to a pseudocode. A typical flow-chart consists of special geometric symbols connected by arrows in a logical order. These symbols are standardized and therefore uniformly understood and
interpreted by all computer programmers. The shape of the symbol represents the type of operation that is to occur. Within each shape is a short phrase indicating the activity to take place at that particular step. A summary of the symbols used in the flowchart for this project are shown below (Schneider, 1999, p. 35):

<table>
<thead>
<tr>
<th>Terminal (Stop Circle)</th>
<th>Represents the events/tasks that start or end a particular flow-chart.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input/Output Block</td>
<td>Represents the block where the output of a process (such as reading or printing) will be stored.</td>
</tr>
<tr>
<td>Process Block</td>
<td>Represents a set of arithmetic and data-manipulation operations that occur one after the other.</td>
</tr>
<tr>
<td>Decision Block</td>
<td>Is like a junction where a path can be chosen depending on the user’s demand. Unlike the Input/Output and Processing symbols, which have only one entry and one exit flow line, the Decision Block has one entry and two exit paths. The path chosen depends on whether the answer to a question is “Yes” or “No.”</td>
</tr>
<tr>
<td>Flowline</td>
<td>Is used to connect all the above symbols and indicates the flow of logic.</td>
</tr>
</tbody>
</table>

**STEP 3: Choosing the interface:** The interface of any computer-based test consists of the actual objects the test-takers see and deal with during a testing session. These objects may include text boxes, command buttons, toolbars, videos, icons, animations, progress bars, date/time indicators, and so on. The key to designing a successful user interface is to have a complete understanding of the target user (Luther, 1992). This includes not only the language ability of the users but also their computer familiarity and experience with various user interfaces as well as what exactly they are supposed to do with the interface of this particular application. The above goals may be achieved by a process called *user task analysis* in which the developer assumes himself or herself as the target user and identifies a series of potential scenarios in order to come up with the most convenient one.
Obviously, the more complex the intended application, the more important task analysis becomes. In the literature on language testing there is no published article about the process of interface design in computer-based tests other than the one by Fulcher (2003). The nature of the relationship between the interface of an application and test-takers is important because, as Fulcher (2003, p. 384) argues, “only through following a principled [logical] approach to interface design can the threat of interface-related construct irrelevant variance in test scores be avoided.” As to the importance of the use of graphics in user interface design, Zak (2001, p. 111) believes that “the human eye is attracted to pictures before texts.” Therefore it is advisable to include a graphic only if it is necessary to do so. The following is the flowchart of the SAM Interview software package:
FIGURE 3.2 Flowchart of the SAM Interview software
It is also advised that in designing the user interface it is better to use a small graphic in a location which will not distract the user. Because test-takers will be directly dealing with the interface of the test, the inclusion of audio and video in the interface is a delicate issue and at the same time a technologically demanding operation. Therefore, every effort was made to make sure that the audio-video option is not only helpful but necessary. A growing body of literature in the field of human-computer interaction design attests to the importance of utilizing the auditory as well as visual channel of communication in a user interface (Baltova, 1994; Brett, 1997).

Also, if sound recordings are added to the interface of a computer-based test, they can possibly make the testing program easier to learn, and to use, and thus can mean examinees make fewer errors in the test-taking process (Parshall & Balizet, 2001). And more importantly, the addition of audio and video can be expected to turn the test-taking process into a more enjoyable experience leading to a stronger positive washback effect on the testees.

Sounds are already playing a very important role in the interface design of most of the contemporary software programs. They can serve to attract the users’ attention (e.g., a beep to indicate that the user is trying to exit an unsaved document or to go past the end of a page), or they can refer to something not visually present (e.g., a tone signalling the arrival of a new email message). Among other potential benefits of multimedia in the interface design of computer-based tests is the “use of audio and video [to] improve the standardization of tests of students with reading deficiencies” (Parshall & Balizet, 2001, p. 9).
Yet another major consideration in the design of a user interface is *consistency* which is part of the overall consistency of the software application. Given the importance of standardization of a new application, the consistency rule must apply rigorously throughout the test-taking process. In addition to consistency, the general simple-is-better principle applies to all types of interface designs including multimedia applications and their interfaces. Maintaining the simplicity of the interface requires a great deal of thinking on the part of the application developer. When we are not sure about the computer familiarity of the potential test-takers, it is very much advisable to be on the safe side and to avoid all ‘bells and whistles’ which generally complicate the user interface unnecessarily. They can often distract the user and may not even be used anyway.

During the design process of the current project, the above principles and details have been observed as far as possible in order to enhance the overall quality of the software package. The quality of an application, defined as “the totality of features and characteristics of a product or service that bear on its ability to meet stated or implied needs,” is largely determined by the quality of the process used to develop and maintain it (Sanders & Curran, 1994, p. 6). Fulcher (2003) identifies three general phases in the development of a good interface design:

**Phase 1: Planning and initial design** which includes hardware and software considerations, navigation options, page layout, terminology, text, colour, toolbars and controls, icons, and the rest of the visible objects on a typical computer-based test.
**Phase 2:** Usability testing which involves activities such as searching for problems and solutions, selecting test-takers for usability studies, item writing and banking, pre-testing, trialling scoring rubrics.

**Phase 3:** Field testing and fine tuning includes trialling the interface with a large sample drawn from the target test-taking population, and also ensuring that the logistics of data collection, submission, scoring, distribution and retrieval, and feedback work as planned. This phase provides an opportunity to test for (possible) variation in the appearance of the interface across sites, machines, platforms, and operating systems. As is the case with any software development project, the instrument for the current research went through all the standard steps for usability (including trial, fine tuning, and usability testing processes with volunteer IT students at Griffith University). The volunteer users were observed by the researcher and their suggestions were taken into consideration in a debriefing interview for the final fine tuning of the program. These IT students were asked to put themselves in the shoes of target test-takers and highlight any undue difficulty in navigating the interface during interaction with the program.

**STEP 4:** Coding: Coding, the translation of the algorithm into a programming language, was the most complex and time-consuming step in the development of the current project. Once the planning of the application and the building of the user interface were complete, Visual Basic instructions were written to direct the objects in the interface on how to respond to events.
As mentioned above, the program codes are a set of detailed instructions that tells Visual Basic how to manipulate data, perform input and output, and respond to the user’s commands such as clicking or scrolling. The controls are mapped on a particular page to allow users to trigger the event procedures when they interact with a certain control.

**STEP 5: Test and Debug:** Due to the complexity and intricacy in the coding stage of the programming, as well as the possible persistence of syntactic anomalies in the programming language, testing and debugging become an inevitable part of the operation. Caution must be exercised at this stage not to compile or distribute the application before the most stringent debugging for possible problems. The most common error that the Visual Basic will report is that of a “Run time error.” A circular method debugging was used to locate correct errors and to “develop a reasonable assurance that the program is working correctly and that it stays correct as it evolves” (Kernigham & Pike, 1999, p. xi). Also, to establish the reliability and consistency of the program, (after successful compilation and distribution) it was installed and run on a number of platforms with different versions of the Windows operating systems.

**STEP 6: Completing the documentation (or Distributing the Application):** This stage involved creating a setup (executable/installable) file along with all its components for new users and new platforms. To this end all the files of the application along with all the projects were translated by VB into a single executable file. As part of supporting documentation and trouble shooting safeguards, all the materials that describe the program were compiled to allow other people, including test users, teachers, administrators, organisations, and proctors, to understand the scope of the program.
and what it does. Distributing the application makes it possible to run the application on different platforms and with different operating systems and to secure the compiled files, projects, and codes. See Appendix VII for instructions on how to install the program.

3.3.4 Security and Copyright

The software package, SAM Interview, is copyrighted to the researcher (© Seyyed Abbas Mousavi 2004) and runs under Windows Operating System (NT, 2000, XP). Once installed on a stand-alone platform, the program is operational with the accompanying CD-ROM. The CD-ROM includes the dichotomously-scored items of the self-assessment questionnaire as well as all the test items (audio-video tasks) recorded in .AVI format.

Breaches of test security can have detrimental effects on the validity of the test (Dunkel, 1999). For this reason, upon the installation of the software, the test tasks are not transferred onto the host computer. The video clips of the tasks remain on the CD-ROM and are in the possession of the researcher and his supervisors.

Another precautionary measure taken in favour of the security of this test is that when a particular testee decides to take the test a number of times in close succession, the underlying algorithm makes it almost impossible for the testee to be given exactly the same set of tasks even if their performance on the initial self-assessment is identical. Furthermore, the database of the software package is password secured and therefore access to the bio-data of the participants and their self-assessment results can only be made possible by entering a 4-digit password.
3.3.5 Procedure

The following section describes the step-by-step procedure of the program as well as its functions and components.

**STEP 1:** The program is installed and run under Microsoft Windows, preferably the XP version. To ensure a user-friendly user interface design and the clarity of the procedures, the colours and colour combinations of the standard Windows operating system were used. Research indicates that university level computer users prefer the Windows-based software interface to old versions (Chou & Hsiao, in press). The shape of the icons and the choice of font style and size followed the most commonly used conventions.

**STEP 2:** As soon as the program runs and prior to the commencement of the actual testing phase, the testees are shown a short introductory video clip performed by an experienced English native speaker who introduces the rubrics of the test in straightforward and clear language. In this introductory video, the speaker welcomes the test-takers and introduces the test, giving information about its steps, format, function, procedure and length. For a full text of the instructions, see Appendix IV. To be on the safe side regarding the validity of the procedure, these instructions are given in a simplified language to make sure that the wording is equally understandable to all test-takers of different ability levels. Again, to ensure the clarity of the instructions and procedures, the test-takers have the option of replaying the Introductory Video as many times as they think is necessary. The following is a screen shot of the Introductory Video:
STEP 3: This step is for the test-takers to enter their personal information and an email address for follow-up correspondence with the researcher or the test centre in which the test is being administered. Every effort was made to minimize the amount of typing on the part of the testees. As soon as the testees type their email address in the required slot, a unique ID is automatically generated for each test-taker. This ID includes the testee’s name and a numerical code for privacy purposes and stored in an underlying location in .MDB (Microsoft Access database) format. The following is a screen shot of the Personal Information page:
STEP 4: Once the students have entered their personal information into the appropriate lines, the next screen will be a computer-administered, thirty-item, can-do style, dichotomously-scored self-assessment test. The purpose of this test is for the computer to determine an initial approximate ability level of the testees in order to be able to randomly select tasks from the appropriate pool. The difficulty level of the subsequent tasks will depend on how well the testees perform on this section. As noted by Wainer (1999, p. 10), quoted in Dunkel (1999):

“…this approach stems from the realization that we learn little about an individual’s ability if we persist in asking questions that are far too difficult or far too easy for that person. We learn the most about an examinee’s ability when we accurately direct our questions at the same level as the examinee’s ability.”
The program algorithmically assigns a score of between 0 and 30 to testees, where 1-10 = Pre-intermediate; 11-20 = Intermediate; and 21-30 = Advanced. The testees with lower levels of ability will be presented with tasks of lower difficulty level; and for the testees who have self-assessed themselves as having higher ability levels, items of higher difficulty will be presented. These items are randomly selected from an underlying pool of tasks for each level. The items of the self-assessment test are presented in Appendix III. The following is a screen shot of the Self-assessment Page:

**STEP 5:** This step is where the actual testing begins. Each testee is presented with five video-based questions each requiring the test-taker to speak for three minutes in response to a question posed about the events taking place in the video clip. The testees are given the first three tasks at their own level of ability as determined in the
self-assessment test (see Tasks, 1, 2, 3 in the program flowchart). The tasks are randomly drawn from a large of tasks for each level. Once the testees have answered the first three tasks, two more tasks are presented to them (see Tasks, 4 and 5 in the program flowchart). These two final tasks are one level higher than the testees’ language ability determined by the self-assessment test. That is, if a testee’s performance on the self-assessment has been assessed as Intermediate, he/she will be given three tasks at the Intermediate level and two tasks at the Advanced level. The purpose of this is to check if the testees — especially those who have underestimated themselves in the self-assessment — are still able to perform at one level higher than they have assessed themselves. The testees hear and see the videos once only and have three minutes to answer each question.

While the testees are speaking into the computer’s microphone, their voice is being recorded in MP3 format. The computer is programmed to create a file name which contains the numerical code of the testee, the testee’s name and the task number:

```
44abbas22.mp3
```

This will allow the researcher or the rater to be able to determine which testee has answered which task. The responses given by each test-taker will be a total language sample of about 15 minutes which is thought to be representative enough for the rater to determine the language proficiency level of the testees. The following screens are samples of the test running in Playing and Recording modes:
STEP 6: Once the test-takers have responded to all the five tasks, a final video clip is played in which a native speaker of English expresses appreciation for their help and taking the test. At this point, the program terminates and the testees have the option of taking the test again or exiting the program. These .MP3 files can be accessed by raters for rating or research purposes.

The program has been designed in a way that if the test-takers decide to take the test again, it is highly unlikely that they will be presented with the same tasks as they were administered before. Potentially, with the increasing capacity of the CD-ROMs and DVDs, an enormous number of tasks can be stored in the Task Pool. The more the number of tasks, the less likely it will be for the test-takers to be given the same task when they take the test twice. Also, this gives the test developer or the researcher an opportunity to include more tasks of different variety and a wider range of target language use situations.

STEP 7: Once the testees have exited the program, they are given a post-CBT questionnaire containing fifteen Likert-type items and one open-ended question. This provides the test-takers with an opportunity to record their experiences, attitudes and comments in relation to the usefulness and other aspects of the testing process.

3.3.6 Rating the Testees’ Performance

The rating process in any speaking test involved the use of a rating scale to evaluate a sample of speech or successful performance generated by a specific task. In this study, for scoring purposes, the testees’ performances on the CBT tasks were rated against

4 “That is the end of the testing session. Thank you very much for taking the test.”
the standard criteria of the ISLPR Guidelines by an experienced and trained rater. The use of a standardized criterion is necessary because it assists the assessor to judge the complex performance of an individual in a reliable, fair, and valid way. The opportunity a CBT provides is that the raters have the option of listening to the testees’ audio file in any order or as many times as they wish to. Important factors to be considered in the ratings of language samples include: grammar, fluency, pronunciation, accuracy of speech, richness of vocabulary, as well as cohesion and coherence of speech samples.
CHAPTER 4: RESULTS AND ANALYSIS

In this chapter the qualitative and quantitative data obtained from the administration of the self-assessment test, the computer-based test, and the post-CBT attitude questionnaire are presented along with the students’ IELTS scores. For quantitative data, statistical tests of student t-test, correlation and Chi-square are carried out. The examinees’ feedback to the Likert type statements on the questionnaire are analysed quantitatively, and their responses to the single open-ended question are analysed qualitatively. Correlational analysis is used to determine the presence, as well as direction and strength, of any relationship between variables which are hypothesised to be important in performance on the computer-based test.

This chapter also reports on the inferential data analysis (Chi-square) to determine the significance of the differences between the variables involved in the hypotheses. The results of the subjects’ IELTS/ISLPR scores are also reported in this chapter. Later in this chapter, samples of test-takers’ spoken performance will be analysed in relation to the criteria of assessment used in the ISLPR. A more detailed discussion of the findings will be presented in the following chapter.

4.1 Descriptive Statistics (Self-assessment, SAM Interview, IELTS)

Table 4.1 shows the overall descriptive statistics of the participants in this study. It includes the gender and age of the participants, computer-generated results of the administration of the CBT self-assessment, ranked order results of the participants’ performance on the SAM Interview and the participants’ score on the speaking
component of the IELTS test. As Table 4.1 shows, a total of 30 students, 13 males and 17 females, took part in the study. The participants’ age distribution had a range of 13 with an average of 25.56 years. The maximum age was 33 and the minimum was 20. Age was considered a potentially important factor in the study as it is commonly thought that this variable would negatively correlate with the attitudes towards and anxiety regarding the use of computers especially in testing contexts.

For two reasons, in this study, the participants’ overall band score on the IELTS was not taken into consideration. Firstly, the focus of this research is on the speaking ability only; secondly, a given individual’s overall band score of the IELTS could have been influenced by a higher score on any other component on the test. Therefore, it was decided that performance on the SAM Interview should be compared against the IELTS speaking scores only. It is important to note that not all participants had taken the IELTS test to enter Griffith University. Actually, about 25% of the participants had taken the ISLPR instead. As a result, for ease of analysis of data and interpretation of results, the candidates’ ISLPR scores were converted to the IELTS scale using the standard score conversion table (see Table 2.2) drawn by Wylie (2006).
### TABLE 4.1 Testees’ characteristics, self-assessment results, and performance on IELTS and SAM Interview

<table>
<thead>
<tr>
<th>Testee</th>
<th>Gender</th>
<th>Age</th>
<th>Self-assessment</th>
<th>SAM Interview</th>
<th>IELTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>F</td>
<td>22</td>
<td>Intermediate</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>M</td>
<td>24</td>
<td>Advanced</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>F</td>
<td>23</td>
<td>Pre-intermediate</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>M</td>
<td>32</td>
<td>Pre-intermediate</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>M</td>
<td>20</td>
<td>Intermediate</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>6*</td>
<td>F</td>
<td>30</td>
<td>Intermediate</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>F</td>
<td>25</td>
<td>Intermediate</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>F</td>
<td>28</td>
<td>Intermediate</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>9*</td>
<td>M</td>
<td>30</td>
<td>Advanced</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>10*</td>
<td>M</td>
<td>25</td>
<td>Intermediate</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>11*</td>
<td>F</td>
<td>29</td>
<td>Intermediate</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>12*</td>
<td>F</td>
<td>33</td>
<td>Intermediate</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>13</td>
<td>F</td>
<td>23</td>
<td>Intermediate</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>14</td>
<td>M</td>
<td>30</td>
<td>Intermediate</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>15</td>
<td>F</td>
<td>26</td>
<td>Intermediate</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>16</td>
<td>F</td>
<td>24</td>
<td>Advanced</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>17</td>
<td>F</td>
<td>30</td>
<td>Intermediate</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>18</td>
<td>M</td>
<td>23</td>
<td>Intermediate</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>19</td>
<td>M</td>
<td>23</td>
<td>Intermediate</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>20</td>
<td>F</td>
<td>20</td>
<td>Intermediate</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>21*</td>
<td>M</td>
<td>23</td>
<td>Pre-intermediate</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>22</td>
<td>M</td>
<td>30</td>
<td>Intermediate</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>23</td>
<td>M</td>
<td>33</td>
<td>Intermediate</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>24</td>
<td>F</td>
<td>32</td>
<td>Intermediate</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>25</td>
<td>M</td>
<td>21</td>
<td>Advanced</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>26</td>
<td>F</td>
<td>21</td>
<td>Intermediate</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>27</td>
<td>F</td>
<td>21</td>
<td>Intermediate</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>28</td>
<td>F</td>
<td>21</td>
<td>Advanced</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>29</td>
<td>F</td>
<td>21</td>
<td>Intermediate</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>30</td>
<td>F</td>
<td>24</td>
<td>Intermediate</td>
<td>6</td>
<td>7</td>
</tr>
</tbody>
</table>

* For these participants, there was no IELTS score available. Therefore, their ISLPR scores were converted to IELTS scale.

As can be seen in Column 4, in the initial computer-administered self-assessment test, 73% of the participants evaluated themselves as belonging to the Intermediate group. The remaining 27% assessed themselves either at Pre-intermediate (10%) or at the Advanced (17%) levels. This is schematically shown in Figure 4.1.
To investigate the relationship between performance on the self-assessment and SAM Interview, the three categories of Pre-intermediate, Intermediate, and Advanced were given the codes of 1, 2, and 3 respectively. The values (as shown in Figure 4.2) were then compared with the SAM Interview scores. As it is seen, there is a positive relationship between the two measures.

FIGURE 4.1 Self-assessment results

FIGURE 4.2 Scatter plot representing performance on the self-assessment and the SAM Interview
A further cross comparison was carried out between the participants’ performance on the IELTS, SAM Interview and their self-assessment of their own oral proficiency.

The results of this comparison are shown in Figure 4.3. (Note that A, B, and C refer to Advanced, Intermediate, and Pre-intermediate respectively).

The Spearman \( \rho \) correlation coefficient between the test-takers’ performance on the SAM Interview test and on IELTS or ISLPR was found to be 0.63 which is considered a high moderate index of reliability.\(^5\) This method of calculating

\(^5\) Depending on the purpose and function of different tests, there are different interpretations of the acceptable level of the correlation coefficient used as indicators of reliability or validity of a test. The degree of reliability needed in a measure depends to a great extent on the use that is to be made of the results. If the measurement results are to be used for making a decision about a group or even for research purposes, a lower reliability coefficient (in the range of .50 to .60) might be acceptable. But if the results are to be used as a basis for making important (and irreversible) decisions about individuals, only instruments with the highest reliability are acceptable. Nevertheless, it is a generally accepted
correlation is acceptable if data are available only in ordinal form (as is the case in this study), or if the number of paired rankings is more than 9 or fewer than 30 with not more than a few ties in ranks (Mousavi, 2002). The results of the correlation are displayed in Table 4.2.

<table>
<thead>
<tr>
<th></th>
<th>SAM Interview</th>
<th>IELTS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Spearman’s rho</strong></td>
<td>1.000</td>
<td>.631(**)</td>
</tr>
<tr>
<td><strong>SAM Correlation Coefficient</strong></td>
<td>.</td>
<td>.003</td>
</tr>
<tr>
<td><strong>N</strong></td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td><strong>IELTS Correlation Coefficient</strong></td>
<td>.631</td>
<td>1.000</td>
</tr>
<tr>
<td><strong>Sig. (2-tailed)</strong></td>
<td>.003</td>
<td>.</td>
</tr>
<tr>
<td><strong>N</strong></td>
<td>30</td>
<td>30</td>
</tr>
</tbody>
</table>

**Correlation is significant at the 0.01 level.**

Figure 4.4 shows the scatter plot for the SAM Interview and the IELTS test scores. The Figure shows that the scores are spread out in a positive direction indicating a relationship between the results of the two tests.

![Figure 4.4 Scatter plot for the SAM Interview and the IELTS results](image)

tradition that the reliability coefficients below .50 are considered low, .50 to .75 are considered moderate, and .75 to .90 or above are considered high. (Mousavi, 2002)
The numerical data for Figure 4.4 is reported in Table 4.3 below. This table reports on the subjects’ ranked scores of the SAM Interview and their corresponding IELTS scores.

**Table 4.3 Ranked comparisons of testees’ SAM Interview scores with their IELTS scores**

<table>
<thead>
<tr>
<th>Student No.</th>
<th>SAM Interview</th>
<th>IELTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>21</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>10</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>11</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>12</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>14</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>15</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>17</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>19</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>20</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>22</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>23</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>26</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>29</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>30</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>13</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>18</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>24</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>27</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>16</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>25</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>28</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>7</td>
</tr>
</tbody>
</table>

As can be seen in the above table, the participants who scored low on the SAM Interview test, also scored relatively low on the IELTS test and vice versa.

When the scores of the SAM Interview and the IELTS are displayed on a line graph for easier comparison (Figure 4.5), it becomes clear that very few students scored at each end of the distribution on the SAM Interview. Only one participant scored 4 on
the SAM Interview, and one subject had a score of 9 on the IELTS and the SAM Interview, though not the same subject in each case. However, most of the participants scored 5, 6, and 7 on the IELTS test and a large number of them scored 6 on the SAM Interview.

![Figure 4.5](image)

**Figure 4.5 A comparison of results from the IELTS and the SAM Interview**

<table>
<thead>
<tr>
<th></th>
<th>Range</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAM Interview</td>
<td>5.00</td>
<td>4.00</td>
<td>9.00</td>
<td>6.37</td>
<td>0.9994</td>
</tr>
<tr>
<td>IELTS</td>
<td>4.00</td>
<td>5.00</td>
<td>9.00</td>
<td>6.30</td>
<td>1.022</td>
</tr>
</tbody>
</table>

As shown in the table, the mean performance on the SAM Interview was 6.37 (standard deviation (SD) = 0.99) and that of the IELTS was 6.30 (SD = 1.02). To determine significance of the differences between the two means a student t-test was run. The results of a t-test indicate a t-observed value of 0.79. The t-critical on a t-table with a degree of freedom of 1 (df = 2 - 1 = 1) indicates that a value of at least 3.078 or 6.314 at the 0.01 and 0.05 levels of significance respectively is needed to reject the null-hypothesis of no significant differences between the two means. The
results therefore strongly support the comparability of the participants’ performance on the two tests.

4.2 Post CBT Questionnaire Results

In this section, results obtained from the administration of the post-CBT questionnaire are analysed. This data collection procedure was carried out through the use of a paper-based questionnaire administered to the participants immediately after they had taken the SAM Interview computer-based test. This questionnaire consisted of two sections, a fifteen 1-5 Likert type items and a single open-ended question. The purpose of the Likert type items was to elicit the participants’ reactions to and attitudes towards the CBT compared to their previous face-to-face interview experience on the IELTS test or the ISLPR. The participants were asked to read the items and indicate their extent of agreement or disagreement with a series of statements. These statements included the interface design of the test, the ease of administration, the test-takers’ like or dislike of the computer-based testing experience and attitudes, their level of anxiety during the test, the quality of the software package, and their perceived difficulty level of the test. The fifteen Likert type items had been designed to tap the students’ sensitivities and elicit their responses in relation to the three categories of a) Anxiety; b) Difficulty Level; and c) User Interface, hypothesised in the research questions. Breakdowns of the test-taker responses to this section of the questionnaire are shown in the following 15 graphs (graphs 4.6-4.20). These results are then analysed quantitatively using a series of descriptive and inferential statistical tests and are reported in section 4.2. (Note: SA = Strongly Agree; A = Agree; N = Neutral; D = Disagree; and SD = Strongly Disagree).
Responses given to the single open-ended item seeking the test-takers’ additional comments and feedback about the computer-based testing experience are analysed qualitatively and reported in section 4.2.4 of this chapter. The inclusion of this item in the questionnaire is justified by the fact that it provides the participants with an opportunity to comment on any aspect of the CBT testing experience which may not have been accounted for in the preceding Likert type items. This single item read: “If you have any other comment, please write it in the following box.”

4.2.1 Descriptive and Inferential Statistics (Multimedia and Anxiety)

Table 4.6 summarizes the descriptive statistics for the multimedia effect (independent variable) on the anxiety level (dependent variable) for the participants taking the SAM Interview test. The influence of the multimedia on anxiety levels was hypothesised in items 9, 11, 14 and 15 of the post-CBT questionnaire. The participants’ responses to these items are shown in Figures 4.6-4.9:

![Figure 4.6: Responses to Item 9](image-url)
Results and Analysis

**Item 11:** I felt more confident communicating with the computer than a face-to-face interview.

![Figure 4.7 Responses to Item 11](image)

**Item 14:** I needed the videos to help me understand the questions.

![Figure 4.8 Responses to Item 14](image)

**Item 15:** I found communicating with the computer instead of a real person was less embarrassing for me.

![Figure 4.9 Responses to Item 15](image)
The above four items yield a maximum score of 20 on a 1-5 Likert scale. As shown in Table 4.5, the observed mean of the reported anxiety for female participants was 14.06 (out of 20) and the standard deviation (SD) was 2.89. The mean score of anxiety perceived by male participants, on the other hand, was found to be 14.33 with an SD of 4.03.

<table>
<thead>
<tr>
<th>Group Statistics</th>
<th>M/F</th>
<th>N</th>
<th>Mean</th>
<th>Std Deviation</th>
<th>Std. Error Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANXIETY</td>
<td>F</td>
<td>17</td>
<td>14.0666</td>
<td>2.88001</td>
<td>.75513</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>13</td>
<td>14.3333</td>
<td>4.02201</td>
<td>1.33421</td>
</tr>
</tbody>
</table>

The frequency of responses by the male and female participants on the above four items have a different face value. Any such differences in the number of responses could be the result of chance error (error variance) or a true difference (true variance). Therefore, to determine the significance of the difference between the frequency counts of the responses, a Chi-square test was run. If the differences were found to be statistically significant, then it could be assumed that test-taker gender did have an effect on the anxiety level in taking the computer-based test. However, the results show that there is no statistically significant difference between the frequencies of the responses to the items on the Likert type scale. The Chi-square test results are shown in Table 4.6.

<table>
<thead>
<tr>
<th>Chi-square results for multimedia effect and anxiety</th>
</tr>
</thead>
<tbody>
<tr>
<td>ITEM 9</td>
</tr>
<tr>
<td>Chi-square</td>
</tr>
<tr>
<td>df</td>
</tr>
<tr>
<td>Asymp. Sig.</td>
</tr>
</tbody>
</table>

Another potentially important variable in relation to the use of the computer in administering language tests is the age of the test-takers and their perceived anxiety.
levels during the testing session. In the post-CBT questionnaire, the participants were asked to indicate their level of anxiety in relation to the computer-based test.

To investigate the relationship between these two variables, the average age of the testees and their responses to anxiety-related items of the post-CBT questionnaire was determined. The descriptive results are summarised in Table 4.7.

<table>
<thead>
<tr>
<th>TABLE 4.7 Descriptive statistics for age and anxiety levels</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Descriptive Statistics</strong></td>
</tr>
<tr>
<td>Mean</td>
</tr>
<tr>
<td>AGE</td>
</tr>
<tr>
<td>ANXIETY</td>
</tr>
</tbody>
</table>

As seen in the table, the average age of the participants was 26.56 (max = 33; min = 20) and the mean for their anxiety level was 14.17. Table 4.8 shows the results of a Pearson correlation between age and reported anxiety levels and indicate that there is almost no (0.05) relationship between the two variables.

<table>
<thead>
<tr>
<th>TABLE 4.8 Results of the correlation coefficient between age and anxiety levels</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age</strong></td>
</tr>
<tr>
<td>Age</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td>ITELM 9</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td>ITELM 11</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td>ITEM 14</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td>ITEM 15</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
</tr>
<tr>
<td>N</td>
</tr>
</tbody>
</table>

** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).

4.2.2 Descriptive and Inferential Statistics (Multimedia and Task Difficulty)

Table 4.9 summarises the descriptive statistics for the perceived difficulty level of the CBT hypothesised in the post-CBT questionnaire. Several difficulty-related statements were included in the questionnaire and the participants were asked to
indicate their extent of agreement with these notions as contributing factors to the difficulty level of the computer-based test. The participants’ responses to these items are shown in Figures 4.10-4.15:

**Figure 4.10 Responses to Item 6**

**Figure 4.11 Responses to Item 7**
Item 8: I had enough time to answer each of the five questions.

FIGURE 4.12 Responses to Item 8

Item 10: The program gave me enough time to think about my answers to the questions.

FIGURE 4.13 Responses to Item 10

Item 12: The questions were at the right level of difficulty for me.

FIGURE 4.14 Responses to Item 12
Results and Analysis

Table 4.9 summarises the mean differences of the difficulty level of the test perceived by male and female participants.

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Mean</th>
<th>Std Deviation</th>
<th>Std. Error Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIFFICULTY</td>
<td>F</td>
<td>17</td>
<td>22.9333</td>
<td>3.78845</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>13</td>
<td>24.8889</td>
<td>2.26078</td>
</tr>
</tbody>
</table>

The average of the perceived difficulty by female participants was 22.93 (SD = 3.78) and that of the male participants was 24.88 (SD = 2.26). Again, frequencies of responses to the items are different in their face value. To determine the significance of the differences a Chi-square test was run, of which the results are reported in Table 4.10. The results show that there is no statistically significant difference between the response frequencies to the Likert type items at the $\rho < 0.05$ level of confidence. In other words, male and female participants find the CBT and face-to-face interview items equally difficult.

<table>
<thead>
<tr>
<th>ITEM</th>
<th>ITEM 7</th>
<th>ITEM 8</th>
<th>ITEM 10</th>
<th>ITEM 12</th>
<th>ITEM 13</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-square</td>
<td>20.600</td>
<td>10.800</td>
<td>23.333</td>
<td>16.400</td>
<td>18.333</td>
</tr>
<tr>
<td>df</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Asymp. Sig.</td>
<td>.000</td>
<td>.013</td>
<td>.000</td>
<td>.001</td>
<td>.001</td>
</tr>
</tbody>
</table>
The age of the participants was also hypothesized to be an underlying factor in the perceived difficulty level of a CBT. The descriptive statistics of age as an independent variable and perceived difficulty level as dependent variable are displayed in Table 4.11. As seen on this table, the average age was 25.57 (SD = 4.27) and the average perceived difficulty was 23.66 (SD = 3.38).

**TABLE 4.11 Descriptive statistics for age and difficulty level**

<table>
<thead>
<tr>
<th>Descriptive Statistics</th>
<th>Mean</th>
<th>Std Deviation</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGE</td>
<td>25.57</td>
<td>4.27</td>
<td>30</td>
</tr>
<tr>
<td>DIFFICULTY</td>
<td>23.66</td>
<td>3.38</td>
<td>30</td>
</tr>
</tbody>
</table>

To determine the presence (as well as the direction and strength) of the relationship between age and the perceived difficulty level of the CBT, a Pearson correlation was run. The following table illustrates the correlation coefficient indices of each of the items with age and the participants’ perceived level of difficulty:

**TABLE 4.12 Results of the correlation coefficient between age and difficulty level**

<table>
<thead>
<tr>
<th>Age</th>
<th>ITEM 6</th>
<th>ITEM 7</th>
<th>ITEM 8</th>
<th>ITEM 10</th>
<th>ITEM 12</th>
<th>ITEM 13</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pearson Correlation</td>
<td>.269</td>
<td>.357</td>
<td>-.063</td>
<td>-.367(*)</td>
<td>-.526(*)</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>.151</td>
<td>.053</td>
<td>.742</td>
<td>.046</td>
<td>.003</td>
</tr>
<tr>
<td>N</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
</tr>
</tbody>
</table>

| ITEM6 | Pearson Correlation | -.269  | 1      | .295   | .325    | .718(**)| .289    | .383(*)|
|       | Sig. (2-tailed)     | .151   | .114   | .080   | .000    | .122    | .037    |
| N     | 30     | 30     | 30     | 30      | 30      | 30      |

| ITEM7 | Pearson Correlation | .357   | .295   | 1      | .256    | .396(*)| .450(*)|
|       | Sig. (2-tailed)     | .053   | .114   | .172   | .030    | .791    | .811    |
| N     | 30     | 30     | 30     | 30      | 30      | 30      |

| ITEM8 | Pearson Correlation | -.063  | .325   | .256   | 1       | .553(**)| .108    | -.008  |
|       | Sig. (2-tailed)     | .742   | .080   | .172   | .002    | .572    | .967    |
| N     | 30     | 30     | 30     | 30      | 30      | 30      |

| ITEM10| Pearson Correlation| -.367(*)| .718(**)| .396(*)| .553(**)| 1      | .304    | .258   |
|       | Sig. (2-tailed)     | .046   | .000   | .030   | .002    | .103    | .168    |
| N     | 30     | 30     | 30     | 30      | 30      | 30      |

| ITEM12| Pearson Correlation| -.526(**)| .289   | -.050  | .108    | .304   | 1       | .454(*)|
|       | Sig. (2-tailed)     | .003   | .122   | .791   | .572    | .103    | .012    |
| N     | 30     | 30     | 30     | 30      | 30      | 30      |

| ITEM13| Pearson Correlation| -.329  | .383(*)| .046   | .000    | .258   | .454(*)| 1      |
|       | Sig. (2-tailed)     | .076   | .037   | .711   | .967    | .168   | .012    |
| N     | 30     | 30     | 30     | 30      | 30      | 30      |

* Correlation is significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed).
4.2.3 Descriptive and Inferential Statistics (Gender, Age and User Interface Design)

In addition to each of the above factors as variables in the usefulness of the test, the user interface design was also considered to be a potentially influential factor contributing to the overall quality and usability of a new CBT. The potential effect of the user interface of the new software package was hypothesised in items 1, 2, 3, 4, and 5 on the post-CBT questionnaire. The participants’ responses to these items are shown in Figures 4.16-4.20:

**FIGURE 4.16 Responses to Item 1**
Results and Analysis

Item 2: I need to know a lot about computers to use this program.

FIGURE 4.17 Responses to Item 2

Item 3: The design of the program was confusing for me.

FIGURE 4.18 Responses to Item 3

Item 4: I spent too much time typing my personal details.

FIGURE 4.19 Responses to Item 4
Results and Analysis

Table 4.13 reports on the descriptive statistics for the effect of user interface design on the performance of the participants on the SAM Interview. It demonstrates the mean differences of the male and female participants’ ideas on the user interface related statements. As seen on the table, the average for the male participants was 25.16 (SD = 2.73) and that of the female testees was 26.16 (SD = 1.80).

<table>
<thead>
<tr>
<th>Group Statistics</th>
<th>M/F</th>
<th>N</th>
<th>Mean</th>
<th>Std Deviation</th>
<th>Std. Error Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTERFACE</td>
<td>F</td>
<td>17</td>
<td>25.16</td>
<td>2.73078</td>
<td>.70508</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>13</td>
<td>26.16</td>
<td>1.80278</td>
<td>.60093</td>
</tr>
</tbody>
</table>

To determine whether the differences in the frequency of responses to the post-CBT questionnaire resulted from chance variance or gender difference in relation to perceptions of the user interface design, a Chi-square was run. Analysis of the results indicates that there is no statistically significant differences between the frequency count of the responses at the $\rho<0.05$ significance level. In other words, male and female participants had comparable ideas about the interface design of the CBT test.
Finally, the relationship between the age of the participants and the nature of their attitudes towards the user interface design and usability of the CBT was also investigated. It is commonly thought that older individuals have less favourable attitudes towards the new technology-based instruction and testing. Table 4.15 summarises the basic statistics of the age of the participants and their attitudes in relation to the user interface design of the CBT. As shown on the table, the average age was 26.58 (SD = 4.15) and the average perceived user-friendliness of the user interface was 26.57 (SD = 4.27).

<table>
<thead>
<tr>
<th>TABLE 4.15 Descriptive statistics for age and user interface design</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Descriptive Statistics</strong></td>
</tr>
<tr>
<td>Mean</td>
</tr>
<tr>
<td>AGE</td>
</tr>
<tr>
<td>INTERFACE</td>
</tr>
</tbody>
</table>

To investigate the relationship (and its strength and direction) between the age as one variable and favourable attitude towards the CBT as another variable, a correlation analysis was conducted whose results are shown in Table 4.16.

<table>
<thead>
<tr>
<th>TABLE 4.16 Results of the correlation coefficient between age and approval of user interface design</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
</tr>
<tr>
<td>Pearson Correlation</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td>ITEM 6</td>
</tr>
<tr>
<td>Pearson Correlation</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td>ITEM 7</td>
</tr>
<tr>
<td>Pearson Correlation</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td>ITEM 8</td>
</tr>
<tr>
<td>Pearson Correlation</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td>ITEM 10</td>
</tr>
<tr>
<td>Pearson Correlation</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
</tr>
</tbody>
</table>
Results and Analysis

<table>
<thead>
<tr>
<th>ITEM 12</th>
<th>Pearson Correlation</th>
<th>Sig. (2-tailed)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-.526(**)*</td>
<td>.003</td>
</tr>
<tr>
<td></td>
<td>.289</td>
<td>.122</td>
</tr>
<tr>
<td></td>
<td>-.050</td>
<td>.791</td>
</tr>
<tr>
<td></td>
<td>.108</td>
<td>.572</td>
</tr>
<tr>
<td></td>
<td>.304</td>
<td>.103</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>.012</td>
</tr>
<tr>
<td></td>
<td>.454(*)</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>30</td>
<td></td>
<td>30</td>
</tr>
<tr>
<td>30</td>
<td></td>
<td>30</td>
</tr>
<tr>
<td>30</td>
<td></td>
<td>30</td>
</tr>
<tr>
<td>30</td>
<td></td>
<td>30</td>
</tr>
<tr>
<td>30</td>
<td></td>
<td>30</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ITEM 13</th>
<th>Pearson Correlation</th>
<th>Sig. (2-tailed)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-.329</td>
<td>.076</td>
</tr>
<tr>
<td></td>
<td>.383(*)</td>
<td>.037</td>
</tr>
<tr>
<td></td>
<td>.046</td>
<td>.811</td>
</tr>
<tr>
<td></td>
<td>-.008</td>
<td>.967</td>
</tr>
<tr>
<td></td>
<td>.258</td>
<td>.168</td>
</tr>
<tr>
<td></td>
<td>.454(*)</td>
<td>.012</td>
</tr>
<tr>
<td>N</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>30</td>
<td></td>
<td>30</td>
</tr>
<tr>
<td>30</td>
<td></td>
<td>30</td>
</tr>
<tr>
<td>30</td>
<td></td>
<td>30</td>
</tr>
<tr>
<td>30</td>
<td></td>
<td>30</td>
</tr>
</tbody>
</table>

** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).

4.2.4 Descriptive Comments

The last item in the questionnaire was an open-ended question to which students were given the opportunity to reflect upon their experience with the test and indicate their like/dislike with any aspect of the test. The fifteen Likert type items of the questionnaire were designed to investigate the candidates’ reactions to the three notions of multimedia effect and anxiety, task difficulty, and interface design preferences while the single open-ended item was included in the questionnaire so that the test-takers could have the opportunity to comment on any aspect of the test which the researcher might have failed to include in the preceding items. The descriptive comments given by the test-takers can be viewed as complementary evidence to support the validity and usefulness of the Sam Interview test. The analysis of the comments revealed that in a majority of the cases the reactions were positive and in support of the new mode of presentation. Of special interest and significance was the potential washback effect of the new CBT on the students’ learning of not only the way the test was administered but also the content of the test. For instance, some of the positive comments they made were:

“nice experience”

“The program is really a creative way to test English ability.”

“Well done job.”

“Overall, it was a good test.”
“The questionnaire design is basically satisfactory with me.”

“It is more enjoyable than a face-to-face speaking test.”

Some of the students, however, reported their concerns as follows:

“It seems to be hard for me because I am not used to talk with myself or with computer”

“Sometimes, it is weird to talk to the machine specially for testing.”

4.3 Analysis of Test-taker Language Samples

In this section, transcripts of the recorded responses of nine examinees — three from each level — are presented to illustrate the language features they used while they took the SAM Interview. Each response is preceded with the corresponding task question.

4.3.1 Pre-intermediate

TAST 3: Why do you think smoking is bad for your health? Give us as many reasons as you can?

EXAMINEE A: I think there are there are a lot of disadvantage and dangerous of dangerous of smoking... and I can address that in some point... First point, it is a bad behaviour, second point the smoking case a lot of health diseases, third point it is costing money, fourth point its give you it is give a bad ..eh... smell, and ...eh....

Personally, I hit smoking and I am always thinking that that smoker who smoke how can the smoker kiss his wife and he has a bad smell. (Total time: 81 seconds)
Results and Analysis

TASK 8: Most people in Australia live near the coast where it is usually green. Do you have four seasons in your country? Can you describe them?

EXAMINEE B: *Actually in my country we don’t have four seasons. In my country we have only three seasons. Those are summer, winter and autumn. The summer semester actually start from January up to November followed by winter seasons which eh... which eh... usually start from December and it is finished in March where the autumn season is in the military between them.* (Total time: 61 seconds)

TASK 7: When you were a child, did you ever dream of your future job? What did you want to be and why?

EXAMINEE C: *When I was a child. I wanted to become policeman because they looked strong and kind, so policemen always fight with serious crime so they are they are my hero. When I was a junior high school student, I gave up becoming policeman because I knew its work is very tough because, time is, how can I say..., suddenly someone attack attacks them and twenty four hours they might be called by someone so I think I thought they couldn’t sleep well. That is why I gave up becoming a policeman.* (Total time: 81 seconds)

The above language samples include features which demonstrate language proficiency corresponding to Level 2 on the ISLPR scale. First of all, the most noticeable feature in all the above three samples is the length of the talk. The examinees were allowed up to three minutes to respond on each task. However, the examinees at this level managed to produce an open-ended response of just over one minute for each task. The responses to the tasks at this level involve a great deal of prolonged pauses, delays and hesitations, grammatically incomplete sentences (e.g., Examinee A: …it is
Results and Analysis

give a bad ..eh... smell...), as well as inaccurate vocabulary, lack of register flexibility, incorrect use of part of speech (e.g., Examinee A: dangerous for dangers). Examinee A, for instance, struggles to express his/her opinion about the disadvantages of smoking and seems to lack the required and appropriate vocabulary. This examinee also uses inappropriate collocations (health diseases), and incorrect grammar (…third point it is costing money...).

In the language samples produced by the Level 2 (Pre-intermediate) examinees, there are significant hesitations which indicate that the examinees lack sufficient control over the discourse to maintain an effective connection of meanings throughout their response. In addition to the prolonged delays, there are also inappropriate repetitions of words (e.g., Examinee C: … suddenly someone attack attacks them…) as well as fillers and noises (e.g., Examinee A: …it is give a bad ..eh... smell, and ...eh... ) or even words that are barely comprehensible (…season is in the [military] between them).

The analysis of the responses to Pre-intermediate Level tasks indicates that the examinees at this level may be able to participate in and sustain informal and simple social conversations. They are also able to talk about their past experience (e.g., Examinee C: When I was a child. I wanted to become policeman because they looked strong and kind...). But their responses lack sophisticated or abstract vocabulary. In terms of grammar, as is seen in the above transcripts, the language produced by the Pre-intermediate Level examinees does not involve any complex or less familiar structures, a typical feature of Level 2 on the ISLPR scale. The language produced by Examinee C, for example, may cause confusion in the listener because it lacks
sufficient cohesion (e.g., …suddenly someone attack attacks them and twenty four hours they might be called by someone so I think I thought they couldn’t sleep well…). All the above features displayed by the speakers at this level influenced their fluency and hence the amount of language they produced in response to the task.

**4.3.2 Intermediate**

**TASK 17:** Language teachers believe that socializing helps us improve our English ability and communication skills. To what extent do you agree with this?

**EXAMINEE A:** Ah... I really agree, it doesn’t help a hundred percent, I mean if you don’t know a language, just with socializing it doesn’t make you to know that language but it helps. I think if you go to classes and learned grammar and basic things of that language and then go around and go to the public places and socialize with people, it helps just the verbal language not exactly the grammar, or the basic things of the languages but it helps to learn the verbal things that people say in the street, and your listening will be improved. I really believe in that. But I cannot say it will improve your language because language is not just the listening and speaking. I have seen a lot of people here. They’ve been in Australia may be twenty years, they’ve been outside, they can go shopping, they can do everything, but to me their English is not good at all, it is not perfect and it is not good. So they learn English speaking, understand people, and then can make themselves understood, but they have a lot of mistakes, and no one in the street tells them that this sentence wasn’t right, this verb was wrong, or something like that. Ah.. they just say something and people understand them and they understand people and that is it. So it helps a lot for the verbal language but not for the grammar. And you learn some words that people use in the street in general conversations but when you start reading
things, you see how weak you are in that language because you don’t learn a lot of words in the streets. You just learn some words just general, and and I think its you need it, you have to do it both, you have to learn language through the books and through the grammar to the basic things of the language and in the streets with socializing with the people. You have to have it both to learn a language. And you can call that, when I can call, when I can say I know this language I have to learn it both.

(Total time: 102 seconds)

TASK 18: Some people believe that preserving and supporting wildlife is very important. To what extent do you agree with this?

EXAMINEE B: I personally strongly agree, eh with that kind of attitude. Why? Because we need the wildlife for our environment and actually the other thing is they are entitled to live. They can live in our environment but unfortunately we jut not trying but mainly we destroyed them, we actually killed them, we sometimes eat them, even if we don’t like to eat them, just for fun or you know entertainment we kill them. We hunt them whatever the style is... eh and I believe that is our, you know, eh... these are the human responsibilities, you have to take care of them, we have to support them, we have to look after them, and try to prevent to destroy the.. eh... wildlife, even whether actually they are coming from the ocean the.. you know the sea life, the sea wildlife or you know the wildlife actually the other wildlife, you have to care of them. (Total time: 95 seconds)

TASK 19: Travelling has become quite easy these days. What forms of transport do you enjoy most when you travel and why?

EXAMINEE C: I prefer travelling with the airplane because travelling with airplane is the safest way of travelling, and it is very fast, it doesn’t waste your time, and you
can easily and fast reach your destination. And also, you can enjoy the time that you are in airplane, the flight attendants serve you with food and drink and sometimes in some flights you can see movie and even play game with the computer or with somebody else. And it is very easy, and fast and don’t waste your time. And I prefer that. But some other transportational means, for example ships and bus sometimes they are more enjoyable but they take time to travel and they are not fast; sometimes when you are travelling with a ship, the ship movement and the rapid movement of the sea and waves maybe make you sick and you don’t feel very good. And another transport by bus, it is dangerous. The road accidents are much more than airplane accidents and sometimes their seats are not very good, are not very comfortable, and usually they don’t serve anything for you when you are travelling by bus. And although it’s a bit cheaper than airplane, but it’s so slow and it takes long time to reach your destination. And travelling by car has some advantages but sometimes it is not cheap because you consume a lot of petrol and with this high petrol prices it would be a expensive way to travel by your private car. And therefore I prefer to travel by airplane that is cheaper, safer, and faster. (Total time: 180 seconds)

The language samples produced by the examinees at this level display features which correspond to Level 3 on the ISLPR. As can be seen in the above transcripts, the examinees were able to produce longer talks than those in the Pre-intermediate level during the available three minutes of the response time. Examinee C, for instance, used the whole period of three minutes to sustain the connected discourse of his/her narrative. In addition, topics such as ‘socializing as a means of language learning, supporting wildlife,’ and ‘transportation’ are relatively confidently managed by the examinees. There is sufficient evidence that the examinees at this level were able to
express argumentations in accurate and precise language (e.g., Examinee A: ...I have seen a lot of people here. They’ve been in Australia maybe twenty years, they’ve been outside, they can go shopping, they can do everything, but to me their English is not good at all...). One of the features of performance at this level is that the examinee is able to produce a relatively long narrative or descriptive monologues (Examinees A and C). However, in some cases there is evidence of the influence of the examinee’s mother tongue (e.g., Examinee C: ... when you are travelling with a ship ...); however, this does not hinder successful communication. The examinees employed a range of vocabulary and collocations which assisted them to produce a more fluent language sample (e.g., public places, entertainment, human responsibilities, support, flight attendant, road accidents, destination). All the above examinees were able to make use of high-frequency idioms and colloquial forms such as ‘I really agree, I personally strongly agree, they are entitled to live, just for fun.’ Examinee A was able to elaborate his/her emotional (... I really believe in that...) and intellectual (...but when you start reading things, you see how weak you are in that language...) attitudes. There are also instances of the use of passive structures in this level (e.g., Examinee B: ...they are entitled to live...).

4.3.3 Advanced

TASK 27: Scientists have made a great contribution to improving the quality of human life. Can you give an example with details?

EXAMINEE A: I think the discovery of Penicillin perhaps was the greatest discovery in terms of eh... medical science. And I think have been great discoveries since then like test tube babies for instance, I think that was a major discovery especially for those couples or families who would like to have babies naturally but they cannot. So
that has brought about major drastic changes for their life and I think that also
brought happiness for them in terms of being able to actually get children and have a
family which they have always liked. And there are other things like say for instance
artificial limbs, I think that also is a major achievement. Because eh.. say for instance
if the fact that now you can actually have artificial limbs. Say, for instance, if it was
half a century ago or maybe before that, the scientists had not actually found that.
People who actually have accidents, or even young children step on mines or say
things like that, they would for their whole life be handicapped and would never be
able to enjoy their life to the fullest. So I believe the fact that scientists have been able
to artificially give these children and people limbs have been a great achievement and
I think is a great achievement for the science and also for the whole world because if
not for that they’d be minus one or many limbs and that would mean so many things
for the people. That could be associated with mental problems, psychological
problems and because even if it is just one person in the family who have had the
problem, it could also mean so many social issues for the family, for the society. And
the society would not be able to get the full benefit of the person as an individual
because of that, But now because science has actually been able to give artificial
limbs to people who have lost their limbs means that they are able to fully engage in
life, and enjoy their life to the fullest. So science, it is amazing what it can do and I
think, eh, however, I think there are some ethical issues when it comes to science
especially the way things are going now, for instance, detecting genetical disorders in
the foetal in the foetus before they are born, and discarding the foetus if it has got
genetic deformations. And so there are pros and cons to it but on the whole I think it
is a positive experience and people are benefitting from the science. (Total time: 180
seconds)
TASK 26: In many big cities with a lot of high rise and modern buildings, you also see some very well protected historical places. Now, think of your own country; why do you think it’s important to maintain the historical places of a country?

EXAMINEE B: This is an interesting question because only yesterday I was talking to a friend and telling her about the history of my country and how we used to be a Buddhist country but now we are a hundred percent Muslim country, and how some of the ancient temples that people had used as places of worship are no longer there, and how it would have been better had the government or somebody at least done something to preserve these places because history has got so many lessons for us and it can us so many things about what had happened in the past and so much about our ancestors, and there are lessons for us to learn from those things and it is not for the lessons, there is also the talent and the skill and the passions of the of our ancestors as well because when I talk about the Buddhist temples that we used to have, there were such beautiful craft that had gone into the making the whole temple and there are some remains of the temple now in the museums. And even looking at those things you can see how skilled the people were and how good they were at using the things that were available at that time because obviously now these days you can use machinery and it only take minutes to do something as beautiful as they had done previously and it would have taken them a lot of time, hardship and work, effort to actually do that so apart from the historical, the meaning of what they have actually done, there is also the concept of the skill and things like that that have gone into building those historical places. So I think it is really really important that we preserve them because there is a lot of information there in terms of what they were able to do as individuals, how creative they were,
how intelligent they were, apart from the actual story that comes with the historical places. So it is indeed very sad that these days people are not paying that much attention to it because there is so much to learn from the history. And I for one and somebody who really really likes visiting historical places because that tells a lot about the place, what had happened in the past, because the past obviously has an influence on the future and it tells so many stories about what had happened. So it is, it is important that we preserve historical places and maintain them in a manner that would make sure that would ensure that places are there for the next generation and so on. (Total time: 180 seconds)

TASK 24: This video shows some aspects of life in a modern city. There are advantages and limitations associated with life in big cities. How do you compare it with life in rural areas?

EXAMINEE C: Well, this is an interesting question, because I am a, I am a city person, so I don’t know I am gonna be somewhat biased weighing out the pros and cons of the city life and the rural life so I mean in the country in rural areas you have a lot of free times and you don’t have many options, life is just like that, you have to take it, like it or not. But in the city you have, you have more options, you have more ways of living, you can have different jobs, there are difference range of jobs available in the city, but in the farmers, mainly farming and in the rural areas it is mainly farming… you know, just a limited number of jobs but on the other hand city life can be very stressing and very bushy, people are always working in the city and I guess you feel the scarcity of resources in the city much more than what you would in rural areas because the land, the resources are limited, and there is a lot more competition in the city. People are always working hard and trying to make money
and investing on their future, nothing is for certain, and nothing is guaranteed, but in rural areas you could have like a huge piece of land and kick back and relax in the evenings and work during the day, and be on your farm. And I think there is a lot more tranquillity in rural areas, and I guess people could be more friendly in the country, but city life has its pros, like I said you have a lot of options, you can buy many many stuff, you can choose to live in many different places, you meet many different people, from different countries, from different walks of life, who do, who are totally different from each other, but in the country... ehm... people are all the same and they own the same stuff and they all have the same jobs. I am, I am a city person, I definitely go for a city because it has a lot of advantages, you have a lot more facility, everything is more modern and up-to-date in the city. (Total time: 146 seconds).

The responses produced by the examinees at this level displayed their native-like language ability as well as a profound understanding of the target culture. First of all, the tasks at advanced level had been designed in such a way that they required analytical thinking and the ability to present persuasive argumentation. Apart from the pure linguistic knowledge, topics such as ‘the contribution of modern science to the quality of human life’ clearly require critical thinking, world knowledge, socio-cultural information about the target language and the ability to form and maintain a sustained discourse. The language samples recorded from the advanced level examinees demonstrate the extreme end of the second language developmental continuum for the speaking ability. The samples feature all the qualities of the language spoken by native-speaker. For example, the examinees at this level were very fluent and displayed very few unnatural pauses in their responses. They were
able to handle subtle, complex, and sophisticated concepts such as ‘comparison and contrast of rural and urban lifestyles’ or ‘the protection of historical places.’ In addition to the native-like fluency, the examinees displayed a wide range of vocabulary and comfortable use of English collocations (e.g., Examinee A: …test tube babies, major drastic changes, major achievement, artificial limbs, mental problems, psychological problems, genetic deformations, positive experience, or Examinee B: remains of the temple, Examinee C: pros and cons, kick back and relax, tranquillity in rural areas, the scarcity of resources, from different walks of life …).

The above samples reveal sufficient evidence that these test-takers are able to comfortably handle a smooth discussion about a wide range of complex and abstract topics. In terms of fluency, grammatical accuracy, the use of colloquial language, collocations, fillers, cultural knowledge and all other features, it is difficult to distinguish the examinees at this level from the English native speakers. For example, Examinee A displays a profound knowledge of the medical register and related issues such as genetic deformations, test tube babies, artificial limbs and the contribution of science to the improvement of the quality of human life supported by experiential evidence. Examinee C, on the other hand, demonstrates a comfortable and very fluent use of sophisticated grammatical structures such as … I am gonna be somewhat biased weighing out the pros and cons of the city life… . Listening to the above samples, however, it was possible for the rater to spot traces of slight foreign accent, but these influences were nowhere near enough to affect the successful and effective communication.
CHAPTER 5: DISCUSSION

This chapter discusses the results presented in the previous chapter in light of the research questions and their relevant hypotheses. The research questions are grouped under four headings which deal with the principal concerns of the study. They are:

a) Questions 1 and 2 dealing with the comparability of scores obtained from the administration of the CBT and those from a conventional face-to-face test;
b) Questions 3, 4, and 5 investigating effects of the computer-based test and its components on test performance and rating accuracy;
c) Question 6 concerning the task difficulty in the CBT and the relevant factors affecting this;
d) Questions 7 and 8 exploring the attitudes and anxiety levels of students in relation to taking a CBT.

5.1 Score Comparability and Self-assessment

5.1.1 Research Question 1

Are CBT and face-to-face speaking test scores comparable? Are the tests measuring the same attribute?

The first and most important research question was concerned with the fundamental issue of whether scores obtained from the administration of a computer-based speaking test was comparable (i.e., equally reliable) to those obtained from the conventional face-to-face assessments. Computers have recently been used in the
testing of the speaking ability (Kenyon, Malabonga & Carpenter, 2001). However, the use of digital videos in delivering the tasks in computer-based tests of speaking is a novel development. For the purposes of this research, a software package was designed, developed, and administered along with a conventional test. The above research question also asked whether the two tests actually measured the same underlying attribute. This would provide the researcher with evidence in support of the validity of the new instrument. The study tested the performance of a group of students on both instruments, namely the SAM Interview CBT and an external, independent, and valid criterion, that is the IELTS or the ISLPR tests (see Table 4.1). The study is trying to present criterion-related validity evidence in justifying the usefulness of the new CBT in light of the framework proposed by Bachman and Palmer (1996).

This research is correlational in nature, and the findings are employed to investigate construct validity of the new instrument. The correlation coefficient between the SAM Interview and the IELTS (or converted ISLPR) scores was found to be 0.63 which is considered a high moderate index of reliability (see Table 4.2, p. 147). As Bachman and Palmer (1996, p. 18) stated in their third principle of test usefulness “test usefulness and the appropriate balance among the different qualities cannot be prescribed in general, but must be determined for each specific testing situation.” Given the fact that the SAM Interview is a totally new computer program designed and developed from scratch by the researcher, and the fact that it is a prototype for future development and expansion, within its context, it offers a potential to function as an independent test of speaking for selection and placement purposes at the Australian tertiary institutions.
In relation to Research Question 1, it was hypothesised (Research Hypothesis 1) that the SAM Interview and the IELTS test results would be equally reliable and that the outcomes would reflect the assessment of the same attribute, i.e., speaking trait. In other words, the two tests would yield results with similar rankings (see Table 4.3). The results indicate that those test-takers who did well on the CBT did almost equally well on IELTS or ISLPR tests. Given that the comparability of scores is difficult to achieve, this finding supports much of what is discussed in the Literature Review section of this thesis highlighting the importance and usefulness of CBTs. It also highlights the usability of the newly developed prototype test of speaking using audio and video as alternative modes of test delivery. Although the inclusion of audio and video in a computer-based test adds practical considerations to the challenge of any comparability study (Wang & Kolen, 2001), an obvious achievement in developing this prototype, however, is the utilization of the maximum potential of the new generation of computers for test delivery purposes. The inclusion of videos in the test made it possible to create tasks which approximate the real-life situations using the new technology, rendering the test more authentic.

However, the comparison of the SAM Interview results with those from the IELTS revealed that the new SAM Interview instrument was less sensitive to very low or very high level performances than the IELTS test, but the overall results are comparable. In other words, ratings of the students’ speech samples in the SAM Interview tended to cluster around the centre of the distribution (see Figure 4.4). For example, when we compare the results of the SAM Interview with those of the IELTS, we find that 17 students were assigned the score of 6 in the SAM Interview and very few above or below this. While in the case of IELTS, the number of students
scoring 5, 6, and 7 were almost equally distributed. We can conclude from this observation that the SAM Interview process is probably not very sensitive to low or high abilities and therefore tends to cluster examinees in the middle of the range of scores resulting in a leptokurtic distribution. This was also evident in the differences between the standard deviations of the two tests (SAM Interview SD = .99, and IELTS SD = 1.02)

Research indicates that as long as computer-based test scores are reported on the same scale as the conventional paper and pencil tests scores, comparability between these two types of scores must be established to an acceptable level (Wang & Kolen, 2001). According to Kolen (1999-2000) cited in Wang & Kolen (2001) there are four aspects of differences between computer-based tests and their paper and pencil counterparts which may result in incomparability of the scores. These aspects include: “(1) differences in test questions, (2) differences in test scoring, (3) differences in test conditions and (4) differences in examinee groups (Wang & Kolen, 2001, p. 21). As to the test questions of the SAM Interview, they reflect the same domains as those used in the IELTS test. The tasks on the SAM Interview reflect the target language use situations of university studies in Australia. The scoring methods used the same scale, which made comparability easier and more meaningful. Test conditions for the two tests were different, however, in the sense that the IELTS was a group test and was administered under more formal conditions, but the SAM Interview was an individual test administered in less formal settings. The examinee groups were the same and there were no missing data. Overall, there are sufficient grounds to claim the comparability of the two sets of scores which confirms the usability of the SAM Interview prototype as a test delivery instrument.
5.1.2 Research Question 2

What additional information can we learn from the provision of self-assessment alongside “normal” assessment, and how consistent are the results of self-assessment with the actual ratings of the language performance?

The second research question was concerned with whether the provision of a pre-CBT self-assessment questionnaire would give the researcher opportunities to obtain any additional information in relation to the quality and accuracy of such assessment. First of all, within the context of this research, the main function of providing a self-assessment prior to the actual spoken CBT was to let the computer determine the approximate level of the test-takers’ language proficiency. This is a major function of any computer-adaptive test in that the computer requires some form of initial input from the examinees in order to be able to adapt the level of the selected tasks to that of the testees. In this respect, the SAM Interview is an adaptive test as the examinees are presented with tasks tailored to their ability level according to their own self-assessment outcomes.

In addition to making the test adaptive, the inclusion of a self-assessment is a positive aspect of this project because it increases test-takers’ confidence in their own judgement (Alderson & Banerjee, 2001) while preparing to take the actual test. This area requires more research as the findings in the current study only partially support it. For example as it is shown on Figure 4.7, out the total of 30 test-takers seven of them still reported that they do not feel confident in communicating with the computer. At the same time, four participants strongly disagreed that the communicating with the computer enhances their confidence. On a similar note, as it
is shown on Figure 4.9, more than fifty per cent of the participants reported that they do not feel less embarrassed communicating with the computer. These findings are indicative of the need for more research into the nature of human-computer interaction. This could also have been the test-effect present in any assessment. However, with the availability and use of online communication software programs like Skype (www.skype.com), it is highly likely that the future test-takers will be more confident in communicating with computers.

On yet another positive side of the use of self-assessment as a complementary instrument in second language assessment, it can be argued that self-assessment is thought to promote learning. This can be regarded as evidence of positive washback in language testing (Larson, 2000). At the same time, self-assessment may raise the level of test-takers’ awareness allowing them to take more responsibility in the assessment process (Oscarson, 1989; Ross, 1998).

The researcher’s observations during the test administration and analysis of the captured screens from the tests demonstrated that the test-takers were actively involved in the self-assessment process and considered it as a serious step in the whole test-taking experience. That is, the test-takers were aware that their honest and appropriate performance in the self-assessment would be reflected in the quality and accuracy of their performance on the main section of the test. Besides, verbal assertions made by the testees immediately after the administration of the CBT as well as their qualitative feedback on the open-ended item of the post-CBT questionnaire indicated that they were actively involved in reflecting upon their own language ability when answering the items in the self-assessment. From a technical point of
Discussion

view, as mentioned above, the inclusion of a self-assessment component was necessary for the computer to tailor the tasks to the test-takers so that it could provide them with the most appropriate task as a starting point. Analysis of the results showed that the examinees who assessed themselves as belonging to the advanced level also scored high on the IELTS test (see Figure 4.2), but a large number of the test-takers (see Figure 4.1) believed that they were in the intermediate level (59%). In addition to this, there were only three students who self-assessed themselves as pre-intermediate; these testees also scored low on the IELTS test. These results demonstrated that there is a high degree of positive relationship between the test-takers’ perception of their own ability level and their test results. The second part of this research question was concerned with how consistent the results of the self-assessment were with those found in the actual assessment (i.e., CBT and IELTS or ISLPR tests). In the SAM Interview prototype, a majority of the test-takers assessed their own speaking ability as at intermediate level. This was also evident in the results of the CBT and more or less the same in the results of the IELTS and ISLPR assessments. The findings of this research, therefore, reinforce previous findings that self-assessment is a useful tool in providing a satisfactory approximation of those same abilities as assessed through more objective measures such as tests (Adair-Hauck & Pierce, 1998; Williams, 1992). Arguably, the self-assessment administered prior to the CBT provides the examinees with opportunities to identify their abilities more accurately instead of solely relying on the results of tests. The most obvious positive washback effect of using self-assessment along with normal assessment is that it allows testees to gain self-confidence in judging their own performance. However, a less positive washback effect of the use of self-assessment is that higher ability students may tend to underestimate themselves because they have a sense of all that remains to be learned.
while students who have arrived at a plateau tend to overestimate their ability because, believing they have stopped learning, cannot perceive a need for improvement.

5.2 Multimedia Effect

5.2.1 Research Question 3

Does the multimedia delivery satisfy the interactiveness criterion of test usefulness? Are there significant differences between the perceptions of male and female test-takers in relation to the interactiveness criterion?

The third group of questions in this research were concerned with different aspects of the multimedia effect on students’ test performance. These aspects include a) a facilitating role; b) interface design effect; and c) quality of assessment.

Regarding the facilitating role of the multimedia as an independent variable, and performance on CBTs as a dependent variable, previous research indicates that the mode of presentation of test tasks does have a potential effect on the quality of test-takers’ performance (Carr, Pan, Vongpumivitch & Xi, 2002, Lee, 2004). Comparisons of presentation modes can be more meaningfully made when the two tests have similar questions and scoring scales as was the case with the SAM Interview and the IELTS/ISLPR interviews. However, as McDonald (2002, p. 302) admits, “statistical investigation of equivalence have largely ignored the fact that presenting a test on a computer, a qualitatively different experience is created.” For this reason, it is imperative that researchers also obtain qualitative feedback from the test-takers in order to produce validity evidence in support of the CBTs. In this research, a post-CBT briefing and an open-ended item on the questionnaire were designed to elicit
test-taker reactions in relation to the effect of the multimedia. To analyse examinee reactions to the CBT testing experience, a qualitative approach is used. The use of qualitative approaches in test validation is useful because they provide first hand judgement, introspective, and retrospective verbal reports, and observations on the part of the test-taker.

Furthermore, these approaches help researchers refine the criteria by which they can evaluate the consistency or replicability of their observations and the validity of the inferences they make on the basis of other quantitative analysis (see section 4.2.4). In fact, qualitative approaches are not meant to be a total replacement for their quantitative counterparts, but rather they can be used as complementary devices in validating the test use, highlighting the qualities of test usefulness from the test-takers’ point of view. In the current research, qualitative analysis of the results of the post-CBT feedback questionnaire confirmed that the test-takers had, in general, highly positive attitudes towards the computer administered testing experience. For example, one of the test-takers reported his/her experience as follows:

“I felt much relax when I spoke to computer. I don’t need to smile and just do the way I like to. The contents of the program is easy to understand and very general subjects. The length of the program is OK, not very long, people will not get bored.”

The above description is a first-hand account of how a given test-taker felt during the test. It corresponds to the researcher’s initial assumption that the delivery of a general proficiency test through the computer would be highly efficient, hence meeting the
impact and practicality components of test usefulness criteria. As a matter of fact, the use of the computer in delivering the speaking test allowed the researcher to take full advantage of the multimedia mode and to present the tasks in video and audio format in a more realistic and authentic manner. This potential assists other researchers and test developers to simulate reality in second language testing contexts, thus maximizing authenticity of the assessment and therefore enhancing the overall validity of the test use and at the same time creating a positive washback effect in the test-takers. In relation to the washback effect of the CBT, one of the test-takers commented that:

“*I think this test has a good potential to be used even for practicing for language exams. This could also be a good attraction for students who feel intimidated talking to ‘real’ people especially for exams when they already feel nervous about doing an exam.*”

A word of caution must be added here. Researchers and test developers need to bear in mind that, in addition to the effect of the mode of presentation, reactions to tests might also be the result of individual difference between test-takers. Therefore, these reactions may take many forms at different levels of proficiency and with test-takers from different language and cultural backgrounds. These discrepancies, however, were reflected in only a few of the comments made by the examinees. For instance, two of them displayed some degree of concern which is thought to have resulted from their lack of experience in dealing with computers as mediums of learning or assessment. For example:
"Sometimes, it is weird to talk to the machine specially for testing."

Or,

"It seems to be hard for me because I am not used to talk with myself or with computer."

The discrepancy in the above comments may indicate the influence of individual differences in relation to taking a computerised test. What is certain, however, is that within the next decade or so familiarity with computers will be less of a concern than it is currently, especially when the use of voice recognition devices becomes more popular. Finally a word of caution needs to be added here. The quantitative analysis of the SAM Interview test-takers’ response to the question of whether they felt more confident and less embarrassed with the computer than with a person in a face-to-face interview (see Figures 4.7 and 4.9) reveals that, at the present stage of the use of computers in assessment, the responses are split almost 50/50. This division reflects levels of acceptance concerning speaking to a computer. However, as McDonald (2002, p. 303) claims “…whilst the implicit assumption has been that paper and pencil assessments are the ‘benchmarks’ against which computer-based assessments are equated, this position may be reversed in the near future when computer-based assessment becomes the norm.” The language use situations in which individuals verbally communicate through the computer is expanding. For example, the use of programs like Skype (www.skype.com) in which a person speaks through the computer is becoming increasingly popular.

On another positive note to the use of computer-based speaking tests, research in the area of tape-mediated oral tasks indicates that test-takers tend to find this format more
difficult and/or more stressful than the live interview situations (Clarke, 1985; Stansfield, 1992); a major factor is related to the preparation time (Elder, Iwashita & McNamara, 2002; Hill, 1998) prior to performing on a task. Therefore, the use of the computer-based test allows the researcher to provide extra pre-task planning time to the test-takers to minimize stress, with the aim of reducing the perceived difficulty of the tasks. In addition to the pre-task planning time, Brown (1993) attributed perceived difficulty of the tape-mediated format to a range of different factors including inadequate response time, speed of voices on the tape, lack of clarity in instructions, unclear prompts, excessive input material to process and lack of familiarity with the task type. The design and development of the SAM Interview CBT allowed the researcher to account for all the above features using the capabilities of the computer.

5.2.2 Research Question 4

Do both male and female students find the user interface design useful and effective for the required tasks in the CBT?

Research Question 4 was concerned with the potential effect of the interface design as an independent variable on the performance of male and female test-takers. The scarcity of published literature in relation to the influence of the interface design *per se* on test performance across genders prompted the researcher to include gender differences and interface as variables in the study. The only published article in language testing literature about interface design is that of Fulcher (2003, p. 384) in which he offers a three-phase model for the development of the interface for a CBT and looks at a good interface as a contributing factor to “the mix of validity evidence presented to support the use of a CBT.” There is no published literature in language
testing investigating the effects of the user interface design on the performance of male or female test-takers.

To fill this gap and gain insight into the potential influence of the user interface on the quality and nature of performance on oral CBTs, it was hypothesized that the interface design will be equally conducive for both genders in their reactions to the CBT experience. The quantitative results of the effect of the interface design on male and female students indicated that there was no statistically significant difference between the reactions of the two groups to the CBT experience (see Table 4.16). This finding could be indicative of the fact that male and female students are equally familiar with computers within university contexts. At the same time, the qualitative analysis of the testees’ post-CBT reports indicates that the inclusion of digital videos to reinforce meaning and contextualize the tasks had been highly successful for both groups. The application of digital video in oral second language testing is technologically and methodologically a novel development and has not been well researched. Clark and Hooshmand (1992), for example, conducted a study on the use of modern media techniques in long distance oral testing that preserved the possibility of interaction by using video conferencing. Based on their quantitative analysis, they found that high agreement between both modalities (face-to-face and video teleconferencing) could be achieved. No other study investigated the relationship between oral CBT user interface design and gender differences or performance on such tests.

An added value to the SAM Interview is the simplicity and clarity of its user interface. This feature potentially contributes to the practicality, interactivity and the ultimate validity components of test usefulness, in that it allows the examinees’ attention to be focused on their performance rather than being distracted by factors such as unusual
objects, confusing or ambiguous design, unconventional colour combinations, unnecessary buttons, disturbing flashing lights, low quality audio or video, and the like. Such distracting features may influence the performance of the examinees, invalidating the test use and the interpretations we make on the basis of test scores. During the design stage of the SAM Interview interface, all the above issues were taken into consideration to minimize the effect of such construct-irrelevant factors in inadvertently influencing test-taker performance or invalidating the results. In the analysis of the results no dissatisfaction or undue interference was reported by the test-takers. This indicates that there was no construct-irrelevant factor, at least in the interface design, on test-taker performance in the CBT. The above measures follow the advice offered by Howlett (1996, p. 24) that “clarity of presentation enables users to find what they want to complete the task at hand with the minimum of cognitive load.” The videos in the SAM Interview design were used as a complementary channel of communication to reinforce the meaning and clarity of the tasks. Some of the comments made by the students in this respect are as follows:

“The video clips are great. It is a good idea to help the students to understand the question.”

“The questions are very clear, so in my case the videos are not necessary. There is no confusing part to take this program. Explanations are fantastic.”

“Voice quality was excellent, introduction and explanations were great.”

The above comments can be regarded as first-hand instances of construct validity evidence for the usefulness of the SAM Interview interface. The findings in this study further confirm Fulcher’s view that “interface development and design is extremely
important in computer-based testing, where usability problems may constitute a threat to construct validity” (Fulcher, 2003, p. 384). The clarity of interface design obviously promoted the test administration efficiency so that no student reported any instance of confusion or ambiguity in running the program, understanding the questions, or completing the tasks. The simplicity of the design also obviated the need for any training for the first-time test-takers.

5.2.3 Research Question 5

Does the change in the testing format enhance the practicality of the test and accuracy of proficiency ratings?

This research question was concerned with whether the introduction of a new test delivery format would enhance the overall quality of the test-takers’ performance and the accuracy of rating process. It is assumed that the quality of test-taker performance is evidenced in the expressed examinee satisfaction with a CBT reflected in the post-CBT questionnaire. In addition to this, the comparability of oral CBT and face-to-face test is demonstrated by the correlation between the SAM Interview results and those from an existing external criterion. The argument for the comparability of the results was provided in the discussion under Research Question 1 above. As mentioned earlier in the literature review chapter, instances of the target language use domain in this study include daily activities in tertiary education contexts such as oral presentation of assignment, giving feedback about videos, listening to lectures, communicating with peers through the webcam and the like. Therefore, it is hypothesised that examinee satisfaction with the oral CBT procedure underlines the efficiency of and the need for the use of this delivery medium. This finding proves the assertion made by Dunkel (1999, p. 77) that “proponents of computer-based testing…
view the increasing use of computers in testing to be a natural evolution in assessment practice.” Arguably, then, the introduction of the new medium of testing through the use of video and sound is a legitimate and practical alternative to conventional tests. The standardized delivery instrument and accuracy in the administration procedure of the test, as well as the digital recording of the test-takers’ language samples, made the SAM Interview a uniquely efficient program.

Added to all these provisions are the ease of running the tasks, saving the testee performance sample, as well as replaying the recorded sounds in any order required by the raters. These are remarkable improvements over the conventional use of cassette tapes in recording examinee responses in which forwarding or rewinding imposed a great deal of time pressure on the examiners.

Digital recording of the test-taker spoken language makes it possible for the sound samples to be sent over the Internet to trained raters regardless of their location or distance. The SAM Interview saves five 3-minute MP3 files for every examinee. Each file is about 4.5 megabytes in size. The total performance of a given examinee results in just over 22 megabytes of data which can be emailed to trained raters for assessing.

An underlying algorithm in the program produces names for the MP3 files in a way that every examinee will have a unique MP3 output. In other words, when the examinee speaks to the computer, it combines three features to produce the MP3 file name which consists of Examinee Number + Examinee Name + Task Number.mp3. The resulting file name makes it impossible for raters to confuse the data obtained from a large number of test-takers. More importantly, even if a particular individual
takes the test for many times, the recorded language samples will be assigned
different file names each sitting. For example in Figure 5.1, the rater can see the testee
number (88), his name (Abbas) and task number (22). This was an advanced level
task because it falls between 21 and 30.

The findings of this research demonstrate the computer’s potential in facilitating the
rating process of spoken language which, in the eyes of the sceptics, is the most
difficult skill to assess. The fact that each examinee takes a test that is tailored to their
ability gives them a confidence in their performance. The testees are very much in
control of the testing situation and therefore are aware that questions that have low
information value — being too difficult or too easy tasks — are automatically
avoided. This is the reason why accuracy in extracting the testees’ speech samples is
achieved. In conventional tests, there is every likelihood that high ability students can
receive low difficulty questions and low ability testees can be given excessively
difficult questions. The SAM Interview program eliminates this inefficient and time-
consuming process. In addition to the accuracy in measurement, the SAM Interview
allows the examinees to work at their own pace the result of which is a boosted
confidence and more comfortable performance. At the same time, the testees are more appropriately challenged than discouraged by experiencing tasks at their approximate level of ability.

5.3 Task Difficulty and Contributing Factors

5.3.1 Research Question 6

What aspects of the computer-based tests do students find difficult to handle, and what are the factors contributing to task difficulty in CBT?

As part of the endeavour in providing evidence for the comparability of computer-based tests and conventional assessments, Research Question 6 investigated the issue of test-taker perception of task difficulty in oral CBTs. In other words, this question looks into whether test-takers find oral CBT tasks more difficult to handle than those in face-to-face tests, and if so, what aspects of the test contribute to the difficulty (see Table 4.9 for descriptive statistics). In second language acquisition, learner perceptions are given special consideration as sources of genuine data and therefore it is assumed that the variation in test-taker performance is influenced by task characteristics or conditions and that this variation leads to different scores obtained under different conditions. In language testing, task difficulty is a novel issue and has been investigated from at least two perspectives: test-taker perceptions of task difficulty (Elder, et al., 2002) and perceptions of communicative success (Fulcher & Marquez Reiter, 2003). In the framework proposed by Fulcher and Marquez Reiter (2003, p. 328), they use a ‘pragmatic approach’ in defining task difficulty which “focuses on performance and perception of task achievement (communicative success) rather than using perceptions of abstract psycholinguistic qualities of a task.”
In the current project, the researcher draws upon perception-based evidence provided by test-takers to explore CBT task difficulty. The examinees’ perceptions are reflected in their responses to a series of statements in the post-CBT questionnaire and are used as the basis for determining whether they found the CBT tasks difficult to handle (see Table 4.9). This approach to determining task difficulty is supported by previous research. For example, according to (Elder, et al., 2002) test-takers’ perception of task difficulty is important because a) it has a potential impact on their performance, and b) if test-takers can predict what makes a task difficult, it may be wise for test developers to assess their views in order to look more closely into what makes a task easy or difficult in the light of the current theories. It can be argued that the qualitative feedback from the examinees is a genuine source of data because it reflects first hand experiences of the actual examinees for whom the tasks are designed. The dependability of test-taker perception in validating a new test is analogous to their self-assessment of their own language ability.

As to the significance of task difficulty and its effects on test-taker performance, Bachman (1990) considers this notion as a test method facet and an important factor in quantifying the construct of language ability. In defining task difficulty, the assumption is, according to Fulcher and Marquez Reiter (2003, p. 322), “that the score awarded to an individual on a speaking task or tasks is affected by [test methods facets such as] the speaking proficiency of the individual, the difficulty of the task and the severity of the rater.” Task difficulty is analogous to the notion of item difficulty in selected response items such as multiple-choice or true-false items, where ‘difficulty’ is defined by the level of performance on the item or the task. But the
difficulty of tasks in speaking cannot be solely defined in terms of parameters like task conditions, person ability and rater severity. Although defining task difficulty in terms of the above parameters is a novel and significant breakthrough in language testing, it still fails to account for a set of other potentially influential factors such as differential familiarity levels with and attitude towards the new technology. As a matter of fact, during the validation of a new CBT, test-takers may even be able to identify additional features of the task or additional challenges during interaction with the computer with respect to those visible to the test developer or the rater (Elder, et al., 2002).

Furthermore, a particular task could prove differentially difficult if it is delivered face-to-face or through a computer, and therefore the interpretations drawn from the scores could be different. The pragmatic approach to investigating task difficulty proposed by Fulcher and Marquez Reiter (2003) assumes that the test-takers who carry out a particular task are indeed capable of reflecting on their performance and come up with a judgement on how successful they may have been in performing the task. This seems to be a further extension of the validity of self-assessment in language testing. In a similar study investigating the role of test-takers in estimating task difficulty, Elder, et al. (2002) provided a group of test-takers with a series of tasks whose characteristics were systematically manipulated and the impact of these manipulations on task performance was analysed. Test-takers recorded their perceptions of the relative difficulty of each task and their attitudes to them. Most research studies indicate that there is a relationship between learner perceptions of tasks and the actual task performance. Apart from the ‘perceptions’ of difficulty, ‘attitude’ towards the tasks and the testing conditions is also thought to be related to difficulty. Brooks
Discussion

(1999) cited in Elder, et al. (2002), for example, found a relationship between attitudes towards different assessment types and levels of performance on the respective tasks. But it must be admitted that interpreting the meaning of attitude/score relationship is not always easy because attitudes may be influenced by the experience of test-taking (high achievers enjoying a better attitude towards the test than the low achievers), test-taking conditions, test delivery mode (CBT vs. face-to-face), and so on. But it seems inevitable to include attitudes towards the task as a potential element in the analysis of the task difficulty.

In relation to Research Question 6, it was hypothesized that the examinees would find CBT and face-to-face tasks equally manageable. The researcher further assumes, however, that the novelty component (as a test method facet) in computer-delivered oral tests, as well as the testees’ control over the administration process, would make this delivery mode more interesting and appealing to test-takers than a conventional face-to-face interview. To test this hypothesis, six statements dealing with the notion of task difficulty were included in the post-CBT questionnaire (see items 6, 7, 8, 10, 12, and 13 in Appendix V). The purpose of these items was to elicit the examinees’ reactions, feelings, and experiences about the difficulty aspect of the test. It was assumed that different aspects of test delivery mode as a legitimate test method facet may influence examinee perceptions of task difficulty. For example, speed of delivery can increase the difficulty level of a given task. Therefore, item 6 included a statement which tried to investigate test-taker perceptions of the speed of delivery of tasks in the CBT. Results show that more than 95% of the respondents had a favourable perception of the speed with which the questions were spoken to them. In addition to
Discussion

this, 20.08% agreed that the speed was OK and only about 5% disagreed with the appropriacy of the speed of delivery.

The researcher further assumed that the provision of video clips would make the tasks easier to handle on the grounds that they provide additional context and reinforce the meaning of the questions. The statement in item 7 of the post-CBT questionnaire tried to elicit test-takers’ degree of agreement or disagreement as to whether they found the video clips relevant to the actual questions. Results indicate that 80% of the examinees found the video clips useful and relevant to the task questions. It can be argued from this finding that the provision of digital videos along with the oral question does have an impact on the test-takers’ positive perceptions of the tasks. Furthermore, the amount of time for pre-planning and that for responding to the task were considered to be potential factors in contributing to the overall task difficulty. Items 8 and 10 on the post-CBT questionnaire investigated whether the examinees had *enough time in answering* and enough time in *thinking about the answers* for each of the five tasks, respectively. It was found that about 79% of the testees agreed with the amount of time they were given to complete the tasks. Also, 85% of the examinees thought that the program gave them enough pre-planning time to think about the questions.

Under Research Question 6, one further issue which had not been investigated before was also explored. It was hypothesised that test-taker age and gender may have an influence on their perception of oral CBT task difficulty (see Table 4.11 for descriptive statistics). In order to determine the relationship between perceptions of task difficulty and test-taker age and gender, the latter two factors were taken as
independent variables and perceptions of oral CBT task difficulty as dependent variable. Statistical analysis of the post-CBT feedback questionnaire results showed no significant differences between the perceptions of male and female students concerning oral CBT task difficulty. As to the relationship between test-takers’ age and their perceptions of task difficulty, a correlational analysis showed that there is only a moderate correlation of 0.61 between these two variables. In other words, the older students found the oral CBT tasks only slightly more difficult than their younger counterparts did.

The above findings have important implications for test developers and task designers. First of all, the provision of pre-planning time is highly favoured by the examinees. In addition, a period of three minutes seems to be an optimal time required for completing a typical oral CBT task. However, variations of longer or shorter time periods can be investigated with different groups of test-takers too. Secondly, male and female test-takers have equal perceptions of CBT task difficulty. And finally, test-taker age and perceptions of task difficulty are only moderately correlated (see Table 4.12), which does not pose any serious threat to the validity of such tests for more mature students.

5.4 Attitude Analysis and Anxiety

5.4.1 Research Question 7

Are differences in test-taker attitudes and perceptions of difficulty associated with actual differences in task difficulty as reflected in scores assigned to learner performance?
Research Question 7 investigated the possible association between test-takers’ attitudes towards computers and their perception of the CBT task difficulty. It was, therefore, hypothesized that students with more positive attitudes towards the CBT will display better performance on such tests. The researcher was prompted by previous research to investigate test-taker attitude as a possible test method facet which may influence the performance on CBTs. Previous research findings indicate that “if test-takers have negative attitudes to the test then they are less likely to perform to their best of their abilities. This has obvious implications for test validity. If test attitudes interfere with test performance this may result in unwarranted inferences being drawn from test scores” (Elder, et al., 2002, p. 350). Attitudes towards computers in general and computerized testing experience in particular overlap to a large extent with the construct of computer familiarity and computer anxiety. Also, the importance of positive computer attitudes towards computers has been considered a prerequisite for developing computer skills (Frosini, Lazzerini & Marcelloni, 1998; McDonald, 2002). In a study investigating the testees’ attitudes towards computer-administered ability testing, Burke, et al. (1987) reported the presence of a positive relationship between experience and acceptance of computer-based assessment. In the current project, in order to test the hypothesis of the constructive effect of positive attitude on CBT performance, the test-takers were administered a post-CBT questionnaire to record their feelings and reactions in relation to this new experience. Qualitative analysis of the data shows that some students found the CBT “more enjoyable than a face-to-face encounter.” Some others also maintained that they “felt much relax” when they spoke to a computer. The SAM Interview allowed the test-takers to be in charge of the assessment process and adjust it to their own pace. This feeling of control over the testing process seems to have
enhanced test-taker positive attitudes towards the test. These findings support the previous comparability studies of CBT and conventional oral assessment. Kenyon and Malabonga (2001, p. 60), for example, concluded that “the adaptive nature of the COPI (Computerized Oral Proficiency Interview) allowed the difficulty level of the assessment task to be matched more appropriately to the proficiency level of the examinee.” Kenyon and Malabonga (2001) also attribute a negative attitude towards the testing situation to the lack of control in conventional oral proficiency interviews and even tape mediated oral proficiency tests. Along the same lines Wise, et al. (1994) suggested that perceived control over the testing process is especially important for anxious test-takers. The SAM Interview draws upon the provisions of the multimedia technology, and allows test-takers to gain control over the pace and procedure of the assessment. The feeling of control over the administration process enhances their positive attitude and therefore increases their self-confidence and performance on the test.

Another factor which was found to contribute to the test-taker positive attitudes towards the CBT was the inclusion of a variety of tasks in the video files. The tasks were recorded from different aspects of university life in Australia and are representative of the target language use (TLU) domain of the students who took the test. The apparent simplicity of the software package and the ease of its operation, as well as the appealing nature and quality of the task videos, might have reduced memory load in understanding the meaning of the questions by the testees and therefore reduced anxiety levels among them. Research indicates that tasks that require only moderate levels of working memory may be successfully completed despite an individual being anxious, whereas symptoms of anxiety may interfere with
the more demanding tasks that make greater use of working memory (McDonald, 2002). The video tasks favoured the test-takers in terms of their subject matter, format, length, quality, and speed of delivery. For example, the Australian life style which is closely associated with the protection and preservation of the environment and wild life was reflected on the videos. In an Australian context, test-takers of the SAM Interview did not find such tasks unusual or unreal. It was assumed by the researcher that the inclusion of such tasks would make the assessment a more authentic experience.

To explore the examinee attitudes towards the SAM Interview CBT, three items (see Appendix V) were included in the post-CBT questionnaire which required the test-takers to express their degree of agreement or disagreement with the way this instrument formulated their attitudes towards the CBT. For example the statement in item 9 read “I found assessment by computers was enjoyable for me.” In relation to this statement, 67.5% of the examinees reported that they found the CBT assessment more enjoyable for them than the face-to-face interview. The participants also reported that they found the nature of computer-human interaction less embarrassing than interaction in a live interview. In general, test-taker comments about the usefulness and efficiency of technology-delivered tests corresponded with the researcher’s initial predictions and conceptualizations of the new delivery mode. Overall, the CBT experience seemed more attractive, non-threatening, and innovative for the testees.

5.4.2 Research Question 8

Do male and female students experience test anxiety in taking a CBT?
The final research question in this project was concerned with whether the examinees experience any degree of anxiety in taking the CBT compared to conventional speaking tests. The sheer novelty of computer-delivered oral tests and their possible impact on the anxiety levels of the test-takers (Rosen & Maguire, 1990; Shermis & Lombard, 1998) prompted the researcher to investigate this aspect of the SAM Interview. It is generally believed that computers induce anxiety in examinees and therefore test-takers with little computer familiarity, for example, will be at a disadvantage in taking a CBT (Amiri, 2000; Gressard & Loyd, 1986; Sticker, Wilder & Rock, 2004). However, it can be argued that with the expansion of the use of computers in almost all aspects of our daily lives as well as in most teaching, learning, and assessment contexts, the adverse impact of the computers seems to be diminishing.

In other words, interaction with new technology is now becoming part of the target language use domain as the new generation of students are involved in activities like computer-based exams, teleconferencing, online assignments, web-cam conversations and the like. Furthermore, the use of interactive multimedia has contributed to the development and production of sophisticated software which involves interactions with machines, that resemble interactions in real life. Communicating with computers will become more natural when the use of voice recognition devices become more popular. In this research, therefore, it is hypothesised that the new medium of test delivery, at the present stage, would reasonably reduce the level of anxiety normally experienced during face-to-face encounters; the reason being that the presence of another human being (whether a tester, researcher, teacher, interlocutor, or even a
Discussion

peer) is likely to cause anxiety in the testees and hence inadvertently affect their performance.

To test this hypothesis several statements were included in the post-CBT questionnaire to record test-taker reactions particularly related to their anxiety levels during SAM Interview administration. In item 9 of the questionnaire, the test-takers were asked to express their degree of agreement or disagreement with the statement: “I found that assessment by computer was enjoyable for me.” Of all the test-takers, 22.8% strongly agreed and another 42.5% only agreed with this statement. As the SAM Interview is a new approach to the administration of the speaking test, every effort was made to reduce any ambiguity of the assessment process and enhance the clarity of the tasks as well as the initial introduction to the test. It was thought that clear instructions would give the examinees confidence in taking the test. In item 13 of the post-CBT questionnaire, the testees were asked to express their degree of agreement of disagreement with the statement: “The instructions in the video at the start were easy to understand.” Seventy five percent (78%) of the test-takers strongly agreed with this statement and the remaining 22% agreed with it. Therefore, it can be argued that the SAM Interview procedure did not involve any ambiguity which could cause anxiety in the test-takers. As a matter of fact the lucidity of the instructions as well as the user-friendliness of the procedure had a positive impact on the group of test-takers, satisfying one of the test usefulness criteria.

In this research, the possible gender effects on levels of anxiety during an oral CBT were also investigated. As the use of computers has expanded into the target language use of both male and female students in university contexts, it was hypothesised that
Discussion

there would be no significant differences between the anxiety levels experienced by female and male test-takers. Descriptive results indicated that female and male students had almost identical levels of anxiety (mean 14.06 and 14.33 respectively). The t-test results indicated that the difference between male and female anxiety levels was not statistically significant. To the best of the researcher’s knowledge, there is no published literature on the relationship between gender differences and experiencing anxiety in taking computerized second language oral proficiency tests.

Finally, a further consideration was given to the relationship between the age of the testees and their possible experience of anxiety in taking an oral CBT. The correlation between age and anxiety levels was found to be 0.05 which is indicative of almost no relationship between these two variables. This contradicts other research findings in which younger learners showed lower anxiety and hold more positive attitudes towards computers (Chou & Hsiao, in press). Arguably, with groups of test-takers having a small age range, the magnitude of differential effects of computer familiarity or anxiety also seem to diminish. In the case of test-takers with a large age range, the difference will be more widely spread. But, in general, it has been assumed that computer anxiety results from lack of sufficient familiarity with the computer (McDonald, 2002); therefore, it can be argued that the new generation of students (age mean 26) has very similar or homogeneous degrees of familiarity with the new technology as a medium of learning and seems to consider it as a useful and non-threatening tool for assessment as well. The findings in this research are in line with Levine and Donitsa-Schmidt (1998) who found that greater computer use and experience would increase confidence and therefore lower anxiety. It appears, however, that exposure is not the only factor related to anxiety levels, as some studies
contrast. For example, in a meta-analysis, Chau, et al. (1999) confirmed a reverse association between computer experience and level of computer anxiety, but found the extent of this association to vary considerably between studies. Therefore we cannot conclude that the sole factor of exposure reduces computer anxiety.

In addition to the age and computer familiarity components as potentially anxiety inducing factors, the perceived anxiety level of computer users may also be attributed to the interface design and complexity of the software package. Research findings emphasise the importance of understanding how test-taker characteristics and behaviour interact with computer hardware and software. This phenomenon is called ‘usability’ and is an important issue for cognitive scientists, educational psychologists, user interface designers, as well as ergonomists (Chou & Hsiao, in press). The need for usability is considered to be inherent in human-computer interface because it expresses the relationship between end users and computer applications. A better understanding of usability can lead to better designed computer interfaces resulting in more efficient and less error-prone usability. It can therefore be argued that a better usability can reduce the effect of construct-irrelevant factors and enhance the validity of the CBT. Following the above principles, the SAM Interview was designed and developed with a simple, easy to follow, user-friendly and efficient interface. It does not require any training for the prospective test-takers. Although it is very difficult to come across first-time computer users these days, proper orientation to the structure of the test for novice users can take place within a matter of a few minutes when it is needed. Throughout the trial process and the actual data collection using the SAM Interview, there was no instance of ambiguity, confusion, or questions in relation to the functioning of the software package, adjusting the audio volume,
scrolling or the like. One more factor which could have contributed to the insignificant anxiety level among the test-takers was perhaps the clear video introduction to the test and its procedure. The fact that the test-takers had the opportunity to replay the introductory video might also have given them the assurance they needed about the format and procedure of the test. Yet another contributing factor to test-taker satisfaction was the final video clip in which a native speaker of English thanks the examinees for having sat the test.
CHAPTER 6: CONCLUSIONS AND IMPLICATIONS

The final chapter of this thesis is organized in four major sections: conclusions, implications, limitations of the study, and suggestions for further research. Firstly, the conclusions drawn from the findings of the study are outlined in light of the components of test usefulness criteria. Secondly, three sets of implications will be presented. The first set of implications is aimed at computer-based test developers, which may include task designers, researchers in the use and validation of CBTs, programmers, IT professionals, and interface designers. This is followed by implications for test administrators who would use CBTs in testing centres and education institutions for general proficiency assessment or research purposes. The final set of implications is relevant to the actual test-takers as the principal stakeholders of the SAM Interview program. Thirdly, this chapter reports on some of the limitations of the study associated with each stage: the initial conceptualisation, development, trial, data collection, and validation of the SAM Interview test. Finally, the chapter concludes with a set of recommendations, based on the findings of the study, and raises some practical and theoretical issues for other researchers who are interested in the design and use of CBTs.

6.1 Conclusions

Several important conclusions can be drawn from the findings of this study. First of all, it can be concluded that computerization of language tests in general, and oral tests in particular, offers great theoretical and practical benefits in creating task-based tests. The computerization of a test does not necessarily mean a mere conversion of an
already existing paper-and-pencil, or in this case face-to-face, test into what is delivered through a computer. If conversion fails to bring about visible theoretical and substantial practical benefits in terms of the overall usefulness of the test, the necessity of the conversion cannot be strongly supported. However, if the conversion leads to the establishment of a reasonable degree of equivalence between CBT scores and their conventional counterparts, and if the inferences we make from the results of the two versions are similar, then the computerization of conventional tests is warranted. In this study, it was found that such an endeavour has both theoretical and practical benefits. From a theoretical point of view, the SAM Interview scores were found to be comparable to those of the face-to-face interview for IELTS or its equivalent ISLPR. The ratings of the participants’ performance on the IELTS and the SAM Interview produced similar rankings. It can therefore be concluded that the SAM Interview test taps the same underlying construct as the IELTS speaking component or its equivalent, ISLPR, and therefore meets the construct validity component of test usefulness criteria.

From a practical point of view, the computer technology allows the inclusion and use of a range of innovative, meaningful and realistic task types in the assessment context. These task types are not usually feasible in conventional face-to-face or even tape-mediated tests. The SAM Interview test, albeit a small-scale prototype, is a working example, demonstrating the successful combination of digital video and audio and their application in the context of testing speaking ability. The use of video in delivering the oral interview is, technologically, an important step forward in creating a virtual replica of real-life-like conditions of language use. It is important to note that, compared to tape-mediated tests of speaking, for example, the SAM Interview
makes use of both auditory and visual modes of communication in delivering the test. In natural communicative environments, the auditory channel is almost always (except in telephone conversations) supported by the visual channel as well. In this respect, the traditional tape-mediated speaking tests lack the visual support for the candidate. However, the SAM Interview involves the use of video, and thus creates a more interactive environment and therefore is a more realistic and authentic assessment alternative.

Another key component among the test usefulness criteria is the impact or washback effect that the test can have on the testees. In this respect, the impact of the SAM Interview can be viewed from the test-takers’ perspective. After all, test-takers are the actual stake-holders or end clients of any test development project. In this study, test-taker input collected from the post-CBT briefings, along with the qualitative comments given by the test-takers, provided insight into the positive impact of the CBT as a new mode of test delivery on the participants’ perceptions of the testing experience. For example, it was found that the human-machine interaction is becoming another aspect of real-life communication and therefore approximating an authentic activity. This makes the use of CBTs a favourable and potentially viable approach to testing in the eyes of the real test-takers. In addition to this, test-taker characteristics such as computer anxiety, differential computer familiarity, technophobia, age, and gender differences no longer pose a threat to the construct validity of the CBTs. The effects of these factors, which have traditionally been considered as inherent threats to test validity, are gradually diminishing as a result of the prevalence of the use of new technology in learning, teaching, and particularly
assessment contexts. From these findings, it can be concluded that the careful use of technology in language testing can establish grounds for comparable tests.

In relation to the tasks designed for this study, the results demonstrated that these tasks were highly successful in eliciting representative language samples from the test-takers, especially at higher levels of proficiency. The recorded language samples suitably qualified for the desired purpose of rating. It can be concluded from this finding that, in the process of conceptualising, designing, recording, editing and administering the tasks, a detailed analysis of the target language communicative situations is a crucial requirement. In this study, the tasks were designed to reflect aspects of the target language use domain of the participants’ prospective activities at Australian universities. This domain involves activities such as watching lectures presented through videos, attending teleconferences, watching presentations and reporting, using the web-cam to communicate with peers and classmates and doing online assignments. Establishing this correspondence between test tasks and the real-life activities ensures the authenticity criterion of test usefulness.

This study has extended the understanding of the wide range of variables involved in the design and implementation of a CBT. It has brought to light various factors affecting task difficulty, including the use of multimedia and manipulation of task delivery mode. This research can assist the choice of a suitable range of tasks and innovative items for assessment purposes and test development. Given that today’s language teachers are increasingly concerned with their students’ ability to become more functional in using a foreign language, the use of task-based testing delivered by
the computer provides language teachers with the opportunity to assess the productive skills of their students.

6.2 Implications

6.2.1 Implications for Test Developers

The issues in the design, development, and administration of the SAM Interview have important implications for future CBT test developers. For example, the step-by-step procedures in conceptualization, planning, programming, trialling, using and validating this test can be used as guidelines and a working model for CBT developers and validation researchers. From a theoretical point of view, test developers can look into what exactly contributes to the comparability of CBTs and face-to-face interviews. This concern is reflected in the selection and design of tasks for the new test. Test developers must make sure that the tasks in an oral CBT are representative of the real-life conditions of language use and therefore conducive to the elicitation of the required language behaviour.

Researchers in validation studies can analyse the nature of the language produced in a CBT and compare its characteristics with that of a face-to-face interview. Also, in order to strengthen the qualitative arguments for the validity of new CBTs, researchers can conduct immediate post-CBT interviews with the candidates to investigate the authenticity and impact of the test from the viewpoint of the real stakeholders and record their reflections as to the usefulness of the test. This study gives direction to other test developers trying to include digital video tasks in the oral CBT. This was a first and a very important step in contextualizing the delivery of traditional tasks. The use of videos was found to be very effective in reinforcing the
meaning of the questions and therefore gave the test-takers direction and focus in formulating their responses to given tasks. In this respect, future test developers can benefit from the services of other IT professionals to add to the variety, novelty, and meaningfulness of video tasks.

Apart from the issues related to task design and formulating validity arguments, there are also more practical implications. For example, this study has important implications for interface designers. During the trial and administration of the SAM Interview, it was found that the participants were happy and comfortable with the use of the current standard Windows-based interface in terms of lay-out, font style, font size, clarity, sound quality, and similar characteristics. An important implication for interface designers may be that the default layout, which is almost identical to the popular Microsoft products, can perhaps be counted on to be non-threatening to test-takers. Test developers must make sure that the interface design of a CBT is conducive to the overall quality of the test, facilitating its smooth delivery. It was found in this study that a good interface design can make the testing process easy and efficient. An effective, simple and clear interface design also reduces the potential effects of possible construct-irrelevant factors. These factors can influence the quality and nature of the test-takers’ performance and may alter the inferences we make from the test scores.

6.2.2 Implications for Test Administrators

It is test administrators, testing centres, education institutions and other stakeholders involved in using general English proficiency requirements as gate-keeping mechanisms who would most immediately benefit from the outcomes of this research.
In this respect, the implications fall into the two categories of resources and standardization of assessment procedures. First of all, upon successful installation and administration of the SAM Interview, it significantly reduces the amount of time, budget, human resources, number of proctors, space, and material that are required for group oral tests.

In large educational institutions, where there is a significant number of incoming international students, there is always the issue of time for organizing and administering an entry or a placement test to incoming students in order to group them into different instructional levels. In these cases, the SAM Interview may function as a standardized placement tool. In addition to the advantages of saving time and resources, raters can benefit from the convenience, clarity, and user-friendliness of digitally recorded sounds of the testees’ performance. The tasks in the SAM Interview are randomly drawn from a pool of tasks and can be administered to a large group of students at the same time. In such contexts, this enhances test security and reduces the possibility of cheating.

Finally, in the interests of test validity, especially in high stakes situations, once the SAM Interview software has been installed on multiple platforms, test administrators are advised to run the program on each machine prior to the actual testing session to ascertain the smooth and trouble-free functioning of the program.

6.2.3 Implications for Test-takers

Test-takers will find the computerized oral testing environment a new learning experience helping them to reduce the levels of anxiety traditionally associated with
Conclusions and Implications

tests and assessments. From a pedagogical point of view, the most obvious implication for the test-takers is that they may have the opportunity to assess their own language ability, including their strengths and weaknesses, in the self-assessment section of the test. Oral interaction with computers is a different mode of communication which future test-takers will need to get used to. Testing of oral ability by computers will be more authentic especially when the voice recognition technology becomes more popular.

6.3 Limitations of the Study

In this section, several limitations of the current study are noted and the need for further research into the comparability of scores obtained from computer-based and conventional tests is demonstrated. Firstly, in spite of all the promises offered by computerized testing, the available body of literature on human-machine interaction demonstrates the complexity of the effects of modes of presentation on performance in tests of language ability, particularly when it comes to the speaking construct. The review of the literature indicates that changes in the delivery mode result in changes in the nature of the task, which in turn has an impact on the quality and nature of performance, test completion time, test-takers’ affect, test administration conditions, logistical requirements for the test, degree of control over the testing procedure, and a whole host of other latent variables all of which contribute to the generalizations and inferences we make on the basis of test scores which may not even be directly influenced by the above factors. The scarcity of empirical research on the use of audio-video as a mode of presentation in language tests, on the other hand, makes it difficult to draw conclusions about the nature and magnitude of delivery mode effect on test performance. As a result, to make the most of technology-driven assessment as
a would-be regular component of standardized testing in the coming century, stringent checks and balances must be put into place to avoid potential pitfalls.

Secondly, it must be acknowledged that this research, as a prototype test development project, was conducted on a small scale with a relatively small number of students. Given that the participants had to be international students who had undertaken IELTS or ISLPR tests immediately prior to the administration of the SAM Interview, it was logistically prohibitive to locate a large number of subjects who would agree to take part in the study. This research, therefore, looked into the language samples of a small group of testees under explicitly declared research conditions, in the sense that the subjects were aware of participating in a research project and therefore might have deliberately concentrated on producing more accurate speech. Had the research employed a larger sample, the findings might have been of a different nature. Apart from the issue of sample size, however, the limitation regarding the effect of the declared research conditions on the nature of students’ language samples is not peculiar to this project alone; the notions of authenticity and validity are under threat in any testing project as it is impossible to perfectly simulate all the real-life conditions (target language use situations) during a testing experiment.

In addition, the presence of variations in the perceptions of test-takers as well as their qualitative comments made in relation to the efficiency of the CBT in delivering speaking tests is a warning to researchers that for validation purposes we cannot solely rely on test-taker verbal feedback. Such individually made qualitative comments should, however, be used as supplementary rather than complementary evidence to quantitative data in designing, developing and validating a new testing
Conclusions and Implications

instrument. Test-taker comments and reactions may inadvertently be affected by a range of construct irrelevant factors such as gender, social class, professional experience, personal preference, test-taking strategies, individual differences and language proficiency (Elder, et al., 2002).

Furthermore, a number of logistical issues need to be addressed by a test developer who chooses to utilize the video-audio capacity of the new technology in delivering language tests. Computerized tests of speaking ability are the most difficult to design and administer. In conceptualizing, designing and implementing a novel computer-based test, a sole researcher is faced with numerous challenges including technical issues in locating, recording, and editing video files, computer programming, integrating the audio-video files into the software package, limitations of hardware and software for creating, manipulating, managing, and storing an item-bank, copyright and security issues, cost, time, and the like. Amiri (2000, p. 83), for example, suggests that “we increasingly need qualified language teachers who at the same time are highly skilled in the relevant aspects of IT including design, development and programming of computer-based materials.” This suggestion seems feasible in small-scale classroom testing where decisions made on the basis of test results do not involve highly sensitive decisions such as accreditation or certification for profession. However, for large-scale and high-stakes testing programs, it is suggested that the development of a computer-based test from scratch be undertaken by a team of applied linguists, psychometricians, language testing experts, and computer programming professionals. Such a team would have the resources to explore creative ways to adapt the measurement models and the technology to the
nature of language ability and language use in the interests of higher validity, reliability, and utility (Bachman, 2000).

The researcher would like to strongly support Dunkel’s (1996) position that appropriately realizing the potential applications of new technologies in language assessment will require a team effort. Finally, for future CBT development and validation purposes, it is crucial to take advantage of the advances in voice recognition technology (as well as Natural Language Processing) to rate the testees’ language samples.

6.4 Suggestions for Further Research

This study constitutes an important step forward in language testing because it involved conceptualizing and implementing a small-scale prototype of a computer-based oral test using digital video for the first time. The purpose was to determine its validity, impact and comparability with face-to-face assessments. Several suggestions are in order for other scholars in CBT development and validation to consolidate the methodology and approach in any follow-up research.

First of all, this research employed the IELTS and ISLPR scores as the external independent criteria for establishing the validity arguments of the SAM Interview. As validity is an iterative and developmental process, it is suggested that future CBT developers and validation researchers use other benchmarks in substantiating the validity claims of their new tests. It is possible to compare the results of test-takers’ performances on other CBTs concurrently or predictively with those on the SAM Interview to investigate the reliability and validity of this test. Moreover, it is also
suggested that other researchers employ the services of multiple raters in order to investigate the internal consistency of CBTs.

In terms of task design, it is suggested that test developers should improve the scope and usefulness of oral CBTs by including a larger pool of tasks into the program. With the availability of DVDs, which can accommodate a much larger number of video clips, such tests will be able to represent broader domains of language use contexts and address a wider range of test-taker abilities, and in turn increase reliability and test security. As for the content of the tasks, future researchers may be willing to develop oral CBTs for tests of English for specific purposes by creating and including a variety of tasks that focus on special domains such as science, aviation, engineering, cultural issues and so on.

Furthermore, future researchers can improve on this project by changing the nature of the initial self-assessment in terms of the number of items, type of items, or scoring system. The pre-CBT self-assessment in the SAM Interview comprised thirty dichotomously-scored can-do items to which the candidates were required to give a Yes or No answer. However, it is conceivable that, had the test-takers been given a wider range of possible responses (instead of Yes or No) in evaluating their own abilities, the initial self-assessment results could have given the researcher a more accurate picture of test-taker ability. In other words, we can increase the number of questions and the number of alternatives for each question so they would be more finely tuned to test-taker ability. Obviously, this would make the scoring system more complicated and therefore would require a more sophisticated set of algorithms.
One further possibility to enhance the usability of oral CBTs is to make them available online. The practical implications of web-based (internet-based) tests are enormous. However, with the use of audio-video clips as prompts in the tasks, further consideration needs to be given to the initial task development. Online oral CBTs with visual support will probably impose a further technical constraint. That is, smooth delivery will depend on the download speed of the clients’ internet service. Therefore, smaller size (both in screen display dimensions and megabytes) video clips could be a possible solution. But care must be taken not to sacrifice the quality of the tasks in favour of a speedy download.

6.5 Final Remarks

With the proliferation of selected-response item types in computer-based tests, the use of more interactive, authentic and meaningful task types in such tests is crucial. This study is a small but important step in using more advanced features of technology in language testing and justifying the validity of their use. The dynamic nature of spoken language makes such an endeavour a challenging task. However, the findings of this research provided the researcher with insight into the quality and nature of the spoken language in the test. This thesis provides evidence that the SAM Interview process is not only perceived as a positive experience but also has the ability to discriminate between different levels of proficiency and, more importantly, it is comparable with its traditional counterparts.
REFERENCES


References


Davidson, P. (2003). Why technology had had only a minimal impact on testing in education? *Proceedings from the 2nd Education Technology Conference and Exhibition*. Oman: Sultan Qaboos University.

References


Dunkel, P. (1996). Checking the utility and appropriacy of the content and measurement models used to develop L2 listening comprehension CATS: Implications for further development of comprehensive CATs. Paper presented at the Issues in computer adaptive testing of second language reading proficiency conference, Minneapolis, MN.


materials: Language testing in Japan (pp. 126-133). Tokyo: The Japan Association for Language Teaching.


APPENDICES

Appendix I: Informed Consent

By signing below, I confirm that I have read and understood the information package and in particular have noted that:

I understand that the project will be carried out as described in the information statement, a copy of which I have retained;

I understand that my involvement in this research will include taking a 30-item questionnaire, answering five audio-video questions, as well as a 15-item questionnaire;

I have had any questions answered to my satisfaction;

I understand the risks involved;

I understand that there will be no direct benefit to me from my participation in this research;

I understand that my participation in this research is voluntary;
I understand that if I have any additional questions I can contact the research team;

I understand that I am free to withdraw at any time, without comment or penalty;

I understand that I can contact the Manager, Research Ethics, at Griffith University Human Research Ethics Committee on (07) 3875 5585 (or research-ethics@griffith.edu.au) if I have any concerns about the ethical conduct of the project;

and

I agree to participate in the project.

Name: ----------------------- Signature: -----------------------  Date:    /     /
Appendix II: Information Sheet

Development and Validation of a Multimedia Package for the Assessment of EFL Learners’ Oral Proficiency in English: Implications for Score Comparability

INFORMATION SHEET

Seyyed Abbas Mousavi
School of Languages and Applied Linguistics
Contact Phone: (07) 387 55140
Contact Email: s.a.mousavi@griffith.edu.au

Why is the research being conducted?
This research is being conducted to investigate the usability and usefulness of a new computer-based test. The research also seeks whether the test can be used instead of face-to-face test of speaking proficiency in English. This is a data collection phase of a PhD program at the School of Languages and Linguistics at Griffith University.

What you will be asked to do:
You will need to take a self-assessment test so that the computer will determine your approximate level of English language ability. Then you will watch and listen to 5 short video clips which are accompanied with relevant questions. You will need to listen carefully and answer those questions during the time period provided.

The basis by which participants will be selected or screened:
This package has been especially designed and developed for those students whose first language is not English. Participation is voluntary.

The expected benefits of the research:
Both language teachers and newly-arrived students at Griffith University will benefit from this research. It provides the opportunity for students to test themselves through the use of a computer program. It also helps teachers and assessors save a lot of time and resource in screening students and conducting face-to-face interviews.
Risks to you:
There are no risks involved in participating in this research. However, the students who have never taken a computer-based test before may be slightly nervous which is a natural part of any assessment procedure. But the participants are assured that the results of the assessment are used solely for research purposes and will have no implications for their study or semester results.

Your confidentiality:
The data provided by the participants in this research will all be codified and no identifiable information will be submitted to or published in any journal or report. The data you provide will be stored in the researcher’s computer and will be kept confidential until the research is fully conducted. Then after a period of five years, all the data will be deleted.

Your participation is voluntary:
Participation in this research is voluntary. Your decision to participate in this research will not have any relation with or effect on your studies or semester results. You are free to withdraw from the study at any time you wish.

Questions / further information:
For further information about any aspect of this research, you are welcome to contact the researcher at the above email address or phone number.

The ethical conduct of this research:
Research at Griffith University is conducted in accordance with the National Statement on Ethical Conduct in Research Involving Humans. If you have any concerns or complaints about the ethical conduct of the research project you may contact the Manager, Research Ethics on (07) 3875 5585 or research-ethics@griffith.edu.au

Feedback to you:
The overall results of this research will be available through the researcher. However, it is not possible to provide individual results to participants.
Privacy Statement:
The conduct of this research involves the collection, access and/or use of your identified personal information. The information collected is confidential and will not be disclosed to third parties without your consent, except to meet government, legal or other regulatory authority requirements. A de-identified copy of this data may be used for other research purposes. However, your anonymity will at all times be safeguarded. For further information consult the University’s Privacy Plan at www.gu.edu.au/ua/aa/vc/pp or telephone (07) 3875 5585.
Appendix III: Pre-CBT Self-assessment

Self-Assessment for General Proficiency in English

INSTRUCTIONS

In order for the computer to give you the appropriate questions, we need to know your general ability in English. Before you start the test, please read the following items and indicate your level of ability in English. Tick appropriate buttons and scroll down for more questions. There are 30 questions and all of them must be answered before you can start the test.

<table>
<thead>
<tr>
<th>No.</th>
<th>Question</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>When you face native speakers of English, can you introduce yourself and tell them how you feel?</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>2</td>
<td>On a rainy day, can you describe a rainbow?</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>3</td>
<td>When you meet new friends, can you describe the members of your family in English?</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>4</td>
<td>Can you ask and answer questions about someone’s nationality and job?</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>5</td>
<td>Can you book a hotel room on the phone?</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>6</td>
<td>You have run out of groceries at home. Can you order food on the phone?</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>7</td>
<td>Can you easily and smoothly finish a conversation with a native speaker of English?</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>8</td>
<td>Can you start and carry on a conversation about your hobbies?</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>9</td>
<td>If you have an accident, can you report it to the police?</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>10</td>
<td>If you are sick, can you explain your illness to the doctor?</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>11</td>
<td>In your own culture, can you describe a wedding party in detail?</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>12</td>
<td>Can you give the details of a pilot's job?</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>13</td>
<td>Can you explain what you will be doing in the future?</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>14</td>
<td>Can you describe the geography of your country to a group of tourists?</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td></td>
<td>Question</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>15</td>
<td>Can you fluently describe the political parties of your country?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Can you handle a case in a court and defend yourself?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>In a class discussion, can you describe the education system of your country?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Can you negotiate the charges of a hotel for a lower price?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>At an airport, can you describe the contents of your luggage to the customs officer?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Can you do mental arithmetic without slowing down in English?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>Do you consider yourself a bilingual and bicultural person?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Do you have the ability to defend your values and give logical reasoning for them?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>Can you carry out a public speech in front of a crowd?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>Do you consider yourself a bilingual interpreter who can handle a news conference?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>Do you always understand native speakers when they speak?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>Can you fully participate in humorous and language pun games on TV?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>Are you able to start and carry on a conversation about your job or profession?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>Can you fully comprehend and respond to a conversation on the phone?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>Can you fluently describe the life style of the people of your country and the recent trends and developments?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>Can you handle sensitive professional situations like international conferences and negotiations?</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix IV: Instructions

INSTRUCTIONS

“Hello, Welcome to SAM, a test of your speaking ability in English. I’m now going to give you an introduction to the test. If you would like to watch and hear this introduction again, you can click on the ‘Replay Introduction’ button. When you tell us that you are ready to start the test, we will first ask you for some personal information. Then we’ll ask you to complete a questionnaire about your speaking. This is to help us to give you questions that are at a suitable level. Then the test will start. You will have five questions. As you watch the screen, you will hear the first question. You will have a maximum of one minute to think about your answer. Then you will have a maximum of three minutes to answer the question. Your voice will be recorded into the computer. Then we’ll give you the second question. When you have answered five questions, the test will be over. Your result will be emailed to you within two weeks. If you would like to watch and hear this introduction again, click on the ‘Replay Introduction’ button. When you are ready to start the test, click on the ‘Next’ button.”
Appendix V: Post-CBT Questionnaire of the SAM Interview Program

**Feedback Form on the SAM Interview Program**

Gender:  □ Male       □ Female       Age: ………

What is your first language?  ………………..

How long have you been in Australia?  ………..  Year(s) ……….  Month(s)

**INSTRUCTIONS**

The computerized test you just took was designed to help us improve the way we will do this English language proficiency interview next time. Thinking about your interview, use the scale below to show how much you agree or disagree with each statement. Please tick the correct button.

<table>
<thead>
<tr>
<th>No</th>
<th>Statements</th>
<th>Strongly Agree</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Strongly Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The program was easy to start and follow.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>I need to know a lot about computers to run this program.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>The design of the program was confusing for me.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>I spent too much time typing my personal details.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>The sound quality was satisfactory to me.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>The speed at which the questions were spoken was okay for me to follow.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>I found the video clips related to the questions.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>I had enough time to answer each of the five questions.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>I found assessment by computer was enjoyable for me.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>The program gave me enough time to think about my answers to the questions.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>I felt more confident communicating with the computer than a face-to-face interview.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>The questions were at the right level of difficulty for me.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>The instructions in the video at the start were easy to understand.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
14 I needed the videos to help me understand the questions.

15 I found communicating with the computer instead of a real person was less embarrassing for me.

If you have any other comment, please write it in the following box:

……………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………...
### Tasks for the Pre-Intermediate Level

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>This is one of Brisbane’s libraries. How much time do you spend at the library and what kinds of books do you usually read or borrow?</td>
</tr>
<tr>
<td>2</td>
<td>What would you do if you won $50,000 in a lottery? Would you spend it on travel or spend it in some other ways?</td>
</tr>
<tr>
<td>3</td>
<td>There are many kinds of music. What kinds of music do you like and why?</td>
</tr>
<tr>
<td>4</td>
<td>This is Brisbane City. Do you remember when you first arrived in Australia? Can you describe how you felt and where you went?</td>
</tr>
<tr>
<td>5</td>
<td>This is one of Brisbane’s busy streets. Tell us about traffic in your home town.</td>
</tr>
<tr>
<td>6</td>
<td>Many Australians like to play sports and watch sports on television. What kind of sports are you interested in? If you play a sport, why do you like it?</td>
</tr>
<tr>
<td>7</td>
<td>When you were a child, did you ever dream of your future job? What did you want to be and why?</td>
</tr>
<tr>
<td>8</td>
<td>Most people in Australia live near the coast where it is usually green. Do you have four seasons in your country? Can you describe them?</td>
</tr>
<tr>
<td>9</td>
<td>Tell us about some of your favourite foods. How different is food in Australia from your country?</td>
</tr>
<tr>
<td>10</td>
<td>Why do you think smoking is bad for your health? Give us as many reasons as you can.</td>
</tr>
<tr>
<td>Task</td>
<td>Text</td>
</tr>
<tr>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>11</td>
<td>This video shows one of the streets in Brisbane. Tell us about the transportation you use in your home country. Which form of transport do you think is the most convenient, and why?</td>
</tr>
<tr>
<td>12</td>
<td>The protection of the environment is one of the most important issues in today’s world. What do you think will happen to the wildlife if we do not protect the environment? What are the main environmental concerns in your country?</td>
</tr>
<tr>
<td>13</td>
<td>In many cities, there are no controls over the number of cars on the roads. Some people believe that we need to reduce the number of cars in our cities and make life simpler and safer. To what extent do you agree with this?</td>
</tr>
<tr>
<td>14</td>
<td>In this video, you see a modern athletics stadium. The quality of sports facilities and equipment is said to have a great influence on the success of athletes. To what extent do you think this is true?</td>
</tr>
<tr>
<td>15</td>
<td>This is a scientific laboratory where scientists perform experiments. During your school days, did you have such a place at your school? What kind of experiments did you try? Do you enjoy learning this way?</td>
</tr>
<tr>
<td>16</td>
<td>This is a chemical laboratory where scientists are busy carrying out different experiments to improve the quality of life. What kind of scientific experiments do you think scientists carrying out at these places?</td>
</tr>
<tr>
<td>17</td>
<td>Language teachers believe that socializing helps us improve our English ability and communication skills. To what extent do you agree with this?</td>
</tr>
<tr>
<td>18</td>
<td>Some people believe that preserving and supporting wildlife is very important. To what extent do you agree with this?</td>
</tr>
<tr>
<td>19</td>
<td>Travelling has become quite easy these days. What forms of transport do you enjoy most when you travel and why?</td>
</tr>
<tr>
<td>20</td>
<td>This is a public celebration in Australia. What events are celebrated in your country? Please tell us about one of them.</td>
</tr>
<tr>
<td>TASKS FOR THE ADVANCED LEVEL</td>
<td></td>
</tr>
<tr>
<td>------------------------------</td>
<td></td>
</tr>
<tr>
<td><strong>21</strong></td>
<td>Climbing, as seen in this video, is a popular sport in many countries. What abilities do you think are required to be a successful climber? What do you think are the advantages and limitations of this sport?</td>
</tr>
<tr>
<td><strong>22</strong></td>
<td>These days, computers have become an essential part of our daily lives. Considering all the new computer and mobile technologies, what do you think the world will be like in 2020? Do you think computers are bringing people together or are they reducing social interaction?</td>
</tr>
<tr>
<td><strong>23</strong></td>
<td>This video was taken by a tourist in Australia. If a group of foreign tourists visited your hometown, where would you take them and what places of historical interest would you show them?</td>
</tr>
<tr>
<td><strong>24</strong></td>
<td>This video shows some aspects of life in a modern city. There are advantages and limitations associated with life in big cities. How do you compare it with life in rural areas?</td>
</tr>
<tr>
<td><strong>25</strong></td>
<td>This flower is powered by solar energy. At present we rely on oil for most of our energy needs. However, oil is a fossil fuel and causes air pollution, and it will eventually run out. Solar energy is the only solution to this problem. To what extent do you agree with this?</td>
</tr>
<tr>
<td><strong>26</strong></td>
<td>In many big cities with a lot of high rise and modern buildings, you also see some very well protected historical places. Now, think of your own country; why do you think it’s important to maintain the historical places of a country?</td>
</tr>
<tr>
<td><strong>27</strong></td>
<td>Scientists have made a great contribution to improving the quality of human life. Can you give an example with details?</td>
</tr>
<tr>
<td><strong>28</strong></td>
<td>Smoking has been banned in all public places and is said to be dangerous to health. However, some people still smoke and consider this as an individual right. To what extend do you agree with this?</td>
</tr>
<tr>
<td><strong>29</strong></td>
<td>There are many ways to advertise, through TV, Radio, Billboards, Newspapers, the Internet, etc. Which ways do you think are most effective and why?</td>
</tr>
<tr>
<td><strong>30</strong></td>
<td>Recycling can help us preserve the earth’s resources and limit the amount of waste we produce. To what extent do you agree with this? And what can be done to encourage people to reduce waste.</td>
</tr>
</tbody>
</table>
Appendix VII: How to Install the SAM Interview Software Package

1) Open the CD-ROM and double click on **Setup2.exe** file. This will run the installer and the program will start to install in **C:\Program Files\Sam Interview** as shown below. Click on the “Next” button to continue.

![Setup2.exe installation screen]

2) The Setup window will also need to install the **Microsoft Data Access Component 2.0**. Click “Yes” to continue.

![Microsoft Data Access Component 2.0 license agreement]
3) This will start the Setup window for **Microsoft Data Access Component 2.0**. Click on the “Continue” button to go to the next step.

![Microsoft Data Access Component 2.0 Setup window](image)

4) Click on the “Complete” icon to install the full **Microsoft Data Access Component 2.0**. This will install the program in `C:\Windows\System32` folder as shown below:

![Complete installation window](image)
5) Once the program has been installed, replace the file Project1.exe in the
C:\Program Files\Sam Interview folder with Project1.exe from the CD-ROM as shown below:

The installation is now complete. Click on icon on your Desktop to run the program.

IMPORTANT NOTE: The test-takers’ speech samples are recorded and saved in
C:\Program Files\SAM Interview\SOUNDS folder.