

**Korean university students' attitudes to and performance on a
Face-To-Face Interview (FTFI) and a Computer Administered
Oral Test (CAOT)**

by Mi-jin Joo

THESIS

**Submitted in partial fulfillment of the requirements for the degree of Doctor
in Education in the Institute of Education of
the University of London, 2008**

I hereby declare that, except where explicit attribution is made, the work presented in this thesis is entirely my own.

Word count (exclusive of appendices, list of references and bibliographies but including footnotes, endnotes, glossary, maps, diagrams and tables): 45,718 words

SYNOPSIS

This study intensely investigated Korean university students' attitudes to a Face-to-Face Interview (FTFI) and a Computer Administered Oral Test (CAOT) first and then their performance on the tests, and finally their effects on performance on the two tests in a Korean university context.

The 42 university students participating in the study took part in both the FTFI and the CAOT. After these tests, they completed a questionnaire about their attitudes towards and their perceptions of the tests. Ten of them were interviewed after the questionnaire to understand more deeply their attitudes and performance. Their performance on the two tests was examined using Multi-Faceted Rasch Analysis.

The results of this study indicated that Korean university students showed much more favorable attitudes to the CAOT compared with previous studies on direct and semi-direct tests, but they still preferred the FTFI to the CAOT in spite of significant negative attitudes to the FTFI with respect to aspects such as nervousness, preparation time, and tiredness.

In terms of performance, Korean university students generally had low speaking abilities, but their speaking ability could still be discriminated well by the Rasch model. Their performance was assumed to be affected by many other intervening factors, but

the findings suggested that their performance was not influenced by factors such as test order, bias between raters and test formats, computer familiarity, gender or age differences; however, there was an effect for the severity between raters. The students preferred the FTFI overall, but the study also showed that the FTFI was more difficult than the CAOT, indicating a test format effect on performance.

Finally, the results of the analyses using the ability estimates and compensating for rater severity indicate that the students' attitudes about the FTFI were associated with their performance on the FTFI, while there was no relationship between their attitudes to the CAOT and performance on the CAOT. The students performed better on the FTFI when they had more positive and less negative attitudes toward the FTFI. That is, this study indicates that Korean university students' attitudes to the FTFI could be important sources of construct irrelevant variance on their speaking test performance on the FTFI.

Based on all the findings of this study, I conclude that the use of the CAOT should be considered by teachers and administrators in Korea. The CAOT may be useful for the assessment of achievement during or at the end of the course, or as an alternative test method, in the situation where it is needed to test students' overall oral ability, but hard to conduct the FTFI, especially due to its impracticality (e.g., the lack of skillful teachers and a large number of students).

To my mother

ACKNOWLEDGEMENT

I am deeply indebted to my supervisors, Dr. Catherine Walter and Professor Andrew Brown whose feedback, stimulating suggestions and encouragement helped me in all the time of research for and writing of this thesis. Especially, Dr. Catherine Walter not only provided her own meticulous feedback but also arranged feedback from Dr. Tim McNamara, who is a specialist related to the methodological aspect of my study. Without their invaluable guidance and encouragement, I may not have completed this thesis successfully. I would also like to give my appreciation to my independent reader, Professor Gordon Stobart and internal and external examiners, Professor Constant Leung at King's College London, University of London and Dr. Glenn Fulcher at University of Leicester, who provided helpful comments and suggestions for my thesis improvement.

My special gratitude also goes to all lecturing instructors for taught courses and tutors for my essays. Their insightful lectures and kind comments and assistance allowed me to try out many ideas and improve them.

To Dr. Jerry Larson at Brigham Young University, I am grateful for allowing me to use the Oral Testing Software (OTS) for this study and the information about the software provided at the early stages of this study.

In data analysis of my thesis, I wish to express my gratitude to Professor Jun-il Oh at Pukyung National University for his helpful advice for and interest in my study and putting me in touch with Dr. Hee-kyung Lee at Yonsei University, who provided a personal tutorial, useful comments and encouragement. I must also address my special thanks to Dr. John Mike Linacre, who provided very valuable advice for the analysis.

He has never refused any inquires as regards the FACETS program and gave me prompt responses. His replies and comments were very practical and useful every time.

Finally, I would like to give very sincere thanks to my parents, two sisters and two brothers, who have been providing all sorts of tangible and intangible support all the time. Their blind love and endless support always gave me courage and strength to overcome my long and tough graduate life. Without them nothing would have been possible.

TABLE OF CONTENT

REFLECTIVE STATEMENT.....	12
CHAPTER ONE. INTRODUCTION.....	24
1.1 Research background and rationale	24
1.2 Past studies on test takers' attitudes to direct and semi-direct tests.....	29
1.3 Research questions	38
1.4 Limitations of the study	41
CHAPTER TWO. THEORETICAL BACKGROUND.....	44
2.1 Teaching and Testing	44
2.2 Qualities of good tests	46
2.2.1 Validity.....	46
2.2.2 Reliability	50
2.2.3 Practicality	51
2.3 Communicative language testing	52
2.3.1 Communicative language ability.....	52
2.3.2 Communicative language test.....	56
2.3.2.1 Constraints of a Face-To-Face Interview (FTFI).....	61
2.4 Computerized Oral Tests	64
2.4.1 Semi-direct oral tests	68
2.4.2 Computer Administered Oral Test (CAOT).....	75
CHAPTER THREE. METHODS OF DATA COLLECTION.....	80
3.1 Participants	80
3.2 Oral tests (FTFI and CAOT)	83
3.3 Questionnaire	92
3.4 Interview	93
3.5 Data Analysis	98
CHAPTER FOUR. ANALYSES OF QUESTIONNAIRES, INTERVIEWS AND SCORES.....	105
4.1 Attitudes	105
4.1.1 Attitude Questionnaires.....	105
4.1.1.1 Attitudes to the FTFI and the COAT.....	107
4.1.1.2 Computer familiarity.....	129
4.1.2 Interviews	131
4.1.2.1 High computer familiarity.....	133
4.1.2.2 Interaction.....	135

4.1.2.3 Nervousness.....	139
4.1.2.4 Unfamiliarity with the CAOT.....	141
4.1.2.5 Others.....	143
4.1.2.6 Conclusion.....	144
4.2 Performance.....	147
4.2.1 Performance on the FTFI and the CAOT.....	147
4.2.1.1 Test taker.....	152
4.2.1.2 Rater.....	156
4.2.1.3 Test format.....	158
4.2.1.4 Test order.....	160
4.2.1.5 Item.....	161
4.2.1.6 Bias.....	165
4.2.1.7 Computer familiarity, age and gender.....	166
4.2.2 Effects of attitudes on performance	170
CHAPTER FIVE. DISCUSSION AND CONCLUSION.....	180
5.1 Research Question One: What are Korean university students' attitudes toward a Face-To-Face Interview and a Computer Administered Oral Test?	180
5.2 Research Question Two: What are the students' performance on the FTFI and the CAOT?.....	189
5.3 Research Question Three: Do the attitudes influence performance on the FTFI and the CAOT?.....	197
5.4 Conclusion and implications	201
5.5 Further study	211
BIBLIOGRAPHIES.....	215
APPENDICES.....	230
Appendix One. Basic English Skills Test Sections.....	230
Appendix Two. A sample of ESPT.....	231
Appendix Three. Consent Form.....	233
Appendix Four. Questionnaire for oral testing formats	234
Appendix Five. The structure of 2004 tests	237
Appendix Six. A sample of the CAOT.....	238
Appendix Seven. Rating scale and descriptors	244
Appendix Eight. Comparative results on Section Two of Joo's study (2004)	246

LIST OF TABLES

Table 1.1 Kenyon and Malabonga’s study on test taker affective reactions (2001)...	36
Table 1.2 Joo’s study on test taker affective reactions (2004)	36
Table 3.1 Characteristics of participants	82
Table 3.2 The structure of the FTFI and the CAOT	84
Table 4.1 Comparative results on Section Two	108
Table 4.2 Comparative results on Section three.....	109
Table 4.3 Crosstabulation of Preference * Comfort.....	114
Table 4.4 Crosstabulation of Nervousness * Comfort	115
Table 4.5 Results of the factor analysis for attitude items.....	120
Table 4.6 Reliability of the factors	122
Table 4.7 Wilcoxon tests of attitude scores of the FTFI and the CAOT.....	123
Table 4.8 Wilcoxon and Mann-whitney tests of attitude scores by gender.....	124
Table 4.9 Wilcoxon and Mann-Whitney tests of attitude scores by age.....	126
Table 4.10 Computer familiarity by gender and age.....	129
Table 4.11 Correlations between computer familiarity and attitudes.....	130
Table 4.12 Misfitting test takers.....	153
Table 4.13 Summary statistics for test takers	156
Table 4.14 Rater measurement report.....	157
Table 4.15 Test format measurement report using the edited data	159
Table 4.16 Test Order Measurement Report	161
Table 4.17 Measures and Fair-M averages for items	162
Table 4.18 Statistics Report for items.....	162
Table 4.19 Bias Calibration Report, rater and test format interaction	165
Table 4.20 Bias calibration report, gender and test format interaction	167
Table 4.21 Bias calibration report, age and test format interaction.....	167
Table 4.22 Test taker gender measurement report.....	168
Table 4.23 Test taker age measurement report	169
Table 4.24 Correlations between attitudes and measures.....	172
Table 4.25 Correlations between FTFI attitude items and measures	175
Table 4.26 Correlations between attitude items and measures using edited FTFI data.....	178

LIST OF FIGURES

Figure 3.1 Summary of the procedure.....	86
Figure 4.1 Relationships among attitude items.....	116
Figure 4.2 All facet summary for the combined data.....	151

REFLECTIVE STATEMENT

I have taken the four taught courses (Foundations of Professionalism, Methods of Enquiry 1 and 2 and a Specialist Course) with the submission of a 5,000 word essay after each course, completed the 20,000 word Institute-Focused Study (IFS), and finally the thesis, as parts of the requirements of the EdD programme.

Through discussion with tutors, coursework and feedback on first and final drafts with my supervisors (Catherine Walter and Andrew Brown), the IFS, the thesis, and very detailed comments and suggestions on each section and chapter, each element of the programs has helped and encouraged me to develop my abilities not only as a research student, but also as a professional.

Taught courses

As part of the Foundations of Professionalism course, the first paper was submitted. In this assignment, I explored the professionalism of EFL teachers and the difficulties they experience as acting professionals in the context of a Korean university. I also explored the issue of how one defines *profession* and *professionalism*, encountering in the process numerous definitions of *profession* by different authors, each stressing the essential nature of different characteristics or functions. It thereby emerged that

although it is difficult to distinguish professional work from areas of work that have not achieved professional status, there are some criteria with which many authors are in agreement.

I was subsequently able to identify three core concepts concerning the nature of professionalism: specialized knowledge base, autonomy, and service. These core concepts formed the basis of the discussion relating to the professionalism of EFL lecturers. It emerged that compared with the standards achieved by professors, many lecturers were less effective in the classroom. This was ascribed to the fact that as well as lacking the autonomy that professors enjoy, lecturers also have less developed knowledge and experience because of their relatively shorter periods of education. Also, few have undertaken research which could develop their professional knowledge. I also addressed the issue of difficulties that EFL lecturers experience within the Korean university context, and offered some suggestions for both lecturers and schools.

Before preparing this paper my concept of professionalism was vague, but in the process of exploring the issues above I acquired a clear framework for professional conduct, standards and qualifications. Researching this paper encouraged me to reflect on how I thought and worked in my institution, and was ultimately beneficial to me in establishing my identity, and developing my abilities, as a professional.

Following the completion of the Methods of Enquiry 1 course, the second essay was required for submission. This was a kind of proposal for future research which would be adapted for my IFS and thesis.

In this paper I set out to discuss the research problems frequently faced in my teaching and testing, citing the frequent absence of direct speaking testing, which could result in harmful backwash effects on learning and teaching. As a potential solution to this research problem, I stated the importance of investigating teachers' marking reliability through different forms of speaking tests and teachers' and students' preference for a computerized oral testing or a face-to-face interview. I also examined and discussed research methodologies, research design, data collection and analysis to determine which would be the most appropriate areas for further studies.

This paper provided me with an opportunity to reflect seriously on the nature of the research problems faced in my work, as well as studying various research methodologies and finally selecting the most appropriate ones in terms of collecting reliable and valid research data. This assignment was very beneficial to me in the sense of providing a springboard for my IFS and thesis.

In preparing this essay I was able to increase not only my knowledge of different types of educational research methods and design, but also my professional practice as a

researcher.

After the Methods of Enquiry 2 course, the third piece was required. For this piece, a pilot study for the IFS, I conducted a small research survey on teachers' and students' perceptions of different forms of oral testing that they were currently using in English conversation classes. Two research methods – the questionnaire and the qualitative interview - were used in this study. My rationale for choosing the questionnaire was that I deemed this the most appropriate means of collecting answers from a large number of teachers and students with limited time and financial resources at my disposal. After the questionnaire, I also conducted qualitative interviews with a small number of teachers and students to probe more deeply their opinions and feelings.

The collected data from the questionnaire and the interview were analyzed and interpreted according to the research questions. These were as follows: What testing methods do teachers use in English conversation classes? How do teachers and students feel and think about their testing methods? How do teachers and students perceive a Computerized Oral Test (COT)?

In spite of some limitations such as time restrictions and limited sample size, the findings of this study were very useful in establishing what teachers and students thought of their testing methods. The results indicated that it was desirable to look for a

more practical and valid oral testing method for teachers, especially for those with larger classes.

The findings endorsed and supported my further study on the COT. Moreover, this small-scale research provided me with very useful practical experience for professional research development.

I published this paper in *Language and Literature Research*, 13: 79-99, and also presented the main ideas and findings of this study in the conference of the Korean Association of Language Sciences, in February of 2003. Some of the attendees asked further questions after the presentation, showing strong interest in the findings and the COT program itself.

For the final paper as part of the last taught course, Specialist Course in International Education, I first sought to understand globalization and the role of English in the international community. How English has come to establish itself as the world's *lingua franca* seems to be contentious, I asserted, but one aspect remains beyond contention; there is a vast and ever-growing demand for English language learning and teaching. Based on the concept of globalization, I made three criticisms of two American ready-made tests: the Test of English for International Communication (TOEIC) and the Test of English as a Foreign Language (TOEFL), both of which are widely used in Korea.

The first criticism was that the TOEFL and the TOEIC had a strong cultural bias, and my second criticism was that these tests did not properly assess the proficiency of learners who need to use the language in specific or real-life sociolinguistic contexts. My third criticism was that the tests did not completely cover all four language areas (reading, listening, writing and speaking) at the time I was writing this paper.

I discussed those criticisms and concluded with some suggestions for Korean EFL tests in the global community. The first of these was a recommendation that the local community - in terms of characters, place names, and references – be included in the tests. Another suggestion was that highly localized forms of tests which are descriptive of the language practices of specific individuals functioning in specific social contexts be more carefully constructed. Lastly, I argued for the establishment of a more direct form of an oral component in the tests.

In the process of writing this essay, I was able to identify certain weak points of these tests for Koreans who want to be a part of the global community. The findings were informative and useful for me in terms of improving my own expertise in the area of EFL testing. An additional benefit was that as a result of writing this essay, my views of the English language and its testing changed and broadened.

This paper was published in *Language and Literature Research*, 14: 83-97 in

February 2004.

Institute-Focused Study (IFS) and thesis

Ability in the English language has increasingly become an asset in the highly competitive global market. English speaking skills in particular have become one of the primary concerns in Korean English Education. Instruction has focused on speaking skills more than ever before, and English language teachers have been encouraged to conduct classes only in English. However, the English speaking ability of Korean students has remained far below expectations, and I assume that the failure of English teaching and learning in Korea may be a result of unimproved English testing. The frequent absence of direct oral testing, in particular, has meant that the majority of teachers are not assessing students' speaking skills in a direct way.

I realized this research problem when I started to work on the second paper, and through the discussions of the problem with tutors and peers I have been able to clarify my ideas since. In addition, the findings of my research in the third essay inspired me to study Computerized Oral testing.

In the IFS I explored students' and teachers' perceptions of, and attitudes toward, a Face-To-Face Interview (FTFI) and a Computerized Oral Test (COT) in a Korean

University Context. For data collection, a questionnaire was administered and a qualitative interview was conducted. The results of this study showed that both students and teachers had generally favorable attitudes toward the two tests—the findings of the IFS were published in *Journal of Language Sciences* in June 2007.

The students showed more positive reactions to the COT than to the FTFI in terms of preparation time, nervousness, fairness, difficulty, and tiredness. They felt that the COT was fairer, less difficult, made them less nervous, and less tired, from waiting for their turn, and gave sufficient preparation time before answering questions.

Similarly, the teachers considered that the COT was able to provide more preparation time and produce a longer sample of the test taker's speech with a sufficient number and variety of items without so much constraint of time and efforts, which may lead to greater content and better construct validity.

Whereas the teachers showed clear preference for the COT, the students preferred the FTFI, because they perceived the FTFI was a real communication, interacting with a person.

Whilst an interesting study in itself, I used this paper as preparatory ground for my thesis. Since the IFS was mainly devoted to addressing the main questions of interest outlined, there were several other questions that had not been answered. It was not

known whether there were any meaningful relationships between test takers' attitudes and their actual test performance. Moreover, few studies seem to have examined variables such as computer familiarity, age, and gender, in conjunction with attitudes to, and performance on, the COT in the Korean university context.

Therefore, I developed and extended the IFS in the thesis. First I again intensively explored Korean university students' attitudes to the two test formats again. The results of the study supported and elaborated the findings of the IFS. Then secondly, I investigated students' performance on the tests as it related to computer familiarity, age and gender, and finally attitudes. I hoped that this study would provide useful information to support further study of computerized speaking tests and serve as a useful and practical catalyst for reflection on teaching and assessing speaking abilities, and finally make it possible to select and develop more effective assessment tools for speaking assessment in the Korean context.

Through the studies of the IFS and the thesis, I improved my research skills and professional practice and academic knowledge in my area. In addition, my knowledge about the following computer software programs has been greatly improved and solidified: SPSS, FACETS, N-Vivo, Excel and EndNote.

The professional and academic knowledge gained through research and discussion

with Catherine Walter's and Andrew Brown's comments, suggestions, and advice, were very valuable resources for future studies.

Links between the courses with assignments and professional practice

Before beginning the EdD program I was merely a "general" English lecturer, but I have experienced significant personal changes in terms of academic and professional knowledge since starting the EdD taught courses. The changes have arisen largely as a result of my studies on the EdD to date.

The first course and the assignment were initially difficult for me because I had never seriously considered the concept of professionalism before. As I became familiar with the concept I began to find my identity as an EFL teacher and to reflect critically on my work. This reflection aided my professional development by initiating a clarification of thought and a challenging of beliefs. It was an appropriate moment in my career to consider both my job as a profession and myself as a professional.

The second and third courses with the essays reflected my growing critical perspective in terms of considering and debating issues in EFL, and my improved ability to determine which theoretical frameworks best represent my interpretation of the relevant knowledge. I have found myself more willing to explore research literature,

particularly that addressing the philosophy of educational research. I have also become more comfortable with the vocabulary of research and thus have become a much more critical reader.

The last course and the paper was also valuable in that I was encouraged to see a wider perspective, with the result that I now recognize different perspectives and angles. In short, it transformed my previously narrow perspective of English teaching and testing into a much wider one. This newfound perspective has helped me to see the weaker points of standardized tests such as TOEIC and TOEFL and has improved my knowledge of my main interest, language testing.

Finally, the IFS and the thesis helped me to develop as a researcher by conducting and commissioning research in my work-place. They led me to further develop more research skills and an understanding of the act called research, based on the learning from the taught courses. They were very important parts of my learning process and professional growth as a researcher. All the taught courses and workshops aided the work that I do with my supervisors, and supported various stages of my work and research on the IFS and the thesis.

Through all of the work as the requirements of the EdD programme, I was encouraged to more critically evaluate thoughts and ideas through engagement in self-

assessment, and critical discussions with my supervisors, tutors, and peers. It has been most beneficial not only in terms of my professional practice, but also in changing my earlier tendency to very passively accept the EFL teaching, learning, and testing theories and the policies of schools that I have worked for. In viewing all of my work as a whole, I regard all my work not only as indicative of my progress as a research student but also representative of my personal and professional development. I have been transformed - largely due to having had the space to think and reflect - from a general English teacher into a professional with the responsibility of professional growth as an EFL teacher and researcher, for the field and myself.

CHAPTER ONE. INTRODUCTION

1.1 Research background and rationale

Test developers and decision-makers have primarily considered reliability and (in particular, construct and content) validity in designing and selecting tests. However, it appears that although it is necessary to examine reliability and validity for the evaluation of tests, these may not be enough, because tests often influence more than the assessment they attempt to accomplish. They are known to affect test takers' attitudes and reactions, which may relate not only to their motivation to learn but also to test performance.

Communicative language testing is essentially concerned with making test tasks look as though they are events which could occur in the real world. This is known as 'face validity' or 'test appeal' which is commonly defined as 'the test's surface credibility or public acceptability' (Ingram, 1977: 18, cited in Alderson *et al.*, 1995: 172). This validity appears often to be underestimated or viewed as having only a subsidiary and limited role in designing and developing tests, mainly because of its lack of a basis in empirical data.

However, Skehan (1984: 208) states that 'whereas previously face validity was regarded as merely cosmetic, and of real importance only to the extent that the format of

the test taker antagonise the test taker, now it is regarded as having greater inherent importance'. Considering test takers' feedback will lead to the overall improvement of the test.

If test takers perceive that the test tasks they have been asked to do are not reasonable, or feel that they are engaged in pointless exercises, this may affect the way they perform and hence the scores they obtain. As a result, it may attenuate the test's validity. This is a serious problem both within an educational setting which involves assessing students or evaluating a teaching program, and for testing which is carried out for research purposes, because it limits the usefulness and application of the results.

Hughes (1989) supports this view arguing that if the test which does not look valid is used, the students' attitudes about and reactions to it might mean that they do not perform on it in a way that truly reflects their ability.

Shohamy (1982: 17) also suggests:

Knowing that attitude and anxiety are significant factors in test performance should lead to careful deliberation before applying an evaluation procedure that may have a negative emotional impact no matter how statistically reliable and valid it is. Test developers and users need to consider such factors when constructing and selecting tests.

Therefore, from these perspectives, test taker feedback as well as their performance becomes justifiable, and indeed a valuable source of evidence for test validity. That is,

there should be greater attention on and consideration for, test takers' feedback (attitudes to and opinions about tests) as an important measurement criterion in designing and selecting tests.

Credit for face validity should not be gained by superficially mimicking real communicative events. The test's face validity should be examined and increased by being underpinned with a sound principled basis.

There have been a number of studies examining the affective reactions of students to various types of language tests (Savignon, 1972; Brutch, 1979; Scott, 1980; Shohamy, 1982; Scott and Madsen, 1983; Madsen and Murray, 1984; Clark, 1988; Stansfield *et al.*, 1990; Stansfield, 1991; Brown, 1993; Shohamy *et al.*, 1993).

Despite a great deal of research concerning test takers' attitudes and reactions to a tape-based test compared with those to a live interview, only a few have been undertaken about test takers' attitudes toward '*a computerized oral test*' which may be an improved test format over a tape-based test in terms of its efficiency and affective influence in test-takers (Kenyon and Malabonga, 2001; Norris, 2001; Joo, 2004) and there has been little research within a Korean university context except Joo's study (2004).

The remarkable development of technology over the past few decades has influenced

not only teaching, but also testing. With the introduction of more powerful desktop computers, there is a movement from conventional pencil-and-paper tests to computerized language tests such as the Computer Based Test of English as a Foreign Language (CBTOEFL) or the Test of English as a Foreign Language Internet-based Test (TOEFL iBT) (Educational Testing Service, 2005) and the Computer Based International English Language Testing System (CBIELTS) (the British Council, IDP Education Australia: IELTS Australia and the University of Cambridge ESOL Examinations, 2006).

The attraction for test administrators and teachers of using computerized tests rather than paper-and-pencil tests seems to come mainly from their efficiency in scoring students' performance, storing and retrieving tests and results, and providing a variety of response elicitation prompts, allowing testers to employ sound, graphics, animations, and even motion video as response elicitation techniques (Larson, 1998). Computerized language tests can yield a wide variety of test items and allow test takers to take the tests at their own pace. Specialized computer algorithms which cause the tests to adapt to the level of ability of the students taking the test have made them even more developed.

With the decreasing costs of computers, increased networking capabilities, and new

video equipment and software, an increasing number of academic institutions and companies have begun to use computers to deliver language tests; computerized tests were being used even in 1998 in over 150 academic institutions, and comments from their users were very positive (Larson, 1998). It seems clear that the use of computerized tests for language assessment and other educational and occupational assessment purposes will become increasingly predominant in the near future.

With the growing worldwide interest in using computers in language education, a number of Korean universities appear to have recognized the efficiency of the computer and have been using it for teaching and learning, and further there have been continuous interests in computerized language tests, especially for the assessment of students' speaking abilities (e.g., Sookmyung Women's University, no date; Yonsei University, 2005) since a face-to-face oral interview in a Korean university context, where a large number of students have to take an exam in a short time, and which requires numerous rooms, facilities and a great deal of time to interview individual students, was not practical. Because of this impracticality, even in an English conversation class which is a compulsory course of instruction for all freshmen at most Korean universities (the Korean Ministry of Education), many teachers were unlikely to test students' speaking ability in a direct test format (Joo, 2003).

Students' attitudes and reactions are believed to be one of many important issues involved in computerized tests and are seen as being the first consideration before employing a new type of a test. Therefore, this study intensely investigated Korean university students' attitudes to a Face-to-Face Interview (FTFI) and a Computer Administered Oral Test (CAOT) first and then secondly, their performance on the tests, and finally their effects on performance on the two tests in the Korean university context. The findings of this study will provide useful information to support further study of computerized tests and to serve as a useful and practical catalyst for reflection on teaching and assessing speaking abilities, and finally make it possible to select and develop more effective assessment tools for appropriate oral assessment in the Korean context.

1.2 Past studies on test takers' attitudes to direct and semi-direct tests

There have been a great number of studies which have explored the test takers' affective attitudes and reactions to a variety of types of language tests. Some of the studies which examined the test taker attitudes and reactions to direct and semi-direct tests are briefly referred to below.

Clark (1988) explored the reactions of 27 test takers after the Oral Proficiency

Interview (OPI) and the first tape-mediated Simulated Oral Proficiency Interview

(SOPI) in Chinese. Although the SOPI provided scoring results broadly equivalent to those of the live interview (OPI), 63% of test takers felt more nervous on the SOPI (the tape based test), 19% were more nervous on the OPI and 18% were equally nervous on the two tests. 78% of the test takers perceived the SOPI as more difficult, 7% found the OPI more difficult and 15% found the two tests equally difficult.

On the other hand, 74% of the test takers on the OPI and 70% on the SOPI thought that their speaking ability had been reasonably covered on both tests. While 70% of the test takers felt the tape test was fair, 96% perceived that the OPI was fair.

According to individual test taker comments, the major problem on the SOPI was the presence of several tasks which were too difficult for many lower-level test takers.

Finally, 89% of the test takers preferred the OPI, but only 4% the SOPI and 7% had no preference.

Clark concludes that face-to-face interviews are preferred whenever the necessary resources can be made available, but that when an alternative approach is needed, test takers involved will generally consider themselves effectively tested through semi-direct means, although as a second-choice procedure.

Stansfield *et al.* (1990) in the same way examined the attitudes of 30 test takers who

completed a questionnaire after undertaking the OPI and SOPI in Portuguese. The test takers strongly preferred the live interview in spite of a correlation of $r = 0.93$ between scores on the two tests.

In this instance 69% of the test takers felt more nervous on the SOPI (a tape test), 24% more nervous on the OPI and 7% equally nervous on the two test formats. A total of 90% felt that the tape test was more difficult and 10% felt it was equally difficult.

Individual comments suggested that the tape test format was felt to be an unnatural context in which to be tested. Overall, 86% of the test takers preferred the live interview, 7% preferred the tape test and 7% had no preference. However, most of the test takers perceived that both of the tests were fair. All test takers felt the questions on the live interview were fair and 83% felt those on the tape test were fair. Moreover, 73% of the test takers believed that their maximum level of Portuguese had been explored under both test conditions.

In a subsequent study, Stansfield (1991) investigated feedback from the OPI and SOPI test takers in a range of languages. The study found that while most preferred the live interview because of the human contact, about a quarter either preferred the tape format or felt there was no difference. Some test takers preferred the tape test because they were not as nervous talking to a tape-recorder as to an unfamiliar and highly

competent speaker of the target language. In general, Stansfield comes to the conclusion that the SOPI may cause excessive stress because speaking into a tape recorder is a less natural situation than talking to a person directly.

Somewhat different results from the previous studies, which reported that test takers generally favour the semi-direct format for speaking tests, were found in Brown's study (1993). Brown (1993) explored test taker feedback in the development of the Occupational Foreign Language Test (Japanese), a tape-mediated test of spoken Japanese for the tourism and hospitality industry in Australia.

The study found that 57% of the test takers (n=53) liked the tape-based format, 25% disliked it and 18% were neutral, even though many of the test takers had had little or no experience in recording their voices in a language laboratory, particularly in a test situation. However, the test taker attitudes to the tape test might not have been so favorable if the test takers had been given the chance to compare their attitudes to this test with a live interview. Negative reactions to the test seemed to be mostly the result of insufficient response time, even though the rater used in the study felt that the responses, were in general, sufficient to rate the test takers' speaking levels precisely.

Although there have been numerous studies concerning test takers' attitudes about live and tape based tests, as mentioned earlier, there seem to be only a few which

examine computerized oral tests (e.g., Kenyon and Malabonga, 2001; Joo, 2004).

Kenyon and Malabonga (2001) examined 55 test takers' attitudinal reactions to taking tests assessing oral proficiency in different formats across three languages (Spanish, Arabic, and Chinese): OPI, SOPI, and COPI (Computerized Oral Proficiency Instrument) (Center for Applied Linguistics, 2002). The study showed high correlations between COPI, SOPI, and OPI ratings for the same test takers: a correlation of $r = .95$ between the COPI and the SOPI; $r = .92$ between the COPI and the OPI; $r = .94$ between the SOPI and the OPI.

Kenyon and Malabonga (2001) focused primarily on differences between the SOPI and the COPI. Test takers showed generally positive reactions to the three test formats, but they preferred the COPI to the SOPI.

It was found that 56.4% test takers experienced more difficulty on the SOPI, 29.1% found it more difficult on the COPI and 14.5% equally difficult on both. 20% of the test takers felt the SOPI was fairer, 47.3% that the COPI was fairer and 32.7% that the tests equally fair. While 47.3% of the test takers felt more nervous on the tape based test (the SOPI), 32.7% were more nervous on the COPI and 18.2% equally nervous on the two tests (1.8% were missing). 10.9% of the test takers thought the tape test (the SOPI) had a clearer direction, 34.5% thought the computer (the COPI) had a clearer direction and

54.5% perceived a clear direction on both tests. Finally, 60% of the test takers perceived that they had more opportunity to adequately demonstrate both their strengths and weaknesses in speaking on the COPI, but only 25.5% felt so on the SOPI and 14.5% felt so on both.

Kenyon and Malabonga conclude that overall, the COPI functioned better in the area of its influence on test taker affect than the SOPI, despite the high correlation between the two tests.

More recently, I examined 43 Korean university students' and 13 teachers' attitudes and reactions toward a Face-To-Face Interview (FTFI) and a Computer Administered Oral Test (CAOT) in a Korean university context (Joo, 2004).

Compared with the attitudes to a tape-based test in the previous studies, Korean university students relatively had much more positive attitudes to the CAOT. 63% of the test takers participating in the study felt more nervous on the FTFI (the live test), 21% on the CAOT (the computer test) and 16% equally nervous on both tests. 54% of the test takers perceived that the FTFI was more difficult, 35% found the CAOT more difficult and 12% found the two tests equally difficult. 29% of the test takers perceived the live test was fairer, 49% fairer on the CAOT and 22% equally fair on both.

On the other hand, 52% of the test takers wanted to take the FTFI again, but 38%

wanted to take the CAOT again and 10% wanted to take both again. 51% of the test takers felt more comfortable talking with a person and similarly 47% felt more comfortable with a computer, and 2% felt equally comfortable on both tests. Finally, 63% preferred the FTFI, but 35% the CAOT and 2.3% the both tests.

On the contrary, teachers showed a clear preference for the CAOT. 62% of the teachers felt more comfortable on the CAOT, 15% on the FTFI, and 23% were equally comfortable on both. 69% felt the CAOT was easier to use for assessing test takers' performance, 15% felt the FTFI was easier, and 15% felt both were equally easy. 92% of the teachers that the CAOT as more practical, but only 8% thought both were equally practical. 62% thought that the CAOT was easier to use for eliciting the test takers' speech sample, 31% thought the FTFI was easier and 8% thought that they were equally easy. 54% of the teachers felt the CAOT was easier to administer, 15% felt that the FTFI was easier to administer, and 31% felt both were equally easy. 69% thought that the CAOT was better for using as their exam, 15% thought that the FTFI was better, and 15% thought both were equally good. 77% of the teachers perceived that the CAOT was fairer and 23% perceived they were equally fair. Finally, 54% preferred the CAOT and 31% preferred the FTFI and 15% equally preferred the two tests.

In brief, students and teachers had positive attitudes toward both the FTFI and the

CAOT, but overall, teachers preferred the CAOT, while students preferred the FTFI even though some negative attitudes to the FTFI were expressed.

The tables summarizing the findings of the two studies (Kenyon and Malabonga, 2001; Joo, 2004) are shown below.

Table 1.1 Kenyon and Malabonga’s study on test taker affective reactions (2001)

Question	SOPI	COPI	Same	Missing
1. Which test did you feel better allowed you to demonstrate both your current strengths and weaknesses in speaking?	25.5%	60%	14.5%	0.0%
2. Which test did you feel was more difficult?	56.4%	29.1%	14.5%	0.0%
3. Which test did you feel was fairer?	20.0%	47.3%	32.7%	0.0%
4. Which test did you feel more nervous taking?	47.3%	32.7%	18.2%	1.8%
5. Which test did you feel had clearer directions?	10.9%	34.5%	54.5%	0.0%
6. Which test do you feel would give someone listening to your performance a more accurate picture of your current ability to speak in real life situation outside the classroom?	29.1%	45.5%	21.8%	3.6%

Table 1.2 Joo’s study on test taker affective reactions (2004)

Question	FTFI	CAOT	Same
1. Which test made you feel more nervous?	63%)	21%	16%
2. Which test do you want to take again as your exam?	52%	38%	10%
3. Which test was more comfortable for you-talking to a computer or to a person?	51%	47%	2%
4. Which test was more difficult for you?	54%	35%	12%
5. Which test was fairer?	29%	49%	22%
6. Which test do you prefer?	63%	35%	2.3%

The two studies are important in the light of the concern that has been raised that features of the computerized context may negatively influence test-takers’ attitudes,

which may then affect their performance on the test and finally, reduce the overall test validity.

However, neither study examined the relationship between test takers' attitudes toward the two test formats and their actual test performance. They have addressed the effects of testing from the perspectives of the students (their attitudes) without linking these to their test performance. This is despite the fact that there could be a link between test takers' affective attitudes and their performance, as suggested by several researchers (e.g., Bachman and Palmer, 1996; McNamara, 1996). Moreover, few studies seem to have examined variables such as computer familiarity, age or gender in conjunction with attitudes and performance in the Korean university context.

Therefore, in this study, I focused primarily on examining test takers' attitudes toward and performance on a Face-To-Face Interview (the FTFI) and a Computer Administered Oral Test (the CAOT) in the Korean university context. I examined both their affective attitudes and their test performance in relation to computer familiarity, gender and age.

In examining test performance, test takers' scores were explored and their performance on the two tests was compared to see whether the tests were equivalent, as the comparability of the performance on direct and semi-direct tests is one of the most important matters in terms of (construct) validity. The scores were further examined to

see whether the scores and performance were associated with other possible intervening factors such as rater severity, test orders, and bias, to give a better understanding of the relationships between the scores and these factors. If the scores are seriously compromised by such intervening factors the analysis of the relationship between the scores and attitudes might not be accurate and valid.

The findings of this study will be worthwhile for assessing the tests' construct validity as a unitary concept including face validity (see more detail in 2.2.1) and will contribute to knowledge about the implications of adopting computerized oral test formats.

1.3 Research questions

In order to explore the attitudes and performance of Korean university students to a Face-To-Face interview (FTFI) and a Computer Administered Oral Test (CAOT), and the effects of attitudes on performance within a Korean University Context, three main research questions were formulated.

Research Question One: 'What are Korean university students' attitudes toward a Face-To-Face Interview (FTFI) and a Computer Administered Oral Test (CAOT)?'

For this research question, attitude questionnaires completed by test takers after the

two tests were analyzed, and the test takers' attitudes were examined and compared to see if there were any significant attitude differences between the tests. Test-takers' attitudes were further explored through qualitative interviews in order to understand their feelings and thoughts about the tests more deeply. Their attitudes were also explored and linked to computer familiarity, age and gender. The students' computer familiarity was explored first and to see if there were any differences in attitudes according to the level of computer familiarity, and further examinations were conducted to discover whether attitudes and computer familiarity were influenced by age and gender. Therefore, the sub-research questions addressed to the main Research Question One were:

1. How do Korean university students perceive and react to the two test formats?
2. To which test format do the students have more positive attitudes overall?
3. Are there any significant attitude differences between gender and between age groups?
4. Are there any significant computer familiarity differences between gender and between age groups?
5. Is there any significant relationship between computer familiarity and attitudes?

Research Question Two: 'What are the students' performance on a FTFI and a CAOT?'

For this research question, using FACETS program I examined how test takers performed on the two tests and the effects of other possible intervening factors which may have impacted upon the performance or scores. (e.g., raters' different severity, the effect of test taking order, and bias between raters and test formats). Furthermore, the effects of computer familiarity and age and gender differences on performance were also examined. Thus, the sub-research questions addressed for main Research Question Two were:

1. Are the test takers' speaking abilities discriminated well?
2. Are the raters' ratings reliable and is there a difference in their severity?
3. Are the two tests equally difficult?
4. Is there a test order effect?
5. What is the item difficulty of the two tests?
6. Is there a bias between raters and test formats?
7. Is there a relationship between computer familiarity and test performance?
8. Do gender and age differences impact on performance on the two tests?

Research Question Three: 'Do the students' attitudes influence their performance on a FTFI and a CAOT?'

Finally, for this research question, the relationship between attitudes and performance were examined and whether positive or negative attitudes affected test performance

were discussed.

This thesis consists of five chapters: Chapter One introduced the research background and rationale, emphasizing the importance and limitations of the research. Chapter Two presented the theoretical background of characteristics of a good test; washback, validity, reliability, practicality and the testing of communicative skills were defined, and several types of semi-direct oral tests including Computer Administered Oral Test (CAOT) were examined and discussed. Chapter Three described the research methods used to explore the research questions. The design, data collection procedures and analysis were also described and discussed. In Chapter Four, the data were analyzed to find the answers to the research questions. Finally, in Chapter Five, there was a discussion of the results according to the research questions and conclusions were drawn and suggestions for further study were made.

1.4 Limitations of the study

Some limitations affect the generalisability of the study results; thus some caution is warranted when considering the current findings.

First, as the sample size which was drawn only from two classes at two Korean

universities was small, it might not have been representative or provided sufficient data for the study. Moreover, it should be noted that the sample only included Korean university students who were first year undergraduates having a similar educational background and high computer familiarity. Thus, in a more diverse and larger sample, a more convincing and distinct or even different set of results might have been arrived at.

Secondly, a simulated test setting was used for the research. Ethically, it was not possible to use the results of the tests utilized for the research as the test-takers' grades because the test tasks and items did not sufficiently focus on and cover all the content of the textbook they were using for the English conversation class during that semester. Instead, the tests were cautiously constructed to match each other for valid results for this current study. Therefore, the students might have been less motivated to perform well on the two tests for the research than on the actual course test for their grades, even though I attempted to motivate them by using an incentive and encouragement. There is a possibility that different forms of motivation cause varying attitudes and performance.

Thirdly, there was, as mentioned above, an artificial attempt to equate the two test formats in order to get solid and valid data about participants' attitudes to the two oral tests by using the same/similar questions and the same elicitation techniques. Thus, if the study had conducted in a more natural Korean university setting where teachers give

fewer questions mostly with an 'ask and answer' elicitation technique in the FTFI (Joo, 2003), the students' attitudes to the CAOT might have been more positive.

Lastly, with respect to the attitude questionnaire, there is some doubt whether or not the items covered all the topics such as test validity, anxiety, perceptions, and test procedures. Furthermore, in view of a possible halo effect, the validity of the items as true reflections of students' attitudes and perceptions may be questioned.

CHAPTER TWO

THEORETICAL BACKGROUND

In this chapter, I first discussed the relationship between teaching and testing and defined and then examined each component of the qualities of good tests - validity, reliability and practicality, which test developers and teachers have considered for developing their tests. I then progressed to exploring and discussing communicative language ability and communicative language testing, including the difficulties of implementing the FTFL. Following these, computerized oral tests including Computer Administered Oral Test (CAOT) which was used for this study, were described.

2.1 Teaching and Testing

New language-teaching methods introduced during and immediately following the Second World War led teachers to change their order of priorities, and the resulting present-day emphasis on the spoken form of the language is now reflected in teachers' testing as well as in their teaching of second languages. Teaching and testing are interrelated. If a test is regarded as valid and important, then preparation for it can come to dominate all teaching and learning activities. This effect of testing on teaching and learning prior to the assessment itself, is known as washback (or backwash) (Brown,

2004). It can be either harmful or beneficial. For example, if the skill of speaking is tested only by multiple choice items, which appear to be frequently used in Korean university classes (Joo, 2003), students will be taught and learn such items, rather than the skill of speaking itself. This is a case of harmful washback, since these endeavors would probably not promote the student's abilities to use the target language for their day-to-day communication needs; whereas if an oral interview is administered to assess that skill, there may be greater practice of it. This would constitute a case of beneficial washback.

The difficulty of assessing speaking ability is, however, something quite common that we teachers have encountered in Korean universities particularly because of its impracticality. The impracticality was discussed more in 2.2.3.

Tests have often tested what is easy to test rather than what is important to test (Heaton, 1975). That is, the tests were often for teachers marking rather than for students taking the tests, and this may be the reason for the frequent absence of a performance-based speaking test (especially, a face-to-face interview), which requires subjective scoring and takes considerable time and resources to administer and mark (Weir, 1988).

However, I propose to demonstrate in section 2.2.1 that performance-based oral tests

are desirable unless the reasons for not administering them are stronger than the washback effects that they tests can have on the teaching and learning that occur before the tests.

2.2 Qualities of good tests

Test developers and teachers should consider the extent to which what they are doing meets a number of general principles that should underlie all good test design: validity, reliability, and practicality (Weir, 1993). These general principles are briefly examined below.

2.2.1 Validity

The most complex criterion of a good test is validity: in general, a test is said to be valid if it measures accurately what it is supposed to measure (Bachman, 1990). Alderson *et al.* (1995: 170) assert that validity issues should be of central concern to all testers, since if a test is not valid for the purpose for which it was designed, then the scores can not mean what they are intended to mean. Through assessing the validity of the test, the test will be made sounder and more efficient.

Messick (1989) regards all different types of validity (e.g., content validity, criterion-

related validity, or consequential validity) as aspects of a unitary concept of validity (construct validity) that includes all of them, but I still find it necessary to look at face validity for this current study.

The test is said to have **face validity** if ‘it appears to test what it is designed to test’ (Brown, 1994: 256). According to Underhill (1983), face validity is used to refer to the reactions of test takers and to the extent to which the test jarred their expectations. In face validity we do not necessarily accept experts’ judgments (Brown, 1994). Face validity involves an intuitive holistic judgment about the test’s content by people such as students, administrators, and non-expert users (Alderson *et al.*, 1995). For example, a test which is intended to measure pronunciation ability but which does not require the student to speak may be thought by the students to lack face validity.

As stated earlier, face validity, appears to be frequently dismissed by teachers and test developers because it is unscientific. Yet its importance lies in the degree to which it may have a significant effect on test performance and results, thus possibly decreasing the overall test validity. Hamp-Lyons and Lynch (1995) also assert that language testing researchers should give great consideration to the voices of test takers, interviewers, raters, test users and other interested groups in the validation process (i.e., face validity), rather than simply listening to themselves.

On the other hand, **construct validity** assumes the existence of the theoretical construct as it has been defined (e.g., communicative language ability) (Heaton, 1975). Traditionally construct validity has been distinguished from face or other types of validity in that it attempts to take a scientific approach to measurement, for which empirical evidence can be presented. Messick (1989), however, views the term construct validity as a dominant term which includes all aspects of validity. Messick argues that each separate type of validity is not enough by itself and each of them is ‘a complementary form of evidence to be integrated into an overall judgment of construct validity’ (Messick, 1998: 37). Messick (1989: 13) describes validity as ‘an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores’. In other words, construct validity is related to the meaningfulness, appropriateness, and usefulness of the interpretation or use of test scores (Bachman and Palmer, 1996).

Test scores should be a meaningful indicator of a particular individual’s ability we want to measure and very little else, but they seem to be easily affected by the factors that are irrelevant to the ability being tested (i.e., construct irrelevant variance) (Messick, 1989; McNamara, 2000). McNamara (2000: 53) asserts that ‘there may be factors in the

test which will cause performances to be affected, or to vary in a way which is not relevant to the information being sought about candidates' abilities.'

In this current study, apart from test takers' attitudes, a number of sources of variability in the oral performance assessment situation may influence the outcomes for the test takers, with potentially serious results. For example, particular raters may react in different ways to a particular test (e.g., a Face To Face Interview (FTFI) or a Computer Administered Oral Test (CAOT)). Or it may be that the test situation has an influence: an unfamiliar and uncomfortable test situation may influence the test takers' chances of a particular score. In the CAOT, the test takers interact not with a teacher but with a computer. The stimulus for speaking is not a face-to-face interaction but is presented to the test takers as a computer recording, in a computer room. The interaction is computer recorded and the performance is rated from the recording.

All of these aspects of the test situation may influence the outcomes of the test takers and as a result, the scores may not be valid and reliable as a guide to test taker ability. Therefore, it is also necessary to consider these aspects to make this current study sound and valid.

2.2.2 Reliability

Reliability is the extent to which we can depend on the results that the test produces (Brown, 1987). The type of reliability that we are most concerned with in an oral or writing test is rater reliability. Rater reliability refers to the consistency of scoring by two or more different raters (inter-rater reliability) or by the same rater on different occasions (intra-rater reliability) (Brown, 1994; Alderson *et al.*, 1995).

Rater reliability is likely to be low when subjective scoring is required. For example, the scoring of an oral test which requires subjective scoring would be more difficult than that of an objective test comprising chiefly multiple choice items because raters are required to make judgments which are much more complicated than objective scoring. The difficulties of scoring of an oral test were discussed more in 2.3.2.1.

One of the best ways to make the subjective assessment more reliable is rater training. Rater training helps raters to understand and familiarize themselves with the rating scales that they must apply competently and confidently (Alderson *et al.*, 1995). The training reduces rater-related score variance. Hence, in this study a short one day rater training was carried out before rating actual performance.

Fulcher (2003), however, insists that it is not possible to remove completely the differences between raters as different rater severity may still exist because of different

interpretations of scales and other reactions which are not relevant to those scales. Rater differences can be reduced by training, but substantial severity differences between raters may still persist (McNamara, 1996). This rater severity is also one of the factors which may influence the outcomes for the test takers stated above. For that reason, this current study attempts to examine not only inter and intra-rater reliabilities but also the severity differences between raters.

2.2.3 Practicality

Practicality is concerned with matters of 'financial limitations, time constraints, ease of administration, and scoring and interpretation' (Brown, 1994: 253). A good and effective test is practical. If the test takes a long time, and/or is expensive to construct or difficult to mark, it is said to be impractical. It would be hard for teachers to administer such a test. For instance, a FTFI, which takes a few minutes for a student to take and several hours for a teacher to administer and score, is impractical. For this reason, many English conversation instructors in Korean universities appear to be reluctant to administer oral tests such as a FTFI. Instead, they seem to give listening tasks or multiple-choice items for assessing speaking skills. In this case teachers can maximize practicality and reliability but reduce validity.

Teachers are always faced with limited resources for developing, administering, and marking tests, and they need to be offered pragmatic test methods that can be used within the constraints of classroom time and space and school curriculums. Practicality was examined further in 2.3.2.1.

2.3 Communicative language testing

2.3.1 Communicative language ability

If language tests are to be adequately designed and implemented, then it is necessary to have a theoretical framework on which to base an understanding of language proficiency.

The term ‘communicative competence’ was coined by Hymes in order to contrast a communicative view of language with Chomsky’s theory of competence (Hymes, 1972). For Chomsky, ‘the focus of linguistic theory was to identify the theoretical abilities speakers possess that enable them to produce grammatically correct sentences in a language’ (Richards and Rodgers 1986: 159). Hymes insisted that such a view of linguistic theory was barren, that ‘linguistic theory needed to be seen as part of a more general theory including communication and culture’ (Richards and Rodgers 1986: 159). For Hymes, communicative competence included the ability to use the language, as well

as having the knowledge which underlay actual performance. His statement, 'there are rules of use without which the rules of grammar would be useless', expresses his view of language clearly (Hymes, 1972).

One of the most comprehensive models of communicative competence is Canale and Swain (1980) and Canale (1983)'s model. They identified four different components of communicative competence: grammatical competence, sociolinguistic competence, discourse competence, and strategic competence. According to them, grammatical competence is concerned with mastering the linguistic code (knowledge of the rules of grammar and lexis). Sociolinguistic competence deals with the appropriateness of utterances and takes into account the appropriateness of what is said, as well as the social context in which the utterance would be deemed appropriate. Discourse competence is the ability to create unified texts, either spoken or written, in different genres, such as a personal letter, a speech, or a narrative. Strategic competence is the use of verbal and non-verbal communication strategies that are called into action to 'compensate for breakdowns in communication due to performance variables or to insufficient competence' (Canale and Swain, 1980: 30).

Canale and Swain argue that in order to have such communicative competence the student must 'have the opportunity to take part in meaningful communicative

interaction with highly competent speakers of the language, i.e. to respond to genuine communicative needs in realistic second language situations' (Canale and Swain, 1980: 27).

Canale and Swain's definition of communicative competence has been modified over the years. In particular, Bachman (1990) contributes to it, using the term 'communicative language ability'. Bachman's framework is consistent with models of Canale and Swain, and Canale's later version. Bachman asserts that:

communicative language ability can be described as consisting of both knowledge, or competence, and the capacity for implementing, or executing that competence in appropriate, contextualised language use (Bachman, 1990: 84).

This is essentially the same as earlier models of communicative competence, and Bachman chose to use the phrase 'communicative language ability' rather than the more common term, 'communicative competence' in order to stress the fact that he is not only interested in 'competence' but in 'performance'.

Bachman describes 'communicative language ability' as being composed of language competence, strategic competence and psycho-physiological mechanisms. Bachman regards language competence as including 'organizational competence', which is itself made up of 'grammatical' and 'textual' competence. The first of these, grammatical competence, includes vocabulary, morphology, syntax and phonological/graphical

competencies, while the second, textual, includes cohesion and rhetorical organization (Bachman, 1990). Bachman argues these competencies are responsible for the ability to control the formal features of a language and which allow the formation of correct sentences and texts. The other major component of language competence is 'pragmatic competence', which is made up of 'illocutionary competence' and 'sociolinguistic competence'. The former he defines as 'the knowledge of the pragmatic conventions for performing acceptable language functions', whilst the latter is defined as 'knowledge of the sociolinguistic conventions for performing language functions appropriately in a given context' (Bachman, 1990: 90). Bachman's model also includes recognition of the importance of strategic competence. He acknowledges the importance of the work of Canale and Swain, but suggests it is limited because it does not describe how strategic competence operates (Bachman, 1990). Finally, Bachman includes psychophysiological mechanisms as a necessary component of communicative language ability. These are the neurological and physiological mechanisms that allow the execution of language ability, such as sight, hearing and speaking.

Bachman is keen to stress that it is not really possible to isolate the individual factors and that much inter-relatedness exists and a great deal of interaction occurs. Talking specifically about language competence, he notes that 'indeed, it is this very interaction

between the various competencies and the language use context that characterizes communicative language use' (Bachman, 1990: 86).

Although Bachman's model is not very different from earlier models of communicative competence, it seems innovative in its attempt to bring earlier components of communicative competence together into a coherent whole and represent the processes by which the various components interact with each other and also with the context in which language use occurs.

2.3.2 Communicative language test

An essential element of communicative testing is the shift in emphasis from the purely linguistic to the communicative dimension. As seen above, Canale and Swain (1980), Canale (1983), and Bachman (1990) all emphasize the importance of both language competence and performance in their models. Accordingly, communicative language tests must be concerned not only with what the student knows about the form of the language and how to use it appropriately in contexts of use (competence), but also the extent to which the student is actually able to perform this knowledge in a meaningful communicative situation (performance) (Canale and Swain, 1980).

Brown (1987:31) characterizes this distinction between competence and performance

by saying that competence is 'one's underlying knowledge of the system of language such as rules of grammar, vocabulary, all the pieces of a language, and how to fit those pieces together', whereas 'performance is concrete manifestation or realization of competence': it is the actual doing of something (e.g., speaking, reading, listening and writing).

In other words, communicative language tests emphasize measuring the ability or capacity to use language communicatively, which involves 'both competence and demonstration of the ability to use this competence' (Weir, 1988:9). This implies that communicative tests should be as direct (or performance based) as possible.

While indirect testing attempts to measure the abilities which underlie the skills in which teachers are interested, direct or performance-based testing requires students to perform accurately the skill which the teachers wish to assess (Hughes, 1989).

Indirect tests may be more generalisable than direct tests because 'they offer the possibility of testing a representative sample of a finite number of abilities which underlie an indefinitely large number of manifestations of them' (Hughes, 1989: 16). They are often favored for reasons of practicality. For example, in composition tests, indirect tests may include grammar, vocabulary, handwriting, punctuation, etc., but will not be able to predict the students' real composition abilities, even if teachers make sure

that the composition scores are reliable by taking many samples. On the other hand, because a direct or performance-based test involves actually performing the criterion behavior in a real-life situation, teachers can straightforwardly assess and interpret students' abilities which they wish to measure (Weshe, 1981). In other words, teachers can expect more accurate estimates of students' language abilities which they want to measure through direct tests rather than through indirect tests. The main problem of the indirect tests is that the relationship between performance on them and performance of the skills in which teachers are usually more interested tends to be rather weak and uncertain (Hughes, 1989).

In speaking tests, a more communicative and direct test form may be a FTFI which involves the student in communicative interaction which takes place under pressure of time and therefore necessitates rapid language processing and production. It is possible, more than with tests in other language skills, to build the dynamic circumstances of real-life communication into the FTFI format and there appears to be scope for meeting some of the criteria for communicative testing. Some such criteria are set out by Weir (1990: 9):

The performance tasks that students are faced with in communicative tests should be representative of the type of task they might encounter in their own real-life situation and should correspond to normal language use where an integration of communicative skills

is required with little time to reflect on, or monitor language input and output.

This indicates that the language production elicited by the tasks should be in some measure unpredictable and demand real-time processing, and the context in which they are set should be realistic.

Authenticity can be seen 'in terms of the extent to which a test task relates to the context in which it would normally be performed in real life', but there are no intrinsic and absolute criteria for an authentic or inauthentic test task (Leung and Lewkowicz, 2006: 214-215). Ultimately, it is a matter of perception and some may perceive a particular task as authentic, while others view as inauthentic (Lewkowicz, 1997).

Many researchers (e.g., Clark, 1979; Underhill, 1987; Van Lier, 1989) view that direct tests authentically reflect the conditions of the most common form of real world communication, face-to-face interaction, but Clark (1979) also acknowledges that the FTFI format does not fully reflect real-life conversational situations. He argues that 'in the interview situation, the test taker is certainly aware that he or she is talking to a language assessor and not to a waiter, taxi driver, or personal friend' (Clark, 1979: 38). Besides, it is difficult to have true interaction between the interviewer and the interviewee and elicit the language in real life conversation because the interviewer normally asks questions and controls the conversation, and the interviewee plays the

subservient role answering the questions and responding to prompts initiated by the interviewer (Stansfield, 1991).

Van Lier (1989) supports this view noting that the FTFI format aims to elicit language rather than to make successful conversation. Van Lier (1989: 495-496) distinguishes the essential features of the FTFI format from the conversation: 'the interviewer has a plan and conducts and controls the interview largely according to that plan,' while a conversation is characterized by 'face-to-face interaction, unplannedness (locally assembled), unpredictability of sequence and outcome, potentially equal distribution of rights and duties in talk, and manifestation of features of reactive and mutual contingency'.

In spite of the lack of authenticity of the FTFI format, the FTFI has been considered to be a more authentic test form than any other types of oral tests in terms of its validity, and the FTFI format has been preferred and recommended to measure students' speaking ability (Clark, 1979; Underhill, 1987; Van Lier, 1989). Clark (1979: 38) asserts that 'the face-to-face interview appears to possess the greatest degree of validity as a measure of global speaking proficiency' and is clearly better than any other forms of tests.

There are, however, some constraints and problems for implementing the FTFI,

especially if it is undertaken on a large scale. These are considered in the following section.

2.3.2.1 Constraints of a Face-To-Face Interview (FTFI)

The problem of extrapolation

In a FTFI, a teacher may assess a student's oral performance in terms of a particular task, for example, the speaker simulated performance whilst at a train station. The difficulty arises thereafter of how the teacher can know whether or not the speaker can perform in other situations, such as a pub or a restaurant? Here, we can see the problem of extrapolation. One might end up describing an impossible variety of situations, which one cannot encompass for testing purposes (Alderson *et al.*, 1981); the lack of content validity.

It seems that the greater the sample of tasks the teacher has, the more reliable his or her predictions of the student's performance will be. But we can immediately see conflicts with the demands of test practicality. As Weir (1988) has pointed out, the larger the sample of tasks and the more realistic the test items, the longer the test will have to be.

From a predictive point of view, traditional language tests which aimed at measuring

mastery of the language code (e.g. phonology, syntax) would seem to be far more powerful because the grammatical and phonological systems of a language are limited and controllable and the lexical resources can be delimited (Morrow, 1977). Thus, grammar appears to have much more generalisability from test results than any other language feature. It is, however, argued that the final goal of one's mastery of English grammar is to be able to communicate in English in a real-life situation. In short, English grammar will be useless if there is no actual use of it.

Communicative language ability involves more than a mere manipulation of certain syntactic patterns with certain lexical items. For this reason, it seems that there should be more attempts to devise measuring instruments which can assess performance ability as much as possible.

The problem of assessment

A FTFI gives rise to subjective scoring, but teachers can see that subjective assessment is much more difficult than objective assessment in terms of reliable scoring. For instance, it is very questionable whether the scoring of students' performance can be considered satisfactory if different raters assess their abilities in the same performance differently, or if a different mark is given when the same performance is re-marked at a

different time (Hughes, 1989).

Many testers seek objective scoring, not for itself, but for the greater reliability that it brings. Hughes argues that generally the less subjective the scoring, the greater consistency there will be between two different markers.

Subjective assessment is, however, desirable if teachers want to measure the ability to integrate different language skills in ways which more closely estimate the actual process of language use.

The problem of practicality

Tests should be easy and cheap to construct, administer, score and interpret because a valid and reliable test is of little use if it does not prove to be a practical one (Weir, 1990). Plenty of evidence, however, shows that performance tests such as a FTFI take much longer, are much more expensive to administer and require greater skills on teachers' part (Breen *et al.*, 1997; Clarke and Gipps, 2000). In the FTFI, the individual administration and the marking of the individuals' performance will take a great deal of time. In addition, the duration of the test may affect its successful operation in other ways (e.g., a fatigue effect on the students and the teachers), and further the training of teachers, which enables them to assess students' performance reliably, will also be

costly (Hughes, 1989). Weir (1990: 35) also mentions the problem of practicality:

Tests such as an oral test are difficult and time consuming to construct, require more resources to administer, demand careful training and standardisation of examiners and are more complex to mark and report results on. The cost of such tests in large scale testing operations may severely restrict their use.

For these reasons, there have been efforts to seek and develop test formats and assessment criteria that provide an overall balance of reliability, validity and practicality in the assessment of communicative speaking skills.

In one of the efforts to reduce these difficulties of implementing the FTFI, semi-direct oral tests which are designed to approximate as closely as possible the linguistic content and manner of operation of a live interview have been developed. These are discussed in the following section.

2.4 Computerized Oral Tests

Since the price of computers has decreased and their cost keeps on falling, computers are being used more and more in language classrooms for Computer Assisted Language Learning (CALL) (Ahmad *et al.*, 1985). It seems that reading, grammar, and vocabulary exercises have been devised to allow students to focus upon areas relevant to their own needs: computers have been able to be made sensitive to the students' pace, pattern of

responses, and so on, and can adjust the linguistic material to the needs of the individual (Ahmad *et al.*, 1985).

As machines and software have become more sophisticated, exercise formats have become more varied. Computers can communicate with the students visually and auditorily by employing text, sound (e.g., speech, music or other audio output), graphics, animations, and even video images (Larson, 1998). Computers have been seen to have a significant role to play in language learning and teaching (Ahmad *et al.*, 1985).

Computers have been employed for language testing since 1935 (Fulcher, 2000). Teachers use computers for a variety of testing purposes such as for testing classroom achievement, deciding course placement levels, diagnosing language problem areas, assessing performance and proficiency, as well as assessing students' grammatical structures, vocabulary, and cultural knowledge (Larson, 1998).

Computers have come to play a central role in test design and construction, item banking, test administration, scoring, data analysis, report generating, research and the dissemination of research (Fulcher, 2000). It seems that the current ease of access of low-cost but powerful microcomputers has made test delivery by computer both useful and attractive.

In spite of a plethora of computerized reading and listening tests, there seem to be

relatively few computerized oral and writing tests, perhaps because of the difficulties in developing oral and writing language tasks such as role plays, interviews or compositions for computerized testing. Computer capability of judging whether or not a particular answer is correct limits the kinds of tasks which students can be asked to complete.

Therefore, the computerized oral tests existing at present appear to be delivered but not marked by computer except for PhonePass, which is a speaking test conducted over the telephone with a computer. In this example, performance is scored by computer using a speech recognition and analysis program (Douglas, 2000; Fulcher, 2000).

For this study on computerized oral testing, the Oral Testing Software (OTS) (Computer Assisted Language Instruction Consortium, 1996-2006), which is designed to test speaking proficiency by Brigham Young University, was used. The test is self-administered but not marked by computer as are other general computerized oral tests. Consequently, it is not a Computer Based Test (CBT) defined as a test which is capable both of delivery and scoring by computer (Alderson, 1990; Tuffin, 1990). Nor is it a Computer Adaptive Test (CAT) or a Web Based Test (WBT) because it is not based on Item Response Theory (IRT), which is 'a measurement system that takes account of both test taker and item characteristics, based on probability theory showing the

probability of a given person getting a particular item right' (Alderson *et al.*, 1995: 291), and not delivered via the Internet: Luecht *et al.* (1998: 30) differentiated the CAT from the CBT stating that a CAT is fundamentally 'a CBT with extensive algorithms added to the computer delivery software to govern the selection of items for each test taker'; a WBT is defined by Roever (2001: 84) as 'a computer-based language test which is delivered via the World Wide Web (WWW)'.

Therefore, the computerized oral test using OTS can be called a computerized or computer administered test, which is a very general and broad term, or a computer-assisted test, which is defined as 'the test which is administered at computer terminals, or on a personal computer' (Brown, 1997: 46). For convenience, it has been called a Computer Administered Oral Test (CAOT) in this thesis.

Firstly, I briefly explored some other types of semi-direct oral tests which had been or would be used: the Simulated Oral Proficiency Interview (SOPI) and the Computerized Oral Proficiency Interview (COPI), and some other types of computerized oral tests, and finally I examined the CAOT more fully.

2.4.1 Semi-direct oral tests

A number of semi-direct tests of foreign language speaking ability have been developed to reduce the difficulties of implementing a FTFI, especially the impracticality of the FTFI. Such instruments include, for example, the early MLA-Cooperative speaking tests, the tape-and-booklet mediated speaking tests produced under the support of the ETS Advanced Placement Program, the Test of Spoken English (TSE), developed by ETS as an oral complement to the listening comprehension and reading sections of the Test of English as a Foreign Language (TEFL), and the Simulated Oral Proficiency Interview (SOPI), developed by the Center for Applied Linguistics (CAL) with support from the National Capitol Language Resource Center (Clark and Swinton, 1979, cited in Clark, 1986).

One of the most widely known tests is the SOPI, which is a performance-based, tape-mediated speaking test. In general, the SOPI involves picture-based tasks requiring test takers to demonstrate their ability to ask questions about pictures (e.g., giving directions to someone using a map or narrating a sequence of events in the present, past, or future using drawings in the test booklet) (Malone, 2000).

Other tasks the SOPI includes requiring test takers to speak on selected topics and perform in simulated real life situations. The goals of these tasks are to assess the test

takers' ability to handle speaking functions such as stating advantages and disadvantages, supporting an opinion, apologizing, or giving an informal talk.

The SOPI simulates the Oral Proficiency Interview (OPI), which is used by government agencies and the American Council on the Teaching of Foreign Language to measure speaking proficiency, as closely as possible in a tape-recorded format. While the OPI is a face-to-face interview, the SOPI draws out speech by a tape recording and printed test booklet: SOPI administration materials are composed of a master test tape involving the audiotape of all test instructions and tasks; a test taker response tape used to record the test taker's responses; and the test booklet including test instructions and all test tasks.

Both the test booklet and the test tape entail directions to the tasks in the test taker's native language. In the test, after listening to and reading the direction, test takers are provided a short pause to arrange their thoughts. Next, test takers hear a native speaker of the target language make a statement or ask a question relevant to the task described. Then test takers respond to the native speaker prompt. After the SOPI is finished, response tapes are marked by trained raters who apply the ACTFL Guidelines.

The SOPI appears to make an attempt to contextualize all tasks to make sure that they seem as authentic as possible. Because this SOPI format is flexible, it is conveniently

tailored to the desired level of test taker proficiency and for specific test taker age groups, backgrounds, and professions (Stansfield and Kenyon, 1996).

Thus the SOPI format has been used by various institutions in the development of tests to meet their specific needs. According to Stansfield and Kenyon (1996: 4), the SOPI format is beneficial and convenient for foreign language teachers for a number of reasons: (1) No additional training is necessary to administer SOPIs: it can be administered to many test takers by a single administrator whereas a live interview must be administered individually by trained interviewers. (2) They are cost-effective: a group of test takers can be tested at the same time. (3) They are reliable and valid instruments for assessing test takers' oral proficiency: there is research confirming that the SOPI is a valid and reliable surrogate to the OPI. These studies showed the high correlation (over or near .90) between the scores of the SOPI and the OPI scored by two raters (Clark and Li, 1986; Shohamy *et al.*, 1989; Stansfield *et al.*, 1990).

In spite of the advantages of the SOPI, there have been criticisms. The major criticism appears to be the lack of authenticity, as Clark (1986: 133) points out:

The SOPI has the lack of a variety of discourse management strategies of major importance to effective face-to-face communication, including following appropriate turn-taking conventions, requesting clarification as necessary, repairing instances of miscommunication, and so forth.

To improve the SOPI, CAL carried out a 2-year study on the development of a computer-assisted oral test, the Computerized Oral Proficiency Interview (COPI). It is a multi-media computer-administered oral proficiency test and an adaptation of the SOPI. Like the SOPI, the COPI depends on taped and written directions to elicit language from test takers. The COPI, however, gives test takers more control of various aspects of the testing situation and increases raters' efficiency in scoring the test. It allows the test taker a choice in the following aspects (Center for Applied Linguistics, 2002):

‘Thinking and response time’: the COPI provides preparation and response time, but the test takers still have the choice to use up all the time allotted or to click on a button when they are finished preparing or are ready to respond.

‘Speaking functions and topics’: test takers are provided choices throughout the program, and an algorithm makes certain that test takers get each speaking function (e.g., narrating in the past) and topic (e.g., food) only once.

‘Level of task difficulty’: test takers' self-assessment scores and the choices of levels during the practice task decide their level of difficulty for the first task. However, after the first task, the program alternates from giving test takers choices or not. During the times that test takers are not given a choice, the algorithm pushes test takers to higher level tasks to ensure that they are provided with an opportunity to reach their ceilings.

‘Language of the directions’: While lower-level Spanish speakers have English directions, higher level Spanish speakers are provided a choice to get the directions in English or Spanish. However, a later version of the COPI plans to include both English and Spanish directions for lower-level speakers.

Malone (2000) and Kenyon and Malabonga (2001) argue that the COPI can be seen as an advanced oral testing instrument of the SOPI, but the COPI has not been operationalized to testing speaking proficiency because it was developed principally as a research study.

Other important computerized oral tests which may be introduced here are: the computer-assisted Basic English Skills Test (BEST) (Center for Applied Linguistics, 2007), the Test of English as a Foreign Language Academic Speaking Test (TAST) (Educational Testing Service, 2005), and the English Speaking Proficiency Test (ESPT)(English Speaking Proficiency Test, 2002-2005).

CAL has been awarded a contract by the U.S. Department of Education to produce an updated, computer-assisted oral assessment based on the Basic English Skills Test (BEST) oral interview. The BEST was developed during the early 1980s by CAL as a means of assessing the English language proficiency of immigrants and refugees who were entering the United States at that time.

The BEST contains two parts, an oral interview section and a literacy section. The oral interview section requires a FTFI administering individually, which takes approximately 15 minutes per test taker. The section consists of simulated real-life listening comprehension and speaking tasks, such as telling time, asking for directions, and following directions. The literacy skills section is administered in one hour. The section includes two parts: reading tasks (e.g., dates on a calendar or labels on food and clothing) and writing tasks (e.g., addressing an envelope or writing a rent check) (see Appendix 1). At present, hundreds of adult ESL programs across the United States use the BEST for assessment purposes.

The new computer-assisted BEST is designed to allow test takers to show what they can do in English in a shorter time frame. It is still administered as a FTFI, but the test administrator enters scores into a computer. The computer then selects the next test question to be administered, and continues to adapt the difficulty level of the questions according to the scores entered for each question. This computer-assisted oral test has been available on CD-Rom since September 30, 2002, along with test administration instructions and test administrator training materials.

The TAST was developed to measure test takers' ability to communicate orally in an academic environment (Educational Testing Service, 2005). In September 2005, when

the new TOEFL (TOEFL, Internet-Based Test) was scheduled to launch, the TAST was included as the speaking section. A revision of the TAST is currently available as a stand-alone test used by individuals and institutions to practice for the new TOEFL. The TAST will be delivered via the Internet at secure, official test centers, and test takers cannot take the official test at home on the Internet. Test takers have to speak into a microphone, and responses are recorded and scored by trained raters.

Finally, the English Speaking Proficiency Test (ESPT) in Korea provides a measure of English speaking proficiency of those whose native language is not English.

It has been in development by Kim, Jongnam at Kangnam University and his researchers for the assessment of generalized speaking ability within English context for 8 years (English Speaking Proficiency Testing Academy). The test consists of 25-26 questions and takes approximately 20 minutes to administer (see sample questions in Appendix 2). A test taker starts by making a self-introduction and moves on to describing descriptions of pictures (e.g., weather, a map or a person) and acting to situations (e.g., a restaurant, a bank or a hospital) within an allotted time. The test also asks test takers to read passages to assess their pronunciation and fluency. All questions are given on a computer monitor and responses are recorded with a digital camera fixed on top of the monitor and a microphone placed in front of test takers. Recorded

responses are evaluated objectively by trained native speakers. The use of the ESPT has been increasing in Korea.

2.4.2 Computer Administered Oral Test (CAOT)

Like the COPI, TAST, and ESPT, the Computer Administered Oral Test (CAOT) is a self-administered computerized oral test. It employs computer technology to allow test takers and computer to work together to produce ratable speech samples. The Oral Testing Software-Enhanced (OTS-E) consists of five separate modules (Larson and Smith, 2001):

1. Question Creation Module – In this module, teachers can create test items. It allows teachers to determine the type of item response elicitation prompt (e.g., text, pre-digitised sound, graphic/picture, video, or combination) required for a given item; decide how much time is allowed for preparing (or thinking) to answer each item; and determine how much time is allowed for test takers to respond to each item.
2. Task List Editor – A program allows teachers to decide which tasks will be included on the test and in what order and how many questions will be given related to each task.
3. Test Installation Module – In this module, teachers can decide in which directory the test will be located (either on an individual computer or on a network server).
4. Test administration module – This module is the ‘administration engine’ that administers the test and records test takers’ responses. It allows teachers to decide in which directory the test takers’ responses will be saved (either on an individual computer or a network server); involves a sound check to ensure that the recording of the computer is working right; carries out test takers’ IDs check; gives test items to test takers; and ends the test, filling the test takers’ responses in the designated directory for subsequent assessment.
5. Test assessment module – In this module, a list of test takers and individual responses are shown. The module allows teachers to take assessment notes in a notepad while listening and assessing each

test taker's oral responses; have immediate access to any response of any test takers who took the test (no fast forwarding or rewinding trying to locate the beginning of a given response); be able to take notes on-line about individual test takers' responses and assign a grade to that individual; and print or save the assessment notes regarding test takers' performance for a later time.

In the CAOT, all contents and methods have to be authored and selected by teachers.

While taking the test, test takers hear test directions with the accompanying text and pictures in English and/or in a target language. All responses are to be spoken into the microphone and the responses are recorded either on the hard drive or a removable zip disk on the computer (see more details about the procedure of the CAOT in 3.2).

Some of the advantages of the CAOT are stated by Larson and Smith (2001) and on its web site:

First, many people can be assessed at the same time by utilizing computers and digital equipment like other semi-direct oral tests (e.g., the SOPI, COPI, TAST and ESPT). This would be a major advantage, allowing teachers with large classrooms to implement an oral test.

Secondly, the CAOT can increase teachers' efficiency in scoring the test. The scoring program allows raters to hear the test takers' responses for each task and listen to test takers in any order. They would then be able to listen to test takers' responses in whole or in part, as many times as they wanted to for a particular task, and likewise go back to previously rated tasks. The program also allows raters to write notes to test takers so

that, aside from providing a global rating, raters are able to give overall comments and task-specific feedback to each test taker. These activities would allow teachers to give more reliable scores to test takers.

Thirdly, a variety of response elicitation prompts are possible (e.g., text, sound, graphics, video, or a combination).

Lastly, it enables tests and results to be easily and economically stored in and retrieved from the hard drive or a removable zip disk on the computer.

Because of the advantages above, the CAOT may be an improvement over the SOPI format in terms of its efficiency in scoring test taker performance, storing and retrieving tests and results, and a variety of response elicitation prompts.

Together with the significant advantages which may be associated with the CAOT, there are some limitations stated in Larson and Smith (2001) and on its web site. The limitations stated here are also the limitations of other computerized oral tests.

Firstly, response elicitation information is restricted to the area of the computer screen. Secondly, for some test takers, taking a test on the computer may cause additional anxiety as compared with a live interview. Lastly, multimedia-capable computers with headsets or speakers and microphones are required to administer the tests, unlike the case of the live interview, in which such a computer facility is not

needed.

Brown (1997) and Roever (2001) also state similar problems with using computerized language tests. Amongst the main problems they identify are:

Firstly, the presentation of a test on a computer may lead to different results from those that would be obtained if the same test were administered in a paper-and-pencil format. Some research indicates that there is little difference between the performance on computer and pencil-and-paper tests: studies on math or verbal items presented on computer as compared with pencil-and-paper version (Green, 1988) or on a medical technology examination (Lunz and Bergstrom, 1994), but much more research should be done on various types of language tests and items.

Secondly, the introduction of construct-irrelevant variances could be a problem as also mentioned in the OTS booklet (2001). For example, test takers' different computer familiarity and anxiety may lead to differences in their performances on computerized tests (Hicks, 1989; Henning, 1991; Kirsch *et al.*, 1997).

Thirdly, some physical considerations such as the cost of establishing computer equipment and networks and the possibility of sudden and unaccountable computer breakdowns could pose potential problems.

In spite of its numerous functions, the CAOT does not use very high level technology compared with more modern testing programmes such as the COPI, because it does not allow test takers to choose either speaking topics or the difficulty level of the tasks presented to them.

However, the CAOT shares features with other computerized oral tests despite its simple technology, in that it is possible to self-administer; for many people to be assessed at the same time; to use a variety of response elicitation prompts (e.g., text, sound, pictures or video); and that the responses are recorded either on the hard drive or a removable zip disk on the computer. Therefore, I believe the findings of the present study will be an aid for the researchers who want further study on test performance on other computerized oral tests.

CHAPTER THREE

METHODS OF DATA COLLECTION

3.1 Participants

Because the aim of this study is to examine Korean university students' attitudes and performance on the FTFI and the CAOT and finally the effects of attitudes on their performance, the research was conducted at two universities in Korea in the fall of 2005.

The Universities were Busan National University of Education (BNUE) and Korea Maritime University (KMU).

The participants were drawn from two English conversation classes, where English conversation was a compulsory course for all freshmen at both universities. At the time of the study, the participants at both universities were taking one 2-hour long English conversation class per week.

The students in each class were given a written consent form (Appendix 3) and asked to volunteer to the FTFI and the CAOT for the research. They were assured that the information collected would not be used towards their course grades. They were also informed that they were not forced to take the tests and told they would not have any disadvantages even if they did not participate in the research.

A total of 27 of 40 students from BNUE and 20 of 43 students from KMU agreed to participate in the study. I promised to give them a gift token, which was worth about 5 pounds as an acknowledgment for their participation.

Both oral tests were administered to the 47 participants, and their performance on the tests were taped and computer recorded so that they could be rated and analyzed later.

The CAOT recorded the students' responses in 'wav' file on the hard drive on the computer. After the tests, they completed the questionnaire about their attitudes and reactions to the FTFI and the CAOT (Appendix 4).

Three computer-based recordings were unsuccessful due to technical faults in the computer recording equipment in the computer room, and these participants were therefore excluded. Two additional participants were also excluded because their performance was not tape recorded in the FTFI due to a technical error. Thus, the analysis was based on the performance from a total of 42 participants on both FTFI and CAOT: 26 from BNUE and 16 from KMU. The characteristics of the participants were in Table 3.1.

Table 3.1 Characteristics of participants

Variables	Categories	N	%
Gender	Male	15	35.7%
	Female	27	64.3%
Age	20 and 19	22	52.4%
	Between 21 and 25	9	21.4%
	Over 26	11	26.2%
Previous experience	CAOT (or any types of computerised oral tests)	0	0%
	FTFI	14	35.9%
	None of CAOT and FTFI	28	64.1%

Out of 42 participants, all 26 freshpersons from BNUE majored in Elementary Education and minored in Science Education; 13 freshpersons from KMU didn't have a major yet because the university permits them to decide upon this after the first semester; two seniors from KMU were studying mechanical engineering and a junior from KMU was studying naval architecture.

As shown in Table 3.1, more participants (14 of 42 participants) had previously experienced the FTFI than I expected, but none of the participants had experienced both CAOT (or any kind of computerised oral tests) and FTFI.

After the tests and questionnaires, the participants were invited to a semi-structured interview. Out of these 21 volunteered and ten of were then chosen and interviewed; ten interviewees representing a spread of various groups such as age, gender, school, and experience of oral tests were selected in order to get unbiased data.

The characteristics of the interview participants were as follows. Four of them were from KMU and six were from BNUE. Five of the interviewees were female; five were male. All of them were freshpersons, but their ages were varied because some of them had crammed to repeat a college entrance exam or re-entered the university in order to get a better job; elementary school teacher is a very popular and secure job in Korea: four were 20 and under; three were between 21 and 25; and three were over 26. Six were majoring in Elementary Education and four from KMU did not yet have a major. Four had previously experienced the FTFI: one from KMU and three from BNUE, but the rest of had experienced neither of the two tests.

3.2 Oral tests (FTFI and CAOT)

In order to examine test takers' attitudes to the tests and their performance on the FTFI and the CAOT, I first asked the participants to take the two oral tests. The structure of each test is shown in Table 3.2.

Table 3.2 The structure of the FTFI and the CAOT

<p align="center">Face-To-Face Interview (approximately 15-20 minutes' duration)</p>	<p align="center">Computer Administered Oral Test (approximately 20-25 minutes' duration)</p>
<p>Warm up (unassessed) : <i>What's your name? How are you?, etc.</i></p>	<p>Warm up (unassessed) : <i>What's your name? How are you today?.</i></p>
<p>Questions and answer (6 items): 1. <i>What day (of the week)/time is it?</i> 2. <i>When/what time does your class finish?/ when do you go to bed/have lunch?</i> 3. <i>What do you usually do on Saturdays/on your birthday/in the morning?</i> 4. <i>What are you going to do after this test/this weekend?</i> 5. <i>How long have you been studying English/ how long have you been studying in this school?</i> 6. <i>Please tell me about your family/ parents.</i></p>	<p>Questions and answer (6 items): 1. <i>What's the date today?</i> 2. <i>What time do you usually leave your house for school? And how long does it take to get to school?</i> 3. <i>What do you usually do in your free time?</i> 4. <i>What are you going to do tomorrow?</i> 5. <i>How long have you lived in Busan?</i> 6. <i>Please introduce yourself/ please tell me about yourself.</i></p>
<p>Picture description (1item) 7. <i>The person in the picture is your friend. Describe him/her.</i></p>	<p>Video description (1item) 7. <i>The person in the video is your friend. Describe him.</i></p>
<p>Role play (1 item) 8. <i>You are supposed to meet your friend at 7:00 today, but you may arrive at the appointment place half an hour late because of a sudden part-time job interview. Now explain the reason you are late and apologize.</i></p>	<p>Role play (1 item) 8. <i>You made an appointment with your friend this weekend, but you have a sudden job interview on the same day and want to cancel the appointment. Your friend is not answering her cell phone, and you have to leave a voice message. Apologize and explain why you can't meet your friend this weekend.</i></p>
<p>Narration (1 item) 9. <i>Make up a story based on a cartoon: a 'birthday' task</i></p>	<p>Narration (1 item) 9. <i>Make up a story based on a cartoon on the screen: a 'restaurant' task</i></p>

The test tasks were the same as those used in Joo's 2004 study, with certain changes: the order of tasks was changed: 'role play' and 'narration' was reversed because 'narration' was assumed to have greater difficulty than 'role play'. The tasks were arranged in the order of the level of difficulty from the easiest task to the most difficult. Then a few questions (Q 3 in the FTFI and Qs 2, 4 and 7 in the CAOT of the 2004 study) were modified to make them clearer. Some were also removed because they had been shown to cause problems in Joo's study (2004), and a few questions (Q 5 in the FTFI and Q 6 in the CAOT) were added to balance the number of questions; in the CAOT a short video was used for the description task to make the task more authentic and interesting; the cartoons were revised to make the story clearer; the same number of questions in each task were used in both tests (see Table 3.2 and the structure of the 2004 tests of Appendix 5). In order to avoid content bias between the FTFI and the CAOT, there was a more careful attempt than in the 2004 study to construct the two oral tests in such a way that both the elicitation tasks and test takers' responses would match each other as closely as possible.

Before taking the tests, the participants were given instructions about the two tests. The types of tasks they were asked to complete were introduced and the narration tasks were shown to familiarize them with the topic. This was done because in a pilot study, it

was found that the participants felt too much difficulty and even frustration when attempting the tasks. For the CAOT there was a more detailed explanation about its operation and administration in order to reduce participants' discomfort and ease their tension.

Each participant group in BNEU and KMU was divided into two groups. The first group in each university was assigned to take the CAOT first, and the second group to take the FTFI first. 22 participants took the CAOT first and 20 the FTFI first – one participant could not take a test on the appointed date due to a personal reason.

There were seven-day intervals between the tests. The tests were conducted over 5 weeks in each university. The procedure is shown in Figure 3.1.

Figure 3.1 Summary of the procedure

Group 1	Group 2
<p data-bbox="300 1330 485 1361"><u>Take the FTFI</u></p> <p data-bbox="316 1429 533 1460">After seven days</p> <p data-bbox="300 1527 507 1559"><u>Take the CAOT</u></p> <p data-bbox="284 1908 609 1939">Fill out the questionnaire</p> <p data-bbox="363 1953 491 1984">Interview</p>	<p data-bbox="810 1527 1018 1559"><u>Take the CAOT</u></p> <p data-bbox="842 1626 1059 1657">After seven days</p> <p data-bbox="842 1720 1034 1751"><u>Take the FTFI</u></p> <p data-bbox="826 1908 1152 1939">Fill out the questionnaire</p> <p data-bbox="922 1953 1050 1984">Interview</p>

The FTFI was individually administered to each test taker in a quiet classroom. The procedure began with a warm-up phase to reduce test takers' anxiety and help them feel comfortable. Following the warm-up, test takers were presented with tasks: 'question and answer', 'picture description', 'role play' and 'narration' (see Table 3.2). Before and after each task or item, test takers were provided with sufficient preparation and response time by saying that they could use as much time as they wanted for preparing for and responding to each task or item.

The FTFI procedure ended with a wind-down phase to ease the test takers out of the testing situation. Each interview was tape recorded for rating and analyzing. The time taken for each interview was approximately between 15 and 20 minutes.

The CAOT was group administered to all participants in the computer rooms of BNEU and KMU. Each department or building of BNEU and KMU had two or three computer rooms, and approximately 45 computers with speakers and headsets were fitted in each room. A teaching assistant in a computer room set up the oral testing software on each computer for the test. I had the test takers take every other seat so that they were not disturbed by other test takers' voices.

When taking the CAOT, the test takers passed through five phases (see Appendix 6). The first was the test taker registration phase. Before beginning the test, the test takers

had to complete the required registration information. They entered their first and last names and test taker number on the keyboard. After the test takers' registration information had been entered, they were asked to confirm that it was all correct. Once they confirmed that the registration information was correct, they were allowed to proceed to the next phase.

The second phase included general instructions regarding the purpose of the test, instructions for proceeding through the test, and information on the time allotted for preparing to respond and for recording responses.

The third phase was the test recording and playback functions phase. In this phase the test takers were directed to click on a Sound Test button and record their names. After about four seconds went by, their names were played back. If the recording of the sound test was loud enough, they were instructed to click on the Begin button to begin the actual test. If not, they were required to click the Stop button and ask me for assistance.

The fourth phase was the actual test. Once the test takers confirmed that they were ready to begin the test, the first written and audio item appeared. The test included written instructions in Korean, written and audio instructions in English, and a video and a strip cartoon that accompanied the tasks.

In a pilot study, participants had a fixed amount of time to think about their response

and to respond to the task. There were about 15 seconds for preparation time and 1 to 2 minutes for response time. The participants said they felt so nervous because of the limited time, and this appeared to prevent them from giving their best performance. Therefore the preparation and response time given for the current study was sufficiently long so that the students could have control of the time that they took to prepare for and respond to a task.

After each question, a Ready prompt was displayed. This was to allow the test takers to pause for a moment and relax from the pressure and anxiety of the test. To continue on to the next question, the test takers clicked the OK button.

The last phase was to end the test. After all items in the test had been answered, an Exit Text was displayed. The test takers clicked on the End button to complete the test.

Depending on the test taker, the administration of the test took approximately between 20 and 25 minutes.

Scoring of performance on the FTFI and the CAOT

In order to examine the test taker performance on the FTFI and the CAOT, each test taker's computer and tape recording was rated by me and a Canadian English teacher, both of whom had more than five years experience teaching English conversation

classes and rating oral performance. We independently rated 42 test takers' responses from two settings: 42 taped face-to-face interview recordings and 42 computer recordings.

Before the actual rating, there was a short one day training session in a classroom to familiarize myself and the Canadian teacher with the rating scale and descriptors and to encourage us to compare and critically reflect on our ratings for a more reliable assessment. The steps we took were: listening to a few tape and computer recordings, marking independently, collating the scores and discussing them, refining the marking scale and arriving at a consensus with our marks. Each of us needed to listen to each recording two or three times before making an independent judgment.

The rating scale and descriptors used for rating were taken from O'Loughlin's study (2001: 217-219). According to O'Loughlin (2001), the scoring criteria were originally chosen by the test development team from the Language Testing Research Center (LTRC) at the University of Melbourne to rate test taker performance on live and tape version tests. After some trial and error, the test development team designed and agreed on the 1 to 7 Likert rating scale criteria using detailed descriptors whose criteria comprise fluency, grammar, vocabulary, intelligibility, cohesion, and overall communicative effectiveness.

The test developers in O'Loughlin's team considered the combinations of the criteria appropriate to assess the performance on the live and tape-based oral tests after listening to sample recordings from the two versions of the tests. Using the recordings, they drew the descriptors. Then they formulated and finally agreed on descriptors of proficiency at the levels for each criterion. The one category which did not have descriptors was the overall criterion 'communicative effectiveness'.

These rating scale and descriptors developed by the test development team by and large match the test tasks on the FTFI and the COAT. Generally these tests focused upon grammar, vocabulary, intelligibility, cohesion, and overall communicative effectiveness. After carefully listening to a sample of tape and computer recordings during the short training session the Canadian teacher and myself agreed that O'Loughlin's rating scale and descriptors were more appropriate to score the participants' performance on the FTFI and the CAOT than other rating scales, such as the Foreign Service Institute (FSI) and the American Council for the Teaching of Foreign Languages (ACTFL), which were originally designed for a live interview. The full list of descriptors for the scoring criteria is included in Appendix 7.

3.3 Questionnaire

In order to investigate the test takers' attitudes to the two tests and computer familiarity, questionnaires were administered. Surveys are probably the most widely used research method in educational research and have been used to measure attitudes, opinions, or achievements - any number of variables – in natural settings (Wiersma, 2000). A survey was believed to be the most appropriate research method for collecting answers from a number of participants in this study. This was done using a set of carefully designed questions about computer familiarity and attitudes to the tests which was constructed for Joo's 2004 study. The questionnaire (Appendix 4) was constructed and revised referring to the questionnaire items used in the previous studies (e.g., Clark, 1988; Stansfield *et al.*, 1990; Kenyon and Malabonga, 2001). It contains four sections: (1) computer familiarity (2) attitudes (3) preferences (4) demographic information.

Section One and Two include six items relating to test takers' computer familiarity (Qs1-6) and eight items measuring their attitudes to the two test formats (Qs7-15). In the two sections, Likert four point scales were used to collect data since they provide a statement that reflects particular attitudes or reactions (Wiersma, 2000). They have the advantage of convenience that makes this method suitable for the study. This convenience seems to lie in obtaining the differences among respondents in a simple

and easy way. To obtain a picture of the characteristics of the test takers' perceptions and attitudes, this method appeared to be suitable.

Section Three contains eight items asking about test takers' preferences and test reliability (Qs16-23). In this section, most items provide the respondents only three alternatives, 'FTFI', 'CAOT' and 'same', to obtain their clear preferences and attitudes.

A few items are open-ended to give freedom to the respondents to make any responses they wish and to prevent eliciting false opinions either by giving an insufficient range of alternatives from which to choose or, by prompting them with apparently acceptable answers (Vaus, 2002).

Finally, Section Four contains six items asking whether they have experienced the two test formats, which was the first test format they had experienced and finally, personal information such as gender, age and major (Qs24-30).

After taking both oral tests the test takers were requested to fill out the questionnaire, which took approximately 10 to 15 minutes.

3.4 Interview

The questionnaire is a good data collection strategy for obtaining a large number of participants' attitudes and perceptions and has definite advantages because it allows the

respondents to control the data-collection process by answering the questionnaire items in any order, taking more time to complete it or skipping some of questions (Gall *et al.*, 1996).

In practice, most studies which deal with test takers' attitudes and perceptions have used a closed or (and) an open-ended questionnaire as their research tools. The researchers construct and develop the items of the questionnaires mostly by reviewing the literature related to their research questions. But there is a hazard to examine the respondents' inside from the researchers' standpoints. The questionnaires may become too simplified or give even wrong or inaccurate the respondents' thoughts and feelings and thus possibly miss important data. Even open-ended items on the questionnaires may not be powerful enough to elicit ample and satisfactory data from the respondents. In other words, the responses to open-ended items may lack depth or may not fully reveal the respondents' feelings and opinions.

Furthermore, a reliance on statistical analyses alone has been criticized, with the argument that quantitative empirical methods are effective, but limited, and qualitative methods are required to complement them resulting in higher quality research (Lazaraton, 2002; Phillips, 1983; Grotjahn, 1986). Biddle and Anderson (1986: 239) also support this view:

It is inappropriate to compare the relative efficacy of quantitative and qualitative research since each has a different purpose; broadly these are the generation of insights on the one hand and the testing of hypotheses on the other. Although advocates for discovery (qualitative researchers) decry the arid tautologies of confirmationists (quantitative researchers), and the latter express disdain for the sloppy subjectivism of discovery research, the two perspectives have complementary goals. We need them both.

Therefore, it was felt that more valid and richer data could be obtained from semi-structured individual interviews in which the researcher could initiate issues and help the respondents to think about and react more carefully to the given issues during discussion.

The interview was useful to understand the test takers' experience in the two oral tests and their attitudes and reactions to the tests in greater detail. Seidman (1991: 3-4) asserts that 'the purpose of interviewing is an interest in understanding the experience of other people and the meaning they make of that experience' and it allows one to 'put behavior in context and provides access to understanding their action'. It made it possible to get a better understanding of not only test takers' attitudes but also their performance on the two tests, and further it helped interpret the findings of the statistical analyses. It was hoped that this interview study would increase both the validity and the reliability of this study while minimizing the weaknesses of the questionnaire.

All semi-structured interviews involved the same basic questions in the same order in

order to minimize the possibility of bias (Gall, *et al.*, 1996). In this data collection method the same initial questions were first asked that contained the specific questions relating to the purpose of the current study, but different probing questions were asked based on the respondents' answers. The specific questions were mainly related to their computer and oral test experience, and their feelings and opinions of the two different test formats. The questions were carefully designed to encourage the respondents to search their experience and not to restrict their answers only to 'yes' or 'no'.

A pilot study was conducted to yield reasonably unbiased and reliable data. The Korean version of the interview questions were tested with three Korean university students. Through the pilot study, interviews were practiced and interview skills were developed. Some restricting and leading questions and unclear Korean versions of interview questions were also revised and improved. Here are the final revised questions:

1. Could you tell me about your day-to-day use of a computer?
2. Tell me about your past experience of taking oral tests before the two tests.
3. Tell me about the experience of the two oral tests: FTFI and CAOT.
4. What do you think about the two tests?
5. What do you think about your test performance on the two tests?

Before the interview took place, with a brief introduction about the purpose of the interview, it was clarified again, as written on the consent form, that I was an

independent researcher with no connection to any organization or authority; that participants would remain anonymous; that whatever they might answer would not lead to any acts of reprisal; and that they had the right to withdraw from the study at any time during the interview or within a specified time after they were completed.

The interviews were individually conducted for the ten chosen from the 28 students who accepted the interview in quiet classrooms in BNUE and KMU. The interview began with questions about their general background, to encourage the interviewees to feel relaxed and to freely express their ideas in an informal atmosphere. I attempted to create a supportive and cordial atmosphere which felt like talking in an everyday, conversational way (Agar, 1980; Bogdan and Biklen, 1982; Holstein and Gubrium, 1995).

Each interview lasted for between 40 and 50 minutes. All responses from the interviews were tape recorded and immediately transcribed after the interviews for the analysis. Throughout the interviews, interviewees' responses continually provided rich information that enabled the construction of possible categories, and instilled in me a deeper understanding of their attitudes to the two oral tests.

After analyzing the transcripts, in an attempt to eliminate any possible ambiguities or misunderstandings, I e-mailed and telephoned some of the interviewees for further

clarification on certain points and additional information. It was useful to have back-up explanations on certain matters which I needed in order to have more understanding.

To increase the reliability of this research, I also decided to do member checking with the interviewees: member checking is considered as ‘the most crucial technique for establishing credibility’ (Lincoln and Guba, 1985: 314). Qualitative research should convince both the researcher and the readers that what the researcher has concluded is accurate. The member checking took place after transcribing and analyzing the interviews.

3.5 Data Analysis

SPSS version 13.0 was used to conduct descriptive statistics, Wilcoxon, Mann-Whitney and Chi-square tests, Factor analysis, and Spearman correlations. Significant p values from the tests were further investigated by calculating effect sizes using Cohen’s d , phi (ϕ), and Cramer’s V .

The interviewees’ responses from the semi-qualitative interviews were transcribed and coded in order to analyze the test takers’ attitudes to and perceptions of the two oral tests. More details about the analyses of the questionnaire and the interview were done in the next chapter.

Primarily 'multi-faceted Rasch measurement' with the program FACETS, version 3.45 (Linacre, 1991-2003) was carried out to investigate test taker performance on the FTFI and the CAOT. The following section addressed 'multi-faceted Rasch measurement' since it was a new theory and method of measurement for performance assessment (McNamara, 1996). It will be helpful for readers to understand this measurement.

Multi-faceted Rasch measurement

Multi-faceted Rasch measurement (Linacre, 1989) is one of the most encouraging developments in analyzing performance assessments with the computer program FACETS and enables one to investigate numerous aspects of performance assessments.

Performance assessments are very intricate because a number of facets (or factors) can influence the outcomes for test-takers in the performance setting: they include the ability of the test taker, the difficulty of the item, the characteristics of the rater, and other characteristics of the situation in which the performance is drawn out and scored (McNamara, 1996). The facets are connected with each other and increase or decrease the probability of a test taker of given ability getting a given score on a particular task or item (Lumley and McNamara, 1995).

Multi-faceted measurement allows one to predict the probable score (i.e., a logit score) for the combination of the facets such as test takers, items and raters, and to evaluate the accuracy of the prediction. In other words, the measurement can deal with the facets of the test situation which vary from test taker to test taker.

In order to make the data predictable, the data is modeled in the data matrix. While the original Rasch model (a version of Item Response Theory based on a one parameter model) (Rasch, 1969/1980) estimates only two facets (test takers and items) for dichotomous items, a multi-faceted Rasch model can handle non-dichotomous items. The basic Rasch model is extended to the multi-faceted model (Linacre, 1989), so that more facets such as rater severity (or other possible facets) can be added to the equation.

The multi-faceted Rasch model analysis compensates for differences across facets and converts raw scores into logits, creating an equal interval scale. All estimates of facets such as test taker ability, item difficulty and rater severity are presented in logits. For example, a logit score for each test taker expresses the ability of the test taker in terms of his/her probability of obtaining a particular score on any item, given the difficulty of the item, the harshness of the rater and the difficulty of other possible facets such as test orders and test formats.

In the context of the oral tests for the current study, besides the possible effects of

attitudes on test performance, test taker performance might have been assessed with the impact from the intervening facets (or factors) such as rater severity, test taking orders, test formats involved in performance testing context. There might also have been a bias between the facets. The multi-faceted Rasch model makes it possible to grasp such intervening facets, estimate their impact and measure test takers' potential ability irrespective of the facets. Therefore, through using the multi-faceted Rasch model the analysis attempted to examine not only the test taker performance, but also the impact of those facets whilst making necessary adjustments on the test takers' speaking ability. If the scores awarded by two raters (i.e., raw scores) were seriously affected by the intervening facets, then the findings from the analysis with the raw scores may not be very valid and accurate.

The output from a multi-faceted Rasch measurement analysis provides several useful statistical indices. Practical ones are Separation, Reliability, and Fixed (all same) chi-square. The separation index is the ratio of the adjusted standard deviation of element measures to the root mean-square standard error. The reliability index is equivalent to the KR-20 or Cronbach's alpha coefficient. It represents the ratio of true variance to observed variance. Although separation and reliability indices are presented in a different metric, the two indices indicate similar results for a particular facet because

they are calculated from the same information. Both indices describe the amount of spread (or variability) in the measures estimated by the model for the various elements in the specified facet relative to the precision by which those measures are estimated (Sudweeks *et al.*, 2005: 245). The separation index (ranging between 0 and 1) close to 1 means that the variability of the measures of the elements in a facet is small while the reliability of separation index (ranging from 1 to infinity) close to 1 indicates that all the elements in a facet are well separated into different levels and thus represent high levels of reliability. However, the interpretation of these statistics is different for different facets. While high indices in the student facet are welcomed, low indices in other facets are desired because variability among the elements of these other facets normally represents construct irrelevant variances in the ratings.

Finally, the fixed chi-square tests the null hypothesis that all the elements of the facet are equal. The chi-squared test compares the estimates of the elements of the facet taking into account the relative difficulty of tasks/items. (O'Loughlin, 2001: 75). The significant chi-square test indicates that all the elements in the facet are not equal in the measure of estimates.

A multi-faceted Rasch measurement analysis also provides information about how well the performance of each person, rater, or item matches the expected values

predicted from the model generated in the analysis. This is reported as the fit of the data to the model: infit and outfit statistics, each expressed in two alternative ways, as mean square and t (McNamara, 1996). The infit is the weighted mean-squared residual which is sensitive to unexpected responses around the scale where a decision is being made. On the other hand, the outfit is the unweighted mean-squared residual and is sensitive to extreme scores. The infit and outfit mean square values are commonly used because they are less sensitive to sample size and are weighted by the information in the response, unlike standardized values (t -distribution) (Smith *et al.*, 1995). Only infit mean square values are reported in this paper because the infit is usually considered more informative because it focuses on the degree of fit in the most typical observations in the matrix (McNamara, 1996). There seem no clear-cut rules for determining what degree of fit is acceptable, but a common rule of thumb for acceptable mean-squared residual difference ranges from 0.5 to 1.5, which has been found to be useful for practical purposes by several researchers (e.g., Lunz and Stahl, 1990; Weigle, 1998). A fit statistic greater than 1.5 indicates a misfit (too much unpredictability in an element of a facet) whereas less than 0.5 indicates an overfit (not enough variation in scores).

Another useful form of information from the multi-faceted measurement is fair M-average. The fair M-average is logit measure to raw score conversion. It gives the logit

measure as an expected average raw response value in a standardized environment in which all other elements interacting with this element have zero logit measures. This fair M-average permits one to compare elements in the raw score metric.

Lastly, multi-faceted measurement allows for analysis of bias or instances of interactions between the facets. Bias analysis identifies unexpected but consistent patterns of behavior which may occur from an interaction of a particular rater with respect to some component of the rating situation such as a particular person, format or task. For example, it can indicate whether there is a tendency that one particular rater scores a certain group of persons differently from other raters or, a certain group of persons performed differently on one task than other tasks. Each interaction between the facets specified in the analysis is given a bias score on the logit scale, and its significance is designated by a standard z-score. The z-score values with an absolute value falling below 2 indicate no significant interactions between facets while the values which are equal to or greater than 2 indicate significant interactions.

CHAPTER FOUR

ANALYSES OF QUESTIONNAIRES, INTERVIEWS AND SCORES

4.1 Attitudes

4.1.1 Attitude Questionnaires

Data from the attitude questionnaire were analyzed using SPSS version 13.0 to examine the 42 Korean university students' attitudes to the FTFI and the CAOT. Wilcoxon tests were carried out first to examine if there were any significant differences in the responses to the nine statements regarding the FTFI and the CAOT.

The percentage of the test takers who chose each option for the questions asking their preferences and attitudes were then compared to explore their attitudes. Using cross tabulations, the relationships among their attitudes were examined in more detail and the answers to the open questions in the questionnaire were also explored.

After analyzing the test takers' attitudes, factor analysis was conducted to make the nine statements about attitudes a more manageable subset of factors, for clear interpretation, as well as for convenience in the analysis. The factor analysis, even

though there were not many items, was useful not only to get a better understanding of the relationships among the items, but also to evaluate the structural validity of the measurements. Utilizing factors from the factor analysis, Wilcoxon tests were conducted once again to interpret the test takers' attitudes to the FTFI and the CAOT more clearly and substantially.

Computer familiarity and additional variables (gender and age) obtained from other information provided by the test takers were also analyzed using the factors (via Wilcoxon and Mann-Whitney tests and Spearman correlations) to consider if there were any differences in attitudes and computer familiarity by age or gender, and if the variables influenced the test takers' attitudes to the tests.

In order to help to determine whether a statistically significant difference or association was a difference or association of practical concern, effect sizes were calculated using Cohen's d , which is the appropriate measure to use in the context of Wilcoxon and Mann-Whitney tests, and phi (ϕ) or Cramer's V , which is considered as the best measure of association for the chi-square test.

4.1.1.1 Attitudes to the FTFI and the CAOT

In order to investigate the test takers' attitudes to the FTFI and the CAOT, the items of Sections Two (Qs7-15) and Three (Qs16-23) were analyzed (Tables 4.1 and 4.2).

Medians and minimum and maximum values were computed for the items of Section Two (Qs7-15) of the attitude questionnaire (see Appendix 4) on Likert four point scales ranging from (1) strongly disagree to (4) strongly agree, and the FTFI and the CAOT medians were compared using Wilcoxon tests.

The percentage of the test takers who chose each option in Section Three (Qs16-23) was compared, and for the responses to Qs16-21 chi-square tests were carried out to see if there were any significant differences in their choices between the first (FTFI) and second (CAOT) options: for the tests, the third option 'same' was excluded from the data. The results are presented in Tables 4.1 and 4.2. Results of significance at the 0.05 level are highlighted in bold.

Table 4.1 Comparative results on Section Two

Questionnaire item	Format	Median (Min-Max)	Z	d
7. It is a good test format as an exam.	FTFI	3 (2-4)	-2.27*	0.51
	CAOT	3 (1-4)		
8. I could do justice to my ability.	FTFI	2 (1-4)	-0.17	-0.03
	CAOT	2 (1-4)		
9. The test had enough questions to assess my speaking ability.	FTFI	3 (1-4)	-1.94	0.28
	CAOT	3 (2-4)		
10. I had sufficient time to think about the questions before I spoke.	FTFI	2 (1-4)	-3.90**	-1.07
	CAOT	3 (2-4)		
11. I had sufficient response time.	FTFI	3 (2-4)	-0.93	0.19
	CAOT	3 (2-4)		
12. I was nervous while I was taking the test.	FTFI	2 (1-4)	-3.94**	0.85
	CAOT	2.5 (1-4)		
13. Questions asked in the test were fair.	FTFI	3 (2-4)	-0.75	0.04
	CAOT	3 (2-4)		
14. I had to wait for a long time before taking the test, and it made me tired.	FTFI	3 (1-4)	-3.62**	0.69
	CAOT	3 (2-4)		
15. The test was easy to use.	FTFI	3 (1-4)	-2.81**	-0.43
	CAOT	3 (1-4)		

*P<.05 **p<.01

Table 4.1 shows that the test takers tended to give positive responses regarding the two tests except to Qs8 (median 2, disagree) and 14 (median 3, agree). The test takers generally thought that they could not do justice to their abilities on both tests and felt tired due to waiting for the tests although they felt significantly more tiredness in the FTFI. All items did not have the same polarity. In the cases of Qs12 and 14, lower median means that the test takers answered more positively, unlike the other items.

Table 4.2 Comparative results on Section three

Questionnaire item	N	FTFI	CAOT	Same	χ^2	ϕ
16. Which test made you feel more nervous?	42	67%(28)	21%(9)	11%(5)	9.76**	0.48
17. Which test do you want to take again as your exam?	42	55%(23)	35%(15)	10%(4)	1.68	0.20
18. Which test was more comfortable for you-talking to a computer or to a person?	42	48%(20)	50%(21)	2%(1)	0.02	0.02
19. Which test was more difficult for you?	42	48%(20)	38%(16)	14%(6)	0.44	0.10
20. Which test was fairer?	42	19%(8)	57%(24)	24%(10)	8.00**	0.44
21. Which test do you prefer?	42	67%(28)	33%(14)	0%(0)	4.67*	0.33
	N	YES		NO		
22. During CAOT, were the voices on the computer clear?	41	91%(38)		7%(3)		
23. While undergoing CAOT, was it disturbing for you to have other test takers talking at the same time?	42	19%(8)		81%(34)		

*P<.05 **p<.01

Tables 4.1 and 4.2 indicate that the test takers showed more positive attitudes to the CAOT in questions related to preparation time (Q10), nervousness (Qs12 and 16), tiredness due to waiting (Q14), ease of use (Q15), and fairness (Q20).

The significantly variable responses to Q10 ($Z=-3.90$, $p=.00$) indicate that the test takers thought that they had more preparation time for the questions in the CAOT even though I attempted to give sufficient time to think about each question before they answered in the FTFI. The effect size was also large ($d=-1.07$): Cohen's (1992) guidelines for interpreting Cohen's d suggest that effect sizes of .2 are considered small, .5 medium, and .8 large. On the other hand, Phi (ϕ) or Cramer's V is described as

ranging from .1 (small effect) through .3 (medium effect) to .5 (large effect) (AcaStat Software, 2007).

The responses to Q12 ($Z=-3.94$, $p=.00$, $d=.85$) and Q16 ($\chi^2=9.76$, $p=.00$, $\phi=.48$) clearly indicate that the test takers felt less nervous during the CAOT. The responses to Q19 reveal that more test takers might have felt difficulty in the FTFI than in the CAOT although the difference was **not significant**. In the open question, 'Which test was more difficult for you? Why?' Some test takers ($n=7$) responded that they felt the FTFI was more difficult because talking to a teacher made them nervous. There was a significant association between difficulty (Q19) and nervousness (Q17) ($\chi^2=17.41$, $p=.00$, $V=.46$). Perception of the difficulty was also related to comfort in talking (Q18) ($\chi^2=17.37$, $p=.00$, $V=.46$). 81% of the test takers who found the CAOT more difficult answered that they felt more comfortable talking to a person. Furthermore, comfort and preference were connected to each other. This was discussed more in Table 4.3.

The test takers perceived that the CAOT was fairer than the FTFI (Q20: $\chi^2=8.00$, $p=.01$, $\phi=.44$). The test takers' responses to Q13 (related to fairness) showed no significant differences, but when they were asked to choose only one as the fairer test (Q20), the CAOT was chosen by over half of the test takers (57%) and the FTFI by only 19%. The test takers seem to have thought the questions given in both tests were

generally fair (their responses for both tests were over 3 (agree) on average), but the CAOT was significantly fairer. The perception of fairness (Q13) was not associated with responses to Qs7 and 8 related to perceptions of test validity ($p > .05$), but the reasons were found in the open question, 'Which test was fairer? Why?' Most test takers ($n=21$) also answered that they thought the CAOT was fairer because the test offered exactly the same questions and the same amount of time for preparation and response. Thus it seemed more objective because it excluded the teacher's subjective view.

In Qs14 and 15 the test takers agreed more with the statements 'I had to wait a long time to take the FTFI and it made me tired' and 'the CAOT was easy to use'. Before the FTFI I let them know the estimated FTFI time for each of them so that they wouldn't have to wait long, but the FTFI still seemed to make them wait longer and feel more tired than the CAOT ($Z=-3.62$, $p=.00$, $d=.70$). It was assumed that there might be some relationship between the two statements, but no significant connection was found ($p > .05$). The effect size of the responses to Q15 was small ($d=-0.43$), and further the validity of Q15 was uncertain since some test takers in the interviews seemed to interpret the item as 'the test was easy'.

Despite the negative attitudes to the FTFI, the responses to Q7 (a good test format) indicate that significantly more test takers perceived that the FTFI was a better test format as an exam ($Z=-2.27$, $p=.02$, $d=.51$). In addition, as the percentage of the test takers' responses to Q21 ($\chi^2=4.67$, $p=.03$, $\phi=.33$) revealed, they showed a preference for the FTFI over the CAOT. The test takers who responded that they preferred the FTFI also answered that they wanted to take the FTFI again ($\chi^2=22.98$, $p=.00$, $V=.74$).

These results are very similar with those of my 2004 study (see Table 1.2 and Appendix 8). A few differences are: the responses to Q7 did not show significance in 2004, suggesting that the test takers perceived both tests as good. Instead, the responses to Q13 (fairness) in the 2004 study revealed that the test takers significantly felt that the CAOT was fairer than the FTFI, even though more test takers felt that the voices on the CAOT were not as clear and that other test takers talking near them disturbed them in 2004. The test takers in the 2004 study seem to have had slightly more favorable attitudes to the CAOT than those in this current study, but they also clearly showed more preference for the FTFI and still wanted to take the FTFI more than the CAOT.

In responses to the open question 'Which test do you prefer? Why?', the responses given by most test takers were relevant to test validity (e.g., it was a real conversation interacting with a person, real speaking ability can be proven, and so on) (see the

summary of the responses to the open questions on Pages 116-118). Some test takers in the interviews mentioned that they liked the FTFI because they felt more comfortable and natural talking with a person.

From the results of the crosstabulation analyses of the responses to Qs7 (related to test validity) and 8 (doing oneself justice) (FTFI: $\chi^2=11.30$, $p=.08$, $V=.37$, CAOT: $\chi^2=27.32$, $p=.00$, $V=.47$) and Qs7 and 21 (related to preference) (FTFI: $\chi^2=7.50$, $p=.02$, $V=.42$, CAOT: $\chi^2=9.63$, $p=.02$, $V=.48$), it was found that the test takers perceiving the FTFI was a better test preferred the FTFI and tended to think they could do justice to their ability in that test (the responses to Qs 7 and 8 about the FTFI were not significantly associated, but a medium effect size was discovered); likewise, the test takers perceiving the CAOT was a better test format liked the CAOT more and felt they could therein better do themselves justice. Moreover, the test takers thinking they did justice to their ability also felt that the test had enough questions (FTFI: $\chi^2=17.14$, $p=.05$, $V=.37$, CAOT: $\chi^2=22.72$, $p=.00$, $V=.52$).

There was also a significant correlation between the test takers' comfort in talking and their test preference ($\chi^2=21.00$, $p=.00$, $V=.71$). Table 4.3 which indicates the correlation more clearly is shown below.

Table 4.3 Crosstabulation of Preference * Comfort

		Which test do you prefer?			Total
		FTFI	CAOT	Same	
Which test was more comfortable for you- talking to a computer or a person?	Computer	33.3%(7)	66.7%(14)	0%(0)	100%(21)
	Person	100%(20)	0%(0)	0%(0)	100%(20)
	Same	100%(1)	0%(0)	0%(0)	100%(1)
Total		66.7%(28)	33.3%(14)	0%(0)	100%(42)

100% of the test takers who answered that they felt more comfortable with a person preferred the FTFI, and 66.7% of the test takers who felt more comfortable with a computer preferred the CAOT. However, it was found that there was a contradiction here. 67% test takers (n=28) said that the FTFI made them feel more nervous (Q16). But 48% test takers (n=20) said that talking to a person was more comfortable for them. This means that some of the test takers said that the FTFI made them feel more nervous and that talking to a person was more comfortable. In order to investigate this contradiction, a crosstabulation of the responses to Q16 (nervousness) and Q18 (comfort in talking) was carried out and a significant association was found ($\chi^2=14.37$, $p=.01$, $V=.41$). 40% of the test takers who answered that talking to a person was more comfortable also responded that the FTFI made them feel more nervous.

Table 4.4 Crosstabulation of Nervousness * Comfort

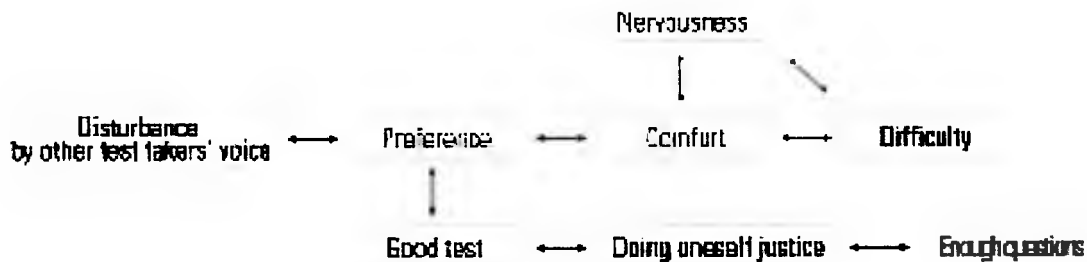
		Which test made you feel more nervous?			Total
		FTFI	CAOT	Same	
Which test was more comfortable for you- talking to a computer or a person?	Computer	90.5%(19)	0%(0)	9.5%(2)	100%(21)
	Person	40%(8)	45%(9)	15%(3)	100%(20)
	Same	100%(1)	0%(0)	0%(0)	100%(1)
Total		66.7%(28)	21.4%(9)	11.9%(5)	100%(42)

The reasons for these contradictory responses and for their nervousness in the FTFI were not found through the analysis of the questionnaire, but some possible causes were discovered in the semi-structured qualitative interviews, which were analyzed and discussed later in 4.1.2.

Most of the test takers (91%) ticked 'yes' for Q22, 'During the CAOT, were the voices on the computer clear?' and 81% of the test takers ticked 'no' for Q23, 'While undergoing the CAOT, was it disturbing for you to have other test takers talking at the same time?' There was a significant relationship between the responses to Qs23 and 21 related to preference ($\chi^2=4.94$, $p=.03$, $V=.34$). 100% of the test takers (n=8) who felt that it was disturbing to have other test takers talking at the same time in the CAOT preferred the FTFI.

As seen above, the test takers' attitudes were not related with only one aspect of their attitudes, but they were linked with one another. The figure below shows their relationships more clearly.

Figure 4.1 Relationships among attitude items



The test takers' answers to the open questions in the attitude questionnaire are summarized below. In the open questions the test takers were asked to specifically compare the FTFI and the CAOT and to explain their responses.

Which test was more difficult for you? Why?

More test takers (48%, n=20) perceived the FTFI as more difficult than the CAOT.

The test takers who perceived that the FTFI was more difficult than the CAOT generally provided the following reasons: (1) talking to a teacher makes them nervous (n=7), (2) they feel embarrassed when they make mistakes or give poor answers (n=6), (3) they have less preparation time (n=1), (4) unclear (n=1) or missing (n=5).

38% (n=16) of the test takers felt that the CAOT was more difficult while 14% (n=6) of the test takers felt that the two tests were equally difficult. Those who felt that the CAOT was more difficult gave the following reasons: (1) it is a new and unfamiliar test (n=5), (2) there is no interactions and reactions, and it is not natural (n=4), (3) talking to

a computer is less comfortable (n=2), (4) the timer on the screen makes one more nervous (n=1), (5) the test gives less response time (n=1), (6) the test is not flexible (n=1), (7) unclear (n=1) or missing (n=1).

Which test was fairer? Why?

More test takers (57%, n=24) considered the CAOT to be fairer. The test takers who felt the CAOT was more fair generally gave the following reasons: (1) it provides the same questions and amount of time under the same conditions (n=16), (2) it is objective because it excludes a rater's subjective view (n=5), (3) the test makes one less nervous (n=1), (4) unclear (n=1) or missing (n=1).

19% (n=8) of the test takers perceived the FTFI to be fairer (25% viewed both tests as equally fair). They provided the following reasons: (1) it assesses their real speaking ability (n=2), (2) it is flexible (n=2), (3) there is more possibility to cheat in the CAOT (n=1), (4) there is no fixed time (n=1), (4) unclear (n=1) or missing (n=1).

Overall, which test do you prefer? Why?

More than half of the test takers (67%, n=28) reported that they preferred the FTFI, for the following reasons: (1) real conversation is talking directly to a person (n=7), (2) through the FTFI, real speaking ability can be improved (n=4), (3) they can do themselves justice (n=2), (4) there is a listener (n=2), (5) they like talking to a person

(n=2), (6) it is possible to clarify the questions and correct answers (n=2), (7) it reflects real speaking ability (n=1), (8) talking to a person makes one feel more relaxed and comfortable (n=1), (9) non-verbal communication such as facial expressions and gestures can be used (n=1), (10) unclear (n=2) or missing (n=4).

33% of the test takers (n=14) preferred the CAOT. Those who preferred the CAOT provided the following reasons: (1) it makes them feel less nervous and less pressured (n=6), (2) they are all under the same conditions (n=2), (3) it excludes a subjective view (n=2), (4) it provides sufficient preparation and response time (n=2), (5) they don't need to wait and it is easier to test (n=2).

In brief, according to the responses to the open questions, more test takers felt that the FTFI was more difficult (though it was not significant) mainly because they felt more nervous as talking in front of a teacher and embarrassed when they made a mistake or gave poor answers. These responses appear to be related to Korean peoples' cultural tendency not to lose their face. It was discussed more in 5.1.

For the test takers, it seems that a fair test is the one which can provide the identical questions under the same conditions allowing raters to assess objectively. In this sense, the FTFI may be an unfair or less fair test for them.

It has been claimed that not only the CAOT but also the FTFI formats are not true performance tests because they are not direct simulations of real-world roles and tasks (i.e., the simulation of real-world tasks in the tests is superficial) (Wesche, 1992), as Clark (1979), Underhill(1987), and Van Lier (1989) also indicate (see 2.3.2). It implies that neither test is able to measure the ability to perform in the real-life. The FTFI was, however, perceived by the test takers to be more authentic reflecting their true speaking abilities due to the face-to-face human contact than the CAOT and for this reason, more test takers preferred the FTFI even though more test takers answered that the FTFI was more difficult and less fair than the CAOT.

In order to make the nine closed statements of Section Two more manageable for analysis and to interpret them more substantially, factor analysis was carried out. For the factor analysis, the data from both test formats were combined into one data set, as it was assumed that the test takers taking both tests would have interpreted the questionnaire items in a similar way.

One of the nine items (Q15) was excluded because the validity of the item was doubtful, as mentioned earlier and the communality variance was increased through the elimination of the item: 59.42 to 63.93. Thus, Q15 was excluded from the analysis and

discussion, although the test takers' responses to the question did show a statistically significant difference (see Table 4.1).

The result of the factor analysis for the eight items with varimax rotation yielded a three-factor solution, as presented in Table 4.5.

Table 4.5 Results of the factor analysis for attitude items

Item	Factor 1	Factor 2	Factor 3	Communality
9 enough questions	0.82			0.69
7 a good test format	0.75			0.57
8 do oneself justice	0.68			0.60
12 nervousness		0.79		0.63
14 tiredness		0.77		0.64
11 enough response time			0.82	0.69
10 enough preparation time		0.54	0.69	0.80
13 fairness			0.61	0.46
Eigenvalue	1.85	1.67	1.59	
% of Variance	23.14	20.87	19.91	
Cum. Variance	23.14	44.01	63.93	

Three items loaded strongly on the first factor: 'the test provides sufficient questions to assess my speaking ability' (.82), 'it is a good test format as an exam' (.75), 'I could do justice to my ability' (.68). Thus the first factor was labeled as 'test validity', which appears to relate to the test takers' perception of the validity of the test.

Two variables loaded strongly on the second factor. They were: 'I was nervous while I was taking the test' (.79), and 'I had to wait for a long time before taking the test, and

it made me tired' (.77). Thus the second factor was labeled as 'negative feelings', which seems to relate primarily to the test takers' feelings of anxiety.

Three variables loaded strongly on the third factor. They were: 'I had enough response time' (.82), 'I had enough time to think about the questions before I spoke' (.69), and 'the questions asked in the test were fair' (.61). Hence, the third factor was labeled as 'test procedures'.

The factor which captures a larger amount of variance for data explanation was placed in a higher priority than the others. Thus the results of this study showed that factor 1 explained 23.14% of the variance, followed by 20.87% and 19.91% for factors 2 and 3, respectively.

Eigenvalues indicate the variance explained by each factor, and they should be greater than 1 if the factor is deemed to be helpful to explain the variances of the variables. In Table 4.5, the eigenvalue of each of the three factors exceeded 1: Factor 1 (1.85), Factor 2 (1.67), and Factor 3 (1.59). Overall, more than 63% of the total variance was explained by the three factors. The other 37% of the variance was not accounted for by the three-factor solution because it was unique variance, which could only be captured by the individual items.

In order to test the factor-reliability of the FTFI and the CAOT, reliability analysis, which is often used in conjunction with factor analysis, was performed. The result is presented in Table 4.6.

Table 4.6 Reliability of the factors

Format	Factor 1 Test validity	Factor 2 Negative feelings	Factor 3 Test procedures	Total
FTFI	0.62	0.55	0.55	0.64
CAOT	0.71	0.58	0.59	0.70

The coefficients of Factor 1 for both FTFI and CAOT were higher than the other two. Generally, a higher coefficient indicates greater consistency of the contribution of the variables in measuring the underlying dimension. Hence, the variables of Factors 1 for the both tests were more consistent than the variables of Factor 2 and 3 in the measurement. Table 4.6 also shows good reliability for the measurement (Cronbach alpha $r=.64$ for the FTFI and $.70$ for the CAOT).

In conclusion, the results of the factor analysis above suggests that the eight items about the test takers' attitudes toward the FTFI and the CAOT could be analyzed and interpreted from the constructs of the three factors. Single quotation marks (' ') were used for the factors to highlight and distinguish them.

Wilcoxon tests and effect sizes were performed again, to investigate the test takers' attitudes to the two oral test formats to examine if there were any significant attitude

differences between attitudes toward the FTFI and the CAOT using the factors. It was assumed that it would be helpful for a more comprehensible interpretation of the test takers' attitudes towards the tests.

Data coding of the items of 'negative feelings' was carried out for the total attitudes by changing scale point 1 to 4, 2 to 3, 3 to 2, and 4 to 1, such that the higher median of the total attitude meant more favorable attitudes. The results are shown in Table 4.7.

Table 4.7 Wilcoxon tests of attitude scores of the FTFI and the CAOT

Attitude factors	FTFI		CAOT		Z	d
	Median	Min-Max	Median	Min-Max		
Test validity	2.67	2.00-4.00	2.67	1.67-4.00	-1.70	0.30
Negative feelings	2.25	1.00-4.00	3.00	1.50-4.00	-4.43**	-0.95
Test procedures	3.00	2.00-4.00	3.00	2.33-4.00	-2.03*	-0.38
Total attitudes	2.75	2.13-4.00	2.75	2.25-4.00	-2.02*	-0.43

*P<.05 **p<.01

The results indicate that the test takers had significantly different attitudes in 'negative feelings' (nervousness and tiredness) ($Z=-4.43$, $p=.00$, $d = -.95$) with more 'negative feelings' toward the FTFI: the higher median of the total attitude means more favorable attitudes. The results also show significant differences in 'test procedures' (preparation, response time, and fairness) ($Z=-2.03$, $p=.04$), with a higher opinion of the 'test procedures' in the CAOT, although the effect size is small ($d=-.38$). The findings suggest that the test takers had fewer 'negative feelings' to the CAOT than to the FTFI,

and were slightly more positive attitudes to the ‘test procedures’ for the CAOT than those of the FTFI.

Subsequently, Wilcoxon and Mann-Whitney tests were carried out to see if there were any significant attitude differences between gender and between age groups and effect sizes were also calculated: age group was divided only into two because of the small sample size. The results are presented in Tables 4.8 and 4.9.

Table 4.8 Wilcoxon and Mann-whitney tests of attitude scores by gender

Attitude factors	Formats	Gender		Z (<i>d</i>)
		Female (n=27)	Male (n=15)	
Test validity	FTFI	2.67	3.00	-1.19 (-0.35)
	CAOT	2.67	2.33	-0.64 (0.05)
	Z (<i>d</i>)	-0.84 (0.18)	-1.74 (0.51)	
Negative feelings	FTFI	2.50	2.00	0.00 (0.06)
	CAOT	3.00	3.00	-0.26 (0.07)
	Z (<i>d</i>)	-3.52** (-0.95)	-2.71** (-0.92)	
Test procedures	FTFI	3.00	3.00	-0.30 (0.30)
	CAOT	3.00	3.00	-0.54 (-0.14)
	Z (<i>d</i>)	-1.09 (-0.23)	-1.79 (-0.66)	
Total attitude	FTFI	2.63	2.75	-0.67 (0.00)
	CAOT	2.75	2.75	-0.05 (0.00)
	Z (<i>d</i>)	-1.51 (-0.38)	-1.42 (-0.50)	

*P<.05 **p<.01

Table 4.8 indicates that both female and male groups had significantly ‘negative feelings’ to the FTFI versus the CAOT (Male: Z=-2.71, p=.00, *d*= -.95, Female: Z=

-3.52, $p=.01$, $d=-.92$). Both groups had fewer 'negative feelings' to the CAOT, like the total group (see Table 4.6). In the male group, the effect sizes for the differences between attitudes to the 'test validity' of and 'test procedures' for the FTFI and the CAOT were in the medium range (test validity: $d=.51$, test procedures: $d=.66$) although the Wilcoxon tests did not show any significances for them. It may imply that the male students tended to have more favorable attitudes to the 'test procedures' for the CAOT than those for the FTFI, but to the 'test validity' of the FTFI than that of the CAOT. However, Mann-Whitney tests didn't show any differences in responses between female and male groups. Thus, it can be concluded that gender differences did not influence attitudes to the FTFI and the CAOT.

Table 4.9 Wilcoxon and Mann-Whitney tests of attitude scores by age

Attitude factors	Formats	Age		Z (<i>d</i>)
		Under 20 (n=22)	Over 21 (n=20)	
Test validity	FTFI	2.67	3.00	-2.33* (-0.70)
	CAOT	2.67	2.67	-0.18 (0.13)
	Z (<i>d</i>)	-0.15 (-0.05)	-2.54* (0.91)	
Negative feelings	FTFI	2.25	2.25	-0.71 (-0.22)
	CAOT	3.00	3.00	-0.21 (0.00)
	Z (<i>d</i>)	-3.45** (-0.97)	-2.89** (-0.93)	
Test procedures	FTFI	3.00	3.00	-0.95 (0.33)
	CAOT	3.00	3.00	-2.75** (0.99)
	Z (<i>d</i>)	-2.15* (-0.60)	-0.26 (-0.13)	
Total attitudes	FTFI	2.63	2.75	-1.85 (-0.31)
	CAOT	2.88	2.75	-1.46 (0.51)
	Z (<i>d</i>)	-2.44* (-0.67)	-0.14 (0.00)	

*P<.05 **p<.01

More interesting than the gender results were the results for age groups. Like the gender groups, each of the age groups showed a significant difference in ‘negative feelings’ and the effect sizes were large (Under 20: Z=-3.45, p=.00, *d*=-.97, Over 21: Z=-2.87, p=.00, *d*=-.93). Table 4.9 indicates that the younger group felt they had better ‘test procedures’ in the CAOT than in the FTFI (Z=-2.15, p=.03, *d*=-.60). This is probably the source of the significantly more favorable attitude in total attitudes to the CAOT (Z=-2.44, p=.02, *d*=-.67). On the other hand, the over 21 group showed no significant differences in attitudes to the ‘test procedures’ on the FTFI and the CAOT; the results of the Mann-Whitney test and the effect sizes indicate the significantly

different responses about the 'test procedures' between age groups ($Z=-2.75$, $p=.01$, $d=.99$).

The older group (over 21) did show significantly more positive attitudes about the 'test validity' of the FTFI ($Z=-2.54$, $p=.01$, $d=.91$), and the results of the Mann-Whitney test also indicate that there were significantly different responses between the two age groups ($Z=-2.33$, $p=.02$, $d=-.70$). The over 21 group showed more positive attitudes to the 'test validity' of the FTFI than the under 20 group did. This suggests that the older group had relatively more favorable attitudes to the FTFI, compared with the younger group (under 20), which had more favorable attitudes toward the CAOT.

Whilst the younger test takers had more positive attitudes in 'negative feelings' and 'test procedure' to the CAOT than to the FTFI, the older test takers had more positive attitudes in 'test validity' to the FTFI than to the CAOT. This shows, in the older group, a higher perception of the 'test validity' of the FTFI than in the younger group, in spite of the same 'negative feelings' to the FTFI as the younger group. It seems that the older test takers had relatively more favorable attitudes to the FTFI compared with the younger group overall.

In conclusion, the test takers had more negative attitudes to the FTFI than the CAOT in terms of preparation time, nervousness, tiredness, and fairness, as the results of the

factor analysis indicating that the test takers had more 'negative feelings' (nervousness and tiredness) towards and slightly less positive attitudes to the 'test procedures' (preparation and response time, and fairness) for the FTFI.

Participants, however, showed a clear preference for the FTFI over the CAOT, despite the negative attitudes to the FTFI. According to the results of the crosstabulations and the responses to the open questions, they preferred the FTFI mainly because it was perceived as having a better test format, included a genuine communication and interaction with a person (thereby seeming perceivably more valid), and allowed them to feel more comfortable and natural.

Males and females didn't have any differences in attitudes, but the older test takers (over 21) perceived that the FTFI had significantly more 'test validity' than the CAOT, showing that they had a higher perception of the 'test validity' of the FTFI than the younger test takers (under 20) did; on the other hand, the younger test takers showed significantly more positive attitudes to the 'test procedures' for the CAOT than the older group did. The test takers' attitudes were explored more in the analysis of the semi-structured interviews.

4.1.1.2 Computer familiarity

In order to investigate if the attitudes toward the oral tests were influenced by computer familiarity, Section One on computer familiarity (Qs1-6) was analyzed to examine the relationship between computer familiarity and attitudes.

The medians of all the test takers' responses to the questions about computer familiarity were taken by summing the response values: the lowest value for a question was 1 and the highest value was 4. Most of the test takers were very familiar with a computer: the average familiarity score of the test takers was high (the median for test takers' computer familiarity=3.02).

Computer-familiarity differences by gender and age were examined before the investigation of the effects of computer familiarity on attitudes, because it was hypothesized that computer familiarity might be related to gender and age. The results are presented in Table 4.10 below.

Table 4.10 Computer familiarity by gender and age

Subgroup		Computer familiarity		
		N	Median (Min-Max)	Z (d)
Gender	Female	27	2.75 (2.25-4.00)	-1.95 (-0.48)
	Male	15	3.00 (2.50-4.00)	
Age	Under 20	22	3.00 (2.50-4.00)	-1.92 (0.59)
	Over 21	21	2.75 (2.25-4.00)	

*P<.05 **p<.01

The table indicates that the female group (M=2.75) had less familiarity than the male group (M=3.00). The Mann-Whitney test did not show a significant median difference, but p-value (p=.051) was just below significance and a very near medium effect size was found. Table 4.10 also shows that there was no significant difference between the groups, but the computer-familiarity median of the under-20 age group was higher than that of the over-21 age group, and the p-value (.054) was close to significance again and the medium effect size was discovered ($d=.59$). The Mann-Whitney tests for gender and age groups might have proven significant had a larger sample been used.

Therefore, findings do indicate that there were tendencies towards a correlation between computer familiarity and gender on the one hand and computer familiarity and age, on the other. The male and under-20 groups in this study tended to be more familiar with computers.

Finally, to examine the relationship between computer familiarity and attitudes towards the FTFI and the CAOT, Spearman correlations were carried out.

Table 4.11 Correlations between computer familiarity and attitudes

Attitude factors	FTFI (n = 42)		CAOT (n = 42)	
	<i>r (p-value)</i>	95% CI	<i>r (p-value)</i>	95% CI
Test validity	-0.06 (0.70)	-0.36-0.25	0.15 (0.34)	-0.16-0.43
Negative feelings	0.17 (0.26)	-0.14-0.45	0.03 (0.81)	-0.28-0.33
Test procedures	0.12 (0.45)	-0.18-0.40	0.49** (0.00)	0.22-0.69
Total attitudes	-0.09 (0.55)	-0.38-0.22	0.21 (0.17)	-0.10-0.48

*P<.05 **p<.01

Table 4.11 shows that there was only one significant relationship between computer familiarity and attitude factors. Computer familiarity and test takers' attitudes to the 'test procedures' (preparation and response time, and fairness) in the CAOT were significantly correlated ($p=.00$). That is, the more familiar the test takers were with a computer, the more they felt they had a better 'test procedures' on the CAOT.

To sum up, the test takers in this study were very familiar with a computer. Male and younger test takers tended to have higher computer familiarity than female and older test takers, although this difference did not reach significance. Computer familiarity seemed not to be associated with the perception of 'test validity' or 'negative feelings' to either test, except 'test procedures' on the CAOT. The higher the test takers' computer familiarity was, the better they perceived the 'test procedures' for the CAOT to be.

4.1.2 Interview

After transcribing and reading the interviews, I found that most of the ten interviewees (seven interviewees) had generally more favorable attitudes to the FTFI rather than to the CAOT (three were more favorable to the CAOT, but none of them appeared to have

very negative attitudes to either test). This was a similar result to that of the questionnaire, which concluded that the test takers preferred the FTFI overall, in spite of their more 'negative feelings' and less positive attitudes to the 'test procedures' for the FTFI (as opposed to the CAOT).

To code the interviewees' responses, I read and lined interesting passages in three interviewees' interviews and then the passages marked as interesting were labeled tentatively in order to conceptualize and create categories regarding the attitudes to the two oral tests.

The categories were carefully then revised through a process of constant comparison as each subsequent interview was analyzed. I paid particular attention to attitudes that stood out and/or regularly occurred throughout the data. The interviewees frequently stated some perceived advantages and/or disadvantages of the tests compared with each other. Through the coding process, four categories were generated from the data: high computer familiarity, interaction, nervousness, and unfamiliarity with the CAOT.

For reliability of the coding, the categories were checked by another coder, who is a Doctor of Education and my coworker in the school. There were a few discrepancies in the coding and interpretation, so we discussed and revised them until a consensus was reached.

4.1.2.1 High computer familiarity

There has always been the issue of computer familiarity in computerized language testing. The lack of familiarity with a computer has been assumed to be a major factor in achieving negative attitudes to, as well as lower scores on, a computerized language test.

Most interviewees in this interview study showed very high familiarity with a computer; the same result as the analysis of the questionnaire.

The interviewees answered that they had generally used a computer about for 2-3 hours on average almost every day for checking e-mails, surfing the net, reading news, chatting, visiting or running friends' or their own mini-homepages, playing a game, using a word processor, and so on. Older interviewees (over 26) who said they didn't play games, chat, or run homepages online, had spent less time using a computer; approximately 1-2 hours every day on average.

Using a computer was identified by all of the interviewees as very familiar, interesting, convenient, or even fun:

- It makes me concentrate 100%, and it is interesting
- Well...I don't know if it is interesting, but it's very convenient and familiar. That's just everyday life.
- I usually use a computer for reading online-news rather than paper-news because it is more convenient. Using a computer is part of my daily life.

It was assumed that computer familiarity may affect test takers' attitudes to the CAOT positively or negatively, but in this interview study the interviewees' high computer familiarity seemed not to affect their attitudes about the CAOT. Regardless of the level of relative familiarity among the interviewees, they seemed to have more or less favorable attitudes to one test or the other. The following responses from the interviewees clearly suggest that their experiences with and feelings about personal computer use are different from computer usage for the purpose of the test. Their computer familiarity seemed to have nothing to do with their attitudes to the tests. Some interviewees stated that:

- I feel very comfortable when I use a computer. I mean it is okay to utilize a computer for personal use, but I don't feel very happy to use a computer for answering and recording for a test. It is very unfamiliar and uncomfortable.
- I am using a computer at least 2 hours every day. It is fun and enjoyable, but it seems that my personal use of a computer and the computer use for taking a test were not the same. I experienced the FTFI several times and it was very familiar, but I have never taken a computer test before. I think the thing that I could do well on the FTFI could not be done well on the CAOT.
- I am using a computer every day and doing many things with a computer...I felt more comfortable to talk to a computer because I had enough preparation and response time on the CAOT. On the other hand, I felt so nervous on the FTFI and felt that I needed to answer quickly because you were assessing and waiting for my answers.

The test takers' positive and negative attitudes seemed to be related to other factors such as familiarity with a test format, preparation and response time, or nervousness, rather than computer familiarity.

This interview study found that the test takers' attitudes to a computer were different from those toward the CAOT and they appeared unrelated to each other.

This finding was similar to that of the quantitative study, which found that computer familiarity was not associated with the two aspects of the attitudes, 'negative feelings' or 'test validity' of the two tests, although it was related to the attitude to the 'test procedures' for the CAOT.

4.1.2.2 Interaction

The main characteristic of the FTFI commented on by the interviewees was the *interaction* with a person.

The interviewees seem to generally think that a conversation is a face-to-face oral interaction between two or more persons. The following responses demonstrate that the interviewees viewed the FTFI as a real conversation interacting with a person, and thus seemed more valid than the CAOT.

- Talking with a person is a real conversation. You made me relaxed and talk freely. In the CAOT I felt a little strange because I had to speak alone. Speaking to a machine is not natural. It is not a conversation.

- We cannot talk well alone. Of course the computer gave us questions, but I didn't feel like I was making a conversation. We can read a book alone, but can not talk by ourselves...A speaking test is not to assess grammar or vocabulary. I think a conversation with others should be assessed. There was no interaction between persons. The FTFI seems better to assess our actual speaking ability.

From the evidence, it can be argued that the FTFI was perceived as a better test format which could measure their speaking ability more properly than the CAOT. Some similarly remarked that how to interact with a person involving nonverbal communication (e.g., gestures, body languages or eye contact) was also an important factor which should be assessed as part of their speaking ability. In the examination of their test performance which was analyzed in 4.2.1, it was revealed that in general the test takers got higher scores on the CAOT, but over half of the interviewees thought that they could perform better on the FTFI because they had more effective communication with the teacher using nonverbal communications. Nonverbal communication may have been part of the linguistic resources they thought they could do more justice to than with their ability in the FTFI. One interviewee stated:

- I could express some difficult words I could not remember well with body languages or gestures. I think they are also important in communication. But in the CAOT I had to speak alone all the time, and it was harder for me.

Another reason the interviewees thought that they performed better on the FTFI was that they could have much more control in an interaction with a teacher (the researcher) in the FTFI than in the limited control of their interactions with a computer on the CAOT. The interviewees mentioned that they could ask the teacher to repeat a question they couldn't understand, could ask about some difficult words they didn't know or about the tasks they were asked to do, and could easily change or correct their answers.

- I think I might have performed worse on the CAOT. I could not ask to repeat a question when I couldn't listen to the question clearly or ask a question I needed, but in the FTFI I could easily correct my answer by simply saying 'Sorry, can you say again?'

From the following responses, it seems that the presence of a person who listens and reacts to them encourages them to talk more and allows for some warm feedback from the teacher such as smiling and nodding, thereby helping the interviewees feel comfortable and relaxed.

- I felt something insufficient and tried to talk more in the FTFI rather than the CAOT. That's because you are looking at me expecting more talk from me. But in the CAOT just answered the question. I answered only what I needed to talk for the question.

- I like the FTFI because there is a person who can react to my answer such as nodding, smiling, or short responses. Those made me feel so comfortable and interested. But the CAOT seems to discourage me to talk more because there isn't anyone who listens and reacts to me. The CAOT just records my answers and speaking to a computer is strange and makes me uncomfortable.

Here, there were contradictory responses, which were also revealed in the analysis of the questionnaire. Although many of the interviewees remarked that they felt more nervous in the FTFI, some of them also said that the feedback from the teacher made them comfortable, whereas talking to a computer is uncomfortable. The following responses seemed to indicate the reasons:

- It is an inconsistent idea, but it was certainly more nervous in the FTFI because I had to answer looking at you, but I was also more comfortable talking to you. I don't know well, but it might be because it was a machine. Talking to a machine is not natural and a little odd.
- Just before and at the beginning of the test, I was more nervous in the FTFI, but I gradually became comfortable because you gave me warm feedback. I think I was nervous, but also felt more comfortable talking to you than to a computer.

The reactions and feedback from the teacher were the main reasons they preferred the FTFI. They liked the feedback from the teacher even though I tried to provide very limited feedback in order to more closely equate it to the tasks of the CAOT. The following responses undoubtedly express their preference for the FTFI over the CAOT.

- I enjoy conversation because of the interaction between persons. Through the interaction, we can have a deeper conversation about a subject when it is needed. But the CAOT gives only questions and I have to simply answer them. That's it. It is awkward to laugh or give facial expressions to a computer.
- The FTFI was more interesting because of the reactions from you, but the CAOT was boring and uninteresting because there weren't any responses.

It seems that test takers generally had more favorable attitudes to the FTFI because they perceived that it to be a more valid test format, in providing a real interaction with a person, thereby better reflecting actual speaking ability. They liked the feedback and reactions from the teacher. Some assumed that the feedback helped them to perform better by making them feel more comfortable and allowing for some nonverbal communication.

4.1.2.3 Nervousness

Less nervousness (or anxiety) was identified by many interviewees as a major advantage of the CAOT. The interviewees remarked that they could do themselves justice and perform well in the test where they felt comfortable. In particular, the interviewees favouring the CAOT felt that they performed better on the CAOT because they felt less nervous in the CAOT than the FTFI. The interviewees answered:

- I felt more comfortable in the CAOT. In the FTFI you were looking at me waiting for my answers. You were assessing my speaking ability. I felt so uneasy and it seemed I had to answer quickly. But in the CAOT there was nobody listening to me, and I could spend as much time as I want to think and answer without much nervousness.
- In the FTFI once I felt I made a mistake, it appeared to negatively influence my next answer because I worried about the mistake I previously made, but in the CAOT I could have enough time to get ready for the next question.

- Talking directly to you made me more nervous. You know, the computer does not know if I give a wrong answer...I didn't feel embarrassed in the CAOT but felt embarrassed and uneasy in front of you when I gave a poor answer.

However, a few of the interviewees thought that too little tension and too much comfort in the CAOT led them to perform worse on the CAOT than the FTFI. They claimed that causing pertinent tension and nervousness helped them to perform better on the FTFI. One of them said that he seemed not to do the test to the best of his ability since the CAOT did not make him nervous, to the point of being bored, and he didn't feel like he was even taking a test. A certain amount of tension and nervousness in the FTFI seems to have encouraged some of the test takers to perform better.

Contrary to the interviewees favoring the CAOT, a few interviewees who had strong positive attitudes to the FTFI stated that they felt more nervous and uncomfortable in the CAOT because of a timer on the screen which told them how much time they had remaining:

- I felt I had less time in the CAOT because of the timer I could see on the computer screen. It seemed to force me to answer quickly in succession without breaking.

- You know, the timer showing how much time I had used made me more nervous and anxious, but in the FTFI you led me to answer questions more appropriately when I didn't do well.

One interviewee commented that the CAOT was uncomfortable because it made her more concerned about her pronunciation and grammar, due to the idea that her

performance would be assessed only with the recording. There seemed to be more pressure to make grammatically correct sentences.

In brief, many interviewees felt less nervous in the CAOT because of the absence of a teacher or person and the allowed preparation and response time, which they could control. They assumed that they performed better on the CAOT. However, a few of them thought perhaps over-comfort in the CAOT negatively influenced them, by not compelling them to do their very best. For the interviewees who felt more anxious in the CAOT, the timer on the computer screen and recording mainly caused them to feel nervous and uncomfortable.

4.1.2.4 Unfamiliarity with the CAOT

Unfamiliarity seemed to be one of the primary reasons the interviewees had less favorable attitudes to the CAOT. Four of the seven interviewees favoring the FTFI mentioned familiarity with the FTFI format. In particular, older interviewees (over 21) had previous experiences in talking with native speakers or in taking the FTFI in schools, companies, or private academies. On the other hand, some interviewees said that they were worried whether they were operating the computer program properly and if the recording was working correctly. It was feared that teachers might not be able to

assess their performance well because their answers might not be recorded or recorded poorly. Further, one or two interviewees stated that they found it difficult to concentrate on their speaking on the CAOT. Because the test was too comfortable they felt more curious about the interesting features of the new test, such as pictures and videos, rather than their performance. They also stated that the use of the mouse and the screen seemed to disturb their concentration such that sometimes they forgot what they had to say next, unlike the FTFI, which required them only to talk. Two of the interviewees mentioned that:

- Before this university, I used to go to company where I needed to work with foreigners. I was accustomed to talk with them there and became unafraid of talking to a native speaker in English. I was not a good English speaker, but I could communicate with them using gestures or body languages. But I have never used the program of the CAOT. It was totally new for me. Therefore, I doubted and worried if I was handling it properly and if it was recording my speaking well.

- The CAOT was new and interesting. I was very curious about it. You know, the video, the pictures, and the colorful screens and so on. But those seemed to disturb in my concentration on speaking. In addition, I had to click and see the computer screen. But in the FTFI it could make me more concentrate only on my speaking.

The interviewees added that if they became familiarized with the CAOT, they would probably like the CAOT more, because they would be familiar with the format, would have learned how to use it properly, and could pay more attention to the testing, rather than the test. In addition, they thought that the CAOT would be easier to prepare for.

They might have performed better on the CAOT had they prepared before taking it; the interviewees said they had not prepared much for the tests because the assessment had nothing to do with their credits.

As shown from this interview and questionnaire study, test takers were very familiar with a computer, but more test takers preferred the FTFI. One of the reasons seems to be that they have never taken a test on a computer and thus have the fear of 'doing old things in new ways' as stated in Fulcher (1999: 296). This might have affected their attitudes toward the tests. By increasing familiarity with the CAOT, it might be possible to increase favorable attitudes towards it.

4.1.2.5 Others

Some of the interviewees mentioned fairness, objectivity and practicality as the advantages of the CAOT. The following responses suggested that the CAOT was perceived to be fairer, more objective, and more practical compared with the FTFI:

- The CAOT seems fairer because it gives exactly the same questions and the same number of questions. It may also be more objective because a rater could only hear the voice.

- I think the CAOT is practical because many test takers can take the same test at the same time. It will be much easier for teachers to use for assessing their test takers' speaking abilities. But in the FTFI teachers will need quite long time and much more efforts to administer it. It will be so hard for teachers, especially when there are a large number of test takers.

Some interviewees stated a disadvantage of the CAOT. They found it a little disturbing to have other test takers talking at the same time in the CAOT, because they could hear other test takers' voices and the other test takers might also hear their answers. Further, they mentioned that their test process became increasingly comparable to others near them.

4.1.2.6 Conclusion

In conclusion, this interview study, which aimed to more deeply understand Korean university students' attitudes towards the two oral tests, and their performance on them, identified test takers' 'high computer familiarity' (which seemed to be unrelated to the test takers' positive or negative attitudes to the CAOT) - although there was a relationship between computer familiarity and the attitude to the 'test procedures' for the CAOT in the quantitative study earlier - 'interaction', and 'familiarity', as the main causes of the test takers' favorable attitudes to the FTFI. In other words, 'unfamiliarity with the CAOT' was one of the causes of less favorable attitudes to the CAOT, and 'comfort (or less nervousness)' was the major cause of favorable attitudes to the CAOT.

It seems that the test takers had more favorable attitudes to the FTFI overall because it was a real conversation interacting with a person, unlike the CAOT interacting with a

computer. It is notable that some still had more favorable attitudes to the FTFI, even though they felt obviously more nervous in the FTFI than the CAOT. It was for the most part because the FTFI was perceived as a better test format for assessing real speaking ability. Furthermore, it needs to be noted that in the interview and the questionnaire some of the test takers who felt more nervous in the FTFI also answered that they felt more comfortable talking to a person than to a computer. They felt talking to a person was more natural and the feedback from the teacher made them gradually feel more comfortable and created a more relaxed atmosphere.

The test takers liked the interaction involving reactions to their answers, gestures and body languages. Familiarity with face-to-face interaction also caused the test takers to have more favorable attitudes to the FTFI. Conversely, unfamiliarity with the CAOT resulted in less favorable attitudes towards it.

Test takers who generally had more favorable attitudes toward the CAOT seemed to feel very nervous during the interaction with a teacher. When talking alone, without a teacher, they felt more relaxed and comfortable. Because of the comfort of speaking they preferred the CAOT to the FTFI. Some advantages (fairness, objectivity, practicality) and a disadvantage (disturbance of other test takers' voices) of the CAOT, stated by some interviewees, appeared not to heavily influence overall attitudes towards

the CAOT or the FTFI. The questionnaire, however, found a significant relation between the disturbance of other test takers' voices and their test preference. The interviewees tended to think that they performed better on the test which they had more favorable attitudes toward, regardless of their actual performance. Though higher scores were awarded on the CAOT, 67% of the test takers had greater preference for the FTFI.

The perceived reasons the test takers performed better on the FTFI were: comfort and naturalness in talking with a person, the interaction with a person involving reactions to their answers and open questions (e.g., clarification and modification), nonverbal communication (e.g., gestures, body languages or eye contact) and encouragement to talk more. On the contrary, the central reasons the test takers believed they performed worse on the FTFI were: less nervousness and the convenience of controlling preparation and response time.

Based on the findings of this interview study, it can be asserted that it may be possible to make the test takers have more positive attitudes towards both tests by constructing the FTFI to make participants less nervous and by adding more interactive functions to, and raising familiarity with, the CAOT.

4.2 Performance

This section examines the test takers' performance and the effects of other factors which might have affected their performance and scores: test takers' performance, raters' severity, test format effect, test-order effect (i.e. the FTFI first and the CAOT second or vice versa), item difficulty, and bias between rater and test format were investigated. Furthermore, the effects of computer familiarity, gender and age on performance on the FTFI and CAOT were also examined. Finally, the relationship between attitudes and performance in the two tests were investigated and discussed. For this examination, 'multi-faceted Rasch measurement' with the computer program FACETS, version 3.45 (Linacre, 1991-2003) was mainly used.

4.2.1 Performance on the FTFI and the CAOT

The data were entered into the Rating Scale Model (RSM; Andrich, 1978), which operates under the assumption that the rating scale associated with each item functions similarly since the step difficulty of the category of the rating scale used in this study was assumed to be equivalent across all items; on the other hand, the Partial Credit Model (PCM; Masters, 1982) constructs a separate scale structure for each item (i.e., the step difficulty varies from item to item).

Three FACETS analyses were conducted: one for the FTFI, one for the CAOT, and one for the combined data. For the FTFI and the CAOT taken separately, the three facets of test taker, rater, item were included in the analysis. For the combined data the five facets were included in the analysis: (1) test taker (2) rater (3) test format (4) order (i.e., the order in which each test was undertaken by a test taker) (5) item. A graphic ruler for the combined data is the only one presented here because of space considerations. The five-facet Rasch model can be formulated as follows:

$$\text{Log} (P_{nijhmk}/P_{nijhmk-1}) = B_n - D_i - C_j - E_h - G_m - F_k$$

Where

P_{nijhmk} is the probability of test taker n being rated k on item i of test h in a test order m when rated by judge j ;

$P_{nijhmk-1}$ is the probability of test taker n being rated $k-1$ on item i of test h in a test order m when rated by judge j ;

B_n is the ability of test taker n ;

D_i is the difficulty of item i ;

C_j is the severity of judge j ;

E_h is the difficulty of test h ;

G_m is the difficulty of test order m ;

F_k is the difficulty of being rated in category k rather than category $k-1$.

Figure 4.2 shows pictorially the differences across the different elements of each facet. It visually displays the relative abilities of test takers, the relative toughness of raters, the relative difficulty of test order, the relative difficulty of test format, and the relative difficulty of items on average in terms of the scale of probability developed in the analysis.

There are seven columns in the figure. All elements in each facet are located at a certain point on the scale in column 1. This acts as a 'ruler' and allows us to compare the ability of the test taker, the severity of raters, and the difficulty of other elements across and within facets. This measurement scale is used to report estimates of probabilities of test taker responses under the various conditions of measurement such as ability, difficulty, rater severity and other possible factors which have been entered into the analysis: the ability estimate is called a 'measure' to differentiate it from the original raw score (McNamara, 1996).

Test takers are shown in the second column. Test takers are ranked by ability, with high ability at the top portion of the column and low ability at the bottom. The third column shows rater severity with the more severe rater at the top of the column and the

more lenient at the bottom. The difficulty of the test taking order is presented in the fourth column with the more difficult test order at the top of the column. Likewise, the fifth and sixth columns show the difficulty of test format and items, with more difficult at the top of the column. Finally, the last column graphically describes the rating scale for the raw score.

The test taker facet was allowed to 'float' in the model (the mean of measures was not constrained to equal zero) in order to compare the abilities of the test takers relative to elements of other facets, and all other facet elements forced means of zero logits. On the other hand, the mean value of each group (ftfi-caot, caot-ftfi) in test order facet was anchored at zero logits in the analysis as a means to a solution to the subset connection problem. This still allows a determination of the extent to which the test order differed in difficulty allowing the two test orders to float relative to that zero mean.

Figure 4.2 All facet summary for the combined data

measure	+taker	-rater	-format	-order	-item	S.1
+ 6	+	+	+	+	+	+(5)
+ 5	+ *	+	+	+	+	+
+ 4	+	+	+	+	+	+ 4
+ 3	+ *	+	+	+	+	+
+ 2	+	+	+	+	+	+
+ 1	+	+	+	+	+ flu com	+
*	rater 1	ftfi				
* 0	* *	* *	* caot-ftfi	ftfi-caot	*	* *
*	rater 2	caot				3
+ -1	+ ***	+	+	+	+	+
*					int	
+ -2	+	+	+	+	+ coh	+

**						
+ -3	+ ***	+	+	+	+	+
**						
*						
+ -4	+ *	+	+	+	+	+ 2
**						

+ -5	+	+	+	+	+	+

*						
+ -6	+ *	+	+	+	+	+
*						
+ -7	+	+	+	+	+	+
**						
+ -8	+ *	+	+	+	+	+
+ -9	+	+	+	+	+	+
*						
+ -10	+ *	+	+	+	+	+(1)
Measure	* = 1	-rater	-format	-order	-item	S.1

4.2.1.1 Test taker

As shown in Figure 4.2, test takers' ability estimates range from a high of approximately 6 logits to a low of -10 logits, a spread of 16 logits in terms of test taker ability. This unusually wide distribution seems to stem from some outliers who received distinctively higher scores than the others and other outliers with extremely low scores.

Two of the test takers had already got MA and (or) BA degrees from other universities and experience in working for large corporations: one person at +5 and the other at +3. They showed much greater ability than other test takers. In addition, although the distribution of the lower half of the scale looks much more normal, a few test takers received bottom scores through all items because they had extremely low English ability: in particular, the persons between -10 and -8. They could speak only isolated words and a few formulaic phrases and barely made a sentence.

Most of the test taker ability estimates are below zero, indicating that there was a mismatch between the difficulty of the tests and the ability of the test takers. The tests were very challenging and hard for them. The raw scores ranging only from 1 to 5 on a rating scale of 7 (mostly used 1 to 3) also suggested the test takers' low speaking abilities.

Furthermore, several test takers on each test had infit mean square values outside of a range of 0.5 to 1.5. Overfitting test takers (below 0.5) are those test takers whose scores did not vary that much across all items, across raters, or across test formats. For example, receiving the same or similar ratings on all items is very likely to produce overfitting values. Hence, investigation of the overfitting test takers would not provide much information about the way things such as raters and items influence outcomes. Rather, it would certainly be useful to examine the misfitting test takers (above 1.5) since they could give clues of some effects of factors on the test takers. The misfitting test takers are presented in Table 4.12.

Table 4.12 Misfitting test takers

FTFI		CAOT		Combined	
Test takers	Infit MS	Test takers	Infit MS	Test takers	Infit MS
11	1.9	8	1.5	14	1.7
14	1.5	15	2.2	22	1.7
22	2.3	17	1.9	26	1.5
26	2.8	26	1.8	28	2.0
27	2.9	28	2.6	32	1.8
		32	2.5	36	2.2
		35	1.6		

In Table 4.12, the many values above the criterion of 1.5 suggest too much variation in performance or inconsistent performance. Within the current research context, as Bonk and Ockey (2003) mentioned, the misfits would not result from guessing, careless

responses or sleeping test takers as the nature of the oral test situations prevent them.

The interpretation of the misfit is not easy and simple, but some possible causes may be that:

First, the misfits might have occurred when *Fluency, Grammar, Vocabulary*, or *Communicative effectiveness* scores were higher than *Cohesion* or *Intelligibility* because the model determined that they were more difficult than *Cohesion* and *Intelligibility*.

For example, Test taker 32's *Cohesion* on the CAOT was given lower scores than other items by the two raters: Rater One-3,3,3,4,2,3 and Rater Two-3,3,3,5,2,3.

Another cause for the misfits might be due to the error of raters. The raters might have scored the test takers inconsistently through the items or the scores for the same performance might have been too different between raters. For example, Test taker 14's performance on the FTFI was rated by Rater One- 2,2,2,2,3,2, but by Rater Two- 1,2,1,4,3,2; Test taker 8's performance on the CAOT was rated by Rater One- 1,1,2,3,2,1, but by Rater Two-2,2,2,3,3,2.

In the case of the combined data, other possible causes for the misfitting test takers can be assumed. The misfits might have occurred when the test takers' scores were higher on the FTFI than on the CAOT or when the scores on the CAOT were far greater than on the FTFI than the model expected. For instance, Test taker 36 was awarded

higher scores on the FTFI than the CAOT by the raters: Rater One- FTFI: 2,2,2,2,3,2;
CAOT: 1,1,1,1,1,1 and Rater Two- FTFI: 2,2,2,2,2,2; CAOT: 1,1,1,1,1,1.

Lastly, test takers' nervousness (or anxiety) or other aspects of attitudes can also be possible causes for the inconsistent performance. They were examined and discussed in 4.2.2.

Two or three of these possible causes might have worked together and resulted in an inconsistent scoring pattern when compared to the general pattern of scoring for the test cohort overall. They certainly impacted the test taker facet most directly.

In spite of the test takers' low speaking ability and the misfits, the wide distribution of the test takers' ability in the combined data resulted in a big separation index of 6.18 with an adjusted SD of 2.67. This means that the variance among the test takers was about six times the error of estimates. The reliability index of this combined data test taker facet is .97, demonstrating that the analysis separates test takers into different levels of ability with a high degree of reliability. Furthermore, the chi-square of the test taker facet, 1290.2 with 40 degree of freedom was significant at $p=.00$, indicates that the test takers did not have equal ability. Table 4.13 also shows wide distribution of test takers' ability on each test: separation index 5.03, 5.00 with reliability index .96 respectively.

Table 4.13 Summary statistics for test takers

Statistics Indices	FTFI	CAOT	Combined
Measure	-5.39	-3.43	-3.29
Fair-M average	2.11	2.23	2.19
SD	4.02	3.77	2.71
Error	0.72	0.73	0.43
Separation	5.03	5.00	6.18
Reliability	0.96	0.96	0.97
Chi-square	791.5	1017.2	1290.2
Degree of freedom	40	39	40
Significance	0.00	0.00	0.00

Taking these statistical indices and the wide band of the estimates of the test takers together, it can be concluded that, although the test takers had low speaking ability and some of them were misfits, the current tests generally discriminated among the test takers reliably in terms of their speaking ability.

4.2.1.2 Rater

Table 4.14 shows that the two raters had high observed and expected exact agreement rates through all tests. The fair-M average score of each rater had a small span, but there were higher observed and expected exact agreement rates on the CAOT than on the FTFI: observed exact agreement; 73.0% and expected agreement, 67.2%.

Their inter-rater reliability, which was computed for the total raw scores of the analytic scores produced from the FTFI and the CAOT, also showed reasonably strong Spearman correlations of $r=0.77$ for the FTFI and $r=0.87$ for the CAOT.

Table 4.14 Rater measurement report

Test format	Statistics indices	Rater 1	Rater 2	M	SD	Statistical indices
FTFI	Measure	0.31	-0.31	0.00	0.31	Separation: 1.93 Reliability: 0.79 Chi-square: 9.40 (p=0.00) Exact agreement: 62.7% Expected: 63.4%
	Error	0.15	0.14	0.14	0.00	
	Fair-M average	2.08	2.21	2.14	0.06	
	Infit MS	0.90	1.00	1.00	0.00	
CAOT	Measure	0.49	-0.49	0.00	0.49	Separation: 2.89 Reliability: 0.89 Chi-square: 18.8 (p=0.00) Exact agreement: 73.0% Expected: 67.2%
	Error	0.16	0.16	0.16	0.00	
	Fair-M average	2.12	2.29	2.20	0.08	
	Infit MS	0.90	1.10	1.00	0.10	
Combined	Measure	0.28	-0.28	0.00	0.28	Separation: 2.95 Reliability: 0.90 Chi-square: 19.30 (p=0.00) Exact agreement: 67.9% Expected: 60.7%
	Error	0.09	0.09	0.09	0.00	
	Fair-M average	2.11	2.25	2.18	0.07	
	Infit MS	0.90	1.00	1.00	0.00	

The fit statistics show that neither of raters was misfit, indicating that their ratings were fairly consistent. However, as shown in Figure 4.2 and the combined data in Table 4.14, Rater 1 (researcher) was slightly more severe than Rater 2 (native speaker teacher), and the analysis showed a significant chi-square given the separation index, 1.93, 2.89,

and 2.95 with reliability .79, .89 and .90 respectively, suggesting that the two raters differ in their severity. If the raters are one logit apart, this means that a given test taker has a 50% chance of getting a given score from Rater 1 and that the test taker only has a 25% chance of getting that score from Rater 2.

Therefore, it can be concluded that the two raters slightly differed in their relative severity even though they had quite reasonable inter and intra reliability. So there was a possibility a test taker gained a different score for the same performance when he or she had a different rater in this current study.

4.2.1.3 Test format

Equivalence of direct and semi-direct tests is a most interesting focus for researchers because if it can be established that the tests measure different speaking abilities, then the interchangeability of direct and semi-direct tests is in doubt.

The Spearman correlation between the logit scores obtained from the two tests was $r = .71$, which was not very high compared with previous studies on direct and semi-direct tests (e.g., Stansfield *et al.*, 1990; Kenyon and Malabonga, 2001). Traditionally, if the scores from the two different tests strongly correlated with each other then it has been assumed that they were measuring the same ability. However, McNamara (1991) argues

that the correlation test is only a measure of the linearity of the relationship and a more accurate test, a chi-square test, is needed for the test of the equality of ability estimates in order to overcome the limitations of the correlation test.

The chi-square test which compared the test taker ability estimates obtained from the scores on the FTFI and the CAOT, taking into account the relative difficulty of the two test formats, indicates that the two tests were not the same in terms of their difficulty: chi-square 17.1 ($p=.00$) with separation 2.75 and reliability .88.

In order to obtain a more accurate and clearer result of the test scores from the two tests, the six misfitting test takers in the combined data were excluded from the data and the analysis was conducted once again. The result of the second analysis (Table 4. 15) pointed more apparently to the existence of a test format effect.

Table 4.15 Test format measurement report using the edited data

Test Format	Measure	Error	Infit MS	Fair-M Average
FTFI	0.46	0.14	1.00	2.03
CAOT	-0.46	0.13	0.90	2.17
M	0.00	0.09	1.00	2.10
SD	0.46	0.00	0.00	0.07

Separation: 3.29 Reliability: 0.92 Chi-square: 23.60 d.f.: 1 Significance: 0.00

Consequently, the two tests were not the same in their difficulty even though there were attempts to make the tasks in each test as close as possible. Possible reasons for the difference in difficulty were discussed in Chapter 5.

4.2.1.4 Test order

The test order was counterbalanced to deal with a possible positive practice effect operating on their second test performance. But there was still a possibility that the test takers who took the FTFI first and the CAOT second had more advantages than the test takers who completed the CAOT first and the FTFI second or vice versa because the CAOT was a new form of an oral test using a computer, which had not been commonly used in other oral tests and would thus be very unfamiliar to the test takers. Under such circumstances, the order of test taking (i.e., the FTFI first and the CAOT second or the CAOT first and the FTFI second) could influence the test takers' performance more severely (Lee, 2004). Making it clear whether there was an order effect or not could give this study more validity and confidence in making inferences about the test takers' performance and in interpreting the results of this study. The test order measurement report is presented below.

Table 4.16 Test Order Measurement Report

Test Order	Measure	Error	Infit MS	Fair-M Average
CAOT-FTFI	0.00	0.09	1.10	2.18
FTFI-CAOT	0.00	0.09	1.00	2.18
M	0.00	0.09	0.80	2.18
SD	0.00	0.00	-2.00	0.00

Separation: 0.00 Reliability: 0.00 Chi-square: 0.00 d.f.: 1 Significance: 1.00

Figure 4.2 and Table 4.16 clearly show that the order of taking tests did not affect the test takers' performance. It can be seen that the separation and reliability indices are, .00 and .00, with chi-square index .00 and d. f. index, 1 ($p=1.00$). That is, the order of test taking (i.e., the FTFI first and the CAOT second or vice versa) was not different from each other in the measure of difficulty. No test takers had advantages or disadvantages from the test-orders.

4.2.1.5 Item

As shown in Figure 4.2 and the combined data in Table 4.17, the items of the combined data were ranked in order of difficulty from hardest to easiest, as follows: *Fluency-Vocabulary-Grammar-Communicative effectiveness-Intelligibility-Cohesion*. Namely, the most difficult items demonstrated to be *Fluency* and the easiest item *Cohesion*. As presented in the all-facet rulers (Figure 4.2), the items were not well targeted at many of the test taker populations in terms of difficulty, suggesting that the rating scale and

descriptors were not very satisfactory since a few of the score levels (higher level scores) were not utilized at all.

Table 4.17 Measures and Fair-M averages for items

Rating item	Measure			Fair-M Average		
	FTFI	CAOT	Combined	FTFI	CAOT	Combined
Fluency	1.40	1.42	1.06	1.88	2.03	1.93
Vocabulary	1.19	1.24	0.93	1.92	2.04	1.96
Grammar	1.40	0.98	0.93	1.88	2.07	1.96
Effectiveness	1.05	1.07	0.83	1.95	2.06	1.99
Intelligibility	-1.77	-2.75	-1.71	2.59	2.83	2.67
Cohesion	-3.25	-1.95	-2.05	2.97	2.65	2.77
M	0.00	0.00	0.00	2.20	2.28	2.22
SD	1.83	1.69	1.33	0.43	0.33	0.36

Table 4.18 Statistics Report for items

Statistics	FTFI	CAOT	Combined
Separation	7.18	5.90	8.32
Reliability	0.98	0.97	0.99
Chi-square	344.20	241.10	444.2
Degree of Freedom.	5	5	5
Significance	0.00	0.00	0.00

The fair-M averages indicate that all items of the CAOT were easier than those of the FTFI except *Cohesion*. Only FTFI *Cohesion* was easier than CAOT *Cohesion* (possible reasons were discussed in 5.2).

However, the items were similarly ordered in each separate analysis of their relative difficulty on both tests. The measures in Table 4.17 suggest that differences in measures

of *Fluency, Vocabulary, Grammar, Communicative effectiveness* did not constitute a massive difference in difficulty from item to item, but *Cohesion* and *Intelligibility* were far easier than other items on each test. The fair-M average indices also show the difference of item difficulty: the test taker fair-M averages of *Intelligibility* and *Cohesion* were distinctively higher than those of the other items. The high separation and reliability indices with significant chi-square values ($p=.00$) support the conclusion that the variability of item difficulty was big. The probable reasons are discussed below.

The test takers' pronunciation was much better than other aspects of language skills and their speech was not generally hard to understand. It appears to have led to the giving of higher scores of *Intelligibility* than those of other items. The test takers' good pronunciation might have arisen from recent approaches in Korean English education which emphasize pronunciation, speaking and listening.

However, the higher ability estimates of *Cohesion* were an unexpected result. This might be because the cartoons for the narrative tasks in both tests were previously shown to the test takers for about 30 seconds a week before the actual test to let them become familiar with the task, as stated in the test procedures. The test takers might have benefited from their familiarity with the narrative task. Some of them might even

have prepared for the narration task prior to the test and thus have been familiar with the process of formulating the meanings (Bygate, 2001).

The last probable reason might be that the step difficulty in logits might be different. That is to say, the step difficulty of *Cohesion* and *Intelligibility* scales might be easier than other items. However, as Bonk and Ockey (2003) also mention, the different step difficulty of the analytic rating scales does not mean that the rating scale is useless because they simply make the interpretation of combined scores from different scales more difficult. In the assessment, each item was used separately and their use is still acceptable.

Therefore, even if the rating scale and descriptors were not very suitable for the test takers in terms of their difficulty, the scores could be expected to be given fairly to the test takers since all test takers took the same test taking process and were rated on all items.

Intelligibility was identified as misfitting on both tests: Infit MS 1.6 on the FTFI and 1.7 on the CAOT. As mentioned earlier, an item is said to be misfitting where performance on the item is inconsistent with the pattern of performance on the test overall. *Intelligibility* might have been interpreted inconsistently by raters or did not

form part of a set of items which together define a single measurement trait. These might also be one of the reasons for the higher scores of *Intelligibility* than other items.

4.2.1.6 Bias

The results show that the FTFI was generally more difficult than the CAOT. However, it was still possible that individual raters judged performance on one format significantly more harshly or leniently than the other. This hypothesis was investigated using an additional facility within the FACETS program known as *bias analysis*. Table 4.19 shows the results from the bias analysis between raters and formats.

Table 4.19 Bias Calibration Report, rater and test format interaction

Rater	Format	Bias measure	Error	Z-score	Infit MS
1	FTFI	-0.03	0.13	-0.22	1.10
2	CAOT	-0.03	0.13	-0.22	1.00
2	FTFI	0.03	0.13	0.23	1.00
1	CAOT	0.03	0.13	0.23	0.80
Mean		0.00	0.13	0.01	1.00
S.D.		0.03	0.00	0.23	0.10

Chi-square: 0.20 d.f.: 4 Significance: 1.00

As stated earlier, where the Z-score values fall between -2.0 and +2.0, the rater is considered to be scoring the test without significant bias. Accordingly, the figures indicate that the two raters were not significantly more severe or lenient on a particular

format. The infit mean square values ranging between 0.5 and 1.5 indicate that the raters' scoring was similar and consistent for the formats overall. Furthermore, the chi-square .2 with d.f. 4 ($p=1.00$) also points out that there was no effect of rater bias in relation to the average difference it made to the raw scores allocated by the raters.

4.2.1.7 Computer familiarity, age and gender

For the investigation of the relationship between computer familiarity and test taker performance, correlations between computer familiarity and the test taker ability estimates from the FTFI and the CAOT were carried out. The results of the analysis were that none of the correlations showed significance. Thus, it can be concluded that computer familiarity was not associated with the test performance on the tests, at least in this study.

In order to examine whether the gender and age differences impact upon test performance on the FTFI and the CAOT, the bias analysis of the FACETS program was utilized again. Bias analyses are appropriate to investigate the impact because as explained above, the bias analysis in multi-faceted Rasch measurement can diagnose unexpected but consistent patterns of behavior from an interaction between a particular test taker or group of test takers and some factors of the rating situation. The gender and

age difference might have helped their performance on one format more or less than the other format. Table 4.20 presents the results of the bias analysis between test taker gender and test format below.

Table 4.20 Bias calibration report, gender and test format interaction

Gender	Format	Bias measure	Error	Z-score	Infit MS
Male	CAOT	-0.08	0.16	-0.52	0.80
Female	FTFI	-0.03	0.11	-0.27	0.90
Female	CAOT	0.06	0.11	0.56	0.90
Male	FTFI	0.12	0.16	0.75	1.40
Mean		0.02	0.14	0.13	10.00
S.D.		0.08	0.03	0.54	0.20

Chi-square: 1.20 d.f.: 4 Significance: 0.87

Since all the z-scores are within the range of -2 to +2 it can be concluded that gender was not significantly biased in favor of neither of the test formats. The second bias analysis examined whether there was a significant interaction between test taker age and test format.

Table 4.21 Bias calibration report, age and test format interaction

Gender	Format	Bias measure	Error	Z-score	Infit MS
Under 20	CAOT	-0.04	0.13	-0.32	1.10
Over 21	FTFI	-0.05	0.13	-0.35	1.00
Over 21	CAOT	0.05	0.13	0.35	0.70
Under 20	FTFI	0.04	0.13	0.34	1.10
Mean		0.00	0.13	0.01	10.00
S.D.		0.04	0.00	0.34	0.20

Chi-square: 0.50 d.f.: 4 Significance: 0.98

Table 4.21 shows again that the interaction between test taker age and test format was not significant. Therefore, test takers' performance on each test was not significantly affected by their age on either test.

Moreover, in Tables 4.22 and 4.23 the same fair-M averages and measures for the age and gender groups on each test format and the combined data clearly indicate that the difficulty of the tests was the same for the test takers regardless of their gender and age.

Table 4.22 Test taker gender measurement report

Test format		Male	Female	M	SD
FTFI	Observed -average	1.90	2.30	2.10	0.20
	Fair-M average	2.12	2.12	2.12	0.00
	Measure	0.00	0.00	0.00	0.00
	Error	0.18	0.12	0.15	0.03
	Infit MS	1.20	0.80	1.00	0.20
CAOT	Observed -average	2.10	2.40	2.30	0.10
	Fair-M average	2.21	2.21	2.21	0.00
	Measure	0.00	0.00	0.00	0.00
	Error	0.20	0.13	0.17	0.04
	Infit MS	0.90	0.90	0.90	0.00
Combined	Observed -average	2.00	2.40	2.20	0.20
	Fair-M average	2.18	2.18	2.18	0.00
	Measure	0.00	0.00	0.00	0.00
	Error	0.11	0.08	0.10	0.02
	Infit MS	1.10	0.90	1.00	0.10

Table 4.23 Test taker age measurement report

Test format		Under 20	Over 21	M	SD
FTFI	Observed -average	2.10	2.20	2.20	0.10
	Fair-M average	2.12	2.12	2.12	0.00
	Measure	0.00	0.00	0.00	0.00
	Error	0.14	0.14	0.14	0.00
	Infit MS	1.00	0.80	0.90	0.10
CAOT	Observed -average	2.30	2.30	2.30	0.00
	Fair-M average	2.20	2.20	2.20	0.00
	Measure	0.00	0.00	0.00	0.00
	Error	0.15	0.16	0.16	0.00
	Infit MS	1.20	0.70	0.90	0.30
Combined	Observed -average	2.20	2.30	2.20	0.00
	Fair-M average	2.18	2.18	2.18	0.00
	Measure	0.00	0.00	0.00	0.00
	Error	0.09	0.09	0.09	0.00
	Infit MS	1.10	0.80	1.00	0.20

From the above analyses it appears that there was no impact of either gender or age difference on test scores on the FTFI and the CAOT and the difficulty of each test was not different according to gender or age, i.e., gender and age differences did not get more advantages or disadvantages from the test formats.

4.2.2 Effects of attitudes on performance

Finally, in exploring the effects of attitudes on the test taker performance on the FTFI and the CAOT, Spearman correlations were carried out to examine whether two variables, 'attitudes' and 'performance' on the FTFI and the CAOT, were associated, and further whether the misfits of the test taker facet were caused from the test takers' attitudes.

According to the findings from the Rasch analysis, there were no particular intervening factors which affected the test takers' performance and scores, i.e., no effects of test order, bias between raters and test formats, computer familiarity, and age and gender differences except the different rater severity despite adequately high inter and intra-rater reliabilities. Thus, the use of measures (or logit scores) which can compensate for the rater severity difference and are interval data, is more appropriate to increase measurement precision than the original raw scores: one of the major advantages of using the multi-faceted Rasch measurement is that it can transform the original ratings which have the properties of ordinal data into interval measures (Sudweeks *et al.*, 2005; Wang *et al.*, 2006).

One may doubt whether the scores created by the Rasch analysis in this study are valid because the scores might have been distorted due to the mismatch between the

ability of the test takers and the difficulty of the tests. When there is a mismatch between test takers' ability level and test difficulty, it often causes the test takers to guess or misbehave in other ways and this would distort raw scores and Rasch measures as indicators of the intended latent variable. As argued earlier, the nature of an oral test situation is, however, unlikely to cause the test takers to guess or respond carelessly and distort the Rasch measures. According to Linacre (personal communication, 2007), in this situation if the Rasch measures are distorted, then the raw scores will be more distorted. Therefore, for the analysis both the measures for the test takers on each item, corrected for differences in rater severity and the measures for the test takers, corrected for differences in item difficulty and rater severity (i.e., total measures) were used. The results are given in Table 4.24 below.

Table 4.24 Correlations between attitudes and measures

Rating item	Test validity		Negative feelings		Test procedures		Total attitudes	
	FTFI	CAOT	FTFI	CAOT	FTFI	CAOT	FTFI	CAOT
Fluency	0.27	0.10	-0.41**	0.13	-0.09	-0.05	0.31*	0.09
95% CI								
Lower	-0.04	-0.21	-0.63	-0.18	-0.39	-0.35	0.01	-0.22
Upper	0.53	0.39	-0.12	0.41	0.22	0.26	0.56	0.38
Grammar	0.27	0.06	-0.38*	0.10	-0.07	-0.01	0.31	0.07
95% CI								
Lower	-0.03	-0.25	-0.61	-0.21	-0.36	-0.31	0.00	-0.24
Upper	0.53	0.36	-0.08	0.39	0.24	0.29	0.56	0.37
Vocabulary	0.28	0.11	-0.38*	0.16	-0.12	-0.08	0.29	0.09
95% CI								
Lower	-0.02	-0.20	-0.62	-0.16	-0.40	-0.38	-0.01	-0.22
Upper	0.54	0.40	-0.09	0.44	0.19	0.23	0.55	0.38
Intelligibility	0.07	0.19	-0.18	0.15	-0.02	0.15	0.12	0.25
95% CI								
Lower	-0.24	-0.12	-0.46	-0.17	-0.32	-0.16	-0.19	-0.06
Upper	0.37	0.47	0.14	0.43	0.37	0.43	0.41	0.51
Cohesion	0.42**	0.07	-0.32*	0.10	-0.18	0.03	0.31	0.10
95% CI								
Lower	0.13	-0.24	-0.57	-0.21	-0.46	-0.28	0.00	-0.21
Upper	0.64	0.37	-0.02	0.40	0.13	0.33	0.56	0.39
Effectiveness	0.31*	0.08	-0.37*	0.11	-0.07	-0.01	0.32*	0.09
95% CI								
Lower	0.01	-0.23	-0.61	-0.20	-0.37	-0.31	0.02	-0.22
Upper	0.56	0.38	-0.08	0.40	0.24	0.29	0.57	0.38
Total measures	0.34*	0.12	-0.29	-0.16	-0.04	-0.03	0.32*	0.12
95% CI								
Lower	0.04	-0.19	-0.55	-0.44	-0.34	-0.33	0.02	-0.19
Upper	0.58	0.41	0.02	0.15	0.27	0.28	0.57	0.41

*p<.05 **p<.01

As illustrated in Table 4.24, no significant correlations were found between the attitudes to the CAOT and any measures of the CAOT items suggesting that attitudes to the CAOT did not influence performance on the test.

On the other hand, several statistically significant correlations between the attitudes to and performance on the FTFI are presented in the table. In particular, 'negative feelings' affected the measures of many FTFI items. The perception of the 'test validity' of the FTFI were significantly correlated with the measures of FTFI *Cohesion*, *Communicative effectiveness*, and total measures, and 'negative feelings' about the FTFI were significantly and negatively correlated with almost all measures of the FTFI items except those of *Intelligibility* and total measures while there were no significant correlations between 'test procedures' and any measures of the FTFI items. In addition, it can be noticed that the measures of *Cohesion* and *Fluency* were more strongly affected by the attitudes: 'test validity' affected the measures of FTFI *Cohesion* ($p=.00$) more than those of FTFI *Communicative effectiveness* and 'negative feelings' influenced the measures of FTFI *Fluency* more than those of the other FTFI items ($p=.00$).

These relationships resulted in a meaningful relationship between total attitudes to the FTFI and the measures of FTFI *Fluency*, *Communicative effectiveness*, and total

measures. P-values of some items showed no significant correlations with total attitudes, but they were very close to the level of significance: *Grammar* ($r = .31$, $p = .051$), *Vocabulary* ($r = .29$, $p = .063$), and *Cohesion* ($r = .31$, $p = .051$), even though all confidence intervals for the correlation coefficients of *Grammar*, *Vocabulary*, and *Cohesion* contain 0, indicating that the relationships may be due to sampling error.

The results indicate that 'test validity' affected performance positively and especially 'negative feelings' strongly influenced the test takers' overall performance harmfully, indicating that the test takers' attitudes to the FTFI were significantly associated with their performance on the test.

In order to explore the relationship between the attitudes and performance further, Spearman correlations were run for each of the items of the attitude factors.

Because there was no relationship between the attitudes to the CAOT and its measures, only the results of the correlations between the attitude items regarding the FTFI and the measures of the FTFI items are presented in Table 4.25.

Table 4.25 Correlations between FTFI attitude items and measures

Rating item	Test validity			Negative feelings		Test procedures		
	Item3	Item1	Item2	Item6	Item8	Item5	Item4	Item7
Fluency	0.04	0.04	0.45**	-0.49**	-0.19	-0.16	0.08	-0.16
95% CI								
Lower	-0.27	-0.27	0.18	-0.69	-0.47	-0.44	-0.23	-0.45
Upper	0.34	0.34	0.67	-0.22	0.12	0.15	0.37	0.15
Grammar	0.04	0.07	0.43**	-0.39*	-0.23	-0.10	0.05	-0.12
95% CI								
Lower	-0.27	-0.24	0.15	-0.62	-0.50	-0.40	-0.26	-0.41
Upper	0.34	0.37	0.65	-0.10	0.08	0.21	0.35	0.19
Vocabulary	0.02	0.09	0.45**	-0.41**	-0.22	-0.11	0.02	-0.21
95% CI								
Lower	-0.29	-0.22	0.17	-0.64	-0.49	-0.40	-0.28	-0.48
Upper	0.32	0.38	0.66	-0.13	0.09	0.20	0.32	0.10
Intelligibility	-0.12	-0.05	0.25	-0.27	-0.02	-0.06	0.14	-0.20
95% CI								
Lower	-0.41	-0.35	-0.05	-0.53	-0.32	-0.36	-0.17	-0.47
Upper	0.19	0.26	0.52	0.04	0.28	0.25	0.43	0.12
Cohesion	0.17	0.12	0.57**	-0.39*	-0.14	-0.17	0.00	-0.24
95% CI								
Lower	-0.15	-0.19	0.33	-0.62	-0.42	-0.45	-0.30	-0.51
Upper	0.45	0.41	0.75	-0.10	0.18	0.15	0.31	0.06
Effectiveness	0.07	0.08	0.47**	-0.42**	-0.20	-0.11	0.03	-0.11
95% CI								
Lower	-0.24	-0.23	0.20	-0.64	-0.47	-0.40	-0.28	-0.40
Upper	0.37	0.37	0.68	-0.13	0.11	0.20	0.33	0.20
Total measures	0.19	0.06	0.45**	-0.37*	-0.11	-0.00	0.09	-0.16
95% CI								
Lower	-0.12	-0.25	0.18	-0.60	-0.40	-0.31	-0.22	-0.44
Upper	0.46	0.36	0.67	-0.07	0.20	0.30	0.38	0.15

*p<.05 **p<.01

Item 2 ('I could do justice to my ability') and item 6 ('I was nervous while I was taking the test') were significantly correlated with the measures of almost all the rating items except only that of *Intelligibility*, indicating that these two attitude items were particularly influential on the test takers' performance on the FTFI. The measures of *Cohesion* and *Fluency* were more strongly correlated with attitudes again as in Table 4.24. These results imply that the test takers performed better when they felt they could do justice to their ability and were less nervous in the FTFI.

It should be remembered that *Intelligibility* was a misfitting item. There was a possibility that the raters interpreted it inconsistently and some of the test takers' performance on that item might have scored inaccurately. Thus, if there had been more consistent ratings of the performance on that item, the results regarding *Intelligibility* might have been different.

In order to investigate if the misfitting test takers' inconsistent performance found in 4.2.1.1 resulted to some extent from their positive or negative attitudes toward the tests, the misfitting test takers on each test (five from the FTFI and seven from the CAOT) were excluded from the data and the correlations with the attitudes were carried out again. No different results were found from the re-analysis with the edited data from the

CAOT, but the results of the correlations using the edited data of the FTFI (Table 4. 26) clearly showed that the significances were distinctively reduced compared to Table 4.25.

Table 4.26 Correlations between attitude items and measures using edited FTFI data

Rating item	Test validity			Negative feelings		Test procedures		
	Item3	Item1	Item2	Item6	Item8	Item5	Item4	Item7
Fluency	0.07	-0.15	0.33	-0.41*	-0.22	0.01	0.19	-0.14
95% CI								
Lower	-0.26	-0.45	0.01	-0.65	-0.51	-0.32	-0.15	-0.44
Upper	0.38	0.18	0.59	-0.10	0.11	0.33	0.48	0.20
Grammar	0.07	-0.06	0.36*	-0.33	-0.25	0.09	0.15	-0.10
95% CI								
Lower	-0.26	-0.37	0.04	-0.59	-0.53	-0.24	-0.18	-0.41
Upper	0.39	0.27	0.67	-0.00	0.08	0.40	0.45	0.28
Vocabulary	0.05	-0.05	0.37*	-0.33	-0.24	0.08	0.13	-0.17
95% CI								
Lower	-0.28	-0.37	0.05	-0.59	-0.52	-0.25	-0.20	-0.46
Upper	0.37	0.28	0.62	-0.00	0.09	0.40	0.44	0.17
Intelligibility	-0.02	-0.17	0.21	-0.18	-0.04	0.05	0.19	-0.20
95% CI								
Lower	-0.34	-0.47	-0.12	-0.48	-0.36	-0.28	-0.14	-0.49
Upper	0.31	0.16	0.50	0.15	0.28	0.37	0.48	0.14
Cohesion	0.09	0.04	0.52**	-0.33	-0.15	-0.06	0.16	-0.31
95% CI								
Lower	-0.24	-0.29	0.23	-0.59	-0.45	-0.38	-0.17	-0.58
Upper	0.40	0.35	0.72	-0.01	0.18	0.27	0.46	0.01
Effectiveness	0.09	-0.09	0.37*	-0.34*	-0.26	0.10	0.17	-0.07
95% CI								
Lower	-0.24	-0.40	0.05	-0.60	-0.54	-0.23	-0.16	-0.39
Upper	0.40	0.24	0.61	-0.02	0.07	0.41	0.47	0.26
Total measures	0.21	-0.04	0.34*	-0.19	-0.17	0.11	0.04	-0.20
95% CI								
Lower	-0.12	-0.36	0.02	-0.48	-0.47	-0.23	-0.29	-0.49
Upper	0.50	0.29	0.60	0.15	0.16	0.41	0.36	0.13

*p<.05 **p<.01

From the analyses, it can be concluded that the test takers' attitudes to the FTFI did affect their performance and made them perform inconsistently on the test and this led to the misfits.

In summary, while there were not any meaningful relationships between attitudes to the CAOT and performance on the CAOT, more positive perceptions of the 'test validity' of the FTFI related to higher performance on the FTFI and particularly, more 'negative feelings' to the FTFI to lower performance on the FTFI. The performance on the FTFI was especially associated with the test takers' perceptions (doing justice to their ability) and with their feelings about nervousness. Eventually the effects of these attitudes appeared to lead to inconsistent performance of the test takers on the test. More specifically, the misfits seemed to be caused to some extent by the test takers' positive or negative attitudes. It was also noted that performance on *Cohesion* and *Fluency* were more strongly influenced by the perception of 'test validity' (or doing justice to their ability) and by 'negative feelings' (or nervousness) respectively.

CHAPTER FIVE

DISCUSSION AND CONCLUSION

5.1 Research Question One: What are Korean university students' attitudes toward a Face-To-Face Interview and a Computer Administered Oral Test?

For this research question, the following sub-research questions were addressed:

1. How do Korean university students perceive and react to the two test formats?
2. To which test format do the Korean university students have more positive attitudes overall?
3. Are there any significant attitude differences between gender and between age groups?
4. Are there any significant computer familiarity differences between gender and between age groups?
5. Is there any significant relationship between computer familiarity and attitudes?

This study revealed that Korean university students had generally favorable attitudes and reactions to the FTFI and the CAOT, but there were some differences between attitudes to the tests. The findings from the analysis of the questionnaire verified that the students had more favorable attitudes to the CAOT than the FTFI in many aspects

(preparation time, nervousness, tiredness, and fairness). The results of the factor analysis also confirmed this. The students had more 'negative feelings' (nervousness and tiredness) to and to some extent less positive attitudes to the 'test procedures' (preparation and response time, and fairness) for the FTFI. There were in addition perceived advantages of the CAOT such as objectivity and practicality through the interviews and the open questions on the questionnaire.

Surprisingly the students, however, preferred the FTFI regardless of the favorable attitudes they showed to the CAOT in many respects. The preference was associated with perceptions of and feelings about test validity, comfort in talking, and disturbance by other students. But the main causes of their preference for the FTFI were, according to the results of the analyses of the interviews and answers to the open questions in the questionnaire, that the students perceived it was a more authentic and valid test format reflecting their real speaking ability and they felt more familiarity with the FTFI format. These findings are opposite to those of Lewkowicz's study (2000) reporting that authenticity is not an important feature for test takers and the test takers tend to be more concerned with the difficulty of the test than with its authenticity; the FTFI was more difficult than the CAOT in this study.

This inconsistent result has also been found in the past studies of face-to-face formats (Savignon, 1972; Brutch, 1979). Savignon (1972) reports that reactions to oral tests of communicative competence investigated in her study were strongly positive, even though the tests were very difficult. Savignon concludes that although most students felt completely unprepared for the testing experience, they regarded the tests as actually testing those foreign language skills they were supposed to be learning.

Brutch (1979) also demonstrated students' concern with the validity and relevance of a testing procedure. This study examined students' attitudes towards two types of written tests: a discrete point proficiency test and a global/communicative test requiring students to write essays and letters. Students preferred the communicative test, reporting that it permitted them to show how they could apply what they had learned, and that it reflected real-life situations.

The students' concern with test validity was often raised in this study as an important factor in their preference. The students preferred the FTFI because they perceived the test was valid. Their preference for the FTFI was revealed more clearly in the semi-structured interview. The students in the interviews stated that they could have more effective communication in the FTFI using nonverbal communication and feedback. The students thought they were allowed to ask to repeat or clarify the questions/tasks

they found understand to understand, ask about some difficult words, and easily change or correct their wrong responses in the FTFI. Further, they felt that a listener encouraged them in talking more and that the warm feedback from a listener helped them feel comfortable and relaxed. It appeared that the students thought that all those elements of the FTFI were necessary in real conversation and helped them do justice to their ability and perform better. It appeared that the students were able to separate their emotions and feelings about the tests from their evaluation of the test validity.

The students liked an interaction with a person. The interaction with a person in the FTFI seems to have made the students feel that the FTFI reflected their actual speaking ability. They liked talking to a person rather than to a computer, believing that the FTFI was a good test which gave opportunities to use the language in a more real life situation. It seems obvious that the FTFI is seen as a more authentic and valid test in the eyes of the students, as Clark (1980: 38) argues that 'the great strength of direct speaking tests of the interview type as measures of global proficiency lies precisely in the highly realistic testing format involved and the direct interview type speaking test enjoys a very high degree of face validity.'

In the communicative testing paradigm, face validity has great importance. If the students do not accept the test as valid, their negative reactions to the test may lead them

to perform in a way which does not rightly reflect their speaking ability (Weir, 1990).

Alderson *et al.* (1995: 173) also insist:

...face validity is important in language testing. For one thing, tests that do not appear to be valid to test users may not be taken seriously for their given purpose. For another, if test takers consider a test to be valid...they are more likely to perform to the best of their ability on that test and to respond appropriately to items.

Studies by Savignon (1972), Brutch (1979) and Shohamy (1982) also show that the perception of test validity is an important factor in students' decisions about how well they like a test. In these studies, students reacted positively to a more communicative test and preferred the test even though it was difficult for them, presumably because of its higher face validity.

One could argue that the difference in students' attitudes towards the two tests might have been influenced by their test scores. But their scores were not reported to the students before they described their attitudes about the tests on the questionnaire and were interviewed, making it impossible for the score feedback to influence the reported attitudes.

All of the findings from the questionnaire and interview study about attitudes suggested that the Korean students showed very favorable attitudes towards the CAOT (even though they preferred the FTFI after all), compared to previous studies showing

much less favorable attitudes to a semi-direct test (i.e., a tape-based test) (e.g., Clark, 1988; Stansfield *et al.*, 1990; Shohamy *et al.*, 1993).

In the interviews the most frequently cited positive comment about the CAOT was that the test made them less nervous. The students, especially those who favored the CAOT, often said that they felt less nervous when they were talking to a computer in the CAOT rather than to a teacher in the FTFI. This might partly be rationalized by the prominent Korean cultural tendency to save face by not making mistakes or by avoiding a situation where mistakes might be made in public (Song, 1994). Some students in the interviews and open questions in the questionnaire stated that they felt more difficulty in the FTFI because they were more embarrassed when they made mistakes and felt a burden when speaking in front of a teacher. Korean students have a strong tendency to be unwilling to speak English to a person or in public when they have little confidence in their English ability because they are afraid of making mistakes (Song, 1994). This cultural factor might have affected the students' attitudes and reactions and helped them feel less nervous and less pressure in the CAOT. The cultural factor might have caused favorable attitudes about the CAOT.

However, the responses to the question, 'I could do justice to my ability' seemed not to be very related to this cultural factor as there was no significant difference between

the two tests on the question. Rather, they were connected with the questions 'the test was a good test format as an exam' ($p < .05$), 'the test had enough questions' ($p < .05$), and other factors such as non-verbal communication and feedback from a teacher found in the interview study, as mentioned earlier.

Aside from the cultural factor, some other reasons for the reduced negative attitudes to the CAOT might be that it has higher face validity than the tape-based tests which were used for comparison with direct tests in the previous studies, and/or the fact that the instructions for the CAOT were written in Korean, helping the students understand the procedure more clearly. Also, that they were introduced to and informed about the CAOT before the test, which probably increased their feelings of familiarity with the new test even though there was no actual practice on the CAOT. O'Loughlin (2001) mentions as possible means for reducing test anxiety the giving of instructions in the native language, or in written form, or by ensuring that all test takers are familiar with the system in advance; James (1988) increased positive attitudes to a semi-direct test through training in that format before taking the test.

In the examination of the effects of gender and age groups on attitudes, male and female groups did not differ in attitudes about the FTFI and the CAOT, but different age groups did. Attitudes about the FTFI appeared relatively more positive in an older (over

21) age group compared to a younger (under 20) age group. The older group significantly had more positive attitudes about the 'test validity' of the FTFI than the younger group did. On the other hand, the younger group had more favorable attitudes to the 'test procedures' for the CAOT than the older group did. One possible explanation might be that the older students probably had more familiarity with the interaction with a person in English (i.e., the FTFI format). In the interview study, especially older (over 21) students who preferred the FTFI, often mentioned about familiarity with the FTFI format. Their greater experience in face-to-face interaction might have led to more favorable attitudes to the FTFI and the significantly more favorable perception of the 'test validity' of the FTFI than the younger group.

With regard to computer familiarity, it was demonstrated by the results of the analyses of the questionnaire and the interview that most of Korean university students were very familiar with a computer, but male and younger (under 20) age groups had a tendency to have more familiarity with a computer than female and older age (over 20) groups; although they were not significant, they were very close to the level of significance .05 and the effect sizes were in the medium range (gender: $Z=-1.95$, $p=.051$, $d=-.48$, age: $Z=-1.92$, $p=.054$, $d=.59$). The relationship between gender and computer familiarity was also found in Taylor *et al.*'s study (1999): a male group in their study

was more familiar with a computer than a female group.

There was a significant correlation between computer familiarity and one aspect of attitudes ('test procedures') toward the CAOT in the quantitative study. However, according to the interview study, it appeared that computer familiarity was not generally connected with the students' attitudes to the CAOT, similarly with the previous studies that found computer familiarity was unassociated with a favorable attitude to a computerized test (Stricker *et al.*, 2004; Joo, 2004).

In the interviews, the responses by the students suggested that the students' attitudes to using the computer were quite different from those to the CAOT. Similarly, some perceived their personal computer use as fun and interesting but the computer use for the test as being strange and uncomfortable.

Their attitudes were more related to their familiarity with the test format. Some students seemed to have favorable attitudes to the FTFI because they were familiar with face-to-face interaction, but less favorable attitudes to the CAOT because it was new and unfamiliar for them. Moreover, some interviewees remarked that they were worried whether they were operating the computer program properly and the recording was working right. There seemed to be a kind of fear of doing old things in new ways: the tendency was for them to like doing things ways they are accustomed to. Specifically,

familiarity seems to contribute to preference for one testing format over another, since students tend to like something they are familiar with and might feel more relaxed in a familiar task (Bygate, 2001). The favorable attitudes may be increased by growing familiarity with the CAOT through the exposure.

Therefore some kind of familiarization might be necessary for the students to gain skills needed to respond to the computerized tasks in the CAOT: in this study, there was a brief introduction of the CAOT before the test, but the introduction was carried out only 'orally' with a description of the CAOT procedure without giving the students an actual trial. Thus, before taking a test, the familiarization with a test format and test item types through actual trials, will certainly be useful and helpful in making sure that the lack of familiarity will not be a confounding variable in test scores. This activity will help the students increase not only their familiarity with the CAOT but also favorable attitudes to the CAOT overall.

5.2 Research Question Two: What are the students' performance on the FTFI and the CAOT?

For Research Question Two, the following sub-research questions were addressed in this study:

1. Are the test takers' speaking abilities discriminated well?
2. Are the raters' ratings consistent and is there a difference in their severity?
3. Are the two tests equally difficult?
4. Is there a test order effect?
5. What is the item difficulty of the two tests?
6. Is there a bias between raters and test formats?
7. Is there a relationship between computer familiarity and test performance?
8. Do gender and age differences impact the performance on the two tests?

The findings from the FACETS analyses indicated that the test takers could be separated fairly well in terms of their speaking ability by the Rasch model although their abilities were quite low. The raw scores ranged mostly from 1 to 3 on a 1-7 rating scale and most of the test takers were below zero logit.

There were several misfits on each test and some possible reasons were provided as to why such misfits were identified. In most performance tests, some misfits seem likely to occur for those reasons. One of the important reasons for the misfits on the FTFI in this current study was that the misfitting test takers' positive or negative attitudes caused inconsistent performance on the FTFI, but it seems that it was not the reason for the misfits on the CAOT in spite of slightly more misfits (more discussion followed about this in 5.5).

From FACETS analyses, it was shown that inter-rater reliability was fairly high and

intra reliability was also adequate, but the raters were different from each other in their severity in spite of rater training prior to the actual assessment. Practically, this result was the same as all the studies that have investigated rater severity on oral performance assessments (e.g., Bachman *et al.*, 1995; Brown, 1995; Lumley and McNamara, 1995; Lynch and McNamara, 1998; Weigle, 1998; Upshur and Turner, 1999). The studies have found significant and meaningful differences among raters.

It seems that the rater training may have made the raters more self-consistent and may even have reduced the severity between the raters, but it could not entirely get rid of the differences in terms of their severity so that raw scores could be used for test performance, as several researchers also assert (e.g., Lunz and Stahl, 1990; Lumley and McNamara, 1995; Weigle 1998). This seems to sufficiently justify the use of FACETS analysis of test scores in this study where two raters were employed in rating. It was able to take the rater difference in severity into account and make adjustments to estimates of test taker ability.

Fortunately, the test takers' performance or scores were not intervened by the factors such as the order of test taking and the bias between raters and test formats. In addition, there was no relationship between computer familiarity and test performance.

There has been an assumption that low computer familiarity can have an impact upon

computerized test scores, and numerous studies have examined the relationship between computer familiarity and performance and many of them usually found that there were to some extent, significant relationships between them (e.g., Kirsch *et al.*, 1998; Min, 1998; Taylor *et al.*, 1999; Wiechmann *et al.*, 2003). However, the test takers' computer familiarity in this study was related to only one aspect of attitudes, 'test procedures' and finally their computer familiarity was not associated with test performance. This may be because the great majority of people in Korea today are very comfortable around computers: especially university students as identified in the questionnaire and interview study, because they can easily access computers at university and home. Access to and experience with computers are likely to increase. Thus, it seems reasonable to expect that some differences shown in recent past studies may decrease over time. When taking computerized language tests in the future, computer familiarity may become less and less of an important issue.

No impact of gender and age differences on performance on the two tests was found.

There was a concern that gender and age differences may produce more advantages from one test format than the other because male and younger groups tended to have more computer familiarity and especially a younger group seem to have a slightly more positive attitude to the CAOT, but the differences appear not to have had a significant

impact on performance. No bias was found between gender and age groups and test formats, and the difficulty of each test was the same irrespective of gender and age differences.

Six of the items as rating categories were found to be poorly targeted at the test taker population. The rating scale was not satisfactory in terms of its difficulty, although the rating scale was carefully chosen for the test takers. The higher level scores were not utilized at all. More appropriate rating scales in difficulty might be more sensitive to register some effects on performance addressed in this study. In addition, *Intelligibility* was identified as misfitting on each test suggesting that scores on *Intelligibility* were inconsistent with the general pattern of performance on the test. This may suggest that *Intelligibility* be excluded from the rating scale in the future to make a more consistent pattern of performance overall. Thus, there seems to be a need to produce a new (or revised) rating scale for lower level speaking performance such as those which are more common amongst Korean university students instead of copying a pre-existing rating scale. In relation to excluding *Intelligibility*, a rating scale containing fewer upper and more lower levels and shorter descriptions of each grade in the scale may be far more suitable for Korean university students and also much easier for raters to score.

In the case of test formats, the ability estimates obtained from the FTFI and the

CAOT did not correlate very strongly ($r=0.71$) compared with the past studies, and the result of the chi-square test suggested the existence of a test format effect. In addition, the result of the analysis of the edited data excluded the misfits of the test taker facet for the combined data more clearly pointed out the difference of the two tests in difficulty.

The FTFI was more difficult than the CAOT.

In the analyses of the items of the two tests, the difference in difficulty was also identified in their items. All items of the FTFI were more difficult than those of the CAOT except the FTFI *Cohesion*, but the items were similarly ordered in each separate analysis of their relative difficulty on both tests. On each test, *Cohesion* and *Intelligibility* were much easier than the other items, and possible reasons were provided in the previous chapter.

Taking account of the findings, it can be concluded that along with its items, the FTFI was overall more difficult than the CAOT even though the tests were designed to match each other as closely as possible. Some possible reasons for the difference in difficulty were stated as follows:

First, the tasks in the FTFI might have been simply more difficult than those of the CAOT because they were only superficially matched. In order to avoid content bias between the FTFI and the CAOT, there was a careful attempt to construct the two oral

tests in such a way that both the elicitation tasks and test takers' responses would match each other as closely as possible. But slightly different sets of tasks and items were used in each test to avoid the practice effects caused by taking the same tasks or items twice, and the level of difficulty of the narration task in each test might have been different: one is a 'restaurant' task for the FTFI, and the other is a 'birthday' task for the CAOT (see Figure 3.1). The 'restaurant' task might have been easier for the test takers, so that they performed it better on the FTFI and finally got higher scores on FTFI *Cohesion*.

Secondly, the test takers showed significantly more 'negative feelings' (especially, nervousness) to the FTFI and their significantly more 'negative feelings' about the FTFI might have partly influenced the test takers' overall performance on the FTFI. The results of this study showed that there was a relationship between 'negative feelings' to the FTFI and performance on the FTFI: the more 'negative feelings' to the FTFI the test takers had the lower scores on the FTFI they tended to have. Likewise, it is presumed that the perception of the 'test validity' of the FTFI might have partly affected the performance in FTFI *Cohesion*, since it tended to be influenced more strongly by the perception of the 'test validity' than other aspects of attitudes.

Lastly, the interaction in the FTFI, which the test takers frequently mentioned in the interviews and open questions in the questionnaire, might have caused different

performance on the test. Shohamy (1994) and O'Loughlin (2001) also support this view, arguing that direct and semi-direct tests appear to tap different speaking abilities (i.e., interactive versus monologic speaking ability).

In spite of a careful attempt to minimize interaction in the FTFI, the FTFI appeared to consist of two-way exchange tasks which involved considerable negotiation of meaning through verbal and non-verbal communication. On the other hand, like other semi-direct oral tests, the CAOT tasks seemed to be monologic consisting of a series of one-way exchanges.

The monologic one-way tasks might have been easier for some of the Korean university students because it seems that the students have few opportunities to practice two-way face-to-face interactive tasks in the classroom and their interactive competence may be naturally low. In the interviews, some of the test takers also stated that they felt difficulty because of the interaction with a teacher in the FTFI.

Higher scores on FTFI *Cohesion* might also somewhat support this view. The test takers might have received the higher score of the FTFI *Cohesion* because there was equally no interaction between a teacher and a test taker in the narration tasks of both tests. Under the same condition, the task might have been easier in the FTFI.

Any one or a combination of these probable reasons outlined above may account for

the difference of the two tests in difficulty.

5.3 Research Question Three: Do attitudes influence performance on the FTFI and the CAOT?

The results of the analyses using the measures (or logit scores) generated by the Rasch model dealing with the rater difference in severity demonstrate that the attitudes to the CAOT were not related to performance on the test, as no single correlation between the attitudes to and the measures of the items on the CAOT were found in the analyses.

Contrary to the results of the CAOT, there were significant connections between the test takers' attitudes to the FTFI and their performance on it. Good performance on the FTFI tended to correspond to the perception of the 'test validity' about the FTFI and in particular, poor performance on the FTFI corresponded to the 'negative feelings' about it. Specifically, doing oneself justice and nervousness were the important factors affecting test performance. Nervousness (or anxiety) has been claimed as a major factor affecting ability to perform on tests in previous studies (e.g., Sarason, 1960; Daffenbacher and Peitz, 1978; Mathews, 1996), but this study demonstrates that doing oneself justice was also an important factor as much as nervousness.

From the re-analysis using edited FTFI data which excluded misfitting test takers, it indicated more clearly that their attitudes to the FTFI influenced performance and

hindered them in performing consistently on the test. In other words, nervousness and doing oneself justice were important reasons for the misfits of the test taker facet for the FTFI.

However, it should be noted that the attitudes were linked to one another, as has been shown in 4.1.1.1, although doing oneself justice and nervousness were revealed as the main effects on performance. Thus, in order to lessen nervousness and improve confidence to help the test takers do themselves justice in a test, overall positive attitudes might be increased and negative ones reduced. The results imply that the test takers' levels of positive and negative attitudes influence the quality of their performance on the FTFI, finally this may threaten not only the reliability but also the validity of the test.

In the FTFI, teachers should try to ease the test takers' anxiety by encouraging and comforting them during the test so that negative attitudes can be reduced. The test takers in this study felt more nervousness during the FTFI and they also thought they performed less well because of this.

However, the teachers should take care to be consistent in a FTFI. They should avoid treating the test takers in different ways because of their different personalities and backgrounds (O'Loughlin, 2001). Otherwise, they may give advantages for some over

others. The impact of teachers/interlocutor on test taker performance has been recognized by some researchers (e.g., McNamara, 2000; O'Loughlin, 2001; Fulcher, 2003). Skillful teachers will be needed for effective and consistent interaction with test takers in the FTFI, and the teachers will be able to develop skills through training which helps them not only to assess test takers more reliably by understanding and becoming familiar with the rating scale, but also to learn how to interact with test takers consistently. In a way this will make test takers feel more confident and lower their anxiety. The training will be far more helpful if it is based on individual feedback using Rasch analysis which is likely to reduce variances among teachers/raters.

In this sense the CAOT may be fairer and it may be easier for teachers to administer because it does not need an interlocutor. Further to this the performance on the CAOT is, according to the findings from this current study, barely related to attitudes towards the test. The lack of relationship between attitudes about the CAOT and performance on the test may suggest that the attitudes to the CAOT are not important factors in improving the test takers' performance. But this does not imply that the teachers don't need to take into account the test takers' attitudes to the CAOT.

As testing and learning are so closely connected to each other, attitudes to the test may also have a strong effect on learning. A number of studies and experiments in

human learning have shown that motivation is a key to learning (see Crookes and Schmidt, 1991; Brown, 1994), and students are often motivated by the tests they are going to take. Students will benefit from positive attitudes since these lead to higher motivation to study for the test. Likewise, negative attitudes may lead to decreased motivation and to failures to reach proficiency before the test (Brown, 1994). Zeidner and Bensoussan (1988: 113) also claim that 'if test takers actually perceive a test as valid and fair, the likelihoods of teacher-student cooperation and optimal test motivation will be enhanced, whereas aversive emotional reactions and debilitating motivational dispositions will be minimized'. The attitudes to the test may wash back on language learning positively or negatively.

In conclusion, the current study suggests that attitudes may represent an important source of construct irrelevant variance in test performance. For example, if test takers' attitudes to the test seriously affect their test scores, the scores may not reflect their real language ability, which the test is intended to measure. So the test would lack (construct) validity. Therefore, it seems obvious that there should be greater consideration of test takers' attitudes to tests in research. By increasing positive attitudes and reducing negative attitudes, test takers will perform to the best of their ability, and teachers, institutions or education authorities will be able to avoid unintended test

effects on test performance, improving the validity of the tests. Further, they will promote beneficial washback effects on learning by increasing test takers' motivation.

5.4 Conclusion and implications

In summary, Korean university students preferred the FTFI even though they had significantly more 'negative feelings' about the FTFI and less favorable attitudes to its 'test procedures'. The significantly more favorable attitudes to the CAOT, compared to the previous studies, which reported relatively more negative attitudes to semi-direct tests (i.e., tape-based tests) (Clark, 1988; Stansfield *et al.*, 1990; Shohamy *et al.*, 1993) might to a certain extent be due to the Korean students' strong cultural tendency to be unwilling to speak English to a person or in public when they have little confidence in their English ability. This might have helped the students feel less nervous and less pressured on the CAOT. Another reason might be that the CAOT *looks* better than the tape-based tests, and lastly the instructions in Korean on the CAOT and prior introduction of the test before an actual test might have led to more positive attitudes to the test.

The findings of the interview study suggested that the students preferred the FTFI mainly because it looked more like an actual performance involving non-verbal

communication, feedback and the reactions necessary for real life communicative situations and looked more valid and a better learning tool for improving speaking ability. Consistently with some previous studies (Savignon, 1972; Brutch, 1979; Shohamy, 1982), the students seem to have been able to separate their feelings and perceptions (e.g., nervousness and tiredness) about the tests from their appraisal of the validity of the test.

From the findings of the FACETS analyses, Korean university students' speaking abilities were generally low, but they could be clearly discriminated in their speaking ability by the Rasch model. The rating scale used in this study was not very satisfactory especially in terms of difficulty and the construction of a new rating scale was recommended for Korean university students who have in general lower level abilities.

The test takers' performance and scores were not affected by other possible intervening factors (the effects of test order, bias between raters and test formats, computer familiarity, and gender and age differences), but severity between raters was different although inter and intra-rater reliabilities were fairly high.

In terms of computer familiarity, the Korean university students had very high computer familiarity and male and younger (under 20) students tended to have higher computer familiarity than female and older (over 21) students. The younger students'

relatively more favorable attitudes to the CAOT compared to older students, might have arisen in part from their higher computer familiarity than the older students.

The students' attitude to the 'test procedures' for the CAOT was associated with their computer familiarity, but through the semi-structured interviews it was found that the students' computer familiarity was not significantly related to their overall positive or negative attitudes to the tests. They experienced using the computers quite differently: the use of personal computer as fun and interesting and the use of computer for the test as strange and uncomfortable. Computer familiarity was unrelated to performance on the CAOT.

The performance on the two tests was also not significantly affected by the students' gender and age, and the difficulty of the tests was the same for them regardless of their gender and age differences.

Although the students preferred the FTFI overall, the students gained higher scores on the CAOT than the FTFI. The examination of item difficulty also showed most of the items of the FTFI were more difficult than those of the CAOT and the chi-square test of the tests suggested a test format effect. There were three possible reasons for the different performances between the tests. Firstly, the tasks in the FTFI might have been more difficult than those of the CAOT because they were only superficially matched;

secondly, the students showed significantly more 'negative feelings' (especially, nervousness) to the FTFI and these significantly more 'negative feelings' about the FTFI might have partly influenced the students' overall performance on the FTFI; lastly, the interaction in the FTFI might have caused different performance on the tests.

Finally, this study found that the students' attitudes toward the FTFI were associated with their performance on the test while there was no relationship between their attitudes toward and the performance on the CAOT. They performed better on the FTFI when they had more positive and less negative attitudes toward the FTFI, and their attitudes were some of the significant causes for the misfits of the test taker facet for the FTFI. This study therefore suggests that the Korean university students' attitudes to the FTFI are an important source of construct irrelevant variance in their speaking test performance on the FTFI.

Taking all the findings of this study into account, it seems that the FTFI is more valid and authentic for assessing the students' broader range of abilities, including interaction ability although it is not able to entirely reproduce real-life conversational settings (i.e., the lack of authenticity) (Clark, 1979; Underhill, 1987; Van Lier, 1989). The FTFI also appears to approximate 'real-life' communication more closely than the CAOT in the eyes of Korean university students. This FTFI seems to be the most proper for a

proficiency test which aims to test students' general global communicative oral ability.

This type of test will attempt to measure the ability to fulfill a variety of language tasks appropriately within a realistic time period under a (simulated) real-life communication situation – typically participating in a face-to-face interactive conversation with a person.

When the students perceive the test as valid, their cooperation and test motivation will be enhanced and lead to positive washback effects on their learning. In this study, the students showed obvious preference for the FTFI over the CAOT, perceiving it as a better test format which actually reflects their real speaking ability, i.e., it has higher face validity than the CAOT. Importantly, Korean university students' positive attitudes to the FTFI were related to their performance on the test.

However, in the situation where it is needed to measure students' overall oral ability, but hard to conduct such a test due to its impracticality (e.g., the lack of skillful teachers and a large number of students), as an alternative method, the CAOT appears a good test format for the teachers to use. Furthermore, the CAOT may be more useful as an achievement test, such as a mid/final exam, than the FTFI, which is likely to measure what the students have learned in their classrooms -including rather discrete aspects of speaking performance such as vocabulary items and syntactic patterns- allowing

untrained teachers to arrive at a decision on the score more easily and quickly (Clark, 1979; Clark, 1988).

The CAOT may be more practical and easier for the teachers to administer and fairer for the students, since Korean university students felt less nervous and tired in the CAOT. The CAOT allows a large number of students to take the same test at the same time making it possible to take more tasks/items in the test without much effort; the more various tasks may give it greater content validity. Also, it does not need an interlocutor, which may also influence the quality of the test performance: for the large part, the validity of the speech sample elicited depends on the skill of the interlocutor/interviewer (Stansfield, 1991). What is better, the teachers in Korean universities preferred the CAOT showing obviously more positive attitudes and reactions to the CAOT (Joo, 2004), and the CAOT appears to have a more positive impact on the students affect compared to the previous studies on tape-based tests.

In the CAOT, the teachers do not need to be concerned about Korean students' level of computer familiarity because their familiarity was very high and the computer familiarity was hardly related to their test performance on the CAOT. Further, the two raters' inter-reliability and exact rating agreement were higher on the CAOT. It may mean that rating performance on the CAOT is easier than on the FTFI for the raters.

The CAOT does not have sufficient empirical evidence to make a valid overall decision about a test taker's oral ability, particularly due to the absence of two-way interaction involving non-verbal communications, but it appears to have the potential to elicit students' sufficient ratable oral performance and make valid inferences about their oral abilities (or communicative language abilities), at least more than a very reliable pencil and paper test consisting of a great number of grammar and vocabulary items. Moreover, the tasks of the CAOT such as talking and listening to a person on a computer or computer screen may be justified in its own right (i.e., simulating real world tasks in the age of the computer); it can simulate test situations more authentically than the tape-based tests.

With all the benefits of the CAOT, some difficulties of implementing oral performance tests in the Korean university context could be reduced (the problems of extrapolation, assessment, and practicality, see 2.3.2.1).

To conclude, if the test is to be used for the assessment of overall oral ability, and a skilled interviewer is available, it may be more appropriate to use the FTFI. On the other hand, if scores are to be used for the assessment of achievement during or at the end of the course, or the test is needed to assess a large group of students in a short time frame, it may be better to administer the CAOT.

The judgment of a good or poor test depends not only on its validity, reliability, practicality, and wash-back effects discussed in 2.2, but also on the particular test purpose and situation. It can be good in one place, but bad in another; that is, no test can be absolutely good or bad. Thus, it may be suggested that neither the FTFI nor the CAOT is better than the other, but each has advantages and disadvantages as found in this study. In the selection of the test, with all the aspects examined in this study (i.e., attitudes, performance, and the various factors which may affect test performance and attitudes) it will be useful to consider a particular test purpose in a given testing situation.

The results of this study are seen as valuable and important for several reasons. First, the observed Korean university students' attitudes to the FTFI and the CAOT are worthy of being considered by EFL teachers and test developers in Korea. The information obtained from the Korean university student feedback will help teachers and test developers to identify and consequently put right certain misunderstandings in designing the FTFI and the CAOT and in their test procedures and situations. The test which the students perceive as valid and fair, and feel comfortable and confident with will beneficially affect their test performance.

Secondly, this study showed the usefulness of multi-faceted Rasch analyses of

FACETS program. As stated earlier, speaking tests are complex because of the number of possible test factors that may occur during test administration, but multi-faceted Rasch analyses made it possible to map all facets (or factors) of the tests, effectively sort out the impact of each facet on scores, take the impact into account, and make adjustments to estimates of test taker ability. In addition, some useful statistical indices (e.g., separation reliability, fit statistics, and chi-square) allowed for the identification and understanding of the characteristics of both tests more clearly.

For the examination of the effects of attitudes on performance, the ability estimates (i.e., measures) compensating for the impact of each facet (in this study, the rater severity) were assumed to be more stable than raw scores, since the use of raw scores impacted by the facet could result in invalid findings and lead to wrong conclusions if they are not managed. Thus, using the measures with the statistics indices, more precise and sound findings could be obtained in the current study. Accordingly, this study may provide a useful methodology for researchers studying performance tests.

Finally, the findings of this study support not only teachers' decision of the types of tests they are going to use for their students, but also the findings of previous studies. They offer further insight into the issues which influence the students' attitudes to and their performance on the FTFI and the CAOT.

As the findings of this study indicated, a speaking test may tell us about something different other than the students' underlying ability. It is apparent that there are numerous possible factors which influence the students' attitudes as well as affecting their test performance. The results of this study demonstrate that although it is indispensable to consider students' attitudes and perceptions before delivering a new test, because it may have negative affective impacts on the students' performance, other issues which influence students' attitudes and test performance seem to be of equal importance when constructing and developing language tests and even when carrying out research involving performance tests.

Teachers and test developers should at least know the factors which may influence their students' test performance and ultimately the scores they obtain. Unfortunately, it may not be possible to completely eliminate the influence of these factors. However, teachers and test developers should make every effort to avoid or minimize these effects.

It is hoped that this investigation will help establish a more complete understanding of Korean university students' attitudes and the effects of possible intervening factors including attitudes on performance and scores on the FTFI and the CAOT, and thus encourage the development of testing and teaching designed to fit the needs of the Korean students in EFL classrooms.

5.5 Further study

The issue of semi-direct versus direct tests of speaking has continued to be a topic of interest and there have been rich debates about the interchangeability of the two types of tests. Although many issues were addressed in this study, teachers and test users still may to some extent have concerns about testing students in the CAOT because of the great number of unexamined areas in this study.

It is becoming more acceptable to teach language and assess its ability using computers in Korea. Hence, far more wide-ranging and in-depth studies with a focus on Korean university/school context and Korean students' needs for English use need to be conducted accompanied by solid theoretical justifications. These justification require support from broadly based empirical data. Thus, I propose several questions for further research.

Since the current study was mainly devoted to addressing the students' attitudes and performance, teachers as an interlocutor and rater and other test users' such as an administrator attitudes and reactions to the FTFI and the CAOT in Korean university context were not investigated in this study.

In the previous study (Joo, 2004), I attempted to explore teachers' feedback about the FTFI and the CAOT, but this study did not closely investigate the perceived advantages

and disadvantages of carrying out the tests and the difficulties in rating performance on the two tests through the actual use of the tests for a certain period of time in the Korean university context.

Thus, further study is required to examine not only teachers but also other test users' attitudes towards the two tests, which may influence the process of rating and administering. This exploration should employ both quantitative and qualitative research methods: the qualitative data gathered in this study was particularly useful in exploring the students' feelings and perceptions to the tests, and other useful qualitative data could be gathered to find out perceived dis/advantages of the FTFI and the CAOT and difficulties in rating on and administering the two tests. This examination would also lead to a greater understanding of the tests and improve testing formats.

Further, follow up studies can be carried out to examine which test condition yields more reliable and valid scores and the effects of attitudes of a teacher as a rater and interlocutor on rating performance and on how students are interviewed.

Any test which is implemented within an educational setting should aim to have a positive washback effect on teaching and learning, and this is an essential factor in all test design and use. In this study, I observed that both tests evoked in general, a positive atmosphere among students, but it is not known what washback effects the tests have on

teaching and learning or how the tests bring about those effects. It would be valuable and interesting to systematically observe the washback effects of the FTFI and the CAOT.

Lastly, correlations used to compare test takers' scores on both tests with their performance, provided certain evidence of concurrent validity of the tests although they were not very high compared to the previous studies. But the result of the chi-square test and the students' comments about the two tests suggested that they might not have measured the same language ability. Moreover, while the misfits of the test taker facet for the FTFI have something to do with attitudes, the misfits for the CAOT seem unconnected with the attitudes to the test, in spite of slightly more misfits on the CAOT. It may mean that the test process worked less effectively on the CAOT than on the FTFI. Thus additional investigations of the misfits on the CAOT is necessary, since some other factors identified in the interviews such as too low tension and nervousness and the disturbance of concentration on performance due to new and curious features (e.g., using a mouse and a video) of the CAOT, might contribute to the inconsistent performance on the CAOT.

Therefore, based on the findings of the current study, there is a need for more studies of the validity of the tests from multiple perspectives. Some qualitative studies (e.g.,

Shohamy, 1994; O'Loughline, 2001) indicated direct and semi-direct tests differed in many respects. Together with the findings from this current study, the findings from the further validation studies using both quantitative and qualitative analyses would allow for obtaining better insight and understanding of what the tests measure and finally help teachers to select the most appropriate test in given context and for given purpose.

BIBLIOGRAPHIES

- AcaStat Software (2007) Applied Statistics Handbook. On-line resource. Available at <http://www.acastat.com/Statbook/chisqassoc.htm> Last viewed Nov 2007
- Agar, M. H. (1980). *The professional stranger: an informal introduction to ethnography*. New York: Academic Press.
- Ahmad, K., Corbett, G., Rogers, M. and Sussex, R. (1985). *Computers, language learning and language teaching*. Cambridge: Cambridge University Press.
- Alderson, J. C. (1990). Learner-centered testing through computers: institutional issues in individual assessment. In Jong, J. H. A. L. D. and Stevenson, D. K (eds). *Individualizing the assessment of language abilities*. Clevedon: Multilingual Matters.
- Alderson, J. C. (1991). Language testing in the 1990s: how far have we come? How much farther have we to go? In Anivan, S. (ed.), *Current developments in language testing: 199-209*. Singapore: Regional Language Center.
- Alderson, J. C. and Clapham, C. and Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.
- Alderson, J. C., Hughes, A. and British Council. (1981). *Issues in language testing*. London: the British Council.
- Andrich, D. (1978). A rating formulations for ordered response categories. *Psychometrika*, 43: 561-57.
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.
- Bachman, L F. and Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

- Bachman, L. F., Lynch, B. K., and Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing*, 12: 238-57.
- Baker, D. (1989). *Language testing: a critical survey and practical guide*. London: Edward Arnold.
- Bennett, R. E. (1999). Using new technology to improve assessment. *Educational measurement: Issues and practice*, 18: 5-12.
- Biddle, B. J. and Anderson, D. S. (1986). Theory, methods, knowledge, and research on teaching. In Wittrock, M. C. (ed.). *Handbook of research on teaching*: 230-252. New York: Macmillan.
- Bogdan, R. C. and Biklen, S. K. (1982). *Qualitative research in education*. Boston: Allyn and Bacon.
- Bonk, W. J. and Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, 20: 89-100.
- Breen, M., Barratt-Pugh, C., Derewianka, B., House, H., Hudson, C., Lumley, T. and Rohl, M. (1997). *Profiling ESL children. Volume 1: Key issues and findings*. Canberra: Department of Employment, Education, Training and Youth Affairs.
- Brown, A. (1993). Test development process: test-takers' reactions to a tape-mediated test of proficiency in spoken Japanese. *Language Testing*, 10: 277-303.
- Brown, A. (1995). The effect of rater variables in the development of an occupation specific language performance test. *Language Testing*, 12: 1-15.
- Brown, H. D. (2004). *Language assessment: principles and classroom practices*. USA: Longman.
- Brown, J. D. (1994). *Principles of language learning and teaching*. New Jersey: Prentice Hall.
- Brown, J. D. (1996). *Testing in language programs*. New Jersey: Prentice Hall.

- Brown, J. D. (1997). Computers in language testing: present research and some future directions. *Language Learning and Technology*, 1: 44-59.
- Brumfit, C. J. and Johnson, K. (1979). *The communicative approach to language teaching*. Oxford: Oxford University Press.
- Brutch, S. (1979). Convergent/discriminant validation of prospective teacher proficiency in oral and written production of French by means of MLA foreign language proficiency tests for teachers (TOP and TWP) and self ratings. University of Minnesota. Unpublished thesis.
- Burke, M. J., Normand, J., and Raju, N. S. (1987). Examinee attitudes toward computer-administered ability tests. *Computers in Human Behavior*, 3: 95-107.
- Bygate, M. (2001). Effects of task repetition on the structure and control of oral language. In Bygate, M., Skehan, P., and Swain, M. (eds). *Researching pedagogic tasks*. Essex: Pearson Education Limited.
- Canale, M. (1983). On some dimensions of language proficiency. In Oller, J. W. (ed.) *Issues in language testing research*. Rowley, MA: Newbury House Publishers.
- Canale, M. and Swain, M. (1980). Theoretical basis of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1: 1-47.
- Center for Applied Linguistics (CAL) (2002). Computerized oral proficiency interview. On-line resource. Available at <http://www.cal.org/project/copi.html> Last viewed December 2005.
- Center for Applied Linguistics (CAL) (2007). Testing/assessment, adult ESL assessment, BEST plus. On-line resource. Available at <http://www.cal.org/BEST/compbest.html> Last viewed December 2005.
- Computer Assisted Language Instruction Consortium (CALICO) (1996-2006). Oral testing software. On-line resource. Available at <http://www.calico.org> Last viewed June 2006.

- Chalhoub-Deville, M. and Turner, C. E. (2000). What to look for in ESL admission tests: Cambridge certificate exams IELTS and TOEFL. *System*, 28: 523-39.
- Chapelle, C. A., Jamieson, J., and Hegelheimer, V. (2003). Validation of a web-based ESL test. *Language Testing*, 20: 409-439.
- Choi, I. C., Kim, K. S. and Boo, J. (2003). Comparability of a paper-based language test and a computer-based language test. *Language Testing*, 20: 295-320.
- Clark, J. L. D. (1979). Direct versus semi-direct tests of speaking proficiency. In Briere, E. J. and Hinofotis, F. B. (eds). *New concepts in language testing: some recent studies*: 35-49. Washington, DC: TESOL.
- Clark, J. L. D. (1980). Toward a common measure of speaking proficiency. In Frith. J. R. (ed.). *Measuring spoken language proficiency*: 15-26. Washington, DC: Georgetown University Press.
- Clark, J. L. D. (1985). Development a tape-mediated, ACTFL/ILR scale-based test of Chinese speaking proficiency. In Stansfield, C. W. (ed.) *Technology and Language Testing*. Princeton, NJ: Educational Testing Service.
- Clark, J. L. D. (1988). Validation of a tape-mediated ACTFL/ILR scale-based test of Chinese speaking proficiency. *Language Testing*, 5: 187-98.
- Clark, J. L. D and Swinton, S. (1979). *An exploration of speaking proficiency measures in the TOEFL context*. TOEFL Research Report 4, Princeton, NJ: Educational Testing Service.
- Clark, J. L. D. and Li, Y. C. (1986). *Development, validation, and dissemination of a proficiency-based test of speaking ability in Chinese and an associated assessment model for other less commonly taught languages*. Washington, DC: Center for Applied Linguistics.
- Clarke, S. and Gipps, C. (2000). The role of teachers in teacher assessment in England 1996-1998. *Evaluation and Research in Education* 14: 38-52.

- Crookes, G. (1989). Planning and interlanguage variation. *Studies in Second Language Acquisition*, 11: 367-383.
- Crookes, G. and Schmidt, R. W. (1991). Motivation: reopening the research agenda. *Language Learning*, 41: 469-512.
- Deffenbacher, J. L. S. and Peitz, S. R. (1978). Effects of test anxiety on performance, worry and emotionality in naturally occurring exams. *Psychology in the Schools*, 15: 446-49.
- Douglas, D. (1994). Quantity and quality in speaking test performance. *Language Testing*, 11: 125-44.
- Douglas, D. (2000). *Assessing Languages for Specific Purposes*. Cambridge: Cambridge University Press.
- Dunkel, P. A. (1999). Considerations in developing or using second/foreign language proficiency computer-adaptive tests. *Language Learning and Technology*, 2: 77-93.
- Educational Testing Service (ETS) (2005). On-line resource. Available at <http://www.ets.org> Last viewed December 2006.
- Ellis, R. (1987). Interlanguage variability in narrative discourse: Style shifting in the use of the past tense. *Studies in Second Language Acquisition*, 9: 12-20.
- English Speaking Proficiency Testing Academy (2002-2005). On-line resource. Available at <http://www.espt.org> Last viewed March 2007.
- Faerch, C. and Kasper, G. (1983). Plans and strategies in foreign language communication. In Faerch, C. and Kasper, G. (ed.). *Strategies in interlanguage communication*. London: Longman.
- Fulcher, G. (1998). Computer-based language testing: the call of the internet keynote address. In Coombe, C. A. (ed.). *Current trends in English language testing*.

Conference proceedings for CTELT 1997 and 1998, 1: 1-14. Arabia: TESOL.

Fulcher, G. (1999). Computerizing an English language placement test. *ELT Journal*, 53: 289-99.

Fulcher, G. (2000). Computers in language testing. In Brett, P. and Motteram, G. (eds). *A special interest in computers: 93-107*. Manchester: IATEFL Publications.

Fulcher, G. (2003). *Testing second language speaking*. Malaysia: Pearson Longman.

Gall, M. D., Borg, W. R. and Gall, J. P. (1996). *Educational research: an introduction*. New York: Longman.

Gay, L. R. and Airasian, P. (2000). *Educational research, competencies for analysis and application*. New Jersey: Prentice-Hall.

Green, B. F. (1988). Construct validity of computer-based tests. In Wainer, H. and Braun, H. I. (eds). *Test validity: 77-86*. Hillsdale, NJ: Lawrence Erlbaum.

Grotjahn, R. (1986). Test validation and cognitive psychology: some methodological considerations. *Language Testing*, 3: 159-85.

Hamp-Lyons, L. and Lynch, B. K. (1995). Perspectives on validity: an historical analysis of the LTRC. Paper presented at the 17th annual Language Testing Research Colloquium. Longbeach, California.

Heaton, J. B. (1975). *Writing English Language tests*. London: Longman.

Henning, G. (1991). Validating an item bank in a computer-assisted or computer adaptive test. In Dunkel, P. (ed.). *Computer-assisted language learning and testing: research issues and practice: 209-22*. New York: Newbury House.

Hicks, M. (1989). *The TOEFL computerized placement test: adaptive conventional measurement*. TOEFL Research Report, 31. Princeton, NJ: Educational Testing Service.

- Hofstede, G. (1986). Cultural differences in teaching and learning. *International Journal of Intercultural Relations*, 10: 301-20.
- Holstein, J. and Gubrium, J. (1995). The active interview. Thousand Oaks, CA: Sage.
In Hughes, A. (1989). *Testing for language teachers*. Cambridge: Cambridge University Press.
- Hughes, A. (1989). *Testing for language teachers*. Cambridge: Cambridge University Press.
- Hymes, D. H. (1972). On communicative competence. In Richards, J. C. and Rodgers, T. S. (1986). *Approach and methods in language teaching*. Cambridge: Cambridge University Press.
- Ingram, E. (1977). Basic concepts in testing. In Allen, J. P. B. and Davies, A. (eds). *Testing and experimental methods*. Oxford: Oxford University Press. In Alderson *et al.* (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.
- James, G. (1988). Development of an oral proficiency component in a test of English for academic purposes. In Hughes, A. (ed.). *Testing English for university study*. ELT documents 127. Oxford: Modern English Publications and the British Council.
- Johnson, K. (1981). Introduction: some background, some key terms and some definitions. In Johnson, K and Morrow, K (ed.). *Communication in the classroom*. Essex, England: Longman.
- Joo, M. J. (1998). Can the Korean English university entrance examination be made more communicative? University of Birmingham. Unpublished MA thesis.
- Joo, M. J. (2003). Teachers' and students' perceptions of testing methods in English conversation classes (pilot study). *Language and Literature Research*, 13: 85-101.
- Joo, M. J. (2004). Teachers' and students' perceptions of and attitudes toward a face-to-face interview and a computerized oral test in a Korean university context. Institute-Focused Study (IFS). Institute of Education, University of London.

Unpublished paper.

Kelly, R. (1978). On the construct validation of comprehension tests: an exercise in applied linguistics. University of Queensland. Doctoral thesis.

Kenyon, D. M. and Malabonga, V. (2001). Comparing examinee attitudes toward computer-assisted and other oral proficiency assessments. *Language Learning and Technology*, 5: 60-83.

Kenyon, D. M. and Tschirner, E. (2000). The rating of direct and semi-direct oral proficiency interviews: comparing performance at lower proficiency levels. *The Modern Language Journal*, 84: 79-99.

Kerlinger, F. N. (1973). Foundations of Behavioral Research. New York: Holt, Rinehart and Winston. In Alderson *et al.* (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Kirsch, I., Jamieson, J., Taylor, C., and Eignor, D. (1998). *Computer familiarity among TOEFL examinees*. TOEFL Research Report No. 59. Princeton, NJ: Educational Testing Service.

Landis, R. S., Davison, H. K. and Maraist, C. C., (1998). The influence of test instructions on perceptions of test fairness: A comparison of paper-and-pencil, computer-administered, and computer adaptive formats. Poster presented at the 10th Annual Convention of the American Psychological Society. Washington, DC.

Larson, J. W. and Smith, K. L. (2001). User's manual for the oral testing software-enhanced, windows edition 1.0. Brigham Young University.

Larson, J. W. (1998). An argument for computer adaptive language testing. *Multimedia-Assisted Language Learning*, 1: 9-24.

Lazaraton, A. (2002). *A qualitative approach to the validation of oral language tests (Studies in Language Testing 14)*. Cambridge: Cambridge University Press.

Lee, H. K. (2004). Constructing a field-specific integrated writing test for an ESL

placement procedure. University of Illinois at Urbana-Champaign. Unpublished doctoral thesis.

Leung, C. and Lewkowicz, J. (2006). Expanding horizons and unresolved conundrums: language testing and assessment. *TESOL Quarterly*, 40: 211-234.

Lewkowicz, J. (1997). Investigating authenticity in language testing. Lancaster University. Unpublished doctoral thesis. In Leung, C. and Lewkowicz, J. (2006). Expanding horizons and unresolved conundrums: language testing and assessment. *TESOL Quarterly*, 40: 211-234.

Lewkowicz, J. (2000). Authenticity in language testing: some outstanding questions. *Language Testing*, 17: 43-64.

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago: MESA Press.

Linacre, J. M. (1991-2003). *User's guide to FACETS: Rasch-model computer programs*. Chicago: MESA Press.

Lincoln, Y. S. and Guba, E. (1985). *Naturalistic inquiry*. Newbury Park, CA: Sage Publications.

Luecht, R. M., Champlain, A. D., and Nungester, R. J. (1998). Maintaining content validity in computerized adaptive testing. *Advances in Health Sciences Education*, 3: 29-41.

Lumley, T. and McNamara, T. F. (1995). Rater characteristics and rater bias: implications for training. *Language Testing*, 12: 54-71.

Lunz, M. E. and Bergstrom, B. A. (1994). An empirical study of computerized adaptive test administration conditions. *Journal of Educational Measurement*, 31: 251-63.

Lunz, M. E. and Stahl, J. (1990). Judge consistency and severity across grading periods. *Evaluation and the Health Professions*, 13: 425-44.

Lynch, B. K. and McNamara, T. F. (1998). Using G-theory and many-facet Rasch

measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing*, 15: 158-80.

Madsen, H. S. and Murray, N. (1984). Retrospective evaluation of testing in ESL content and skills courses. Brigham Young University. Unpublished manuscript.

Malabonga, V. A. (2000). Trends in foreign language assessment: the computerized oral proficiency instrument. On-line resource. Available at <http://www.cal.org/ncirc>
Last viewed Feb 2004.

Malone, M. (2000). Simulated oral proficiency interviews: recent developments. ERIC digest. Washington, DC: ERIC Clearinghouse on Language and Linguistics (ERIC Document No. ED 447 729).

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47: 149-74.

McNamara, T. F. (1996). *Measuring Second Language Performance*. London: Longman.

McNamara, T. F. (1991). Test dimensionality: IRT analysis of an ESP listening test. *Language Testing*, 8: 139-59.

McNamara, T. F. (2000). *Language testing*. Oxford: Oxford University Press.

Messick, S. (1989). Validity. In Linn, R. L. (ed.). *Educational Measurement*: 13-103. New York: Macmillan.

Messick, S. (1998). Test validity: a matter of consequence. *Social Indicators research*, 45: 35-44.

Min, B. C. (1998). A study of the attitudes of Korean adults toward technology-assisted Instruction in English-language programs. *Multimedia-assisted Language Learning*, 1: 63-78.

Morrow, K. E. (1977). *Techniques of evaluation for a notional syllabus*. London: Royal Society of Arts.

- Morrow, K. E. (1979). Communicative language testing: revolution on evolution. In Brumfit, C. J. and Johnson, K. (ed.). *The communicative approach to language teaching*: 143-58. Oxford: Oxford University Press.
- Norris, J. M. (2001). Concerns with computerized adaptive oral proficiency assessment. *Language Learning and Technology*, 5: 99-105.
- O'Loughlin, K. (2001). *The equivalence of direct and semi-direct speaking tests (studies in language testing 13)*. Cambridge: Cambridge University Press.
- Phillips, S. U. (1983). An ethnographic approach to bilingual language proficiency assessment. In Rivera, C. (ed.). *An Ethnographic/Sociolinguistic approach to language proficiency assessment*. Clevedon: Multilingual Matters Ltd.
- Pope-Davis, D. B. and Twing, J. S. (1991). The effects of age, gender, and experience on measures of attitude regarding computers. *Computers in Human Behavior*, 7: 333-39.
- Powers, D. E., and O'Neill, K. (1993). Inexperienced and anxious computer users: coping with a computer administered test of academic skills. *Educational Assessment*, 1: 153-173.
- Rasch, G. (1969/1980). *Probability models for some intelligence and achievement tests*. Chicago: University of Chicago Press.
- Richards, J. C. and Rodgers, T. S. (1986): *Approaches and methods in language teaching*. Cambridge University Press.
- Roever, C. (2001). Web-based language testing. *Language Learning and Technology*, 5: 84-94.
- Sarason, S. B., Davidson, K. D. and Lighthall, F. F. (1960). *Anxiety in elementary school children*. New York: Wiley.
- Savingnon, S. J. (1972). *Communicative competence: an experiment in foreign language teaching*. Philadelphia, Pennsylvania: The Center for Curriculum

Development Inc.

- Schmidt, F. L., Urry, V. W., and Gugel, J. F. (1978). Computer assisted tailored testing: examinee reactions and evaluations. *Educational and Psychological Measurement*, 38: 265-273.
- Schmitt, N., Gilliland, S. W., Landis, R. S., and Devine, D. (1993). Computer-based testing applied to selection of secretarial applicants. *Personnel Psychology*, 46: 149-65.
- Scott, M. L. (1980). The effect of multiple retesting on affect and test performance. Brigham Young University. Unpublished MA thesis.
- Scott, M. L. and Madsen, H. S. (1983). The influence of retesting on test affect. In Oller, J. W. (ed.). *Issues in language testing research*. Rowley, MA: Newbury House Publishers.
- Seidman, I. E. (1991). *Interviewing as qualitative research: a guide for researchers in education and the social sciences*. New York: Teachers College Press.
- Shermins, M. D. and Lombard, D. (1998). Effects of computer-based test administrations on test anxiety and performance. *Computers in Human Behavior*, 14: 111-123.
- Shin, D. (2001). Revising an ACTFL-SOPI formatted speaking test in Korea: problems and suggestions. *English Teaching*, 56: 309-331.
- Shohamy, E. (1982). Affective considerations in language testing. *Modern Language Journal*, 66: 13-17.
- Shohamy, E. (1994). The validity of direct versus semi-direct oral tests. *Language Testing*, 7: 99-123.
- Shohamy, E., Gordon, C., Kenyon, D. M. and Stansfield, C. W. (1989). The development and validation of a semi-direct test for assessing oral proficiency in Hebrew. *Bulletin of Hebrew Higher Education*, 4: 4-9.

- Shohamy, E., Donitsa-Schmidt, S. and Waizer, R. (1993). The effect of the elicitation mode on the language samples obtained in oral tests. Paper presented at the 15th Language Testing Research Colloquium, Cambridge, England.
- Skehan, P. (1984). Issues in the testing of English for specific purposes. *Language Testing*, 1: 202-20.
- Smith, R., Schumacker, R. and Bush, J. (1995). Using item mean squares to evaluate fit to the Rasch model. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.
- Song, M. J. (1994). A study on common factors affecting Asian students' English oral interaction. *English Teaching*, 49.
- Sookmyung Women's University (no date). On-line resource. Available at <http://www.sookmyung.ac.kr> Last viewed March 2007.
- Stansfield, C. W. (1991). A comparative analysis of simulated and direct oral proficiency interviews. In Anivan, S. (ed.). *Current developments in language testing*: 199-209. Singapore: Regional Language Center.
- Stansfield, C. W. and Kenyon, D. M. (1992). Research on the comparability of the oral proficiency interview and the simulated oral proficiency interview. *System*, 20: 347-64.
- Stansfield, C. W., Kenyon, D. M., Paive, R., Doyle, F., Ulsh, I. and Cowles, M. (1990). The development and validation of the Portuguese speaking test. *Hispania*, 73: 641-51.
- Stansfield, C. W. and Kenyon, D. M. (1996). Simulated oral proficiency interviews: an update. ERIC Digest. Washington, DC: ERIC Clearinghouse on Language and Linguistics. (ERIC Document No. ED 395 501).
- Stricker, L. J., Wilder, G. Z., and Rock, D. A. (2004). Attitudes about the computer-based test of English as a foreign language. *Computers in Human Behavior*, 20:

37-54. Available at <http://www.elsevier.com/locate/comphumbh>.

Sudweeks, R. R., Reeve, S., Bradshaw, W. S. (2005). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing*, 9: 239-61.

Tarone, E. (1981). Some thoughts on the notion of communication strategy. *TESOL Quarterly*, 15: 285-95.

Taylor, C. Kirsch, I. and Eignor, D. (1999). Examining the relationship between computer familiarity and performance on computer-based language tasks. *Language Learning*, 49: 219-74.

Torkzadeh, G. and Angulo, I. E. (1992). The concept and correlates of computer anxiety. *Behaviour and Information Technology*, 11: 99-108.

Underhill, N. (1983). ESL writing assessment: subject-matter knowledge and its impact on performance. *English for Specific Purposes*, 9: 123-43.

Underhill, N. (1987). *Testing spoken language: a handbook of oral testing techniques*. Cambridge: Cambridge University Press.

University of Cambridge ESOL Examinations, British Council and IDP Education Australia: IELTS Australia (2006). The International English Language Testing System (IELTS). On-line resource. Available at http://www.ielts.org/article_1.aspx Last viewed October 2006.

Upshur, J. A. and Turner, C. E. (1999). Systematic effects in the rating of second-language speaking ability: test method and learner discourse. *Language Testing*, 16: 82-111.

Van Lier, L. (1989). Reeling, writhing, drawing, stretching, and fainting in coils: oral proficiency interviews as conversation. *TESOL Quarterly*, 23: 489-508.

Vaus, D. D. (2002). *Surveys in social research*. London: Routledge.

- Wang, W. C., Yao, G., Tsai, Y. J., Wang, J. D. and Hsieh, C. L. (2006). Validating, improving reliability, and estimating correlation of the four subscales in the WHOQOL-BREF using multidimensional Rasch analysis. *Quality of Life Research*, 15: 607-20.
- Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, 15: 253-87.
- Weir, C. J. (1988). *Communicative language testing*. New York: Prentice Hall.
- Weir, C. J. (1990). *Communicative language testing*. New York: Prentice Hall.
- Weir, C. J. (1993). *Understanding and Developing Language tests*. New York: Prentice Hall.
- Weshe, M. B. (1981). Communicative testing in a second language. *Canadian Modern Language Review*, 37: 551-71.
- Wesche, M. B. (1992). Performance testing for work-related second language assessment. In Shohamy, E. and Walton, R. (eds). *Language assessment for feedback: testing and other strategies*. National Foreign Language Center Publications, Washington DC: 103-22.
- Wiechmann, D. and Ryan, A. M. (2003). Reactions to computerized testing in selection contexts. *International Journal of Selection and Assessment*, 11.
- Wiersma, W. (2000). *Research methods in education: introduction*. Boston: Allyn and Bacon.
- Yonsei University (2005). On-line resource. Available at <http://www.yonsei.ac.kr> Last viewed March 2007.
- Zeidner, M. and Bensoussan, M. (1988). College students' attitudes towards written versus oral tests of English as a foreign language. *Language Testing*, 5: 100-114.

APPENDICES

Appendix One. Basic English Skills Test Sections

The oral Interview section is an individually administered, face-to-face interview requiring approximately 15 minutes per examinee. Tasks include:

- Telling time
- Asking for directions
- Following directions
- Counting money to buy items
- Verifying correct change
- Conversing socially

Elementary reading and writing tasks are also included in this section; together they may be used as a screening device to identify examinees for whom the Literacy Skills Section may be appropriate.

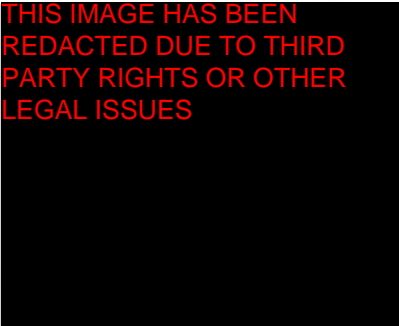
The Literacy Skills Section may be administered in one hour, either individually or to groups. Reading tasks include:

- Dates on a calendar
- Labels on food and clothing
- Bulletin announcements
- Newspaper want ads

Writing tasks include:

- Addressing an envelop
- Writing a rent check
- Filling out an application form
- Writing a short biographical passage

Appendix Two. A sample of ESPT

ESPT - Adult			
Divisions	Type of questions	Answer time	Number of questions
Part 1	Yes/No Question	10sec	2
Part 2	Choice Question	10sec	2
Part 3	Personal Information	20sec	3
Part 4	Picture Identification	25sec	1
Part 5	Giving Direction	25sec	1
Part 6	Basic Survival Situation	30sec	2
Part 7	Persuading	25sec	1
Part 8	Situation Response	30sec	1
Part 9	Reading Passage	30sec	1
Part 1 : Yes/No question			
Are you hungry?			
Part 2 : Choice Question			
Do you like being alone or with your friends?			
Part 3 : Personal Information			
What kind of movies do you like?			
Part 4 : Picture Identification			
What is the weather like in this picture?			
<p style="color: red; font-weight: bold;">THIS IMAGE HAS BEEN REDACTED DUE TO THIRD PARTY RIGHTS OR OTHER LEGAL ISSUES</p> 			

Part 5 : Giving Direction

You are at the coffee shop. How will you get to Tony Romas?

THIS IMAGE HAS BEEN REDACTED DUE TO THIRD PARTY RIGHTS OR OTHER LEGAL ISSUES

Part 6 : Basic Survival Situation

You are very unhappy with the poor service you received at a restaurant. What will you say to the manager?

Part 7 : Persuading

You want to go to Australia for vacation. Explain to your spouse why it's a good place to visit.

Part 8 : Situation Response

THIS IMAGE HAS BEEN REDACTED DUE TO THIRD PARTY RIGHTS OR OTHER LEGAL ISSUES

Talk about what these people are doing.

Part 9 : Reading Passage

예제) This winter I plan on spending a lot of time with my family. I will go to the mountains for some skiing and sledding. We will all enjoy the snow and the clean air together. I want to also go fishing with my dad and my brother. Winter is a great time for vacation.

Appendix Three. Consent Form

Dear All,

I am an English instructor and graduate student specializing in language testing at the Institute of Education, University of London. Under the direction of Dr. Catherine Walter and Professor Andrew Brown, I am working on doctoral thesis entitled 'Korean university students' attitudes to and performance on a Face-To-Face Interview (FTFI) and a Computer Administered Oral Test (CAOT). This study aims to examine the students' perceptions of and attitudes to and their performance on the two tests, and finally the effects of the attitudes on performance on the tests.

I understand that you are all busy and that your time is valuable, but this is an important study for improving the oral assessment tool in terms of its validity and reliability, and finally will result in more beneficial effects on learning and teaching. I would be most grateful if you could help me by participating in the study. The information you provide will be very valuable for the study. You will be asked to take the FTFT and the CAOT and fill out the questionnaires, which will cost you approximately 10 minutes. They are designed to obtain your perceptions of and attitudes toward the FTFI and the CAOT in a Korean university context. Some of you will also be invited to participate in interviews.

Your participation is voluntary, and you may refuse to participate or withdraw from participation at any time and for any reason without penalty. The data collected from you such as tape and computer recordings of your performance will be used only for my research. Your personal information will, however, be kept confidential. Your name and other identifying information will never be released with your express written consent. If you have any questions, please feel free to ask me via e-mail at ing116@yahoo.co.kr or cell phone at 010-6657-7965.

I want to thank you in advance for your kind cooperation.

Sincerely Yours

Mi-jin Joo

Please sign a copy of the letter if you wish to participate. The other copy is for your records.

I fully understand the terms in the letter and agree to participate

Name _____ Signature: _____ Date: _____

Appendix Four. Questionnaire for oral testing formats

Please tick (✓) one box only on each line

	<u>Less than once</u> <u>a month</u>	<u>Between once</u> <u>a week and</u> <u>once a month</u>	<u>A few times</u> <u>each week</u>	<u>Almost</u> <u>every day</u>
1. How often do you use a computer?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	<u>Less than 1 year</u>	<u>1-2 years</u>	<u>3-4 years</u>	<u>More than</u> <u>5 years</u>
2. How long have you used a computer?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	<u>Strongly disagree</u>	<u>Disagree</u>	<u>Agree</u>	<u>Strongly agree</u>
3. I feel comfortable when I use a computer.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4. Using a computer is very interesting.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5. I want to use a computer as much as possible.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
6. It is easy to use a computer.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	<u>Strongly disagree</u>	<u>Disagree</u>	<u>Agree</u>	<u>Strongly agree</u>
7. It is a good test format as an exam.				
Computer Administered Oral Test	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Face-To-Face Interview	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
8. I could do justice to my ability.				
Computer Administered Oral Test	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Face-To-Face Interview	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
9. The test had enough questions to assess my speaking ability.				
Computer Administered Oral Test	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Face-To-Face Interview	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
10. I had sufficient time to think about the questions before I spoke.				
Computer Administered Oral Test	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Face-To-Face Interview	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
11. I had sufficient response time.				
Computer Administered Oral Test	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Face-To-Face Interview	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
12. I was nervous while I was taking the test.				
Computer Administered Oral Test	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Face-To-Face Interview	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
13. Questions in the test were fair.				
Computer Administered Oral Test	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Face-To-Face Interview	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
14. I had to wait for a long time before taking the test, and it made me tired.				
Computer Administered Oral Test	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Face-To-Face Interview	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
15. The test was easy to use.				
Computer Administered Oral Test	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Face-To-Face Interview	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

	<u>CAOT</u>	<u>FTFI</u>	<u>Same</u>
16. Which test made you feel more nervous?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
17. Which test do you want to take again as your exam?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	<u>A computer</u>	<u>A person</u>	<u>Same</u>
8. Which test was more comfortable for you - talking to a computer or to a person?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	<u>Yes</u>	<u>No</u>	
19. During the Computerized Oral Testing, were the voices on the computer clear?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
20. While undergoing the Computerized Oral Testing, was it disturbing for you to have other students talking at the same time?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	<u>CAOT</u>	<u>FTFI</u>	<u>Same</u>
21. Which test was more difficult for you? Why?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
22. Which test was fairer? Why?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
23. Overall, which test do you prefer? Why?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

	<u>Yes</u>	<u>No</u>
24. Have you ever experienced the following tests before?		
Face-To-Face Interview	<input type="checkbox"/>	<input type="checkbox"/>
Computerized Oral Testing	<input type="checkbox"/>	<input type="checkbox"/>
	<u>CAOT</u>	<u>FTFI</u>
25. Which speaking test did you do first?	<input type="checkbox"/>	<input type="checkbox"/>
	<u>Female</u>	<u>Male</u>
26. Gender	<input type="checkbox"/>	<input type="checkbox"/>
27. Age	_____	
	<u>Freshperson</u>	<u>Sophomore</u>
		<u>Junior</u>
		<u>Senior</u>
28. Grade	<input type="checkbox"/>	<input type="checkbox"/>
		<input type="checkbox"/>
		<input type="checkbox"/>
29. Major	_____	

If you agree to my contacting you for an interview after this questionnaire, please tick this box

Yes No

Thank you very much for your kind cooperation!!

Appendix Five. The structure of 2004 tests

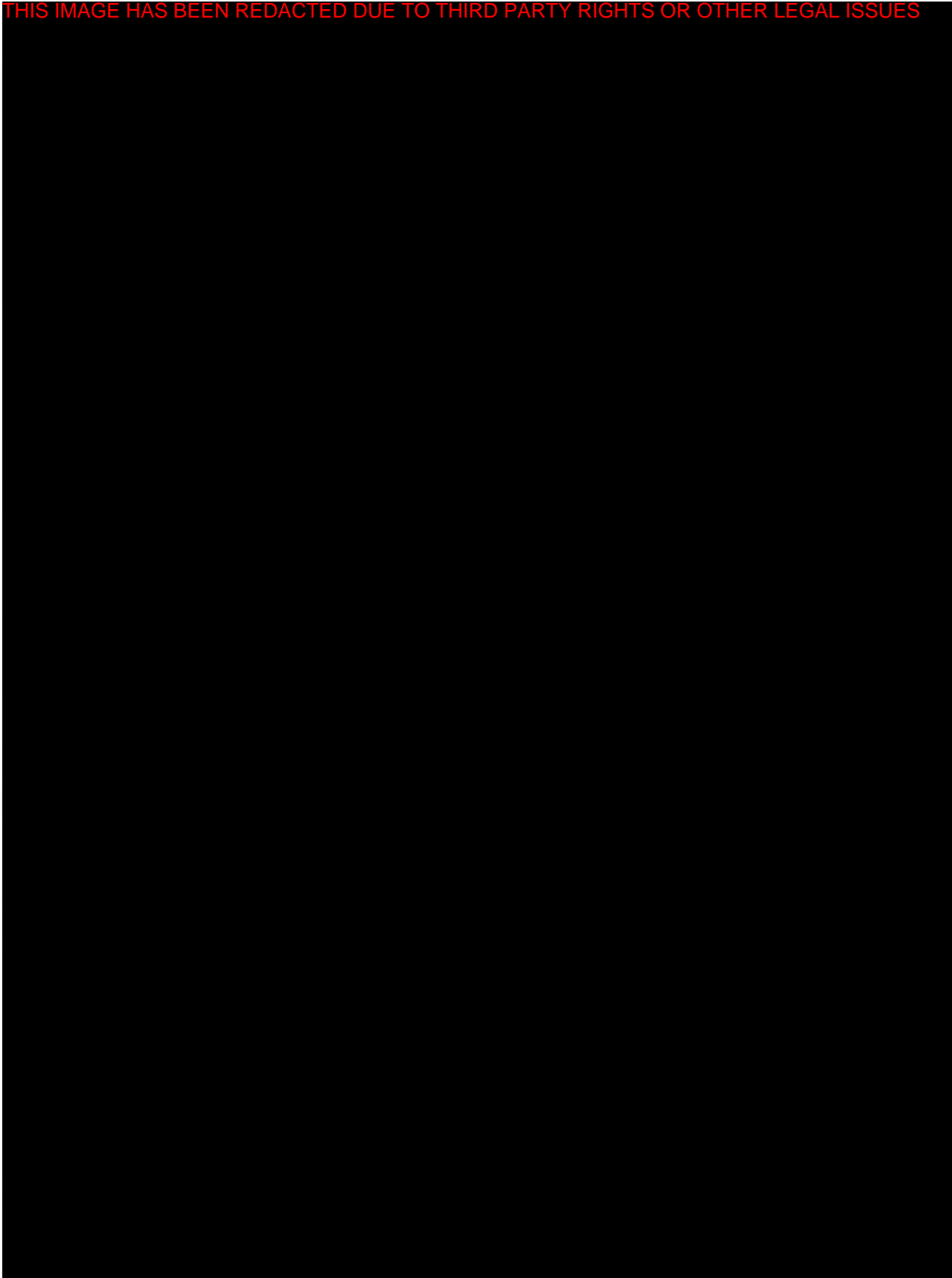
<p align="center">Computerized Oral Test (approximately 20 minutes' duration)</p>	<p align="center">Face-To-Face Interview (approximately 10 minutes' duration)</p>
<p align="center">Warm up (unassessed) <i>: What's your name? How are you today?</i></p>	<p align="center">Warm up (unassessed) <i>: What's your name? How are you?, etc.</i></p>
<p>Questions and answer task (7 items) <i>: What date is it today?</i> <i>What does your father do? Where does he work?</i> <i>What time do you usually leave your house? How long does it take to get to school?</i> <i>How long have you lived in Busan?</i> <i>Please introduce yourself.</i> <i>What do you usually do in your free time?</i> <i>What are you doing tomorrow?, etc.</i></p> <p>Using a picture or a picture story task (2 items) <i>: You are looking for your friend at the party. The person in the picture is your friend. Describe him/her.</i> <i>Make up a story based on a cartoon on the screen.</i></p> <p>Role play (1 item) <i>: You are supposed to meet your friend at 1:00 today, but you may arrive at the appointment place an hour late because of a sudden part-time job interview. Your friend doesn't answer the phone. Leave a message on the phone.</i> <i>Or you made an appointment with your friend this weekend, but you have a sudden job interview on the same day and want to cancel the appointment. Your friend is not answering her cell phone, and you have to leave a voice message. Apologize and explain why you can't meet your friend this weekend.</i></p>	<p>Questions and answer task (3-4 items) <i>: What time/day is it?</i> <i>What does your mother/brother/sister do?</i> <i>How long have you been studying English?</i> <i>What do you do on Weekends?</i> <i>What are you doing on Monday/after this test?, etc.</i></p> <p>Using a picture or a picture story task (1-2 items) <i>: You are looking for your friend at the party. The person in the picture is your friend. Describe him/her.</i> <i>Or describe your mother/father/sister, etc.</i> <i>Make up a story based on a cartoon.</i></p> <p>Role play (1 item) <i>: You are supposed to meet your friend on 7:00 today, but you may arrive at the appointment place half an hour late because of a sudden part-time job interview. Now explain the reason you are late and apologize.</i> <i>Or you want to see a movie with your friend this weekend. Ask several questions in order to make an appointment.</i> <i>Or you are going to phone your friend and invite her to your birthday party. Now make several questions for the invitation.</i></p>

Appendix Six. A sample of the CAOT

THIS IMAGE HAS BEEN REDACTED DUE TO THIRD PARTY RIGHTS OR OTHER LEGAL ISSUES



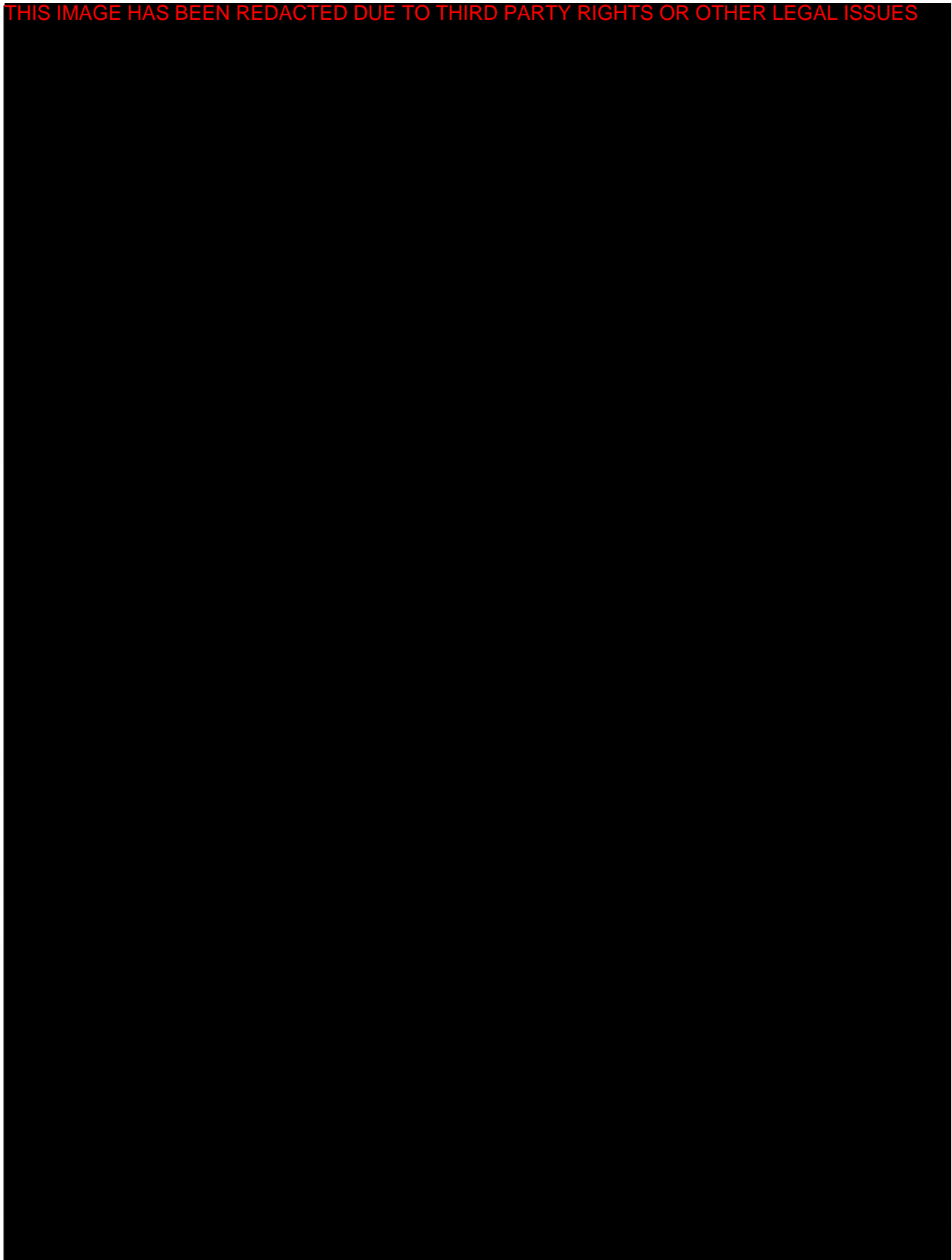
THIS IMAGE HAS BEEN REDACTED DUE TO THIRD PARTY RIGHTS OR OTHER LEGAL ISSUES



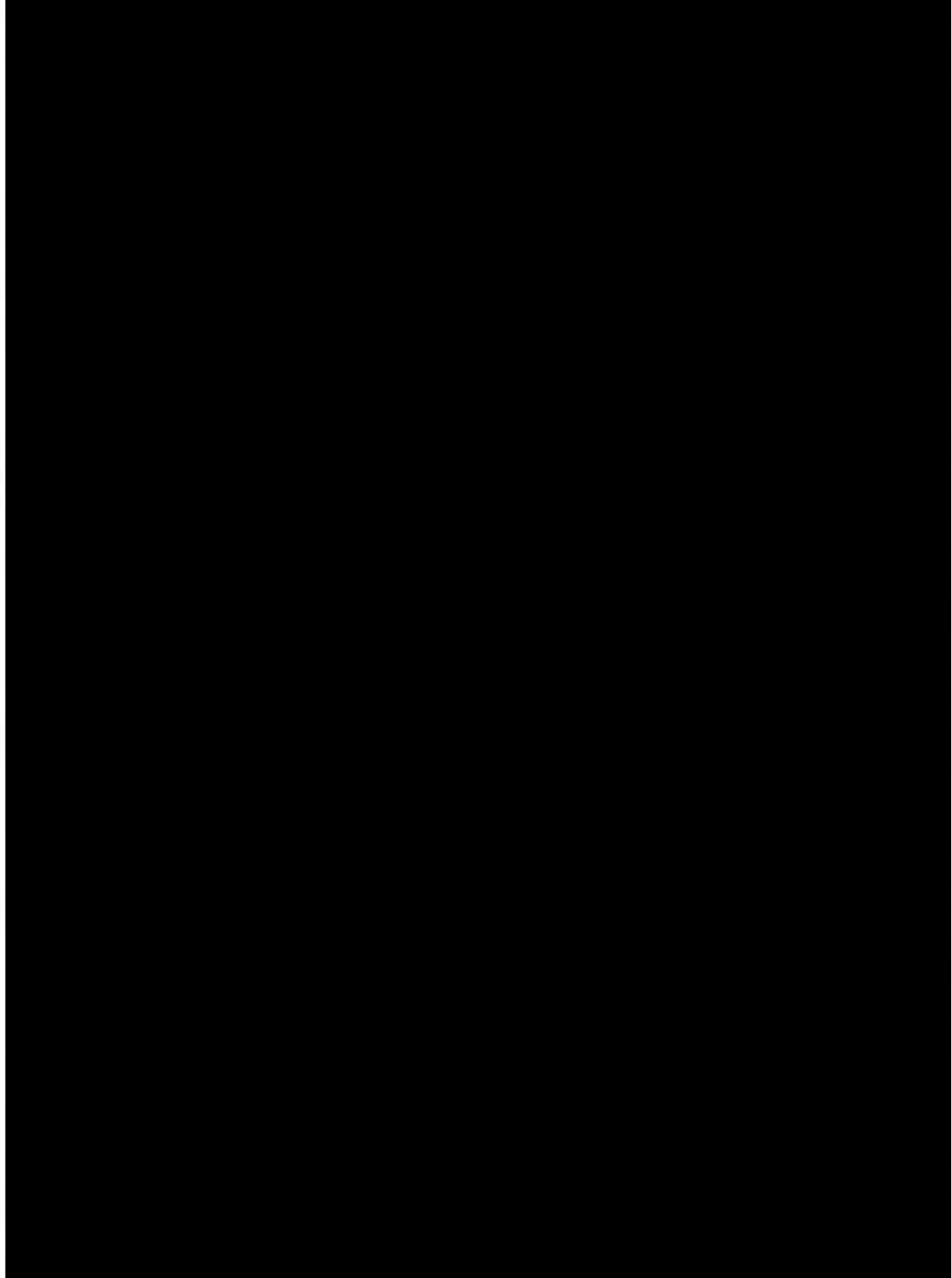
THIS IMAGE HAS BEEN REDACTED DUE TO THIRD PARTY RIGHTS OR OTHER LEGAL ISSUES



THIS IMAGE HAS BEEN REDACTED DUE TO THIRD PARTY RIGHTS OR OTHER LEGAL ISSUES



THIS IMAGE HAS BEEN REDACTED DUE TO THIRD PARTY RIGHTS OR OTHER LEGAL ISSUES



THIS IMAGE HAS BEEN REDACTED DUE TO
THIRD PARTY RIGHTS OR OTHER LEGAL
ISSUES

Appendix Seven. Rating scale and descriptors

Fluency

- 7 Speech is marked by a very high degree of fluency.
- 6 Speech is marked by a high degree of fluency with occasional hesitation.
- 5 Speech is fluent but with some hesitation or deliberation.
- 4 Noticeable hesitation and some groping for words is present, but does not impede communication.
- 3 A marked degree of hesitation, grasping for words or inability to phrase utterances easily impedes communication.
- 2 Speech is fragmented because of hesitations, pauses or false starts.
- 1 Fluency is evident only in the most formulaic phrases.

Grammar

- 7 Range and control of grammatical structures are precise and sophisticated.
- 6 Candidate uses a broad range of structures with only occasional minor errors.
- 5 Communication is generally grammatically accurate with a range of structures; minor errors may be noticeable.
- 4 Satisfactory communication is achieved despite a limited range of structures and/or obvious grammatical inaccuracies.
- 3 Communication is less than satisfactory because of a limited range of structures and/or the presence of frequent errors.
- 2 Limited communication is possible but errors are likely to be frequent and intrusive.
- 1 Severe limitations of grammar prevent all but the most basic communication.

Vocabulary

- 7 Candidate uses a wide range of vocabulary precisely, appropriately and effectively.
- 6 Candidate uses a wide range of vocabulary effectively though occasionally may be imprecise.
- 5 Vocabulary is broad enough to allow the candidate to express ideas well. Circumlocution is smooth and effective, if required.
- 4 Vocabulary is broad enough to allow the candidate to express most ideas. Can usually circumlocute to cover gaps in vocabulary, if required.
- 3 Vocabulary is broad enough to allow the candidate to express simple ideas. Circumlocutions are sometimes ineffective.

- 2 Limited vocabulary restricts expression to common words and phrases. Circumlocution is laborious and often ineffective.
- 1 Vocabulary is very limited.

Intelligibility

- 7 Speech is clear and can be followed effortlessly.
- 6 Speech is generally clear and can be followed with little effort.
- 5 Speech can be followed though at times requires some concentration by the listener.
- 4 Speech can generally be followed though sometimes causes strain.
- 3 Speech can generally be followed though frequently causes strain.
- 2 Speech requires constant concentration to be understood.
- 1 Speech can only be followed intermittently and then only with considerable effort.

Cohesion

- 7 Cohesive devices are smoothly and effectively managed.
- 6 A good range of cohesive devices is used but occasionally these may be inappropriate.
- 5 A range of cohesive devices is used but these may be inappropriate.
- 4 Cohesive devices are limited in range and may be used inappropriately or inaccurately.
- 3 Very simple cohesive devices are used to link sentences but errors are frequent.
- 2 There is some evidence of connected discourse but the overall effect is disjointed.
- 1 Candidate is able to use only isolated words and formulaic phrases.

Overall communicative effectiveness

Scale of 1 to 7 without descriptors (7=near native flexibility and range; 1 = limited).

(O'Loughlin, 2001: 217-219)

Appendix Eight. Comparative results on Section Two of Joo's study (2004)

Questionnaire item	Format	Median (S.D)	Z (p-value)
7. It is a good test format as an exam.	FTFI	1.90(.57)	-.20
	CAOT	2.09(.71)	(.229)
8. I could do justice to my ability.	FTFI	2.69(.74)	-.50
	CAOT	2.63(.69)	(.616)
9. The test had enough questions to assess my speaking ability.	FTFI	2.44(.85)	-.209
	CAOT	2.47(.67)	(.835)
10. I had sufficient time to think about the questions before I spoke.	FTFI	2.47(.70)	-3.43**
	CAOT	1.95(.68)	(.001)
11. I had sufficient response time.	FTFI	2.04(.70)	-.028
	CAOT	2.04(.53)	(.978)
12. I was nervous while I was taking the test.	FTFI	1.59(.82)	-2.96**
	CAOT	2.06(.85)	(.003)
13. Questions asked in the test were fair.	FTFI	2.09(.71)	-2.87**
	CAOT	1.73(.44)	(.004)
14. I had to wait for a long time before taking the test, and it made me tired.	FTFI	2.14(.89)	-4.76**
	CAOT	3.11(.58)	(.000)

* p<.05 ** p <.01