A COMPARATIVE STUDY OF THE EFFECTS OF A COMPUTERIZED ENGLISH ORAL PROFICIENCY TEST FORMAT AND A CONVENTIONAL SPEAK TEST FORMAT

DISSERTATION

Presented in Partial Fulfillment of the Requirement for

the Degree Doctor of Philosophy in

the Graduate School of The Ohio State University

By

Eunjyu Yu, M.A.

* * * * *

The Ohio State University

2006

Dissertation Committee:

Dr. Charles R. Hancock, Advisor

Dr. Ayres D'Costa

Dr. William E. Loadman

Approved by

_____

Advisor

College of Education

ABSTRACT

Despite the increasing use of computer technology in language testing, limited research literature is available about the validity, reliability, and nature of computerized spoken language tests. To date, only mixed results about interactions between test taker characteristics and computerized language test format have been reported.

To add to the body of research on this topic, this study explored the relationship between test taker characteristics and test delivery format during spoken English proficiency assessments. A total of 210 international students whose native language was not English were recruited at a U.S. university in autumn 2005. The main data sources included the results of a computerized spoken English test, an audio-taped SPEAK test, and replies to a questionnaire. For data analysis, this study utilized a 2×2×2 mixed factorial research design with random assignment.

This study found that an interaction among all three independent variables (*i.e.*, self-reported years of English study, self-reported computer use, and test delivery format) was not significant. Self-reported years of English language study and test delivery format, however, cooperatively produced a significant influence on test scores for the spoken English test. Specifically, the computerized speaking test, not the audio-taped SPEAK test, seemed to affect test results more for the group that self-reported less English study than for the group that self-reported greater English study. In addition,

self-reported computer use did not significantly affect test results during oral proficiency assessments.

Although this study was limited in terms of a single research site, a single test, and self-reported data, the study has corroborated previous research that emphasized the appropriate use of different test delivery formats according to the purposes of the tests and the characteristics of test takers. Thus, this study called for further study to ensure that a test functions fairly across various types of test takers regardless of their backgrounds. Also, this study suggested sharing ownership of testing among test makers, test takers, and test users, which might allow all interested parties to receive the benefits of testing. Finally, the findings will be useful to understand both the benefits and disadvantages of using technology in language testing.

Dedicated to my parents

ACKNOWLEDGMENTS

I would like to gratefully acknowledge the support and love of all those people who have made my work possible.

Most of all, my deepest appreciation goes to my parents, Myungik Yu and Kyungja Kim, and my younger bother, Jeyong, who is no longer with us. Their never-ending love and support opened my eyes to the world and continue to sustain me. I have learned so much about life from them.

I would also like to express deep appreciation and respect for my academic advisor, Dr. Charles R. Hancock. His amazing support has been a blessing to me. His encouragement and professional insight has inspired both my personal and professional development. He has made me think even more about how education can make a better world. Indeed, his enthusiasm and professional expertise have provided tremendous guidance throughout my study at The Ohio State University. Moreover, he is a great role-model for all educators, scholars, and educational administrators.

My whole-hearted appreciation also goes to Drs. Ayres D'Costa and William Loadman. I am very lucky to have had these two wonderful people on my dissertation committee. Dr. D'Costa has a warm heart and expert knowledge about measurement and evaluation. His thoughtful care and advice continues to stimulate my work. Dr. Loadman has a big smile and professional expertise in applied research and evaluation methodology. His insightful comments continue to motivate me as well.

I am very thankful to Dr. Robert Hite, NCATE Coordinator. Throughout three years of working experience as his graduate research associate in preparing for the 2005 National Council for Accreditation of Teacher Education On-Site Review at OSU, he promoted my knowledge of teacher education.

I sincerely thank Dr. Keiko Samimy. As an expert on non-native English speakers, she understands international students and her loving care has encouraged me a lot.

My appreciation goes to Dr. Shelley Wong. She has a kind heart and passion for sociocultural theory. She gave me one invaluable lesson: “less is more.”

My gratitude goes to Ms. Susan Sarwark, director of the Spoken English Program at OSU for her support in the process of data collection.

My acknowledgment goes to faculty members in the Department of English Language Education and the Department of English Language and Literature at Pusan National University, ROK.

Also, I would like to give my appreciation to the many others who helped me complete this study.

I will never forget the help I received from so many people, and I will devote my knowledge and passion to helping others.

VITA

September 7, 1969	Born—Pusan, The Republic of Korea
1993	B.A. English Language Education, Pusan National University, Pusan, Korea
1993-1995	Graduate Research Associate, Pusan National University, Pusan, Korea
1995	M.A. English Language and Literature, Pusan National University, Pusan, Korea
1995-2001	Instructor, English, Pusan National University and *etc*., Korea
2000	ABD doctoral candidate, English Language and Literature, Pusan National University, Pusan, Korea
2003-2006	Graduate Research Associate, The Ohio State University

FIELDS OF STUDY

Major Field: Education

Area of Focus: Language, Literacy, and Culture

Foreign/Second Language Education

Minor Field: Language Testing

Quantitative Research, Evaluation and Measurement

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

OVERVIEW OF THE STUDY

1.1 Introduction

Thousands of languages are currently spoken in the world. They, however, have one essential common purpose, communication. In the global age, learning an additional language is encouraged for both international and intra-national communication. In order to have a command of additional languages for various purposes, there is an increase in the number of people who learn a second, third, and even fourth language as well as their mother tongue.

Corresponding to this trend, language educators and researchers have made a multifaceted effort to understand the foreign/second language learning phenomenon and to establish a conceptual framework for language proficiency. For example, Hymes (1972) first published the theoretical framework for the communicative approach to language teaching. According to Hymes (1972), the communicative approach emphasized communicative competence, a functional language ability to use a language appropriately.

This communicative approach to language teaching also made a shift in the field of language testing. Moving the focus from the knowledge about language to the use of language, language testing practitioners have constantly developed scales and tests for

measuring communicative ability in foreign/second languages. A good example was the Foreign Service Institute Oral Proficiency Interview (OPI). This interview-based oral proficiency test was the first spoken language test requiring a test taker to demonstrate his/her functional language ability through a live interview with a trained rater (Sollenberger, 1978; Fulcher, 2000). The FSI developed an 11-point (0-5) rating scale (Appendix A) distinguishing a wide range of general oral proficiency of the US government employees in foreign/second languages (Arnett & Haglund, 2001). Later, the FSI was called the Interagency Language Roundtable (ILR) OPI.

For the need to identify oral proficiency at low level commonly found in academic settings (Liskin-Gasparro, 1987; Arnett & Haglund, 2001), the American Council on the Teaching of Foreign Languages (ACTFL) (1982) published the ACTFL Proficiency Guidelines (Appendix B & C). The ACTFL modified the ILR scale into a 10-point scale over four major levels (novice, intermediate, advanced, and superior) (Swender, 1999). The ACTFL OPI is described as "a standardized procedure for the global assessment of functional speaking ability" (Swender, 1999, p. 1). The ACTFL OPI is conducted either in live face-to-face or over the telephone. The OPI requires a test taker to demonstrate his/her functional language ability to perform specific interactive tasks, described in the ACTFL Proficiency Guideline for Speaking, in a target language (Swender, 1999). An interview is recorded onto an audio tape for evaluation. One or more ACTFL-certified testers evaluate each recorded response sample. The proficiency level of a test taker is determined according to 'what he/she can do with the language' and 'what he/she cannot do with the language' required for a corresponding proficiency level (ACTFL/ILR, 1999).

Despite a reputation as a valid, reliable, standardized measurement of speaking ability, practical issues in administering the ACTFL OPI have been discussed (Stansfield, 1996; Chalhoub-Deville, 1997). For example, this OPI was originally designed for one-to-one administration. For a live face-to-face interview, an interlocutor and a test taker should be simultaneously in the same place. The logistic requirement for its administration limited the use of the live face-to-face interview (Kenyon & Malabonga, 2001). In addition, an interlocutor generated interview questions on the basis of what a test taker said during the interview in order to elicit sufficient ratable response samples (ACTFL/ILR, 1999). This individualized structure of the OPI procedure and variance across interlocutors might affect test takers' performance on the test (Lazaraton, 1996). These concerns called for a more practical surrogate.

In the early 1980s, as an alternative to a live face-to-face OPI, an audiotape recorder began to deliver oral proficiency tests (Kenyon & Malabonga, 2001). Examples of these audio-taped oral proficiency tests included the Test of Spoken English (TSE), the Speaking Proficiency English Assessment Kit (SPEAK), the Simulated Oral Proficiency Interview (SOPI), and the Texas Oral Proficiency Test. The TSE and the SPEAK were administered by the Educational Testing Service (ETS). The SOPI and the Texas Oral Proficiency Test were designed by staff at the Center for Applied Linguistics (CAL).

As semi-direct tests, these tape-mediated oral proficiency tests require a test taker to accomplish particular language tasks. The language tasks are categorized into three main tasks:

> picture-based speaking tasks include giving directions, describing activities in a familiar setting, and telling a story; topic-based speaking tasks include describing a procedure, presenting advantages and disadvantages, explaining and defending

> a point of view, or describing what would happen if a hypothetical situation were to come true; situation-based tasks include giving advice to a friend, apologizing for having offended someone, and making a formal presentation to a group (CAL, 2003).

These tape-mediated oral proficiency tests consist of a test booklet and a master test audiotape including test directions and test items (ETS, 1982a; Stansfield & Kenyon, 1996; Malone, 2000). Each test taker's response is recorded on a separate audiotape for evaluation.

Stansfield and Kenyon (1996) list several practical and psychometric benefits of these tape-mediated tests. Specifically, a proctor can administer a tape-mediated speaking test to a group of test takers and an individual as well. A proctor is not required to be fluent in the language being tested. The variance across interlocutors is reduced by using recorded test items and printed materials. The recorded response samples are evaluated under controlled conditions. Raters can relatively easily distinguish a proficiency level because this audiotape format elicits response samples answering to the same questions.

Despite the widespread use of these tape-mediated speaking tests over two decades, Kenyon and Malabonga (2001) reported that some aspects of typed oral proficiency tests made the test more difficult than a live face-to-face OPI. For example, test takers could not control any of test context/content, item difficulty, and response time. The beep sound in the SOPI often interrupted test taker's response to a question. Besides, Underhill (1987) pointed out insufficient use of visual clues and the lack of interaction with a human being.

With the technological advances of the late twentieth century, computers have become language testing tools as well as instructional tools (Gruba & Corbel, 1997;

Brown, 1997, 2004; Bachman, 2000; Chapelle, 2001). Experiments using computers to teach language began as early as 1960s (Warschauer & Healey, 1998; Chapelle, 2001). A considerable literature base has described the use of computer technologies as a tool for language teaching and learning (Chapelle, 2001). However, it was not until the advent of cheap, powerful 16 and 32 bit personal computers with built in sound capability that computer assisted language teaching and testing became viable as a real alternative to traditional approaches (Jonassen *et al*., 2003). The efficiency and accuracy of computer-mediated tests was compelling and led to develop a variety of computer-mediated language tests such as the Graduate Record Examination, the Test of English as a Foreign Language, and the General Management Admission Test.

Computer-enhanced multimedia tools for language testing have now become widespread. Nevertheless, much less is known about the authenticity, effectiveness and long-term potential of computerized language tests, particularly, computerized speaking tests. This is partly due to the fact that the technologies necessary for a computerized speaking test have only recently become available (Chalhoub-Deville & Deville, 1999; Jeong, 2003). There exists a substantial need among language testers, educators, future employers of foreign nationals, and test takers to utilize computer technologies in ways that will improve the authenticity, cost efficiency, accessibility and ease of administration of language proficiency tests and several such tests have now been developed.

Examples of a computerized speaking test in action include Computerized Oral Proficiency Instrument (COPI), Digital Video Oral Communications Instrument (DVOCI) and Purdue's oral English test. The COPI was developed by staff at the Center for Applied Linguistics in order to measure less commonly taught languages such as

Arabic, Chinese and Spanish. DVOCI has been administered by the LARC at San Diego State University. Recently, LARC has distributed the LARC Speech Test Authoring Resource (LARCStar), software supporting rich multimedia capabilities. LARCStar enables the construction of online speaking tests in foreign/second languages, with a collection of results to a web server. And, Purdue's oral English test was developed for the purpose of screening non-native English speaking teaching assistants.

1.2 Significance of the study

As test items have been developed in new task forms and presented on various delivery formats, language testers have paid growing attention to the comparability and the validity of test scores across various test formats (Mead & Drasgow, 1993; Hambleton, 1994, 2001; Bachman, 2000; Schwarz *et al*., 2003). Moreover, since the late 1980s, the awareness of the social significance of the interpretation and use of test results has called more attention to test taker characteristics. In other words, a test result should be accounted for by the ability of interest, not by other factors, particularly by test taker characteristics (Lord, 1980; Messic, 1989; Bachman, 1990; Angoff, 1993; Raju & Ellis, 2002). This concern called for research on potential effect of test taker characteristics on test results. Examples are studies on test validation with a focus on fairness, test bias, and differential item function.

Currently, the perceived efficiency of computer technology has motivated test developers to take advantage of computer technology as a language testing tool. Nevertheless, much less is known about the authenticity, effectiveness and long-term potential of computerized language tests, particularly, computerized speaking tests. This

is partly due to the fact that the technologies necessary for a computerized speaking test have only recently become available (Chalhoub-Deville & Deville, 1999; Jeong, 2003). This current trend in language testing raised concerns about validity of computer-mediated language tests (Dunkel, 1999; Grabe, 1999; Alderson, 2000a; Bernhardt, 2000).

In response to the concerns, few studies have been conducted on computerized language tests but mixed results were reported. For instance, Taylor, Jamieson, Eignor and Kirsch (1998), and Roever (2001a) reported that there was no significant correlation between examinees' computer familiarity and their test scores on computer-based TOEFL test measuring language ability in listening and reading. Kenyon and Malabonga (2001) investigated examinee attitudes toward face-to-face OPI, taped SOPI and COPI. This study found that the COPI was less difficult than the SOPI. A live face-to-face interview was perceived to measure "real-life speaking skills" (p. 60). On the other hand, Jeong (2003) reported that there was a positive relationship between the electronic literacy and the English oral proficiency of the examinees who took DVOCI in English as a foreign language context.

Considering the limited volume of professional literature and mixed findings about interaction test taker characteristics and speaking test delivery format, particularly computerized speaking test (Chalhoub-Deville & Deville, 1999; Sawaki, 2001), further study is necessary to investigate the validity of a computer-mediated language test.

In order to add to the professional literature base in this area, this present study explores the relationship between test taker characteristics and test delivery format during oral proficiency testing. Specifically, with respect to test taker characteristics, this study focused on two self-identified test taker characteristics: years of English language study

and self-reported computer use. In terms of test delivery format, a conventional taped speaking test (*i.e.*, the SPEAK) and a computerized speaking test were utilized for this study.

This study will contribute to an understanding of both the benefits and disadvantages of the use of technologies in the oral proficiency assessments. The findings of this study will be useful for test developers, language educators and other stakeholders. It will enable them to develop valid testing instruments, to appropriately interpret test results and to customize ESL/EFL speaking programs that will correspond to the need of the particular population. It may also help to provide testing and assessment formats based on students' preferred learning styles and to explore testing and assessment formats to efficiency purposes.

1.3 Purpose of the study

The present study is aimed at exploring the effects of test delivery formats and test taker characteristics (*i.e.*, self-reported years of English language study and self-reported computer use) on performance during oral proficiency assessment. Specifically, this study investigates the following questions: 1) To what extent does self-reported years of English language study affect test results during spoken English proficiency testing? 2) To what extent does test delivery format affect test results during spoken English proficiency testing? 3) To what extent does the self-reported computer use of a test taker affect test results during spoken English proficiency testing? 4) Are there interactions between the two self-reported test taker characteristics and these test delivery formats?

For the purposes of this study, 210 international graduate students whose native language was not English were recruited in a major US university on a volunteer basis. This study utilizes the results of a computerized speaking test and an audio- taped speaking test (*i.e.*, the SPEAK test), and responses to a participant questionnaire. The participant questionnaire was designed to collect pertinent information including participants' experience of taking a spoken English test, their computer use, their English study experience, and comprehensive background.

For the analysis of the data, this study adopts a 2x2x2 mixed factorial research design with use of statistical software SPSS. Specifically, the statistics about reliability and validity for test items and a participant questionnaire are followed by the descriptive statistics presenting central tendency and variability of test scores on an English speaking test. Subsequently, ANOVA statistics tested the significance of any possible effects of factors. Conclusions of the investigations were made according to the results of data analysis.

1.4 Research questions

With the use of the test results on a computerized speaking test and an audio-taped speaking test (*i.e.*, the SPEAK test), and responses to a participant questionnaire, this study investigates the following questions (detailed in 3.4.1):

Q1: To what extent does self-reported years of English language study relate to performance on an oral proficiency test?

Q2: To what extent does test delivery format relate to performance on an oral proficiency test?

Q3: To what extent does self-reported computer use relate to performance on an oral proficiency test?

1.5 Basic assumptions

Language proficiency is defined as "the ability to use the language effectively and appropriately in real-life situations" (ACTFL, 1999, p.1). Under the assumption that language ability is measurable, language testers have developed scales for measuring language proficiency in listening, reading, speaking and writing (Bachman, 1990). Numerous language testing instruments have been developed and administered to determine language proficiency. The communicative approach to language has assumed that oral proficiency is one of the important measurable communicative competencies.

Several testing agencies have developed speaking test items in a tape-mediated format as a practical alternative to live interviews. The Test of Spoken English (TSE) is a good example that "has been administered worldwide" to measure oral proficiency of non-native English speakers in an academic or professional environment (Miles, 2004). The TSE is a criterion-referenced test with high inter-test-form reliability of .91 (Educational Testing Service [ETS], 1982b, p. 21). As the institutional version of TSE, the SPEAK test has been used in US universities for over two decades for the purpose of diagnosing oral proficiency and screening potential international teaching assistants (Sarwark, Smith, MacCallum, & Cascallar, 1995). With identical test items to those of TSE, the SPEAK test also demonstrates high inter-test-form reliability. In this study, both

SPEAK and a computerized speaking test use the same test item sets selected from original forms of the SPEAK test. Thus, the test item sets are assumed to be reliable in measuring constructs (e.g., accuracy of pronunciation).

Since the participants are recruited on a voluntary basis they are assumed to be highly motivated and interested in measuring their oral proficiency. They are assumed to complete tests to the best of their ability and respond honestly to the questionnaire. The spoken English tests of interest will be administered in almost identical testing environments. The testing settings such as the testing room, proctoring, and test equipment are assumed to meet commonly acceptable testing standards. A mandatory tutorial is assumed to familiarize test takers with the procedure of test administration as well as testing equipment.

1.6 Limitations of the Study

This study used quantitative research methodology. A mixed methodology including both quantitative and qualitative components may enrich information about variables related to test takers. In addition, the data were collected at a large US public university with a significant population of international students. Specifically, 3,799 international students were enrolled in undergraduate and graduate programs at the research site as of autumn quarter 2005. Approximately 2,000 international students were eligible for this study. The limitations in time and cost confined the sample size to 210 participants. Therefore, if the results are to be generalized beyond the participants in this study, the findings of this study should be interpreted with caution.

1.7 Definition of terms

The following terms are defined both constitutively and operationally as used in the present study. In each of following definitions, the constitutive definition is presented first and the operational one second.

Self-reported computer use

The extent to which participants are familiar with various computer tasks. In the present study, self-reported computer use is categorized as either more computer use or less computer use. Self-reported computer use is defined in terms of the frequency of computer use reported by participants at various levels. The participant questionnaire (Appendix I) collects information about computer use of participants.

Computerized speaking test

A type of oral proficiency assessment administered via a computer. A computerized speaking test incorporates text, graphics, full-motion video, and sound into an integrated assessment package (Burstein *et al*., 1996). In the present study, a computerized speaking test is a test designed to measure spoken English proficiency of non-native English speakers with the use of digital multimedia.

Internationals

Non-native speakers of English who study at colleges or universities where English is spoken. In the present study, internationals are non-native English speaking international students who are enrolled full-time or part-time in a graduate degree program at The Ohio State University, USA.

Oral proficiency

The ability to communicate appropriately in a target language in real-life situations (ACTFL, 1999). In the present study, oral proficiency is defined as the ability to communicate appropriately in English in a U.S. academic environment.

Partial credit model

An extension of dichotomous (correct/incorrect) scoring (Masters & Wright, 1997). It identifies "one or more intermediate levels of performance on an item and awards partial credit for reaching the intermediate levels" (Bateman & Griffin, 2003, p. 6). In the present study, polytomous scoring is a model that gives partial credit to each item on a computerized speaking test and the SPEAK test with a 0 to 3 score scale.

Performance

Performance is defined as "the actual use of language in concrete situations" (Chomsky, 1965, p.4). In the present study, performance means the test results of a computerized speaking test and/or the SPEAK test.

The Speaking Proficiency English Assessment Kit (SPEAK)

A version of the Test of Spoken English designed to measure general spoken English proficiency for internationals and professionals whose native language is not English (ETS, 1982a). In the present study, the SPEAK is a test used to measure English oral proficiency of international graduate students with the use of an audiotape recorder.

Tape-mediated (or taped) oral proficiency assessment

A type of test using a test booklet and audio-taped directions to elicit spoken response samples from the examinee. In the present study, tape-mediated oral proficiency assessment means the SPEAK test.

Test delivery format effect

Results caused by different test delivery formats. In this study, test delivery format effect means the effect that may be caused by the different test delivery formats of a computerized speaking test and the SPEAK test.

1.8 Organization of the dissertation

This dissertation is composed of five chapters. Chapter one provides an overview of this study, including the significance and purpose of this study, research questions, basic assumptions, limitations, and definition of terms. Chapter two reviews the relevant literature, with a focus on four areas: communicative approach to language teaching, communicative approach to language testing, technology for language testing, and validity in language testing. Chapter three describes the methodological procedures

adopted to design the research and analyze the data. It includes details of participants, setting, instruments, research methods, and procedures for data collection and data analysis.

Chapter four will present the statistics about reliability and validity for research instruments, descriptive statistics, and ANOVA statistics. The findings related to the research questions will be discussed. In chapter five, the conclusions of this study will be made according to the findings and then its implications will be discussed. Limitations will be followed by recommendations for further research. Conclusion will recap main points of this study. After references, the appendices present the key study instruments, and other relevant documents.

CHAPTER 2

REVIEW OF LITERATURE

2.1 Introduction

The theoretical framework for this study is drawn from four areas: 1) the communicative approach to language teaching, 2) the communicative approach to language testing, 3) technology for communicative language testing, and 4) validity in language testing.

2.2 Communicative approach to language teaching

In the early 1970s, a new approach to language teaching, called the communicative approach, drew attention to language learners' communicative ability, particularly, their communicative ability in oral use. The theoretical framework for the communicative approach began with Dell Hymes. With a new concept, 'communicative competence,' Hymes (1972) emphasized the ability to use a language appropriately over the knowledge about a language. Canale and Swain (1980) further developed Hymes' framework by elaborating on three components of communicative competence (Choi, 1999; Sato & Kleinsasser, 1999; Kim, 2001). According to Canale and Swain (1980), communicative competence is composed of grammatical competence, sociolinguistic competence and strategic competence. Grammatical competence refers to "knowledge of lexical items and of rules of morphology, syntax, sentence-grammar semantics and

phonology" (p. 29). Sociolinguistic competence is the ability to "appropriately use a language within a given sociocultural context" (p. 30). Strategic competence indicates an ability to "compensate for breakdowns in communication" (p. 30) through the use of strategies when communicating. Later, Bachman extended the theoretical framework of communicative competence to communicative language ability (Choi, 1999; Sato & Kleinsasser, 1999; Kim, 2001). According to Bachman (1990), communicative language ability is composed of "both knowledge, or competence, and the capacity for implementing, or executing that competence in appropriate, contextualized communicative language use" (p. 84). These studies suggest the development and subsequent refinement of the concept of communicative competence during the 1970's through the early 1990's.

In short, in the communicative approach, communicative competence is defined as "the ability to use language appropriately, both receptively and productively, in real life situations" (K. and S. Kitao, 1996, p. 2). The receptive language ability includes reading and listening ability while the productive language ability refers to speaking and writing skills.

2.3 Communicative approach to language testing

2.3.1 Communicative language testing

Initiated for understanding the second language learning phenomena, the communicative movement was extended to include assessing language learners' communicative ability during the late 1970s (Fulcher, 2000). K. and S. Kitao (1996) pointed out that:

> communicative language tests are intended to be a measure of how the testees are able to use language in real life situations. In testing productive skills, emphasis is placed on appropriateness rather than on ability to form grammatically correct sentences. In testing receptive skills, emphasis is placed on understanding the communicative intent of the speaker or writer, rather than on picking out specific details. In fact, the two are often combined in communicative testing, so that the test takers must both comprehend and respond in real time (p. 2).

To measure the actual use of language, communicative language tests are required to include "new content areas such as sociolinguistic appropriateness rules, new testing formats to permit and encourage creative, open-ended language use, new test administration procedures to emphasize interpersonal interaction in authentic situations, and new scoring procedures of a manual and judgmental nature" through the test process (Canale, 1984, p. 79). To meet these requirements, communicative language tests have been developed in numerous different ways. The communicative language tests, however, share three common features: proficiency-oriented, performance-based, and authentic (Morrow, 1979; Fulcher, 2000). Further details of these characteristics are described in the following sections.

2.3.2 Proficiency Movement

During World War II, the US military addressed the significance of the communicative proficiency to carry out functional language tasks in a target language in real-life situations (Omaggio, 1983; Lowe, 1988; Spolsky, 1995; McNamara, 1996; Fulcher, 2000). As a response to the need to identify functional language ability, language tests have been developed with a focus on general communicative language ability, particularly oral proficiency. For example, in 1956, sponsored by the US government, the

Foreign Service Institute (FSI) published a standardized proficiency definition and scale (Jones, 1975; Spolsky, 1975, 1995; Liskin-Gasparro, 1984a). Later, the FSI proficiency scale was called the Interagency Language Roundtable (ILR) proficiency scale.

The ILR scale defines proficiency in speaking, reading, writing, and listening based on a scale from 0 to 5 (Appendix A). The range of proficiency level includes no proficiency (0), elementary proficiency (1), limited working proficiency (2), general professional proficiency (3), advanced professional proficiency (4), and functionally native proficiency (5) (ACTFL/ILR, 1999). The ILR scale distinguishes intermediate levels of proficiency between six main proficiency levels by including plus levels from 0 to 5 in a scale which has eleven proficiency level descriptions. As can be seen in the subtitles for the IRL scale descriptions, each proficiency level is identified based on the extent of a test taker's functional language ability simulated for real life situations (ACTFL/ILR, 1999). In other words, the intensive focus is on 'what a test taker can do with a target language' and 'what he/she cannot do with the language' (ACTFL/ILR, 1999). In fact, the ILR scale has been extensively used to measure language proficiency of the US government employees (Fulcher, 2000; Arnett & Haglund, 2001).

The ILR scale, however, could not distinguish between the levels within the low level of proficiency often found in academic settings (Arnett & Haglund, 2001). According to the empirical data obtained in academic settings, the proficiency levels of L2 learners spread within a wide range of language proficiency which corresponded to the ILR scale of 2 or below (Liskin-Gasparro, 1984a & 1984b; Omaggio, 1986; Arnett &

Haglund, 2001). The scale also included a very high proficiency level rarely found in academic settings (Liskin-Gasparro, 1984a & 1984b; Omaggio, 1986; Arnett & Haglund, 2001). Thus, a new proficiency scale was needed for use in academic settings.

For use in academic settings, in 1982, the American Council on the Teaching of Foreign Languages (ACTFL) generated the ACTFL proficiency scale (Liskin-Gasparro, 1987; Fulcher, 2000; Arnett & Haglund, 2001). The ACTFL proficiency scale was designed to define the range at the intermediate proficiency levels often found in academic environments (Arnett & Haglund, 2001). The proficiency scale includes ten proficiency levels in total. The main levels are described as novice, intermediate, advanced and superior. Each main level except superior has subscales labeled low, mid and high (ACTFL, 1999) (see Appendices B & C). The superior level of the ACTFL corresponds to level 3 on the ILR scale.

Since the communicative approach to language testing emphasized the importance of functional language ability in real life, the proficiency movement encouraged language tests to measure the ability to use a language functionally rather than the knowledge about a language. According to the proficiency movement in language testing, communicative language tests have adopted various performance-based test items.

2.3.3 Performance-based Assessment

Performance is defined as "the actual use of language in concrete situations" (Chomsky, 1965, p.4). On the basis of this definition, a performance-based language test requires a test taker to accomplish particular functional language tasks in simulated

situations (Khattri & Sweet, 1996; McNamara, 1996; Brown & Hudson, 1998). Accordingly, oral proficiency is evaluated in terms of how well a test taker performs simulated functional language tasks (McNamara, 1996).

By the same token, the logic behind performance-based assessments is that the test results obtained in simulated real-life situations would predict actual performance in the real world (Spolsky, 1985; Bachman, 1990; Hancock, 1994; Khattri & Sweet, 1996; McNamara, 1996; Fulcher, 2000; Chapelle, 2001). Chapelle (2001) suggested that, performance-based tests should utilize authentic materials obtained in real life situations to maximize their predictability.

Brown and Hudson (1998) reported that a performance-based test provides "more valid measures of students' abilities to respond to real-life language tasks, estimates of students' true language abilities than traditional standardized multiple-choice assessments and predictions of students' future performances in real-life language situations" (p. 662). Because of these benefits, according to McNamara (1996, 1997), in the communicative approach to language testing, performance-based language tests have been widely used to measure oral proficiency of non-native speakers of a target language including students and professionals.

In 1956, for example, the FSI OPI was the first oral proficiency test requiring a test taker to perform real-life language tasks during a structured interview (Sollenberger, 1978; Fulcher, 2000). With the success of the FSI OPI (Fulcher, 2000), there has been an increase in the number of oral proficiency assessments requiring a test taker to perform various functional language tasks including interviews and role plays.

First of all, the ACTFL OPI, published in 1982, is reputed to be "a standardized procedure for the global assessment of functional speaking ability" (Swender, 1999, p. 1). The OPI is conducted by a human interlocutor, for approximately 15 to 30 minutes. The time for its administration varies across test takers because the OPI does not have a fixed question set. In other words, the OPI uses an adaptive algorithm that allows an interlocutor to tailor the process of the interview corresponding to the proficiency level of a test taker. Usually, it takes more time to elicit a ratable response sample from a test taker at a higher proficiency level.

In detail, the ACTFL OPI procedure is composed of four phases: the warm-up, the level checks, the probes and the wind down (Swender, 1999). According to *ACTFL oral proficiency interview tester training manual* published in 1999, during the warm-up, an interlocutor starts with easy conversation that makes a test taker feel comfortable with the testing procedure. Moreover, the warm-up helps an interlocutor to collect background information about a test taker, which is a resource for interview topics.

Through the level check phase, an interlocutor asks a test taker to perform various functional language tasks corresponding to his/her baseline ability. At the probes, challenging tasks are given to find the best of his/her oral proficiency, which is called 'the ceiling.' The characteristics of the functional language tasks are aligned with the ACTFL proficiency descriptions (see Appendices B & C). Further, role-play, an essential element of the OPI, requires a test taker to solve a problem within a simulated real-life situation. Finally, the OPI ends with easy questions during the wind-down. The wind-down is designed to let a test taker know the interview is over. The level of oral proficiency is

assigned according to 'what a test taker can do with a language' and 'what a test taker cannot do with a language.'

Despite the solid reputation of the ACTFL OPI, as mentioned in chapter one, concerns have been raised about the logistic requirements for its administration, potential variances of interlocutors and limited use of authentic materials (Stansfield & Kenyon, 1996; Chalhoub-Deville, 1997). As a solution for the practical problems with the live OPIs, recorded stimuli were administered through an audio recorder.

The tape-mediated oral proficiency instruments are reportedly practical and efficient in terms of time and cost (Kenyon & Malabonga, 2001). For instance, a tester saves time and money by administering a test to a group at a time. In addition, potential confounding effects of interlocutors are controlled by using the same recorded stimuli across all test takers (Stansfield & Kenyon, 1996; Kenyon & Malabonga, 2001).

The structure of the tape-mediated oral proficiency instruments (*i.e.*, the SPEAK test) is similar to that of the ILR/ACTFL OPI. Indeed, the SPEAK test, a taped spoken English test, begins with warm-up questions, level-check and probe questions, and ends with wind-down questions. Unlike the flexible structure of the live ILR/ACTFL OPI, however, the SPEAK test is framed in a fixed linear structure. *Guide to SPEAK* (1982) explains that the SPEAK is composed of seven sections (see Appendix F). Section one asks general background information about a test taker as a warming up. Section two requires a test taker to read a paragraph aloud. Section three asks to complete fragmentary sentences. Section four asks to make a story based on a series of pictures. Section five asks to answer questions about a picture. Section six asks to give an opinion on certain topics. Section seven asks to make an announcement of changes in schedule.

When a test is completed, a recorded response sample is sent to one or two trained raters to assign oral proficiency. The ETS developed oral proficiency scale for the SPEAK. The original version of ETS scale defined oral proficiency in terms of pronunciation, grammar, fluency and comprehensibility with a range from 0 to 3, where 3 is the highest rating. Oral proficiency at level 3 is defined as fluent, close to native speakers.

With a highlight on functional language ability in real-life situations, performance-based language tests have been predominantly used in the communicative approach to L2 instruction. Several studies, however, have addressed issues of authenticity of existing performance-based oral proficiency assessments (Savignon, 1985; Raffaldini, 1988; Di Pietro, 1989; Van Lier, 1989; Lazaraton, 1992, 1996, & 1997; Fulcher, 1996; Yoffe, 1997; Riggenbach, 1998; Salaberry, 2000; Johnson, 2001). Further details are described in the next section.

2.3.4 Authenticity in Language Testing

Under the assumption that performance in a simulated real life situation is a good indicator of language ability in real world (Spolsky, 1985; Bachman, 1990; Hancock, 1994; Khattri & Sweet, 1996; McNamara, 1996; Fulcher, 2000; Chapelle, 2001), authenticity became a critical requirement for communicative language tests (Lynch, 1982; Bachman, 1990; Morrow, 1991; Wood, 1993; Bachman & Palmer, 1996; Douglas, 1997; Lewkowicz, 2000).

In the late 1970s when Widdowson addressed the significance of authenticity in language tests, authentic materials referred to intact real life materials, distinguished from

materials modified for a specific purpose such as teaching or testing (Nunan, 1999; Lewkowicz, 2000; Chalhoub-Deville, 2001b). Further, Bachman and Palmer (1996) refined the concept of authenticity by introducing the notion of interactiveness. Specifically, authenticity refers to "the degree of correspondence of the characteristics of a given language test task to the features of a target language use task" (Bachman & Palmer, 1996, p. 23). They posited that interactiveness means "the extent and type of involvement of the test taker's individual characteristics in accomplishing a test task" (p. 24). Thus, the interaction "between the test taker, the test task and the testing context" (Bachman, 1990, p.322) became an important characteristic of communicative language tests as well as authentic test material.

With an effort to promote authenticity in language testing, communicative language testers have constantly addressed issues on authenticity of performance-based oral proficiency assessments in action. For example, first, with respect to rating scale, the ACTFL OPI and the revised SPEAK test use a holistic rating scale under the assumption that a holistic rating scale, not a discrete-point rating scale, appropriately measures communicative language ability in real life situations. Several studies, however, reported that in reality raters evaluated test takers' oral performance heavily relying on discrete linguistic features such as pronunciation (Savignon, 1985; Byrnes, 1987; Salaberry, 2000).

Second, with respect to testing setting, testing situations of the ACTFL OPI and the SPEAK test were criticized as unauthentic (van Lier, 1989; Lazaraton, 1992; He & Young, 1998; O'Loughlin, 2001). Researchers described that, in the ACTFL OPI, a tester controls the entire testing procedure. For this reason, the OPI seemed the unlikely setting for natural conversation in which participants tend to interact with each other in balance.

On the other hand, in the SPEAK test, a test taker does not have any interaction with a human interlocutor but only with recorded stimuli and printed materials.

Because of the particular test delivery format, such as a live interview and recorded stimuli, face-to-face oral proficiency tests and taped oral proficiency tests are very limited in choice of various real life language use discourse (Byrnes, 1987; Raffaldini, 1988; Shohamy, 1988; Kenyon & Malabonga, 2001). In the same sense, Bachman and Savignon (1986) pointed out the limited generalizability of the live OPI beyond the tested contexts. Therefore, language testers are calling for oral proficiency assessment tools that provide rich real life discourses, and maximize its authenticity and predictability.

2.4 Technology for Communicative Language Testing

2.4.1 The use of computer technology in oral proficiency assessments

In the late twentieth century, the advance of computer technology brought the expectation that computer technology may improve the authenticity and efficiency of existing performance-based oral proficiency tests (Brown, 1997; Gruba & Corbel, 1997; Bachman, 2000; Chapelle, 2001). In fact, the efficiency and accuracy of computer-mediated tests led to the development of various computer-mediated spoken language tests such as the Digital Video Oral Communications Instrument (DVOCI), the Computerized Oral Proficiency Instrument (COPI), and Purdue's Oral English Proficiency Test. Figure 2.1 shows test-delivery formats for oral proficiency assessment in foreign/second languages.

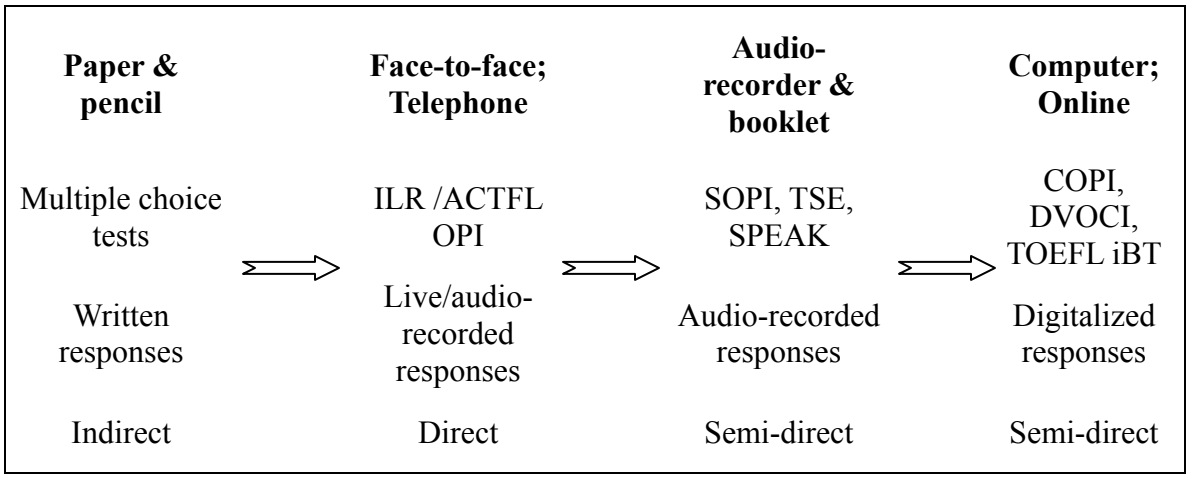


Figure 2.1 Types of test-delivery formats for oral proficiency assessment (modified from Jeong's figure, 2003, p. 34)

With the use of digital multimedia, the DVOCI was developed as a performance-based oral proficiency placement test by the Language Acquisition Resource Center at San Diego State University (LARC, 2003). Digitalized functional language test tasks are saved onto a compact disc in the format of video files. The digital video prompts are delivered on a computer screen. For the administration of the test, a computer lab must be equipped with high speed computers compatible with the multimedia files (Brown, 1997). The test taker's responses are digitally saved after they are recorded via a microphone on the computer. The responses are evaluated against the ACTFL oral proficiency guidelines (LARC, 2003). The use of digital multimedia allows the testing package to utilize various authentic materials (Burstein *et al*., 1996). Furthermore, test designers easily tailor test items to particular populations.

Recently, the LARC at San Diego State University published LARCStar, software that enables instructors to generate language tests online. Using the Internet as a test delivery format, the online test meets practical needs. According to Roever (2001b), test takers complete the test by simply downloading the test items from the server to a computer anywhere the Internet is available and anytime they want. The responses are saved on a server so the raters can access the responses anywhere and any time right after a test taker completes the test (Burstein *et al*., 1996; Chalhoub-Deville, 2001a; Roever, 2001b). As a result, immediate feedback is also available online.

The COPI is "a multi-media, computer-administered adaptation of the tape-mediated simulated oral proficiency interview (SOPI)" (CAL, 2003), developed by the Center for Applied Linguistics (CAL) with the support of US Department of Education (Kenyon & Malabonga, 2001). As a solution to the linear process of a tape-mediated oral proficiency assessment, the COPI incorporates an adaptive algorithm into computerized oral proficiency assessment (Kenyon & Malabonga, 2001). Through the test, according to Kenyon & Malabonga (2001), a human interlocutor and a test taker both face a computer screen. A human interlocutor tailors an appropriate set of test items from a digitalized test item pool to the proficiency level of an individual at the site. The adaptive testing algorithm maximizes the efficiency and accuracy of testing by providing selective test items covering each individual's oral proficiency level (Brown, 1997; Chalhoub-Deville & Deville, 1999). Also, the algorithm allows a test taker to choose topics (Kenyon & Malabonga, 2001) and to respond at his/her own pace (Brown, 1997). The structure of the COPI is as follows:

> welcome, information on the purpose and structure of the COPI, input and correction of personal information, self-assessment of proficiency level, listening to an adequate response to a sample task(s), practice with the same sample task(s), responding to performance tasks (the actual test), feedback about the levels of the tasks that the examinee took, and closing (Kenyon & Malabonga, 2001, p. 65)

The response samples to the COPI are evaluated using the ACTFL rating scale. As of 2005, the COPI is available in several languages including Arabic, Chinese, and Spanish (CAL, 2005).

Purdue's Oral English Proficiency Test (OEPT) was designed to screen international teaching assistants' oral proficiency by the staff of the oral English proficiency program at Purdue University. Like other computerized speaking tests, the OEPT takes advantage of computer technology. The OEPT is composed of 10 categories of test items including "short answer, personal history, read aloud, interpret graph, express an opinion, compare/contrast, offer advice, give information, and summarize casual speech" (Purdue University, 2005). The test items are contextualized in academic environments such as campus life, lectures, and so on. It takes approximately forty-five minutes to complete the test. The response to digital video prompts is recorded on a computer.

As empirical data regarding computerized spoken language tests have been established in the past few years, language testers have reported both advantages and concerns about computerized speaking tests.

2.4.2 Advantages of computer-mediated oral proficiency assessments

The most prominent advantage of use of computer technology for an oral proficiency test seems to be "maximizing the efficiency of test administration" and "improving psychometric qualities of test scores" (Bachman, 1990 p. 336). For example, in terms of test administration, first, the advances in computer hardware and software have solved some of the logistical problems that limited the administration of live OPIs. Specifically, computer-mediated tests can be conducted all over the world through out the year as long as a computer and the Internet are accessible (Educational Testing Service, 1996; Hancock, 1996; Alderson, 2000a; Kenyon & Malabonga, 2001; Norris, 2001). In other words, online testing expands the flexibility of scheduling test administration (Roever, 2001b). Moreover, a computerized speaking test is available for group administration, as well as one-on-one administration. In addition, the ETS (1996) reported that computerized speaking tests controlled variances across human interlocutors that might occur in live OPIs, by standardizing the procedure for testing.

Second, once a response sample is digitalized on the server, a rater can review the sample at his/her convenience in a controlled condition (Kenyon & Malabonga, 2001). This accessibility shortens turnaround time for the test results. Accordingly, immediate test feedback can be available to a test taker (Burstein *et al*., 1996; Chalhoub-Deville, 2001a).

In terms of psychometrics, first, with the use of digital multimedia, computer technology easily simulates various real life (Burstein *et al*., 1996; Hancock, 1996; Warschauer, 1999; Hawisher & Self, 2000; Roever, 2001b). Fulcher (2000) claims that

this capacity of computer technology promotes predictability of oral proficiency test results beyond specific tested contexts.

Second, computer technology can efficiently manage a large amount of test item pools (Alderson, 2000b). The availability of a large database enhances adaptive algorithms for improving the accuracy and efficiency of testing (Brown, 1997, 2004). Also, rich diagnostic information about a test taker' performance is available by "recording multiple aspects of examinee test-taking behavior" (Brown, 1997, p. 47) such as the amount of preparation and response time (ETS, 1996).

In brief, with the use of computer technology, oral proficiency tests have been improved logistically and psychometrically. Examples include efficient administration, immediate feedback, enhanced authenticity, rich diagnostic information, and so on. Nevertheless, concerns about potential influence of computer technology on test takers' performance have been addressed due to the lack of research on this area.

2.4.3 Concerns about computer-mediated oral proficiency assessments

Despite numerous benefits of computer technologies for language testing, Brown (1997) called attention to several concerns identified in two categories: physical and performance considerations.

> Among the physical considerations, 1) computer equipment may not always be available, or in working order. Reliable sources of electricity are not universally available… 2) The amount of material that can be presented on a computer screen is still limited…3) The graphics capabilities of many computers (especially older ones) may be limited, and even those machines that do have graphics may be slow (especially the cheaper machines). Thus, tests involving even basic graphs or animation may not be feasible at the moment in many language teaching situations.

> Among the performance considerations, 1) The presentation of a test on a computer may lead to different results from those that would be obtained [from traditional ways] ...2) Differences in the degree to which students are familiar with using computers or typewriter keyboards may lead to discrepancies in their performances on computer-assisted or computer-adaptive tests (Hicks, 1989; Henning, 1991; Kirsch, Jamieson, Taylor, & Eignor, 1998). 3) Computer anxiety (*i.e.*, the potential debilitating effects of computer anxiety on test performance) is another potential disadvantage (Henning, 1991) (cited in Brown, 1997, p.48).

Among these concerns, performance considerations have also addressed the validity issues of computerized language tests that might be caused by potential test delivery format effects (Kirsch *et al*., 1998; Chalhoub-Deville & Deville, 1999; Kenyon & Malabonga, 2001; Roever, 2001a; Sawaki, 2001; Jeong, 2003). However, few studies have been reported in the professional literature on potential test format effects, particularly for computerized oral proficiency assessment. Thus, further research should be conducted in the area of computerized oral proficiency assessment.

2.5 Validity

2.5.1 Validity in language testing

Test validation is an essential procedure for making "decisions about what constitutes a good test for a particular situation" (Chapelle 1999 p. 254). Hence, constant effort has been made to define the concept of validity in the field of educational and psychological testing (Bachman, 1988, 1990; Brown & Iwashita, 1998; Chapelle, 1999).

The last several decades witnessed the changes in the definition of validity. For example, until the 1980s, the traditional approach defined validity in several subcategories (Messick, 1989). Specifically, according to the AERA/APA/NCME (1985), construct validity was defined as the extent to which a test measured the ability in

question. Content validity focused on the extent to which a test covered the subject content of interest. Criterion validity was the extent to which a test score fit certain criteria for the ability of interest.

With the publication of *Standards for Educational and Psychological Testing*, however, the AERA/APA/NCME (1985) introduced the unitary concept of validity (Chapelle 1999). The new approach explains that validity "refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores" (AERA/APA/NCME, 1985, p. 9). Accordingly, test validation is a procedure to aggregate enough data to support the use of test results (AERA/APA/NCME, 1985). Bachman (1990) supports the claim that the test validation procedure should focus on "the interpretation and use of the information gathered through the testing procedure," not "the test content or even the test scores themselves" (p. 238).

In the same line, Messick (1989) extended validity to the unitary concept with multi-facets. According to Messick (1989), validity is "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 or other modes of assessment" (p. 13). Messick (1989) visualized the multi-facets of unitary validity as outlined in Table 2.1 below.

	TEST INTERPRETATION	TEST USE
EVIDENTIAL BASIS	Construct validity	Construct validity +Relevance/Utility
CONSEQUENTIAL BASIS	Value implications	Social consequences

Table 2.1 Facets of Validity (Messick 1989, p 20)

As can be seen in Table 2.1, Messick emphasized the significance of construct validity by presenting it as an "overarching" feature in his "progressive matrix" (Messick 1989, p.21). He also highlighted the social effect of test use and the interpretation of test results as well.

With the introduction of the unitary but multifaceted notion of validity, test validation became a comprehensive process that is required for "all data yielded by the administration of a test that could serve as legitimate evidence of validity—not only predictive data, but correlational studies generally, factorial studies, studies of differences with respect to groups, situations, tasks, and times, observational studies of change, and studies of experimentally induced change" (Angoff, 1988, p. 30).

Likewise, Messick (1989) described six basic sources of validity evidence as follows:

> We can look at the content of the test in relation to the content of the domain of reference. We can probe the ways in which individuals respond to items or tasks. We can examine relationships among responses to the tasks, items, or parts of the test, that is, the internal structure of test responses. We can survey relationships of the test scores with other measures and background variables, that is, the test's

> external structure. We can investigate differences in these test processes and structures over time, across groups and settings, and in response to experimental interventions—such as instructional or therapeutic treatment and manipulation of content, task requirements, or motivational conditions. Finally, we can trace the social consequences of interpreting and using the test scores in particular ways, scrutinizing not only the intended outcomes but also unintended side effects (p. 16).

This new movement made language testers aware of their social responsibility. Their awareness of the social significance of the interpretation and use of test results has called more attention to test taker characteristics (Messick, 1989; Bachman, 1990). The reason behind the concern is that test takers with an identical ability in question should have an equal probability to gain the same test results (Lord, 1980; Hulin, Drasgow, & Parsons, 1983; Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991; Angoff, 1993; Raju & Ellis, 2002). More importantly, a test result should be accounted for by the ability of interest, not by other factors, particularly by test taker characteristics such as gender (Messick, 1989; Bachman, 1990; Angoff, 1993). Therefore, further research should explore the relationship between test results and test taker characteristics. The present study seeks to add to the professional discourse in this area.

2.5.2 Test taker characteristics

Numerous language tests have been developed without paying sufficient attention to potential effects of test taker characteristics on test results. Further, only a few studies have been reported in the literature with several language tests, mostly tests measuring language ability in reading and listening. For instance, in 1982, Farhady reported a high correlation between ESL students' background and the test score on the

UCLA English as a Second Language Placement Examination (ESLPE). In this study, test taker background referred to sex, university status, academic major, and nationality. The ESLPE measures ability in listening, reading and grammar. Spurling and Ilyin (1985) also found that test taker characteristics (*i.e.*, age, language background, academic major, and educational background) affected performance on a language test estimating reading and listening ability. According to Taylor, Jamieson, Eignor and Kirsch (1998), and Roever (2001a), however, test taker characteristics, particularly computer familiarity, did not significantly affect the test scores of computer-based TOEFL test assessing language ability in listening and reading.

Furthermore, the recent capacity of computer technology extends its use for evaluating language ability in speaking. As oral proficiency has been measured by computerized language tests, research has been conducted on the relationship between test taker characteristics and computerized test delivery formats of oral proficiency assessments. For example, Kenyon and Malabonga (2001) investigated the potential influence of test taker characteristics (*i.e.*, native language) on performance on different kinds of oral proficiency tests such as live OPI, the SOPI, and the COPI. This study found that different language groups did not show statistically significant variance in reactions to different formats of speaking tests. On the other hand, Jeong (2003) reported that there was a positive relationship between the electronic literacy and the English oral proficiency of the examinees who took DVOCI in English as a foreign language context.

These studies revealed mixed results. Furthermore, little reported research in oral proficiency assessment has investigated the interactions between test delivery formats, experience of target language, and computer skill. More importantly, as Messick (1989)

pointed out, the investigation of interactions between background variables would be useful to appropriately interpret and apply test results. Therefore, there is the need for further research such as the present study.

The present study investigates the potential effects of test delivery format and test taker characteristics relevant to the test delivery medium and learning experiences of a language of interest. Of a variety of test taker characteristics, the study focuses on test takers' computer use and English language learning experience (*i.e.*, self-reported years of English language study). Speficically, the present study examines to what extent the number of years of English language study relates to performance on an oral proficiency test; to what extent test delivery format relates to performance on an oral proficiency test; to what extent computer use relates to performance on an oral proficiency test. This study also examines the interactions between these main variables.

2.6 Summary

In the early 1970s, the communicative approach to language teaching moved the focus of language pedagogy from knowledge about a target language to the ability to use a target language appropriately in real life (Fulcher, 2000). Accordingly, language testers have made an effort to develop appropriate assessment tools to measure communicative competence. Various oral proficiency assessments have been developed, such as live oral proficiency interviews and tape-mediated oral proficiency assessments. The limited practicality of these assessments resulted in a call for practical surrogates. As a solution for the problem, computer technology was used to measure oral proficiency. The advance of computer technology improved oral proficiency tests logistically and psychometrically.

As oral proficiency tests have been developed in various test-delivery formats, concern about potential influence of test-delivery format, recently computerized test, was addressed (Mead & Drasgow, 1993; Dunkel, 1999; Grabe, 1999; Alderson, 2000b; Bachman, 2000; Bernhardt, 2000; Schwarz *et al*., 2003). To date, however, few studies have reported on interactions between test delivery formats and test taker characteristics.

Therefore, the present study is intended to investigate to what extent several factors (*i.e.*, self-reported years of English language study, test delivery format, self-reported computer use) account for test scores on a spoken English proficiency test. Also, the interactions between these factors will be examined. The findings of this study are expected to contribute to developing valid testing instruments, to appropriately interpreting test results, and to providing insights for ESL/EFL speaking programs that correspond to the needs of particular populations.

The present chapter had documented the theoretical framework for this study in terms of the communicative approach to language teaching, the communicative approach to language testing, technology for communicative language testing, and validity in language testing. Also, this chapter supports the need for the present study that investigates the relationship between test taker characteristics and test delivery format during oral proficiency testing. The following chapter describes details of methodological procedures for the study including participants, setting, instruments, research methods, and data analysis procedures.

CHAPTER 3

METHODOLOGY

3.1 Introduction

This study investigated the extent to which self-reported English language study experience of test takers (*i.e.*, self-reported years of English language study), test delivery format, and self-reported computer use (three independent variables) related to the test score on a spoken English proficiency test (single dependent variable). For the purposes of the study, a three factor design was used. The number of self-reported years of English language study was self-reported on two levels: less self-reported English study and greater self-reported English study. Test delivery format was represented with two formats: a computerized spoken English test and an audio-taped spoken English test. Computer use was self-reported on two levels: less computer use and more computer use.

This chapter outlines the characteristics of participants and describes the institutional setting where the study was conducted. The instruments used for this study are described. Details of the research hypotheses, research design, data collection and data analysis procedures are explained. Finally, a brief summary recaps the major points of the chapter.

3.2 Participants and setting

The target population of this study was non-native English-speaking international students who were enrolled in graduate programs in a US academic environment. The Ohio State University (OSU) was selected as the sampling site because this research site included a significant population of international students coming from diverse countries. To help international students with English language ability, OSU has provided English as a Second Language (ESL) Programs including American Language Program, ESL Composition Program, and Spoken English Program (OSU, 2005). The American Language Program is an intensive pre-admission ESL program. The ESL Composition Program provides academic writing courses for undergraduates and graduates. The Spoken English Program (SEP) has trained international graduate students to communicate about their academic discipline in English in a US classroom setting. The SEP has developed an assessment system to evaluate an oral proficiency of internationals. For example, the SPEAK test has been used as a placement test. A live face-to-face interview and a mock teaching test have been used to determine whether an international graduate is qualified for teaching in OSU classrooms.

To participate in this study, the volunteers should be international students whose native language was not English. They needed to be enrolled in either full-time or part-time academic study in a graduate degree program at the research site. Approximately two thousand international students were eligible for the present study. They were contacted to ask if they could participate in this present study by email. The first two hundred ten volunteers were recruited from various academic graduate programs. They

were randomly assigned to take either a computerized spoken English test or a taped spoken English test.

The age of most of the participants ranged from 21 years old to 33 years old. In terms of nationality, out of a total 210 participants, 36% of the participants (n=36) came from China, 16% of the participants (n=33) came from Korea, 16% of the participants (n=33) were from India, 20% of the participants (n=41) were from European countries, and 12% of the participants (n=27) were from other countries (*e.g.*, Kenya). In terms of gender, 54% of participants were female (n=113) whereas 46% of participants (n=97) were male.

One hundred thirteen participants self-reported that they have taken a taped SPEAK test once or at most two times before participating in this study. Descriptive statistics and the *F* test were conducted to examine if the test taker's previous experience of taking the test affected test scores in a spoken English test. The previous experience of taking the test was sorted into two levels: never (participants who have never taken the test before) and once or more (participants who have taken the test once or more before).

Test Experience	N	Mean	SD	Min.	Max.	95% Confidence Interval	
						Lower	Upper
Never	97	217.53	53.33	100	300	207.15	227.90
Once or more	113	228.05	50.55	100	300	218.44	237.67
Overall	210	223.19	51.99				

Table 3.1: Summary of Descriptive Statistics of Test Scores by Test Experience.

Table 3.1 presents the data for previous experience of taking the SPEAK test

based on participants' self-reported statements about how many times they took the SPEAK test before participating in this study. Ninety seven participants had never taken the test before while 113 participants had taken the test once or at most two times before. The test scores for both groups spread from 100 to 300 where 300 was the highest. Similar standard deviations were observed. The mean for the group that had never taken the test before was slightly lower than that for the group that had previous testing experience. The confidence intervals for each group were overlapped. An *F* test was run to examine the significance of this mean difference between the two groups. Table 3.2 summarizes the results of the *F* test.

Source	*df*	SS	MS	F	p-value	η_p^2	Power
Test experience	1	5784.51	5784.51	2.15	0.14	0.01	0.31
Error	208	559177.87	2688.36				
Total	209	564962.38					

$p > .05$

Table 3.2: ANOVA Statistics of Test Scores by Test Experience.

As appears in Table 3.2, the result of the *F* test, $F(1, 208) = 2.15$, $p > .05$, suggested that previous experience of taking the test did not produce significant influence on test scores on a spoken English test. Thus, this study focused on only three factors (*i.e.*, self-reported years of English language study, test delivery format, and computer use).

3.3 Instruments

This study utilized two instruments: the SPEAK test and a participant questionnaire. The SPEAK test was used to obtain the dependent variable for this study because the professional literature reported its reasonable reliability and validity (Subkoviak, 1985; Tatsuoka, 1985). Clark and Swinton (1979) reported a high reliability of .91. In addition, the SPEAK was reported to measure oral proficiency, a construct of interest in this study, by showing a high correlation of .73-.77 with the live interview developed by the Foreign Service Institute (Clark and Swinton, 1980).

The SPEAK test was designed to measure the oral proficiency of non-native English speakers (ETS, 1982a). It has been used internationally because of its "flexible and efficient administration and rapid score turnaround" (Sarwark *et al*., 1995, p. 2). It was also used as a placement test at the research site at the time this study was conducted.

For the purposes of this study, the first edition of the SPEAK was administered to the participants. This edition included seven sections, each section required a test taker to perform different language tasks (ETS, 1982a) (see Appendix F). The same edition of the SPEAK test was delivered either on a computer screen or through an audiotape recorder (see Appendix G). Each response sample of the computerized test was electronically saved on a computer hard drive as an audio file, while that of the taped test was recorded on a regular audio-tape.

The participant questionnaire used in this study adapted from the work of the ETS (1982b), Eignor, Taylor, Kirsch, & Jamieson (1998), Hill (1998), Kenyon & Malabonga (2001) and Jeong (2003), (see Appendix I). The participant questionnaire was designed to collect relevant background information about the participants. Specifically,

the questionnaire was composed of four main parts. Part I investigated participants' experience of taking a spoken English test. With focus on computer use, Part II asked participants to self-report the frequency with which they performed various computer tasks. Part III collected in-depth information about participants' English language learning experience both in and out regular classroom settings. Part IV collected participants' comprehensive demographic information.

3.4 Method

3.4.1 Research and statistical hypotheses

The present study hypothesized that self-reported years of English language study related to spoken English proficiency test scores (A main effect); that test delivery format related to spoken English proficiency test scores (B main effect); that self-reported computer use related to spoken English proficiency test scores (C main effect); that self-reported years of English language study and test delivery format jointly affected spoken English proficiency test scores (A x B interaction); that self-reported years of English language study and self-reported computer use jointly affected spoken English proficiency test scores (A x C interaction); that test delivery format and self-reported computer use jointly affected spoken English proficiency test scores (B x C interaction); and self-reported years of English language study, test delivery format, and self-reported computer use jointly affected spoken English proficiency test scores (A x B x C interaction).

The research hypotheses of this study are presented in the following statistical hypotheses:

A main effect:	H_0:	$\overline{T_{le}} = \overline{T_{ge}}$
	H_a:	$\overline{T_{le}} \neq \overline{T_{ge}}$
B main effect:	H_0:	$\overline{T_c} = \overline{T_t}$
	H_a:	$\overline{T_c} \neq \overline{T_t}$
C main effect:	H_0:	$\overline{T_l} = \overline{T_m}$
	H_a:	$\overline{T_l} \neq \overline{T_m}$
A x B interaction:	H_0:	$I_{YxT} = 0$
	H_a:	$I_{YxT} \neq 0$
A x C interaction:	H_0:	$I_{YxC} = 0$
	H_a:	$I_{YxC} \neq 0$
B x C interaction:	H_0:	$I_{TxC} = 0$
	H_a:	$I_{TxC} \neq 0$
A x B x C interaction:	H_0:	$I_{YxTxC} = 0$
	H_a:	$I_{YxTxC} \neq 0$

where $\overline{T_{le}}$ = the mean test score for less self-reported English study group

$\overline{T_{ge}}$ = the mean test score for greater self-reported English study group

$\overline{T_c}$ = the mean test score for computerized test group

$\overline{T_t}$ = the mean test score for typed test group

$\overline{T_l}$ = the mean test score for less computer use group

$\overline{T_m}$ = the mean test score for more computer use group

I_{YxT} = interaction of self-reported years of English language study and test delivery format

I_{YxC} = interaction of self-reported years of English language study and self-reported computer use

I_{TxC} = interaction of test delivery format and self-reported computer use

I_{YxTxC} = interaction of self-reported years of English language study, test delivery format, and self-reported computer use

3.4.2 The statistical model

The statistical model for this study is expressed in the linear model as follows (Keppel, 1991, p. 433):

$$\mathbf{Y_{ijkl} = \mu_T + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \varepsilon_{ijkl}}$$

where
Y_{ijkl} = the score of a single subject
μ_T = the overall mean of the population
α_i, β_j, and γ_k = the average treatment effects at levels a_i, b_j, and c_k, respectively
$(\alpha\beta)_{ij}$, $(\alpha\gamma)_{ik}$, and $(\beta\gamma)_{jk}$ = the average interaction effects at $a_i b_j$, $a_i c_k$, and $b_j c_k$, respectively
$(\alpha\beta\gamma)_{ikj}$ = the three-way interaction effect at cell $a_i b_j c_k$
ε_{ijkl} = experiment error unique to subject "*l*" in group $a_i b_j c_k$

The statistics underlying the model are designed to analyze the variances, namely, the mean squares. The expected values of the mean squares, *E*(*MS*), in this study are expressed as follows (modified Keppel, 1991 p. 432):

$$E(MS_A) = \sigma^2_{error} + \frac{bcn}{a-1}\sum_i \alpha_i^2$$

$$E(MS_B) = \sigma^2_{error} + \frac{acn}{b-1}\sum_j \beta_j^2$$

$$E(MS_C) = \sigma^2_{error} + \frac{abn}{c-1}\sum_k \gamma_k^2$$

$$E(MS_{AxB}) = \sigma^2_{error} + \frac{cn}{(a-1)(b-1)}\sum_{ij}(\alpha\beta)^2_{ij}$$

$$E(MS_{AxC}) = \sigma^2_{error} + \frac{bn}{(a-1)(c-1)}\sum_{ik}(\alpha\gamma)^2_{ik}$$

$$E(MS_{BxC}) = \sigma^2_{error} + \frac{an}{(b-1)(c-1)}\sum_{jk}(\beta\gamma)^2_{jk}$$

$$E(MS_{AxBxC}) = \sigma^2_{error} + \frac{n}{(a-1)(b-1)(c-1)}\sum_{ijk}(\alpha\beta\gamma)^2_{ijk}$$

where $E(MS_A)$, $E(MS_B)$, and $E(MSc)$ = the expected values of the mean squares of factor A, B, and C, respectively

$E(MS_{AxB})$, $E(MS_{AxC})$, and $E(MS_{BxC})$ = the expected values of the mean squares of treatment combination of factors A & B, A & C, and B & C, respectively

$E(MS_{AxBxC})$ = the expected values of the mean squares of treatment combination of factors A, B, and C

a, b, and c = the number of the levels of factor A, B, and C, respectively

n = the number of observations

σ^2_{error} = the population error variance

As can be seen above, the mean square is composed of "treatment effects (main effects or interaction) and error variance" (Keppel, 1991, p. 433). Once the mean squares are obtained, the truthfulness of the null hypothesis can be decided on the basis of the value of the *F* ratio (MS_{effect} / MS_{error}). That is, the null hypothesis is true when the value of *F* is equal to or smaller than 1, whereas the null hypothesis is false when the value of *F* is greater than 1 (Keppel, 1991, p. 45). The results of null hypotheses test will be reported in chapter 4.

3.4.3 Research design

A 2×2×2 mixed factorial design was used to investigate the impact of three factors on test scores on a spoken English test. The three independent variables were self-reported years of English language study, test delivery format, and self-reported computer use.

Self-reported years of English language study was defined as the number of years that a participant has studied and used English both at school and in everyday life. The self-reported years of English language study self-reportedly ranged from 3 years up to 25 years. For the purpose of this study, 12 years, the median, was used as a cut-off point. Participants who have studied English for less than 12 years were categorized as a short-term English language study group. Participants who have studied English for 12 years or longer were categorized as a long-term English language study group.

Test delivery format was represented with two levels: a computerized test and a typed test. The test items of two tests were the same except the fact that a computerized test delivered the test items on a computer screen whereas a typed test used an audiotape recorder and a test booklet (Appendix G).

Computer use was categorized as either less computer use or more computer use. Computer use was defined in terms of the frequency of computer use at various levels. The participant questionnaire asked participants to self-report the frequency with which they performed various computer tasks. The list included routine tasks such as checking email and writing papers, as well as more complex activities such as writing code in HTML or C++ languages. A maximum score of 50 indicated that participants used a computer for all listed tasks on a daily basis. The median score of 31 was used as the cut-

off point in this study. The score of 31 indicated that participants used a computer for all listed tasks monthly or more frequently. Accordingly, an individual, who got more than 31 in total, was categorized as a more computer use group. An individual, who got 31 or less in total, was categorized as a less computer user group.

According to test taker characteristics of interest, participants were categorized into 8 groups. Table 3.3 shows that each group had a slightly different number of participants with a range from 17 to 34. The smallest sample size (n=17) in this study met the minimum sample size (n=17) of each treatment condition required for the reasonable power of the experiment (power = .80 or higher, where α =.05) (Keppel, 1991, p.72).

ABC Cell Sample Size				
	Less Computer Use		More Computer Use	
	Less English Study	Greater English Study	Less English Study	Greater English Study
Computerized test	21	34	33	17
Taped test	23	29	24	29

Table 3.3: Sample Size Matrix by All Three Factors.

The dependent variable was test scores on a spoken English test. The spoken English test used for this study was a performance-based test requiring test takers to demonstrate their oral proficiency. Usually, a performance-based language test utilizes the partial credit model to distinguish a wide range of language proficiency levels. Grown out of Rasch's dichotomous model (correct or incorrect), Andrich's polytomous rating

response model was reformulated as Masters' partial credit model (Toit, 2003). The partial credit model identifies "one or more intermediate levels of performance on an item and awards partial credit for reaching the intermediate levels" (Bateman & Griffin, 2003, p. 6). This model provides richer psychometiric information about the ability in question than the dichotomous model (Zickar, 2002).

This study applied the partial credit model with a range in the scores from 0 to 3 to obtain test scores on the spoken English test. The response samples were rated corresponding to four subcategories: pronunciation, grammar, fluency and comprehensibility, with a 0-3 rating scale (Appendix E). The score of overall comprehensibility, the final score of the test, was obtained using the average comprehensibility score in subcategory. The following is an example to compute overall comprehensibility score (ETS, 1985, p. 15):

> The average comprehensibility score is multiplied by 100 and rounded to the nearest ten unit. Scorers should keep in mind the following overall comprehensibility rounding rules:
> * Scores ending in 5.000 or greater are rounded up to the nearest round ten unit.
> Example: Compute total score = 1.5500
> Total score x 100 = 155.0000
> Rounded O.C. score = 160
> * Scores ending in 4.9999 or less are rounded down to the nearest round ten unit.
> Example: Compute total score = 1.54
> Total score x 100 = 154.00
> Rounded O.C. score = 150

The overall comprehensibility scores ranged from 0 to 300 (Appendix D) where 300 was the highest score. This study used the scores of overall comprehensibility as a dependent variable.

3.4.4 Data collection

The data were collected at the OSU during Autumn Quarter of 2005. Once the present research was approved by the Institutional Review Board of the Office of Responsible Research Practices at the OSU, the recruitment letter (Appendix I) was distributed to current non-native English-speaking international graduates via the campus email account. The letter explained the study, its purpose, procedures, possible benefits and eligibility for participation. The first two hundred ten volunteers were selected and randomly assigned to take either a computerized test or a taped test.

The time for test administration was scheduled for the convenience of each test taker. After signing the consent form (Appendix J), all participants were required to complete a mandatory tutorial before taking a test. The tutorial was designed to familiarize participants with the testing procedures and test equipment. During the tutorial session, each participant had enough time to preview sample test items and to practice how to use testing equipment.

After the tutorial, all the participants took a test individually for consistency and under the same testing conditions. The physical setting, proctoring, and test equipment were all controlled. The test was administered in two rooms typically used for administrating the SPEAK test at The Ohio State University. Each research room was equipped with an audio-tape recorder or an IBM compatible laptop computer for administering a taped test or a computerized test, respectively. It took twenty-five minutes to complete the test. After taking a test, all of the test takers were asked to complete a questionnaire designed to gather information about their background.

The researcher ensured that a participant completely filled out the questionnaire. The whole process took approximately one and a half hours.

The response samples of the spoken English test were individually evaluated by two trained raters. The two raters had teaching experience in the ESL/EFL programs for more than six years. They completed a SPEAK rater training workshop at the same institution. The two raters established a high inter-rater reliability of .99 through calibration process (detailed in chapter 4). Each response sample was scored against the SPEAK scoring key (Appendix D & E) by one rater. Out of a total 210 response samples, a rater scored 158 response samples and the other rater scored 52 response samples.

3.4.5 Data analysis

The data for the research questions included the results of a computerized spoken English test, a taped spoken English test, and replies to a participant questionnaire. For data analysis, this study utilized the computer software SPSS.

Validity and reliability of the instruments (*i.e.*, SPEAK test and a participant questionnaire) were examined. The SPEAK test was developed by the ETS. This test has measured oral proficiency of non-native English speakers world wide since it was first published in the early 1980s. The professional literature reported reasonable reliability and validity of the SPEAK test (Clark & Swinton, 1979, 1980; Subkoviak, 1985; Tatsuoka, 1985). Inter-rater reliability was computed to investigate rater consistency statistically (*i.e.*, Pearson correlation) and practically (detailed in chapter 4).

A participant questionnaire was developed to collect data on independent variables for this study (*i.e.*, self-reported years of English language study, computer use).

Validity and reliability of this participant questionnaire was examined using factor analysis and Cronbach's alpha statistics.

Descriptive statistics were generated for all three factors (self-reported years of English language study, test delivery format, and self-reported computer use) and for their interactions. Descriptive statistics provided central tendency and variability of test scores on an English speaking test. Central tendency indicates the center of the score distribution (Allen & Yen, 1979). This study used mean, a very common measure of central tendency, to describe properties of distribution of test scores on an English speaking test.

Variability is "the extent to which the scores of a group tend to differ or spread above and below a central point in the distribution" (Hopkins, 1998, p. 33). This study used standard deviation, a common measure of variability, to describe spread of test scores on an English speaking test. Standard deviation (σ) expressed in Equation 3.1 is defined as "the square root of variance (σ^2), the mean of the squared deviations of the scores from their mean" (Hopkins, 1998, p. 34):

$$\sigma = \sqrt{\frac{\sum (X - \mu)^2}{N}} \qquad \text{Equation 3.1}$$

Where
X = the test score of a single subject
μ = the overall mean
N = the number of observations

Descriptive statistics also provide a confidence interval of a treatment population mean. A confidence interval is an estimated range of an unknown population parameter based on sample data (Keppel, 1991). A range of confidence interval is defined in a lower limit and an upper limit. The range is computed by the following formula (Hopkins, 1998, p.59):

$$\overline{X} \pm 2 \sigma_{\overline{X}}$$

where $\overline{X}$ = the sample mean
$\sigma_{\overline{X}}$ = the standard error of the mean

The degree of confidence is expressed in (Keppel, 1991, p.99):

Confidence = $100(1 - \alpha)$ %

This study used the 95% confidence interval to investigate if the treatment means were precisely measured. For example, a wide range of a confidence interval suggests that larger sample size is needed to measure treatment population means more precisely. In addition, a confidence interval determines whether the mean differences between treatment groups are significant or not. For example, confidence intervals do not overlap when the mean differences between treatment groups are significant.

The ANOVA assumptions were tested. Corrective actions were unnecessary because all of the assumptions were satisfied. The ANOVA statistics were performed to test the research hypotheses. The ANOVA statistics estimated the *F* value, effect size and

power. The *F* values were computed to test the significance of each main and interaction effect for the present study. The *F* value of 1 or smaller indicates a non-significant treatment effect while a greater *F* value than 1 indicates a significant treatment effect (Keppel, 1991). This study tested null hypotheses against this rule.

Effect size is the proportion of variation accounted for by the treatment manipulation in an experiment (Keppel, 1991). According to Cohen (1988), the effect size of .01, .06 and .15 is small, medium or large, respectively. SPSS provides partial eta squared (η_p^2) as the measure of effect size. Partial Eta squared was computed using the following equation (Becker, 1999):

$$\eta_p^2 = \frac{SS_{effect}}{SS_{effect} + SS_{error}}$$

This study used partial Eta squared, a measure of effect size, to explain the mean differences in test scores on a spoken English test.

Power is "the probability of rejecting a false null hypothesis" (Kirk, 1995. p. 58). Power indicates the degree of sensitivity that the research design draws correct conclusions of investigations. Low power causes a potential type I error (rejecting a true null hypothesis) or a potential type II error (retaining a false null hypothesis). The solutions for this problem are to adopt a stringent α -level or to increase a sample size (Keppel, 1991). The power value of .80 is recommended to reduce the probability of errors (Keppel, 1991; Kirk, 1995). This study made decisions on null hypotheses according to the results of ANOVA statistics, particularly the *F* value, effect size and power.

3.5 Summary

This three-factor research design investigated the relationships between test scores on an oral proficiency test and test taker characteristics. For the purposes of this study, the SPEAK test was administered to 210 volunteers. A questionnaire was used to collect information about test taker characteristics in question. These test takers were non-native English speakers enrolled in a graduate program at a US university.

For data analysis, statistical software SPSS was used. Reliabilities and validities of instruments were examined. Inter-rater reliability was computed. Descriptive statistics showed central tendency and variability of test scores on an English speaking test. The ANOVA assumptions were checked. The *F* test for ANOVA was conducted to test if any significant effects were present.

In response to the main research questions, chapter 4 summarizes the results of the statistics. Chapter five presents findings, implications, limitations, recommendations for further research, and conclusions. After the references, the appendices present the key study instruments and other relevant documents.

CHAPTER 4

DATA ANALYSIS AND DISCUSSION

4.1 Introduction

This study investigated the effects of a test delivery format and test taker characteristics (*i.e.*, self-reported years of English study and self-reported computer use) on test taker performance during oral proficiency assessment. For the purposes of this research, 210 international graduate students whose native language was not English were recruited from a US university. The main data sources included the results of a computerized spoken English test and an ETS-produced audio-taped spoken English test, as well as replies to a participant questionnaire designed for this study.

For data analysis, this study utilized a 2×2×2 mixed factorial research design. The independent variables in this study included self-reported years of English language study, test delivery format, and self-reported computer use. Years of English language study was self-reported in terms of two levels, less years of English language study and greater years of English language study. Test delivery format was also represented in two formats, a computerized spoken English test and an audio-taped version of the English test (*i.e.*, the SPEAK test). Computer use was self-reported in terms of two categories, less computer use and more computer use. The dependent variable in this study was the test scores on the spoken English proficiency test.

This chapter presents the results of the data analysis and related discussion. For example, reasonable reliabilities and validities of the research instruments, including the SPEAK test and a participant questionnaire, are reported. With respect to the three factors of interest, self-reported years of English language study, test delivery format, and self-reported computer use, descriptive statistics were generated to identify central tendencies and variability of test scores on the English speaking test.

First, descriptive statistics by self-reported years of English language study detected a significant mean difference in test scores on the spoken English test between the self-reported less English study group and the self-reported greater English study group. Second, descriptive statistics by test delivery format revealed that the mean difference in test scores on the spoken English test was not significant between the computerized test group and the taped test group. Third, descriptive statistics by self-reported computer use revealed that the mean difference in test scores on the spoken English test was not significant when the less computer use group and the more computer use group were compared.

In order to investigate the presence of main effects and interaction effects, ANOVA statistics were conducted. First, ANOVA assumptions were tested. The test results of the ANOVA assumptions suggested that all of five assumptions were satisfied.

The omnibus F test for ANOVA was run to detect any possible effects by all three independent variables. According to the results of the ANOVA statistics, the three-way interaction among all three independent variables was not significant. With respect to a two-way interaction, the two-way interaction effect between self-reported years of English language study and test delivery format was significant. The two-way interaction

effect between self-reported years of English language study and self-reported computer use was not significant. The two-way interaction effect between test delivery format and self-reported computer use was approaching to non-significant but further analysis was conducted to determine its significance.

Since a significant two-way interaction was detected, a main effect alone could not explain the variance on test scores. As a result, instead of one-way ANOVA statistics, two-way ANOVA and a post hoc comparison were conducted for further analysis. Details of the descriptive statistics and the ANOVA statistics are further explained below. A discussion of the data is presented later in the chapter. Finally, a brief summary recaps the major points of the chapter.

4.2 Overall report of statistics on instruments

4.2.1 Data on the reliability and validity of the test items

The SPEAK test, an institutional edition of the Test of Spoken English, was used for this study because it had published and established reasonable validity and reliability (Subkoviak, 1985; Tatsuoka, 1985). As for validity, a high correlation of .73-.77, with a live interview process developed by the Foreign Service Institute (FSI), was reported (Clark & Swinton, 1980). With support from the US government, the FSI standardized the definition of and scale for language proficiency (Jones, 1975). Since the FSI scale emphasized the language ability needed to function appropriately in real-life situations, the FSI oral proficiency interview (OPI) requires a test taker to perform real-life language tasks during a structured interview. Therefore, a high correlation of .73-.77 with the FSI

OPI supported the claim that the SPEAK test measures a construct of interest in this study, oral proficiency.

With respect to the reliability of the SPEAK test, professional literature reported a high reliability of .91. (Clark & Swinton, 1979). For the present study, two trained raters evaluated the participant response samples from the spoken English test. The two raters had teaching experience in ESL/EFL programs for more than six years. They also completed a SPEAK rater training workshop at the same institution.

Inter-rater reliability between the two raters was tested statistically and practically. For example, sample ratings were compared. Out of a total of 210 response samples, 25 samples were randomly selected to compute the inter-rater reliability between the two raters. The two raters individually evaluated the 25 comparison response samples. A high correlation coefficient of .99 was obtained between the test scores for the twenty-five response samples using the SPSS, computer package.

In addition, rater severity between the two raters was examined in terms of practical agreement. A difference of 20 points in the final test score between raters was deemed to be acceptable (Sarwark, 2006) because a difference of 1 point on a subsection, comprehensibility, made a difference of 20 points in the final score as explained in chapter 3. All of the final scores for the 25 analyzed samples were within this agreement range.

Since an inter-rater reliability of .99 was observed, each of the remaining participant response sample was scored by only one rater using the SPEAK scoring key (Appendix D & E). Of the total 210 response samples, 158 response samples were

evaluated by one rater and 52 were scored by the other rater. The independent ratings are presented in Appendix H.

4.2.2. Data on the reliability of the participant questionnaire

A participant questionnaire was designed for this study to collect information about the number of years of English language study and frequency of computer use. First, the questionnaire asked participants to report how long they had previously studied the English language both at school and in daily life. As shown in Figure 4.1, the self-reported years of English language study ranged from 3 years to 25 years.

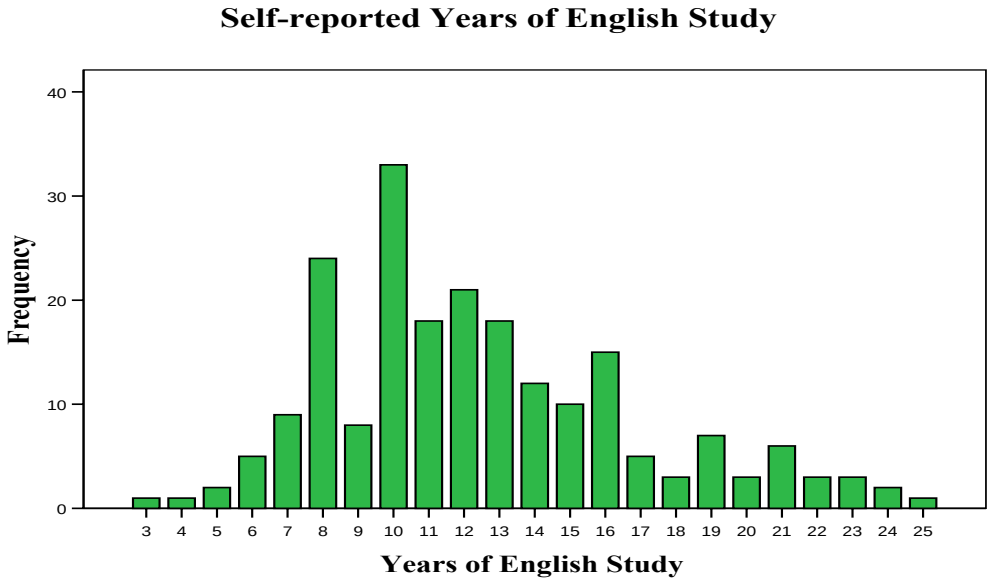


Figure 4.1: The Distribution of Self-reported Years of English Study.

For the purposes of this study, 12 years, the median, was used as a divider to categorize participants. For example, participants who had studied English for less than 12 years were assigned to the self-reported less English language study group. Participants who have studied English for 12 years or more were assigned to the self-reported greater English language study group.

With respect to the data on years of English language study, the data completely relied on self-reported information by participants. It was assumed that the data was reliable because the participants were recruited on a voluntary basis and they were requested to complete the participant questionnaire to the best of their memory and with honesty.

In addition to years of English language study, the questionnaire asked participants to self-report the frequency with which they performed various tasks using a computer. The list included routine tasks such as checking email and writing papers, as well as more complex activities such as writing computer code in HTML or C++ languages (Appendix I).

As appears in Table 4.1, descriptive statistics provide the means and standard deviations to check the central tendencies and variability of the frequency scores on self-reported computer use, respectively. Most participants used a computer both at home and at school on a daily basis, given that means for accessibility were close to 5 and standard deviations were small. Item 6 through item 18 in Table 4.2 listed possible computer tasks. The mean of 5 and a zero standard deviation for item 6 suggested that all the participants daily used a computer for email communication. Item 6 was excluded because this item did not distinguish the frequency of computer use.

	Computer Use	N	Mean	SD
4.	I access a computer at home	210	4.86	0.64
5.	I access a computer at school	210	4.85	0.48
6.	I use a computer to send or receive email	210	5.00	0.00
7.	I use a computer to read or write articles on website bulletin boards	210	4.51	0.98
8.	I use a computer to write academic papers or assignments	210	4.56	0.67
9.	I use a computer to prepare for presentations (*e.g.*, PowerPoint)	210	3.84	1.03
10.	I use a computer to make a database (*e.g.*, Excel or Access)	210	3.73	1.17
11.	I use a computer to listen to music or to watch movies (*e.g.*, DVD)	210	4.30	1.01
12.	I use a computer to participate in online chats	210	3.69	1.43
13.	I use a computer to manage a web page	210	2.50	1.50
14.	I use a computer for advanced webpage authoring using either HTML source code or Java	210	1.81	1.29
15.	I use computer programming languages (*e.g.*, C++, Pearl)	210	2.20	1.56
16.	I use a computer to create multimedia projects using video/audio editing	210	1.79	1.14
17.	I use a computer to create interactive applications or projects similar in complexity to a computerized speaking test	210	1.37	0.91
18.	I use VOIP telephony technologies such as SKYPE	210	1.50	1.08

Table 4.1: Summary of Descriptive Statistics by Self-reported Computer Use.

With respect to reliability of the data on the computer use variable, Cronbach's Alpha statistics showed a reliability of .77 for the initial 15 items. In order to reduce redundant data and effectively analyze the patterns of variance in response to the questionnaire, factor analysis was also conducted. A scree plot displays eigenvalues, "column sum of squared loadings for a factor," representing "the amount of variance accounted for by a factor" (Hair *et al*, 1998, p. 89). Eigenvalues were used to determine how many factors explained the pattern of variance in response to the questionnaire. The scree plot in Figure 4.1 shows that the largest eigenvalue of 3.9 was greater than the

second largest eigenvalue of 1.7. In addition, one clear shift in direction was observed on the scree plot in Figure 4.1. In other words, there was one dominant factor that accounted for the variance in responses to the questionnaire. This one dominant factor, frequency of computer use, explained approximately 26% of the variance in responses on the questionnaire.

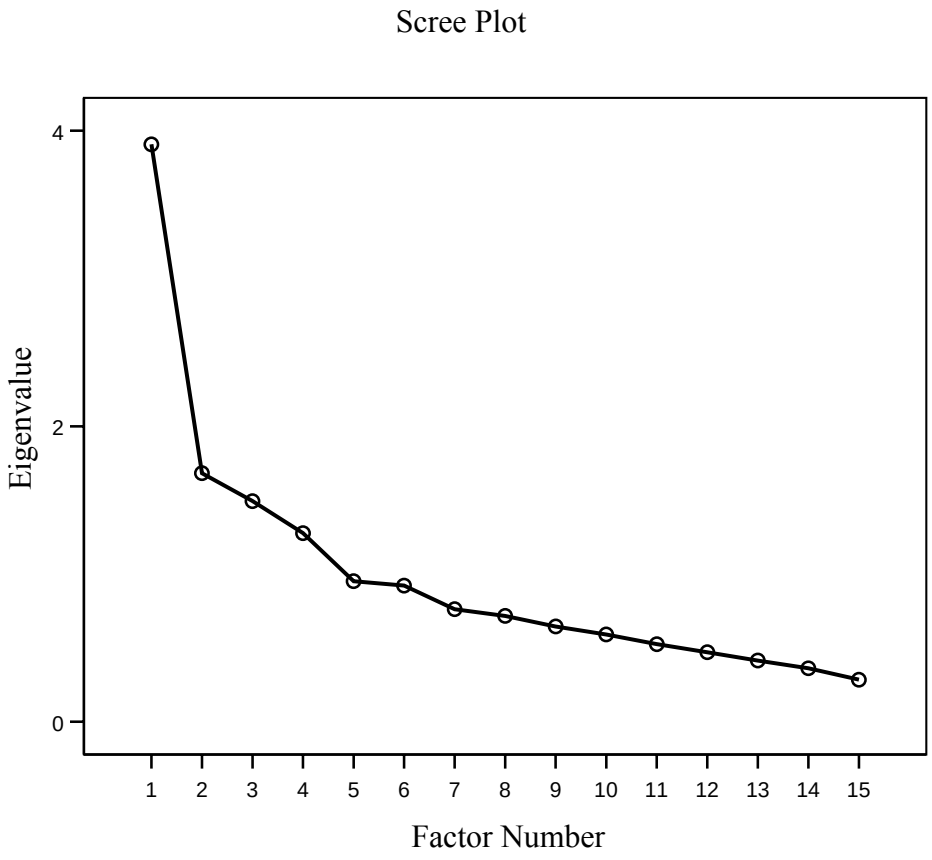


Figure 4.2: Scree Plot of Factor Analysis.

It was important in this study to delve deeper into the data analysis for factors in addition to the dominant finding mentioned above. Principal axis factoring generated a factor matrix to identify data poorly correlated with the main factor, self-reported computer use. Since the questionnaire was unidimensional, only factor loadings under factor 1 were considered. Table 4.2 reveals that factor loadings, particularly the correlation between the items and factor 1, ranged from 0.14 to 0.71. For purposes of this study, a correlation coefficient of .40 was arbitrarily selected as a cutoff to extract items. Of the first 15 items, 10 items (Items 7, 8, 9, 10, 11, 12, 13, 14, 16, and 17) were used to identify the level of self-reported computer use. Cronbach's Alpha statistics computed a reliability of .77 for these 10 items.

Item	Factor1
4	0.23
5	0.14
7	0.43
8	0.40
9	0.62
10	0.54
11	0.51
12	0.43
13	0.71
14	0.64
15	0.34
16	0.57
17	0.43
18	0.24

Extraction Method: Principal Axis Factoring.

Table 4.2: Factor Matrix.

Since a reliability of .77 was found, participants' responses to these 10 items were used to categorize participants as either a less computer use group or a more computer use group. As shown in Figure 4.3, the frequency scores on self-reported computer use ranged from 19 to 50.

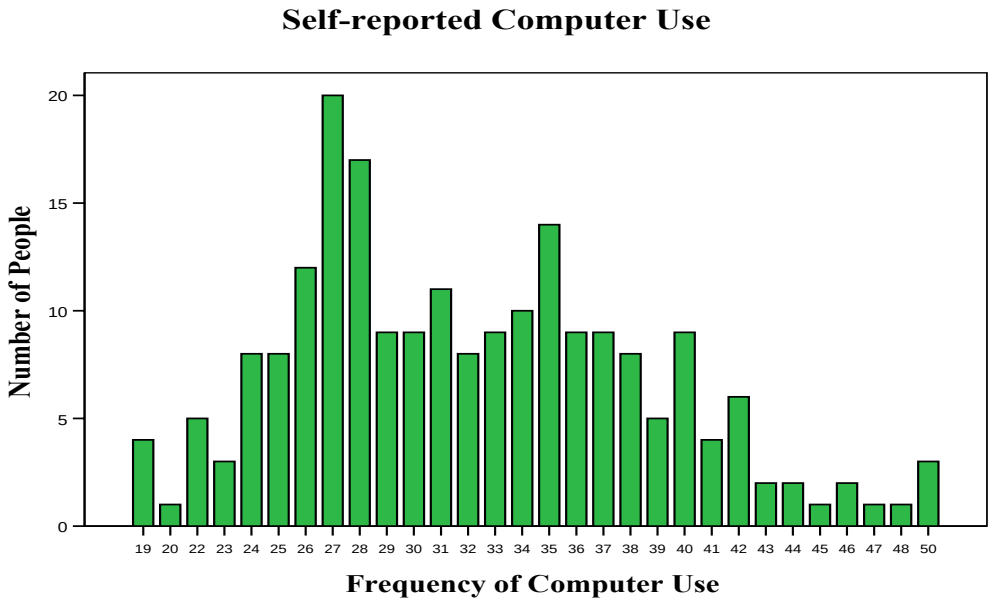


Figure 4.3: The Distribution of Self-reported Computer Use.

This study used the median score of 31 as the cut-off point where 50 was the highest. A maximum score of 50 indicated that participants used a computer for all listed tasks on a daily basis. The score of 31 indicated that participants used a computer for all listed tasks monthly or more frequently. Accordingly, an individual who received 31 or

less as a total score was categorized as a less computer use group. An individual who received a score of more than 31 was categorized as a more computer use group.

4.3 Data on descriptive and ANOVA statistics

4.3.1 Descriptive statistics

First, correlation coefficients were computed for the key variables in the study: self-reported years of English language study, test delivery format, self-reported computer use, and the SPEAK test score.

	SPEAK test score	Computer Use	Years of English Study	Test Delivery Format
SPEAK test score	1	0.01	0.31**	0.09
Computer Use		1	-0.14*	0.03
Years of English Study			1	0.07
Test Delivery Format				1

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

Table 4.3: Correlations for Self-reported Years of English Language Study, Test Delivery Format, Self-reported Computer Use, and the SPEAK Test Score.

As shown in Table 4.3, self-reported years of English study, $r = .31$, $p<.01$, was significantly associated with the SPEAK test score. Given the correlation of $r = \pm .50$ indicates moderate relationship between two variables (Hopkins, 1998), self-reported years of English study was correlated with the test result to a significant degree but with

low correlation. In addition, Table 4.3 showed that self-reported years of English study, r = -0.14, $p < .05$, was significantly associated with self-reported computer use with very low correlation coefficient of -.14. Thus, the data suggested that the three independent variables were not strongly correlated with each other. Self-reported years of English study was significantly associated with the spoken English test score, the dependent variable. Accordingly, descriptive statistics and ANOVA statistics were conducted.

With respect to the dependent variable, descriptive statistics were generated to identify central tendencies and variability of test scores on the English speaking test.

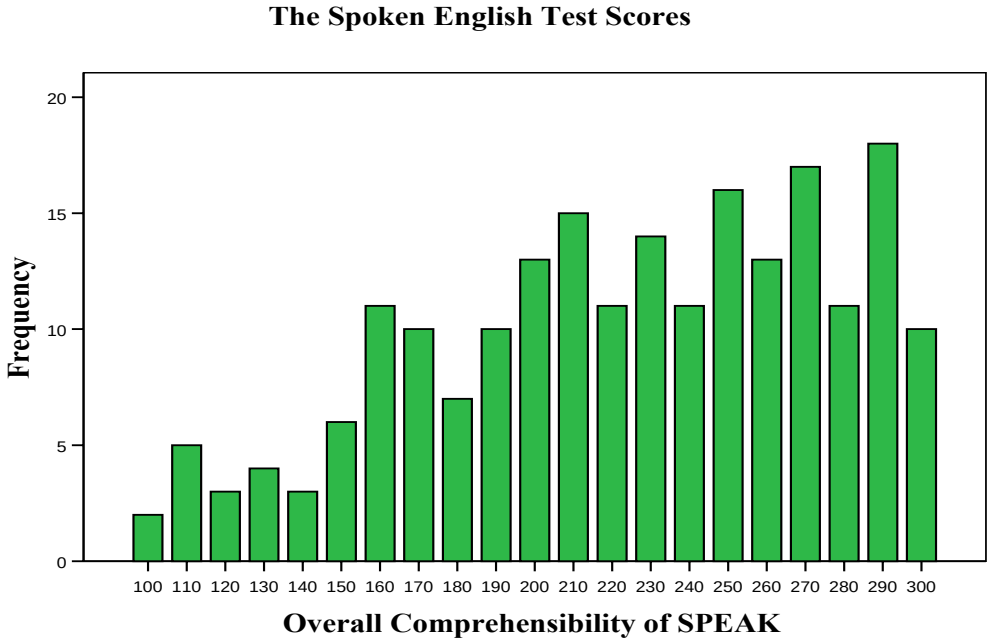


Figure 4.4: The Distribution of the Spoken English Test Scores.

As shown in Figure 4.4, the SPEAK test scores ranged from 100 to 300, where 0 is the lowest and 300 is the highest. The overall mean score for the spoken English test was 223.19 with standard deviation of 51.99. The distribution of the spoken English test scores, in Figure 4.4, seemed to be negatively skewed. Considering, however, the reality that individuals at the low level of language ability were unlikely to take the SPEAK test, the distribution data suggested that the test seemed to distinguish the target level of Spoken English abilities, particularly the 150–300 score range. A test score of 150 indicated the response sample was generally comprehensible with frequent errors whereas a test score of 300 indicated the response sample was completely comprehensible in normal speech.

Subsequently, descriptive statistics were computed for each main factor, two-way interaction, and three-way interaction. The results are presented below based on the key variables in the study: self-reported years of English language study, test delivery format, and self-reported computer use.

4.3.1.1 Descriptive statistics by self-reported years of English language study

Self-reported years of English language study was defined as the number of years that a participant had previously studied English both at school and in daily life. Table 4.4 summarizes the results of the descriptive statistics on test scores in the spoken English test by self-reported years of English language study. One hundred and one test takers were assigned to the self-reported less English study group, whereas 109 test takers were assigned to the self-reported greater English study group. Similar standard deviations suggested that each group was homogeneous. The observed test scores on the

spoken English test for the 210 participants showed a range from 100 to 300. The test scores on the spoken English test for the self-reported less English study group ranged from 100 to 300, while those for the self-reported greater English study group ranged from 110 to 300. The mean score on the speaking test for the greater English study group (238.62) was greater than that of the less English study group (206.53) as well as the overall mean (223.19). The lower limit of the confidence interval for the greater English study group (230.76) was higher than upper limit of the confidence interval for the less English study group (215.69). That is, the data in Table 4.3 suggested that the greater English study group performed better on a spoken English test than the less English study group. This finding, while not unexpected, confirmed that the variable of self-reported years of English language study produced an influence on performance on the spoken English test.

Self-reported Years of English Language Study	N	Mean	SD	Min.	Max.	95% Confidence Interval	
						Lower	Upper
Less English Study	101	206.53	52.26	100	300	196.20	215.69
Greater English Study	109	238.62	46.93	110	300	230.76	249.88
Overall	210	223.19	51.99	100	300		

Table 4.4: Summary of Descriptive Statistics for Test Scores by Self-reported Years of English Language Study.

4.3.1.2 Descriptive statistics by test delivery format

Test delivery format was represented in two versions: a computerized test and a typed test. The same test items were delivered either on a computer screen or via an audiotape recorder with a test booklet. Participants were randomly assigned to take either the computerized speaking test or the audio-taped speaking test.

Table 4.5 summarizes the results of the descriptive statistics on test scores in the spoken English test by test delivery format. A computerized test group and a typed test group had the same sample size of 105. Similar standard deviations for the two groups suggested that each group was homogeneous. The observed test scores on the spoken English test for a total of 210 participants showed a range from 100 to 300. The test scores on the computerized spoken English test ranged from 100 to 300. The test scores on the taped spoken English test ranged from 110 to 300. The mean score on the speaking test for the taped test group (227.81) was slightly larger than that for the computerized test group (218.57) and the overall mean (223.19). The confidence intervals for the computerized test group and that for the taped test group overlapped. Accordingly, the data in Table 4.4 suggested that the mean difference between the two groups was not statistically significant.

Test Delivery Format	N	Mean	SD	Min.	Max.	95% Confidence Interval	
						Lower	Upper
Computerized test	105	218.57	54.96	100	300	209.45	229.12
Taped test	105	227.81	48.67	110	300	217.51	236.45
Overall	210	223.19	51.99	100	300		

Table 4.5: Summary of Descriptive Statistics for Test Scores by Test Delivery Format.

4.3.1.3 Descriptive statistics by self-reported computer use

Computer use was categorized as either less computer use or more computer use according to participants' self-reported frequency of various computer uses. As mentioned earlier, the use of a computer varied from routine tasks (*e.g.*, writing papers) to more complex activities that require advanced levels of computer skill (*e.g.*, writing code in HTML). A maximum frequency score of 50 indicated that participants used a computer for all listed tasks on a daily basis. The median frequency score of 31 was used as a cutoff in this study. The score of 31 indicated that participants used a computer for all listed tasks monthly or more frequently. Accordingly, an individual who received more than 31 in total was categorized as a more computer use group. An individual who received 31 or less in total was categorized as a less computer use group. Table 4.6 summarizes the results of the descriptive statistics on test scores on a spoken English test by computer skill. One hundred seven participants were assigned as a less computer use group, whereas 103 participants were assigned a more computer use group based on their questionnaire responses. Similar standard deviations suggested that the two groups were homogeneous. The observed test scores on the spoken English test for a total of 210 participants spread from 100 to 300. The test scores on the spoken English test both for the less computer use group and for the more computer use group spread in the same range from 100 to 300. The mean score on the speaking test for more computer use group (223.88) was almost equal to that for less computer use group (222.52) and the overall mean score (223.19). In fact, the confidence intervals for the two groups overlapped. In other words, the mean difference in test scores on the speaking test between less computer use group and more computer use group was not statistically significant. Thus,

the data in Table 4.6 indicated that self-reported computer use did not account for the variance in test scores on the speaking test.

Self-reported Computer Use	N	Mean	SD	Min.	Max.	95% Confidence Interval	
						Lower	Upper
Less computer use	107	222.52	51.38	100	300	209.60	228.59
More computer use	103	223.88	52.86	100	300	217.36	236.98
Overall	210	223.19	51.99	100	300		

Table 4.6: Summary of Descriptive Statistics for Test Scores by Self-reported Computer Use.

4.3.1.4 Descriptive statistics by self-reported years of English language study and test delivery format

Table 4.7 shows the results of the descriptive statistics on test scores on the spoken English test by self-reported years of English language study and test delivery format. Four groups were established by years of English language study and test delivery format. Group 1 was composed of 54 participants who reported having studied English for less than 12 years and took a computerized speaking test. Group 2 was composed of 47 participants who reported having studied English for less than 12 years and took a taped speaking test. Group 3 was composed of 51 participants who reported having studied English for 12 years or longer and took a computerized speaking test. Group 4 was composed of 58 participants who reported having studied English for 12 years or longer and took a taped English speaking test.

Similar standard deviations across all groups suggested that the four groups were homogeneous. The mean score on the speaking test for Group 1 was the lowest, while that for Group 3 was the highest. The mean difference (22.40) in test scores on the speaking test between Group 1 (196.11) and Group 2 (218.51) suggested that less English study group performed better on a taped speaking test than on the computerized test. The mean difference (7.01), however, in test scores on the speaking test between Group 3 (242.35) and Group 4 (235.34) suggested that greater English study group performed similarly both on a taped speaking test and on a computerized test.

The confidence interval for Group 1 did not overlap with that for Group 3, nor that for Group 4. This means that the mean difference between Group 1 and Group 3 was statistically significant. In like manner, the mean difference between Group 1 and Group 4 was statistically significant. In other words, years of English language study produced a significant influence on the test scores of the speaking test delivered in different formats.

Source	Source	Group	N	Mean	SD	95% Confidence Interval	
						Lower	Upper
Less English Study	Computerized test	1	54	196.11	51.04	182.93	209.29
	Audio-taped test	2	47	218.51	51.58	204.38	232.64
Greater English Study	Computerized test	3	51	242.35	48.97	228.79	255.91
	Audio-taped test	4	58	235.34	45.24	222.63	248.06

Table 4.7: Summary of Descriptive Statistics for Test Scores by Self-reported Years of English Study and Test Delivery Format.

The mean scores on the speaking test by self-reported years of English language study and test delivery format are presented in Figures 4.5 and 4.6. The two figures provide the same information except the fact that values for *x*-axis and *y*-axis in Figure 4.5 were flipped around in Figure 4.6 in order to check an interaction between years of English language study and test delivery format from every possible angle. As shown in Figures 4.5 and 4.6, the pattern of mean differences reflected by years of English language study was not the same for the two types of test delivery format (*i.e.*, a computerized test and taped test). Particularly, these disordinal mean plots in Figures 4.5 and 4.6 revealed that years of English language study and test delivery format interacted and jointly affected test scores on the speaking test. However, further statistical analysis (*i.e.*, ANOVA statistics) is needed to determine whether or not the interaction is significant. Therefore, these data showed that years of English language study and test delivery format interacted and jointly affected test scores on the speaking test.

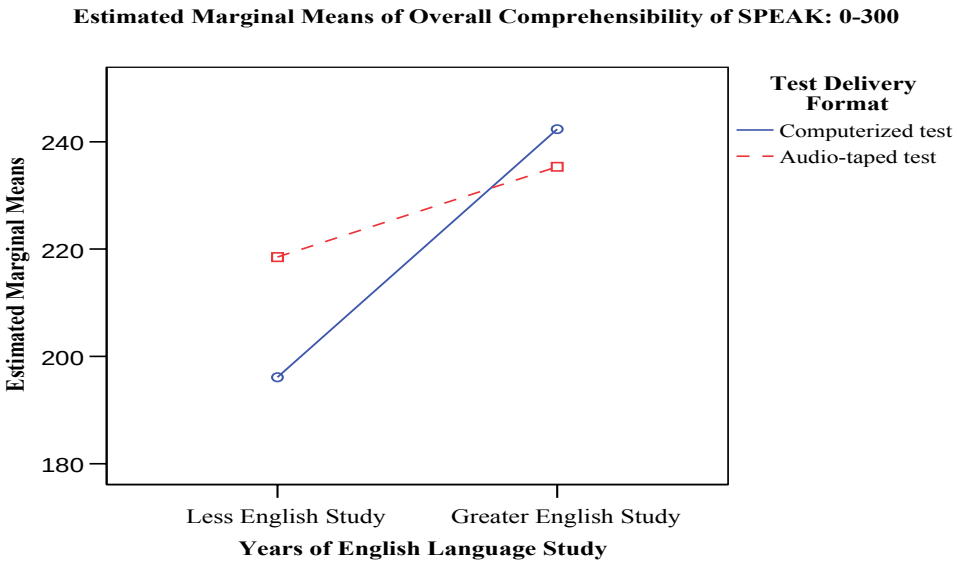


Figure 4.5: Mean Plots by Years of English Language Study with Test Delivery Format.

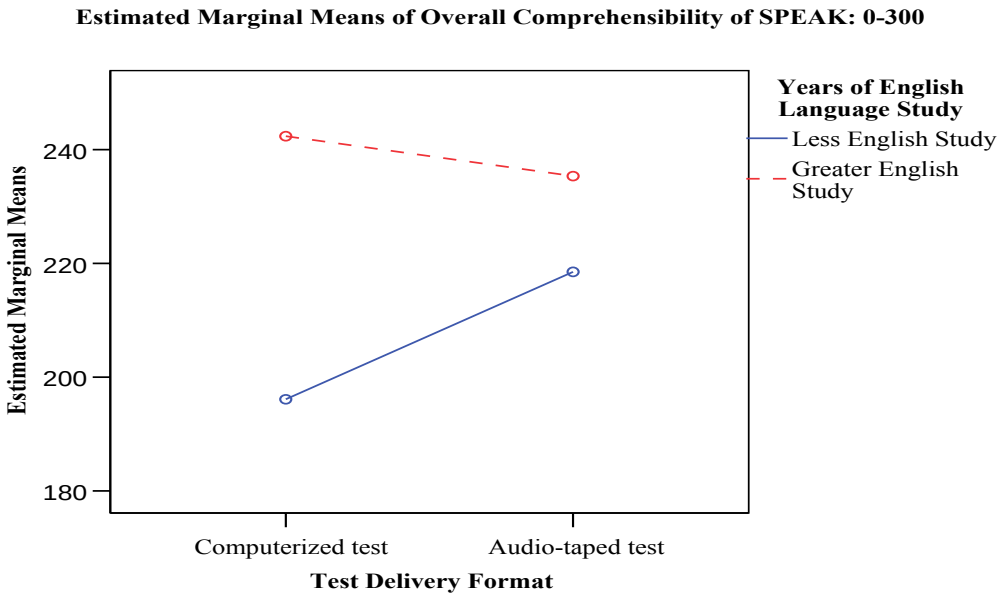


Figure 4.6: Mean Plots by Test Delivery Format with Years of English Language Study.

The following section presents data that further compares two key study variables: self-reported years of English study and self-reported computer use.

4.3.1.5 Descriptive statistics by years of English language study and computer use

Table 4.8 shows the results of the descriptive statistics on the test scores on the spoken English test by self-reported years of English language study and self-reported computer use. Four groups were generated by self-reported years of English language study and self-reported computer use. Forty four participants were assigned to Group 1 that had studied the English language for less than 12 years and self-reported less computer use. Fifty seven participants were assigned to Group 2 that had studied the English language for less than 12 years and self-reported more computer use. Sixty three participants were assigned to Group 3 that had studied the English language for 12 years or longer and self-reported less computer use. And 46 participants were in Group 4 that had studied the English language for 12 years or longer and self-reported more computer use.

Standard deviations in Table 4.8 suggest that the four groups were homogeneous. The mean score (203.18) on the speaking test for Group 1 was the lowest, while that (242.17) for Group 4 was the highest. The mean difference (5.94) in test scores on the speaking test between Group 1 (203.18) and Group 2 (209.12) was minimal. Similarly, the mean difference (5.14) in test scores on the speaking test between Group 3 (236.03) and Group 4 (242.17) was not significant. In addition, the confidence interval for Group 1 overlapped with that for Group 2. The confidence interval for Group 3 overlapped with that for Group 4. These results suggested that self-reported years of English language

study were not significantly associated with self-reported computer use and that computer use did not account for the variance in test scores on the speaking test.

A large mean difference (32.85) in test scores on the speaking test was obtained between Group 1 (203.18) and Group 3 (236.03). Similarly, a large mean difference (33.05) in test scores on the speaking test was computed between Group 2 (209.12) and Group 4 (242.17). In addition, the confidence interval for Group 1 did not overlap with that for Group 3. The confidence interval for Group 2 did not overlap with that for Group 4. The data in Table 4.8 suggested that self-reported years of English language study significantly affected test scores on the speaking test.

Source	Source	Group	N	Mean	SD	95% Confidence Interval	
						Lower	Upper
Less English Study	Less Computer Use	1	44	203.18	52.64	188.41	217.96
	More Computer Use	2	57	209.12	52.28	196.14	222.10
Greater English Study	Less Computer Use	3	63	236.03	46.27	223.68	248.38
	More Computer Use	4	46	242.17	48.12	227.72	256.62

Table 4.8: Summary of Descriptive Statistics for Test Scores by Self-reported Years of English Study and Self-reported Computer Use.

The mean scores on the speaking test by self-reported years of English language study and computer use are presented in Figures 4.7 and 4.8. The two figures provide the

same information except the fact that values for *x*-axis and *y*-axis in Figure 4.7 were flipped around in Figure 4.8 in order to check an interaction between years of English language study and computer skill from every possible angle. As shown in Figures 4.7 and 4.8, the pattern of mean differences obtained for self-reported less English study group and self-reported greater English study group was parallel at the two levels of computer use (*i.e.*, less computer use and more computer use). Therefore, these data revealed that self-reported years of English language study and computer use did not significantly interact or jointly affect test scores on the speaking test.

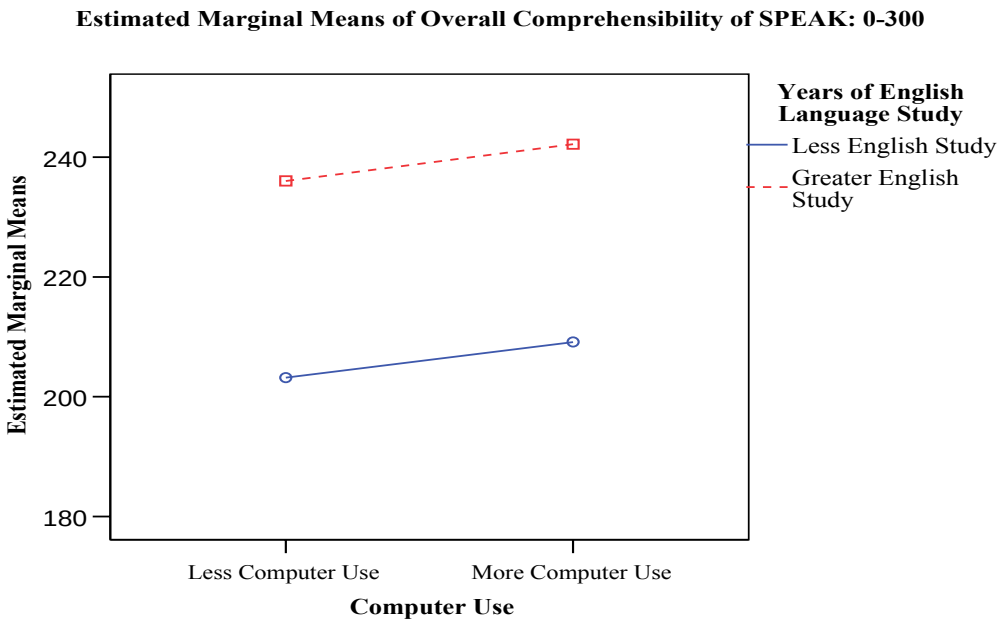


Figure 4.7: Mean Plots by Years of English Language Study with Computer Use.

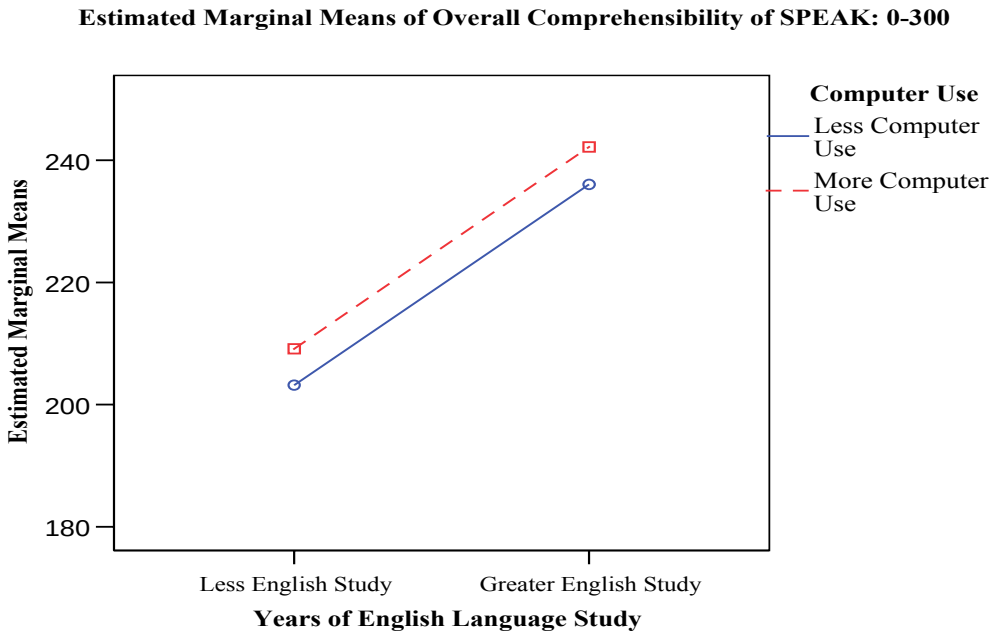


Figure 4.8: Mean Plots by Computer Use with Years of English Language Study.

4.3.1.6 Descriptive statistics by test delivery format and self-reported computer use

Table 4.9 shows the results of the descriptive statistics on test scores on the spoken English test by test delivery format and self-reported computer use. Four groups were generated by test delivery format and computer use. Fifty five participants were assigned to Group 1 that self-reported less computer use and took the computerized speaking test. Fifty participants were in Group 2 that self-reported more computer use and took the computerized speaking test. Fifty two participants were assigned to Group 3 that self-reported less computer use and took the taped speaking test. Fifty three

participants were in Group 4 that self-reported more computer use and took the taped speaking test.

The four groups were homogeneous as evidenced by the similar standard deviations obtained for all groups (see Table 4.9). The mean score (215.09) on the speaking test for Group 1 was the lowest, while that (230.38) for Group 3 was the highest. The mean difference (7.31) in test scores on the speaking test between Group 1 (215.09) and Group 2 (222.40) was minimal. Similarly, the mean difference (5.1) in test scores on the speaking test between Group 3 (230.38) and Group 4 (225.28) was not significant. In addition, the confidence interval for Group 1 overlapped with that for Group 2. The confidence interval for Group 3 overlapped with that for Group 4. Therefore, these results suggested that test delivery format did not significantly influence test scores on the speaking test on the two frequency levels of computer use.

A medium size mean difference (15.59) in test scores on the speaking test was obtained between Group 1 (215.09) and Group 3 (230.38). The confidence interval for Group 1 overlapped with that for Group 3. Unlikely, a minimal mean difference (2.88) in test scores on the speaking test was computed between Group 2 (222.40) and Group 4 (225.28). The confidence interval for Group 2 completely overlapped with that for Group 4. The data in Table 4.9 suggested that the mean differences among four groups generated by self-reported computer use and test delivery format were not significantly different. In other words, self-reported computer use and test delivery format did not jointly affect test results.

Source	Source	Group	N	Mean	SD	95% Confidence Interval	
						Lower	Upper
Computerized test	Less Computer Use	1	55	215.09	53.68	201.25	228.93
	More Computer Use	2	50	222.40	56.62	207.88	236.92
Audio-taped test	Less Computer Use	3	52	230.38	48.10	216.15	244.62
	More Computer Use	4	53	225.28	49.56	211.18	239.38

Table 4.9: Summary of Descriptive Statistics for Test Scores by Test Delivery Format and Computer Use.

The mean scores on the speaking test by test delivery format and computer use are presented in Figures 4.9 and 4.10. The two figures provide the same information except the fact that values for *x*-axis and *y*-axis in Figure 4.9 were flipped around in Figure 4.10 in order to check an interaction between test delivery format and computer use from every possible angle. As shown in Figures 4.9 and 4.10, the disordinal pattern of mean differences was obtained with test delivery format for the two levels of computer skill (*i.e.*, less computer use and more computer use). This means that interaction between test delivery format and computer use was present.

As mentioned earlier, however, confidence intervals between groups were overlapped. Moreover, a difference of 1 point on a subsection made a difference of 20 points in the final score. The difference of 20 points or less in the final test score was considered not significant practically (Sarwark, 2006). In the same sense, the mean difference range of 15.29 from the lowest mean (215.09) to the highest mean (230.38) was not practically significant. Consequently, it seemed that the interaction between test delivery format and computer use was not significant.

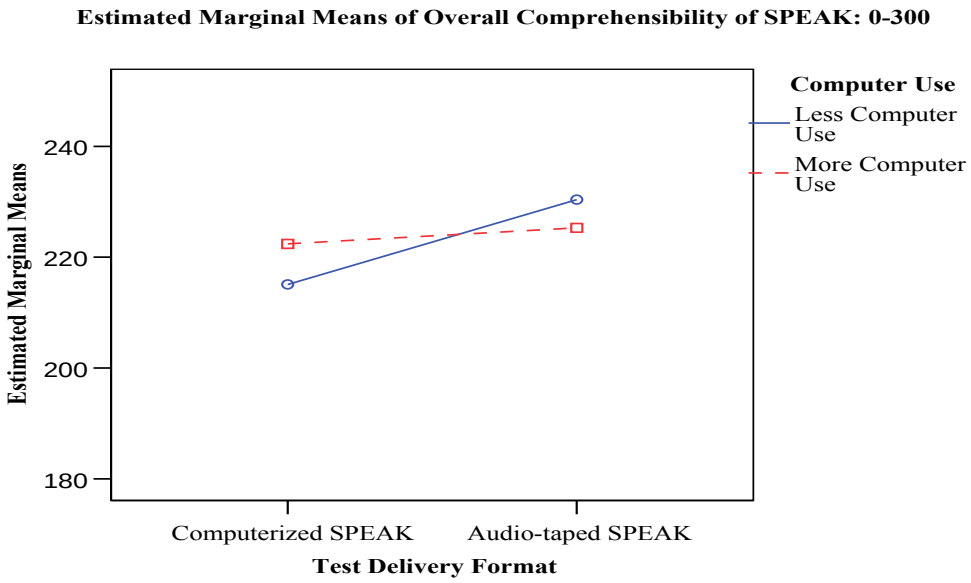


Figure 4.9: Mean Plots by Test Delivery Format with Computer Use.

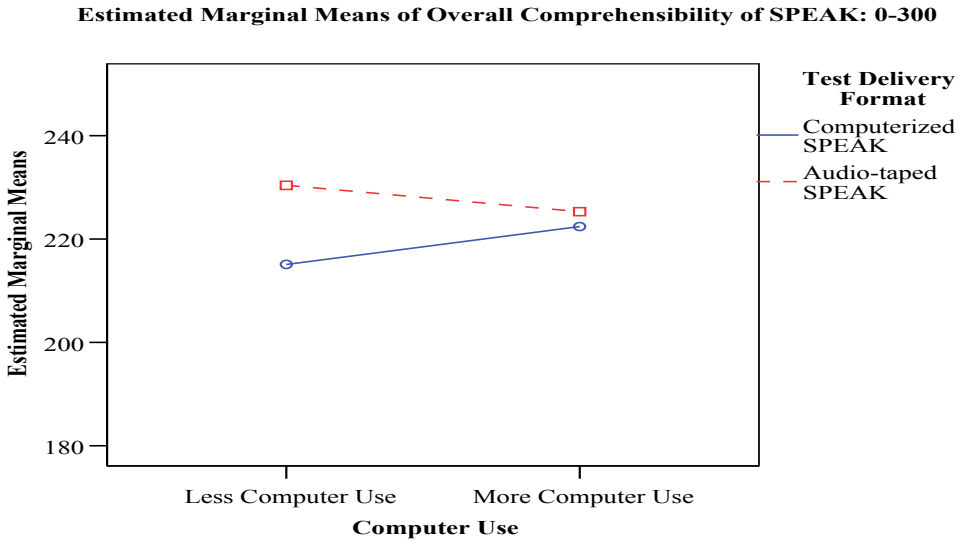


Figure 4.10: Mean Plots by Computer Use with Test Delivery Format.

4.3.1.7 Descriptive statistics by years of English language study, test delivery format and computer use

Table 4.10 and Table 4.11 presented the results of the descriptive statistics for test scores on the spoken English test by self-reported years of English language study, test delivery format, and self-reported computer use. Eight cells were generated according to group membership by self-reported years of English language study, test delivery format, and self-reported computer use.

Each group was arbitrarily assigned a name for convenience of reporting (*e.g.*, Group 1). Specifically, Group 1 refers to 21 participants who self-reported less computer use, studied English language for less than 12 years, and took the computerized speaking test. Group 2 refers to 34 participants who self-reported less computer use, studied English language for 12 years or longer, and took the computerized speaking test. Group 3 refers to 33 participants who self-reported more computer use, studied English language for less than 12 years, and took the computerized speaking test. Group 4 refers to 17 participants who self-reported more computer use, studied English language for 12 years or longer, and took the computerized speaking test. Group 5 refers to 23 participants who self-reported less computer use, studied English language for less than 12 years, and took a taped speaking test. Group 6 refers to 29 participants who self-reported less computer use, studied English language for 12 years or longer, and took a taped speaking test. Group 7 refers to 24 participants who self-reported more computer use, studied English language for less than 12 years, and took a taped speaking test. Group 8 refers to 29 participants who self-reported more computer use, studied English language for 12 years or longer, and took a taped speaking test.

Table 4.10 presents standard deviations for eight groups. Relatively similar standard deviations across the cells were observed. This means that each group was homogeneous. This homogeneity reduces the probability of rejecting a true null hypothesis --no significant treatment effect.

ABC Cell Standard Deviations				
	Less Computer Use		More Computer Use	
	Less English Study	Greater English Study	Less English Study	Greater English Study
Computerized test	40.68	49.78	54.94	46.51
Taped test	54.50	42.64	49.23	48.45

Table 4.10: Standard Deviation Matrix by Three Factors.

Table 4.11 presents the means for each cell. Group 4 achieved the highest mean (254.12) followed by Group 2 (236.47), Group 6 (235.52), Group 8 (235.17), Group 5 (223.91), Group 7 (213.33), Group 3 (206.06), and Group 1 (180.48). A significant mean difference (73.64) was obtained between Group 4, which earned the highest mean (254.12), and Group 1, which earned the lowest mean (180.48).

Not surprisingly, regardless of the other two factors (*i.e.*, test delivery format and computer use), the mean scores on the speaking test for Groups 4, 2, 6 and 8 that self-reported greater years of English language study were greater than those for Groups 5, 7, 3, and 1 that self-reported less years of English language study.

Interestingly, regardless of the level of their computer use, the mean scores on the computerized speaking test for Groups 4 and 2 that self-reported greater years of

English study were greater than those on the taped speaking test for Groups 6 and 8 that self-reported greater years of English study. On the other hand, the mean scores on the computerized speaking test for Groups 3 and 1 that self-reported less years of English study were smaller than those on a taped speaking test for Groups 5 and 7 that self-reported less years of English study.

Consequently, the data in Table 4.11 suggested a significant interaction effect between self-reported years of English language study and test delivery format on test scores during oral proficiency testing.

ABC Cell Means				
	Less Computer Use		More Computer Use	
	Less English Study	Greater English Study	Less English Study	Greater English Study
Computerized test	180.48 (G1)	236.47 (G2)	206.06 (G3)	254.12 (G4)
Taped test	223.91 (G5)	235.52 (G6)	213.33 (G7)	235.17 (G8)

Table 4.11: Mean Matrix by Three Factors.

In addition to mean scores, a box plot is useful to check location and variations in a data set at a glance. As can be seen in Figure 4.11, a box plot identifies the minimum, the lower quartile (25^{th} percentile), the median (50^{th} percentile), the upper quartile (75^{th} percentile), the maximum, and outliers.

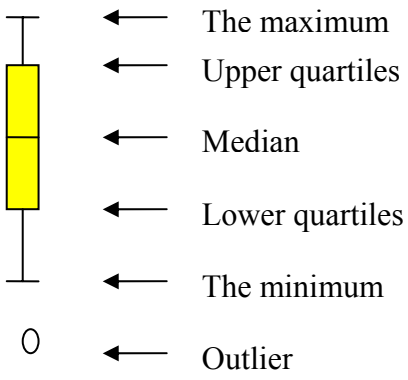


Figure 4.11: The components of box-plot

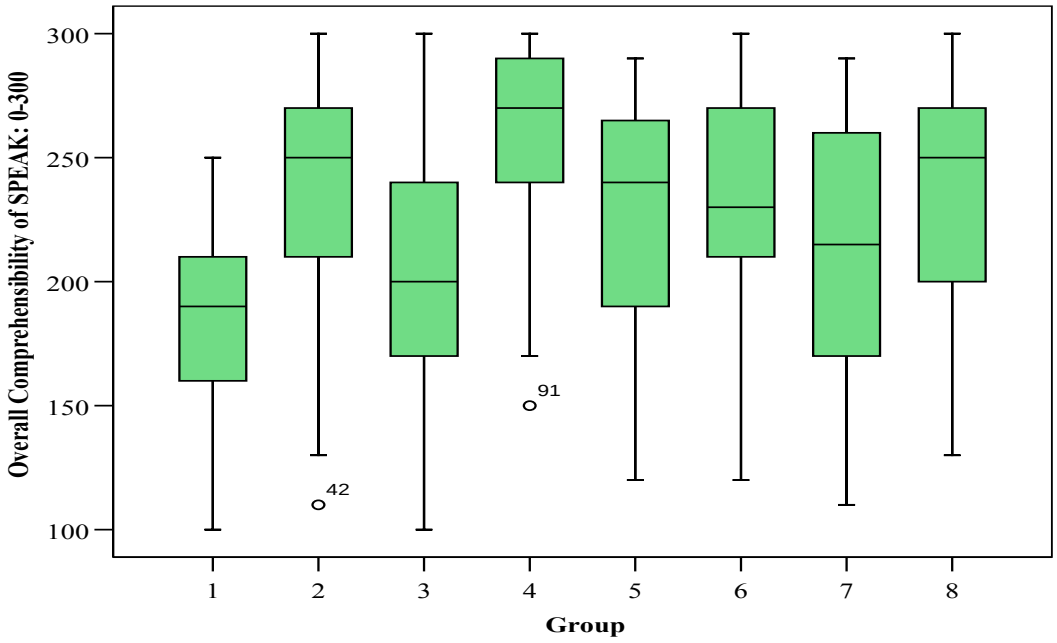
According to the National Institute of Standards and Technology (2003), the minimum is the smallest observation in the data set. The lower quartile is the value at the bottom 25% of the observations in the data set. The median is the value at the center of the data set. The upper quartile is the observation at the 75th percentile. The maximum is the largest observation in the data set. And an outlier is an extreme observation in the data set. As identified as 42 and 91 in Figure 4.12, two minor outliers were observed in this study. Outlier 42 indicated an extreme test score observed in Group 2 whereas Outlier 91 was an extreme test score observed in Group 4. This study did not consider the two outliers not to misrepresent the data set. Moreover, the ANOVA statistics are robust against outliers.

Further, side-by-side box plots are useful to compare several groups simultaneously. Side-by-side box plots were generated in order to examine location of and variations in test scores between the study treatment groups. As shown in Figure 4.12, the values on the horizontal axis indicate each group name (*e.g.*, 1 for Group 1) while the

values on the vertical axis indicate test scores on the spoken English test including the lowest score of 0 and the highest score of 300.

In terms of location of test scores on the speaking test, as shown in Figure 4.12, the middle 50% (*i.e.*, from the lower quartile to the upper quartile) of test scores for the eight groups ranged approximately from 160 to 290 where 300 was the highest. Group 4 (self-reported greater English study, computerized test, more computer use) was at the top of the range, followed by Groups 8, 2, 5, 6, and 7. On the other hand, Group 1 (self-reported less English study, computerized test, less computer use) and Group 3 (self-reported less English study, computerized test, more computer use) were located at the bottom of the score range. The data in Figure 4.12 with respect to the location of test scores on the scale for oral proficiency suggested that significant treatment effects were likely present.

In terms of variations in test scores on the speaking test, test scores on the taped speaking test distributed relatively wider than those on the computerized speaking test. For example, test scores for Group 5 (self-reported less English study, taped test, less computer use) and Group 7 (self-reported less English study, taped test, more computer use) spread wider than those for the other groups. The data with respect to variations in test scores revealed that test takers, especially self-reported less English study group, reacted differently to the different test delivery formats.



where Group 1: Self-reported less English study/Computerized test/Less computer use
Group 2: Self-reported greater English study/Computerized test/Less computer use
Group 3: Self-reported less English study/Computerized test/More computer use
Group 4: Self-reported greater English study/Computerized test/More computer use
Group 5: Self-reported less English study/Taped test/Less computer use
Group 6: Self-reported greater English study/Taped test/Less computer use
Group 7: Self-reported less English study/Taped test/More computer use
Group 8: Self-reported greater English study/Taped test/More computer use

Figure 4.12: Side-by-side Box-plots of Test Scores for All Eight Groups.

In sum, descriptive statistics in this section focused on mean test scores and variance in test scores on the speaking test in order to investigate any significant mean differences between treatment groups. The data on descriptive statistics suggested that self-reported years of English language study and test delivery format influenced

performance on the spoken English test. In the following section, ANOVA statistics were conducted to confirm if the effect was significant.

4.3.2 Assumptions of the analysis of variance

The analysis of variance (ANOVA) is based on five assumptions: random sampling, independent observations, homoscedasticity, normal distribution of population, and interval measure level assumption. Each assumption is discussed below in terms of data collected in the present study.

Random sampling

In ANOVA, sampling should be randomly conducted. That is, each member from the known population should have an equal chance of being selected for the experiment (Keppel, 1991). In addition, the participants should be randomly assigned to an experimental condition. Random sampling and random assignment together reduce systematic experimental bias which allows the researcher to further generalize the findings beyond the particular experimental situation.

Approximately two thousand international students were eligible for the present study at the particular research site. They were contacted via email to ask if they would participate in the present study. A total of 210 volunteers participated in this study, and the rest of students on the contact list either opted not to participate or failed to reply to the email.

Technically, the recruitment of volunteers is not random sampling. In order to reduce any systematic experimental bias that might be caused by volunteer sampling, all

of international students eligible for the present were contacted and asked to participate in this study with an equal chance of being included. In addition, all of the participants were subsequently randomly assigned to take either a computerized version of the SPEAK test or an audio-taped test. All the participants individually took the same test under the same testing conditions. Although volunteer sampling was used for this study, the assumption of randomness was satisfied and random assignment to treatment groups was strictly followed.

Independent observations

Individual observations should not be related with each other in the experiment because a lack of independence confounds variables (Keppel, 1991). Otherwise, the results of the experiment cannot be explained by the treatment of interest. Independence can be achieved "by randomly assigning subjects to conditions and testing them individually" (Keppel, 1991, p. 97).

As mentioned earlier, in this study, all the participants had an equal chance to be selected. They were each randomly sampled and assigned to a particular treatment condition. Each individual participant separately took a test and filled out a questionnaire under controlled conditions. Therefore, the assumption of independence was deemed to have been met.

Homoscedasticity

Within-group variances across groups should be the same. A large discrepancy between the variances increases the probability of Type I error, the rejection of a true null

hypothesis. This assumption was tested twice. First, an *F* test was conducted to examine whether or not this assumption was violated. Standard deviations presented in Table 4.10 were used to compute the *F* value for the present study:

$$F_{max} = \frac{s^2_{largest}}{s^2_{smallest}} = \frac{54.94^2}{40.68^2} = 1.82$$

The F_{max} of 1.82 was much smaller than the F_{max} of 9, a cut-off point representing a severe violation (Keppel, 1991). Thus, within-group variances were assumed to be the same across groups.

In addition, Levene's Test was performed to check for equality of error variances.

F	df_{num}	df_{denom}	Sig.
0.72	7	202	0.65

$p > .05$

Table 4.12: Levene's Test of Equality of Error Variances by Treatment.

The null hypothesis for Leven's test was the error variance (or within-group variance) of the dependent variable and was equal across groups. Table 4.12 reveals that the result of Leven's test was not significant, $F(7, 202) = .72$, $p > .05$. In other word, the error variance of the dependent variable was not different across the groups. Therefore, the results of both the *F* test and Leven's test in the present study suggested that the assumption of homoscedasticity was satisfied.

Normal distribution

The individual observations should distribute normally. The violation of this normal distribution assumption affects the *F* test results and, consequently, the findings would be distorted. The present study utilized a normal Q-Q plot to test if the observed test scores distributed normally. In a normal Q-Q plot, test scores with a range from 0 to 300 on a speaking test were on the horizontal axis whereas the expected normal values were on the vertical axis. The observed test scores were represented by the round. Observations distributed close to a diagonal line representing expected normality, when the data set distributes normally. In other words, substantial deviations from the line indicate that the distribution is not normal.

In this study, eight normal Q-Q plots were generated to test if the observed test scores in each group distributed normally. The normal Q-Q plot for Group 1 was presented in Figures 4.13. The normal Q-Q plot for Group 2 was presented in Figures 4.14. The normal Q-Q plot for Group 3 was presented in Figures 4.15. The normal Q-Q plot for Group 4 was presented in Figures 4.16. The normal Q-Q plot for Group 5 was presented in Figures 4.17. The normal Q-Q plot for Group 6 was presented in Figures 4.18. The normal Q-Q plot for Group 7 was presented in Figures 4.19. The normal Q-Q plot for Group 8 was presented in Figures 4.20.

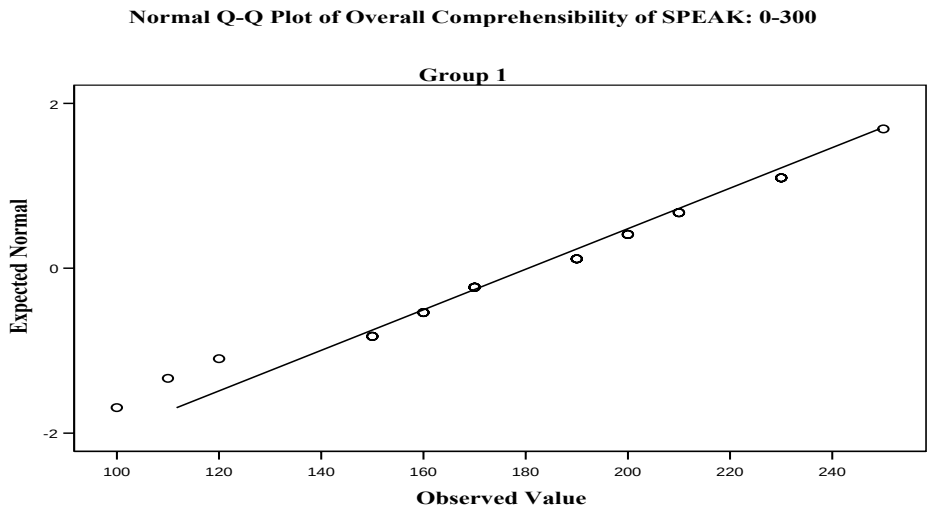


Figure 4.13: Normal Q-Q plot for Group 1.

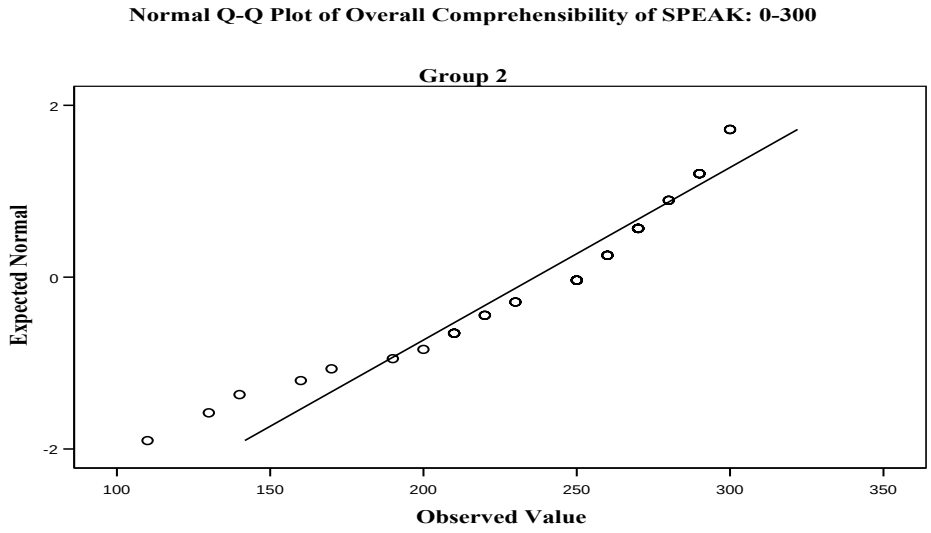


Figure 4.14: Normal Q-Q plot for Group 2.

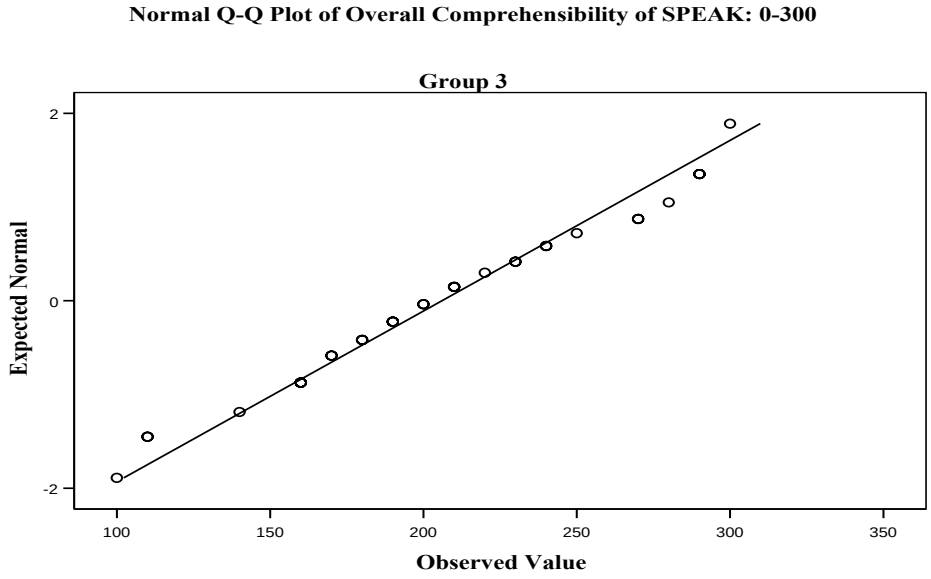


Figure 4.15: Normal Q-Q plot for Group 3.

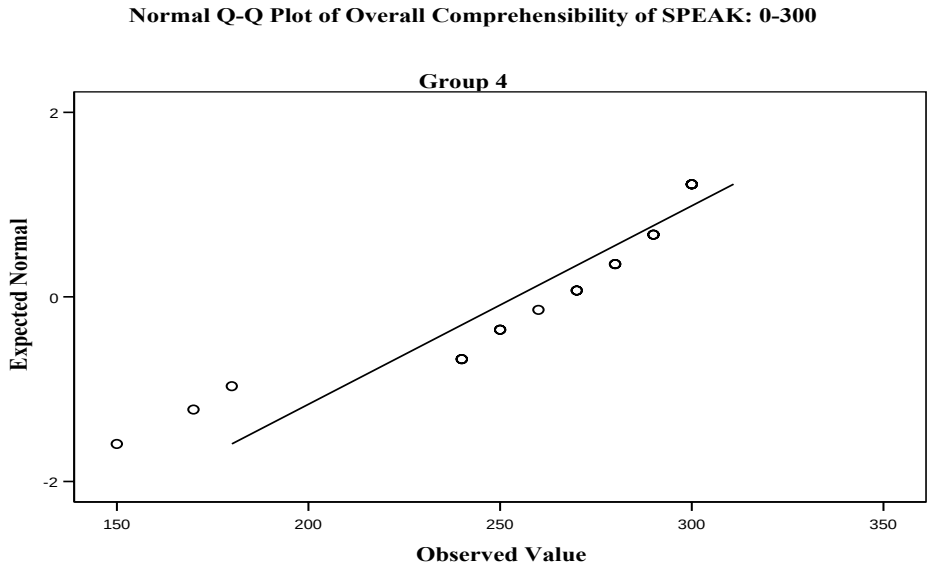


Figure 4.16: Normal Q-Q plot for Group 4.

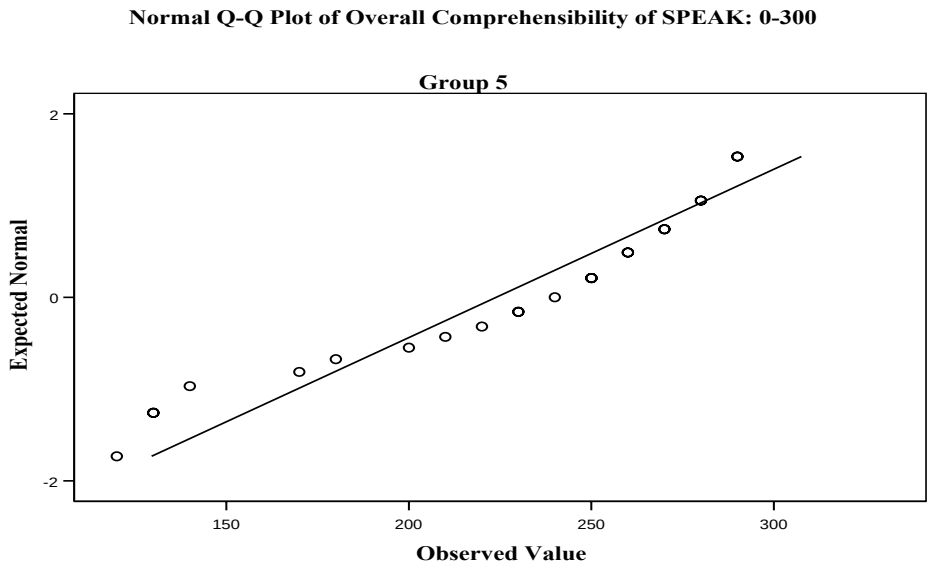


Figure 4.17: Normal Q-Q plot for Group 5.

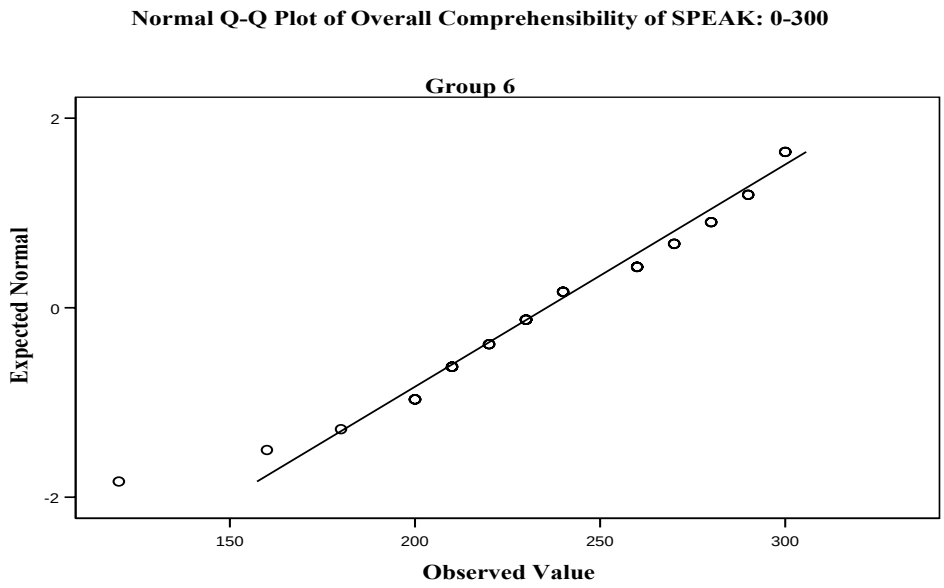


Figure 4.18: Normal Q-Q plot for Group 6.

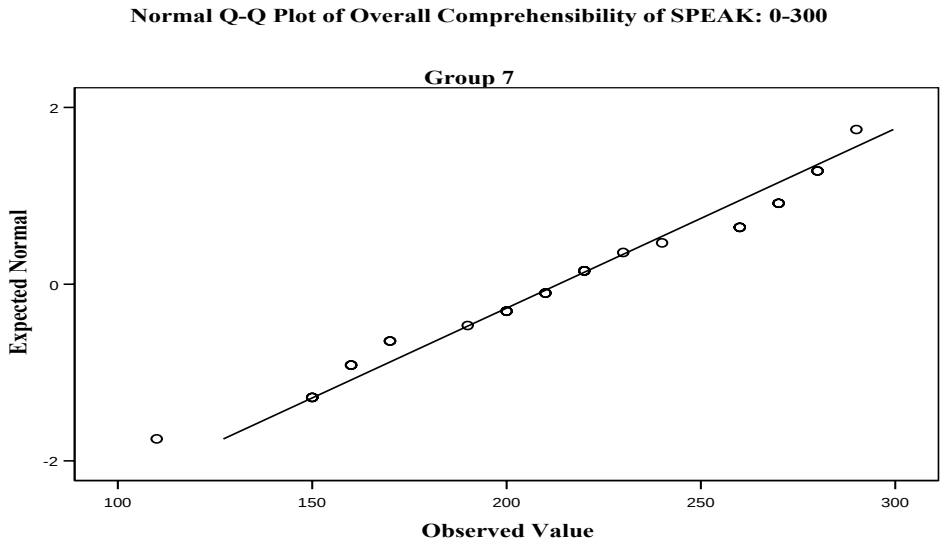


Figure 4.19: Normal Q-Q plot for Group 7.

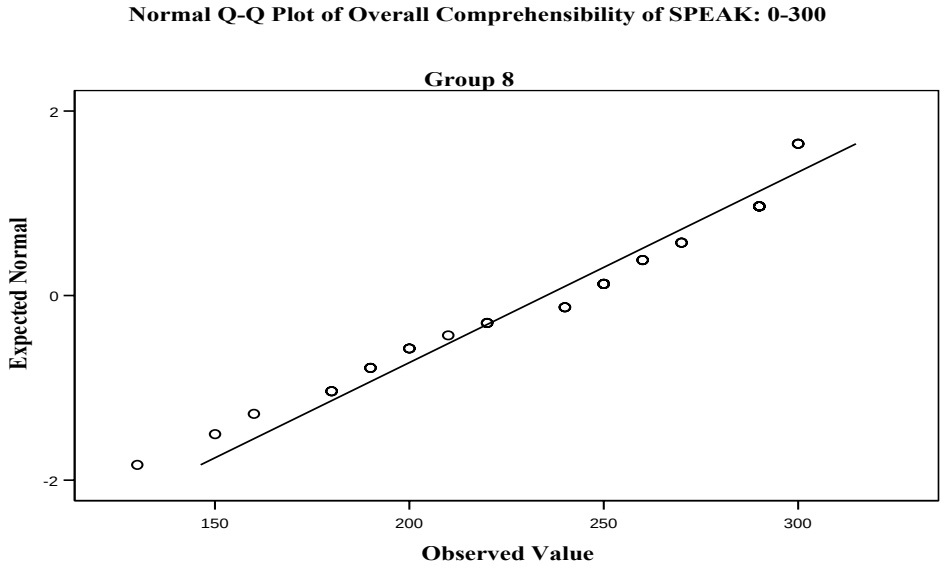


Figure 4.20: Normal Q-Q plot for Group 8.

As can be seen in Figures above, any substantial deviations from the normality line were not observed. Therefore, the normal Q-Q plots suggested that the test scores for each group appeared to distribute close to the normal distribution. In addition, the *F* tests were robust against the violation of this normality assumption. Thus, the data set for this study was deemed to have met the assumption of normality.

Interval measure level

The dependent variable for ANOVA should be either an interval or a ratio measure (Keppel, 1991). The hypothesis test relies on the *F* test. The computation of the *F* test is meaningful only as long as the involved data are at least interval or ratio. This study was attentive to the interval measure level concept. Specifically, this study utilized the scores of a spoken English test, interval data, as a dependent variable. Accordingly, the researcher deemed that this assumption was satisfied.

4.3.3 Omnibus F Test Analysis

A three-way ANOVA was conducted to investigate any possible significant interaction and/or main effects simultaneously. Table 4.13 presents a summary of three-way ANOVA statistics by all factors: self-reported years of English language study, test delivery format, and self-reported computer use.

Source	*df*	SS	MS	F	p-value	η_p^2	Power
Years of English Language Study (Y)	1	59028.09	59028.09	24.646	.000*	.109	.999
Test Delivery Format (T)	1	2964.14	2964.14	1.238	.267	.006	.198
Computer Use (C)	1	3258.95	3258.95	1.361	.245	.007	.213
Y x T	1	15566.70	15566.70	6.500	.012*	.031	.718
Y x C	1	16.48	16.48	.007	.934	.000	.051
T x C	1	9157.57	9157.57	3.824	.052	.019	.495
Y x T x C	1	1031.11	1031.11	.431	.512	.002	.100
Error	202	483793.89	2395.02				
Total	209	564962.38					

* $p < .05$

Table 4.13: Three-way ANOVA Statistics of Test Scores by Self-reported Years of English Study, Test Delivery Format, and Self-reported Computer Use.

First of all, a three-way interaction effect was reviewed to test the null hypothesis: there was no interaction between self-reported years of English language study, test delivery format and self-reported computer use. The null hypothesis was retained according to the F test results, $F (1, 202) = .431$, $p > .05$, presented in Table 4.13. An almost zero effect size was computed with low power of .10 at the interaction. The data in Table 4.13 suggests that all three factors did not jointly produce a significant effect on the test scores of the spoken English test.

Since the three-way interaction was not significant, all possible combinations of two-way interactions were reviewed. First, as shown in Table 4.13, the two-way interaction between self-reported years of English language study and test delivery format, $F (1, 202) = 6.5$, $p < .05$, was significant with a medium effect size (η_p^2) of .03 and a reasonable power of .72. Therefore, self-reported years of English language study and test delivery format significantly affected the speaking test results.

Second, the two-way interaction between self-reported years of English language study and self-reported computer use, $F (1, 202) = .007$, $p > .05$, was not significant with a zero effect size and a low power of .05. The results suggested that self-reported years of English language study and self-reported computer use did not jointly produce any significant effect on the speaking test results.

Finally, as shown in Table 4.13, the p-value of .052 for the two-way interaction between test delivery format and computer skill was approaching non-significance. Considering the results of two-way descriptive statistics presented in section 4.3.1.6, it seemed that this two-way interaction, $F (1, 202) = 3.824$, $p > .05$, was not significant.

In addition to a significant two-way interaction, Table 4.13 presents a significant main effect of self-reported years of English language study, $F (1, 202) = 24.65$, $p < .05$, with a medium effect size (η_p^2) of .11 and a high power of 1. Thus, it seemed that there was a mean difference between the self-reported less English language study group and the self-reported greater English language study group.

However, as a significant two-way interaction was present, it was not meaningful to use the main effect alone to account for the variance in test scores. Instead, an average two-way design was generated to investigate further if the mean difference for self-reported years of English language study groups depended on the different test delivery formats and if the mean difference for test delivery format depended on the different levels of years of English language study.

4.3.4 Further analyses

4.3.4.1 Average two-way design for self-reported years of English language study and test delivery format

Given the significant two-way interaction, it was necessary to test if mean differences for one factor depended on the different levels of the other factor (Keppel, 1991). Specifically, first, the simple effect of years of English language study (Y) was analyzed on the computerized speaking test (t_1) and on the taped speaking test (t_2).

Source	*df*	SS	MS	F	p-value
Y at t_1	1	56084.734	56084.734	23.245	0.000*
S/Y at t_1 (error)	206	497039.828	2412.815		
Y at t_2	1	7357.342	7357.342	3.049	0.082
S/Y at t_2 (error)	206	497039.828	2412.815		

* $p < .05$
Dependent variable: test score on the spoken English test

Table 4.14: Analysis of Simple Effect for Years of English Language Study at Test Delivery Format.

Analysis was conducted on the different test delivery formats by self-reported years of English language study in order to test if the mean difference for years of English language study depended on the different test delivery formats. The results in Table 4.14 indicated that the simple effect of self-reported years of English language study on the computerized speaking test, $F(1, 206) = 23.24$, $p<.05$, was significant. This

finding suggested that the mean difference between less English study group and greater English study group on the computerized spoken English test was significantly different.

On the other hand, the results in Table 4.14 indicated that the simple effect of self-reported years of English language study at the taped speaking test, $F(1.206) = 3.049$, $p>.05$, was not significant. This finding suggested that the mean difference between less English study group and greater English study group on the taped spoken English was not significantly different.

Therefore, the mean difference for self-reported years of English language study depended on different test delivery formats. Specifically, self-reported years of English language study produced a significant effect on the test results of the computerized test, but not those of the taped test.

Second, the simple effect of test delivery format (T) was analyzed at less English language study (y_1) and at greater English language study (y_2).

Source	*df*	SS	MS	F	p-value
T at y_1	1	12608.051	12608.051	5.225	0.023*
S/T at y_1 (error)	206	497039.828	2412.815		
T at y_2	1	1332.827	1332.827	0.552	0.458
S/T at y_2 (error)	206	497039.828	2412.815		

* $p < .05$

Dependent variable: test score on the spoken English test

Table 4.15: Analysis of Simple Effect for Test Delivery Format at Self-reported Years of English Language Study.

Analysis was conducted on self-reported years of English language study by test delivery format in order to test if the mean difference for the test delivery format depended on the years of English language study. The results in Table 4.15 indicated that the simple effect of test delivery format at less English language study, $F(1, 206) = 5.225$, $p<.05$, was significant. This finding suggested the mean for less English study group that took the computerized test was significantly different from the mean for the group that took the taped test.

On the other hand, with respect to greater English language study group, as shown in Table 4.15, the simple effect of the test delivery format at the level of greater English language study, $F(1, 206) = 0.552$, $p>.05$, was not significant. The results of this analysis suggested that the mean for greater English study group that took the computerized test was not significantly different from the mean for the group that took the taped test.

Therefore, the mean difference for the test delivery format depended on the different levels of English study. Specifically, less English study group significantly interacted with test delivery format, while greater English study group did not significantly interact with test delivery format.

With respect to this average two-way design for self-reported years of English language study and test delivery format, post hoc comparisons were conducted to investigate further mean differences among the four groups presented in Table 4.7.

4.3.4.2 Post hoc comparisons

Unlike planned comparisons, post hoc comparisons "extract the maximum amount of information from any given study" (Keppel, 1991, p. 171). Accordingly, the Scheffe post hoc test, the most conservative test with "the family-wise rate at a particular value regardless of the number of comparisons actually conducted" (Keppel, 1991, p. 172), was undertaken in order to analyze further the pattern of interaction between self-reported years of English language study and test delivery format.

The Scheffe statistic exhaustively generated six possible pairs for pair-wise comparisons, presented in Table 4.16, according to group membership (*i.e.*, self-reported years of English language study and test delivery format). The mean scores for two groups in each pair were compared to identify if mean differences in test scores were significant.

Comparison	Mean Difference	p-value
Less English Study on Computerized Test vs. Less English Study on Taped Test	-22.40	.16
Less English Study on Computerized Test vs. Greater English Study on Computerized Test	-46.24	.00*
Less English Study on Computerized Test vs. Greater English Study on Taped Test	-39.23	.00*
Less English Study on Taped Test vs. Greater English Study on Computerized Test	-23.84	.13
Less English Study on Taped Test vs. Greater English Study on Taped Test	-16.83	.39
Greater English Study on Computerized Test vs. Greater English Study on Taped Test	7.01	.91

* significant at family wise alpha = .05
Dependent variable: test score on the spoken English test

Table 4.16: Scheffe Post Hoc Analysis for Self-reported Years of English Language Study and Test Delivery Format.

Table 4.16 presents the results of the Scheffe post hoc analysis. Two significant mean differences were obtained: 1) between less English study group and greater English study group that took the computerized test; and 2) between less English study group that took a computerized test and greater English study group that took the taped test. As shown in Table 4.16, greater English study group performed significantly better on a computerized speaking test than less English study group. Besides, greater English study group that took the taped speaking test performed significantly better than less English study group that took the computerized test. In other words, test delivery format had a significant impact on the test scores for those that had less years of English language study but not for those that had greater years of English language study.

4.4 Discussion

This chapter presented the results of the data analysis and related discussion. The professional literature indicated that the SPEAK test had reasonable reliability and validity. A high inter-rater reliability between two raters of .99 was obtained. A reasonable reliability on the questionnaire was observed (Cronbach's Alpha coefficient = .77).

Descriptive statistics were computed for each main factor (self-reported years of English language study, test delivery format, and self-reported computer use), two-way interactions, and three-way interaction. The result of descriptive statistics by self-reported years of English language study revealed that the mean difference (32.09) in test score on the speaking test between greater English study group (mean = 238.62) and less English stud group (mean = 206.53) was significant. In other words, self-reported years of

English language study seemed to produce a significant effect on test scores on the speaking test.

Descriptive statistics by test delivery format showed that the mean of the taped test group (227.81) was similar to that of the computerized test group (218.57). This means that test delivery format did not account for the variance in test scores on the speaking test. In addition, descriptive statistics by computer use suggested that self-reported computer use did not influence test scores on the speaking test, given that the mean score for the more computer use group (223.88) was almost equal to that for the less computer use group (222.52).

Descriptive statistics by self-reported years of English language study and test delivery format showed that the highest mean (242.35) was observed for greater English study group on the computerized speaking test, followed by 235.34 for greater English study group on the taped speaking test; 218.51 for less English study group on the taped speaking test; and 196.11 for less English study group on the computerized test. The mean plots for the two factors, presented in Figures 4.5 and 4.6, revealed that self-reported years of English language study and test delivery format significantly interacted and jointly affected test scores on the speaking test. These results indicated, perhaps not surprisingly, that self-reported years of English study is a key factor in speaking test results no matter what test format is used. However, further study of this topic is warranted. Still, it is interesting to note that length of English language study is an important factor.

Descriptive statistics by self-reported years of English language study and self-reported computer use revealed that the highest mean (242.17) was observed for more

computer use group that self-reported greater English study, followed by 236.03 for less computer use group that self-reported greater English study; 209.12 for more computer use group that self-reported less English study; and 203.18 for less computer use group that self-reported less English study. However, the mean plots for the two factors, presented in Figures 4.7 and 4.8, confirmed that self-reported years of English language study and self-reported computer use did not significantly interact. Again, it can be seen from these data that self-reported years of English language study and self-reported computer use did not jointly produce any significant influence on performance on the spoken English test.

Descriptive statistics by test delivery format and self-reported computer use showed that the highest mean (230.38) was observed for less computer use group that took the taped speaking test, followed by 225.28 for more computer use group that took a taped speaking test; 222.40 for more computer use group that took a computerized speaking test; and 215.09 for less computer use group who took a computerized speaking test. The interaction between the two factors was plotted in Figures 4.9 and 4.10. The interaction, however, was considered not significant, given that the mean difference (15.29) between the highest and the lowest mean scores was smaller than 20, a practical cutoff for the significance of difference in test score on the speaking test. The results suggested that test takers' computer use did not affect test results on different test delivery formats (*i.e.*, computerized speaking test and audio-taped speaking test).

Descriptive statistics for all three factors found that the mean (254.12) for Group 4 (Greater English study/Computerized test/ More computer use) was the highest, whereas the mean (180.48) for Group 1 (Less English study /Computerized test/Less

computer use) was the lowest. Substantial mean difference (73.64) between the highest mean and the lowest mean implied the presence of a treatment effect. Side-by-side box plots were generated in order to examine location of and variation in test scores between treatment groups. The results of the descriptive statistics and box plots suggested that significant treatment effect(s) were present.

Before ANOVA statistics were conducted, the five assumptions of the ANOVA were checked and they all were satisfied. As there were three independent variables presented for this study, three-way ANOVA statistics were conducted to investigate any significant effects. The results of the three-way ANOVA determined that there was no significant interaction between self-reported years of English language study, test delivery format and self-reported computer use ($F(1, 202) = .431$, $p > .05$). The data, therefore, suggested that the three factors did not jointly affect test scores on the speaking test to a significant degree.

Since a significant two-way interaction effect between self-reported years of English language study and test delivery format was observed, it was not meaningful to utilize the main effect for self-reported years of English language study alone to explain the variance in test results. Instead, further analyses were conducted to investigate the simple effects of self-reported years of English language study on the computerized speaking test and on the taped speaking test. The results of this analysis, in Table 4.14, revealed that self-reported years of English language study produced a significant effect on the test results of the computerized test, but not those of the taped test.

In addition, the simple effects of test delivery format were analyzed at the level of less English language study and at the level of greater English language study. The

results of this analysis, in Table 4.14, suggested that less English study group significantly interacted with test delivery format, while greater English study group did not significantly interact with test delivery format.

Further, a Scheffe post hoc analysis was conducted to investigate how self-reported years of English language study and test delivery format interacted. The results of the Scheffe post hoc analysis revealed that two significant mean differences were obtained: 1) between less English study group and greater English study group that took the computerized test; and 2) between less English study group that took a computerized test and greater English study group that took the taped test. It seemed that test delivery format (*i.e.*, computerized test) affected the test scores for those that self-reported less English language study but not for those that self-reported greater English language study.

Interestingly, less English study group performed better, but not statistically significantly, on the taped test. However, greater English study group performed better, but not statistically significantly, on the computerized test. It seemed that there was no exclusively good test delivery format to fit every testing situation. In other words, various test delivery formats should be developed and be used corresponding to diverse test taker characteristics and specific testing purposes. Thus, further research is necessary to link test delivery format and test taker characteristics.

This chapter presented the data and related discussion of this study. The following chapter will present findings, answers to the research questions, implications, limitations, recommendations for further research, and conclusions.

CHAPTER 5

FINDINGS, IMPLICATIONS, LIMITATIONS, RECOMMENDATIONS FOR FURTHER RESEARCH, AND CONCLUSIONS

5.1 Introduction

In the early 1970s, the communicative approach introduced communicative competence to the field of language education. Communicative competence was defined as the language ability to carry out various functional language tasks appropriately in real life situations (Hymes, 1972; Canale & Swain, 1980; Bachman, 1990). With an emphasis on communicative language ability, the communicative approach moved the focus from knowledge about language to the use of language in real life, including simulated use.

Accordingly, the communicative approach recommended that a language test should include measuring the functional language ability needed in real life. In other words, a communicative language test should use authentic materials to assess the predictability of test results for daily life situations (Bachman & Palmer, 1996). At the same time, a communicative language test is expected to maximize the interaction among the test taker, test task, and testing context (Bachman, 1990). Thus, a communicative language test should contextualize test tasks and testing environments related to real life situations.

With a concerted effort to measure communicative ability appropriately and effectively, various language tests have been developed over the years. For example, the ACTFL developed and disseminated a live interview format to assess functional language ability. Specifically, a test taker is required to perform functional language tasks during a live interview with a human tester. Concerns, however, have been expressed in terms of potential confounding effects of human testers and demanding logistics requirements for administration of the interview.

As a solution for the practical constraints of the ACTFL Oral Proficiency Interview (OPI), an audio-tape recorder was subsequently employed to deliver these types of spoken language tests. Examples were the Test of Spoken English, the SPEAK test, and the Simulated Oral Proficiency Interview. As recorded test items were delivered in the same way, these taped oral proficiency tests reduced the possible confounding effects of a human tester and logistics requirements for test administration. Despite the benefits, these tape-mediated oral proficiency tests have been heavily criticized because of the lack of interaction with a human being during the testing session.

The need for better test delivery formats drew attention to advanced computer technology. Computer technology makes a test available all over the world throughout the year as long as a computer and the Internet are accessible (Educational Testing Service, 1996; Hancock, 1996; Alderson, 2000a; Kenyon & Malabonga, 2001; Norris, 2001). Furthermore, as computer technology simulated various real life situations using multimedia, (Bachman, 1990; Burstein *et al*., 1996; Hancock, 1996; Warschauer, 1999; Hawisher & Self, 2000; Roever, 2001b), a computerized version of a speaking test enhances the psychometric information on the test. For example, test results obtained in

virtual situations have estimated performance in real life more accurately (Spolsky, 1985; Bachman, 1990; Hancock, 1994; Khattri & Sweet, 1996; McNamara, 1996; Fulcher, 2000; Chapelle, 2001).

As oral proficiency tests have been delivered in a various formats, concerns about validity and accountability across tests have been addressed. Particularly, computer technology has been used to evaluate oral proficiency in recent years (Jeong, 2003). However, further research is necessary on the interaction between test taker characteristics and computerized test formats.

In order to add to the professional literature in this area, the present study explored the effects of test taker characteristics (*i.e.*, self-reported years of English language study and self-reported computer use) and test delivery format on performance on an oral language test. This study found that years of English language study, based on self-reports, and test delivery format cooperatively produced a significant effect on test scores of study participants on a speaking test. However, computer use, as self-reported, did not produce any main effect or interaction effect on speaking test scores.

For the purposes of this study, 210 international students were recruited from various graduate academic programs in a major US university during Autumn Quarter of 2005. The native languages of the study participants were other than English. They were all enrolled in a graduate program as either part-time or full-time students at the time of the study. Since the research site included a significant population of international students coming from diverse countries, the pool of the participants was deemed to be representative of the population of international students studying in US universities during this year.

This study utilized the results of a computerized spoken English test, a conventional audio-taped spoken English test, and replies to a participant questionnaire. The first edition of the SPEAK test, developed and disseminated by the Educational Testing Service, was delivered either on a computer screen or via an audiotape recorder. A questionnaire was developed for the study to collect information about participants' self-reported computer use, their self-reported English learning experience and demographic background information.

For analysis of the data, this study used a 2×2×2 mixed factorial research design. The research was conducted using three independent variables (*i.e.*, self-reported years of English language study, test delivery format, and self-reported computer use) and one dependent variable (*i.e.*, test scores on a spoken English test).

Each independent variable had two levels. Specifically, levels for years of English language study were grouped into less English language study and greater English language study. The categories for test delivery format were a computerized test and a taped test. Levels for computer use were grouped into less computer use and more computer use. The study research questions were answered according to an analysis of the results of descriptive statistics and ANOVA statistics.

This chapter discusses the findings, answers to research questions, conclusions, and applications of the data from the study. Conclusions of the investigation were made based on the results of data analyses presented in the previous chapter. Implications of this study are discussed for language testers and language educators. Limitations of the study are followed by recommendations for further research. Finally, conclusion recaps the key aspects of the study.

5.2 Findings and answers to the research questions

Three research questions were developed to study the effects of test taker characteristics (*i.e.*, self-reported years of English language study and self-reported computer use) and test delivery format on an English speaking test. Specifically, this study investigated the following research questions: 1) To what extent do self-reported years of English language study relate to a spoken English test score; 2) To what extent does test delivery format relate to a spoken English test score; and 3) To what extent does the self-reported computer use of test takers relate to a spoken English test score?

With respect to the number of years of English language study, the participants in this investigation self-reported a range from a low of three years to a high of 25 years. In order to investigate the effect of years of English study on spoken English test results, participants were categorized as either less English study group or greater English study group. 12 years, the median, was used as a cut score. For example, less English study group included individuals who had previously studied English language for less than 12 years. Greater English study group, on the other hand, included individuals who had previously studied English language for 12 years or more.

With respect to test delivery format, the purpose of identifying the effect of test delivery format on test scores on the speaking test, the identical test items taken from the SPEAK test were delivered either on a computer screen or through an audio-tape recorder. The SPEAK test is an audiotape-mediated spoken English test that requires a test taker to perform various functional language tasks. This test was originally developed to measure oral proficiency of non-native English speakers by the Educational Test Service. Since its first publication in 1980s, the test has been used world-wide because of efficiency in time

and cost as well as published reasonable reliability and validity statistics (Clark & Swinton, 1979; Subkoviak, 1985; Tatsuoka, 1985). For the same reasons, this test was selected for and utilized for the present study.

Participants were randomly assigned to take either a taped version of the spoken English test or a computerized version of the spoken English test. The time for test administration was scheduled at the convenience of each participant. All participants were required to complete a mandatory study tutorial before taking a test. The tutorial was designed to familiarize participants with the testing procedures and testing equipment (*i.e.*, an audiotape recorder or a computer).

After the tutorial, each participant took a test individually under the same testing conditions. The physical setting, proctoring, and testing equipment were all controlled. The test was administered in two rooms typically used for conducting the SPEAK test at the research site. Each research room was equipped with an audio-tape recorder or an IBM compatible laptop computer with a 15 inch screen for administering a taped version of the test or a computerized version of the test.

Each response sample of the taped test was audio-tape recorded on a regular audio-tape machine while the computerized test was electronically saved on a computer hard drive as an audio file. The data were subsequently analyzed by the researcher.

With respect to computer use of test takers, and in order to explore the extent to which computer use related to performance on an English oral proficiency test, participants were categorized by computer use as either less computer use group or more computer use group based on self-reported data. Computer use was defined in terms of the frequency with which test takers performed various computer tasks. The computer

tasks included simple tasks requiring basic computer skills (*e.g.*, checking email) as well as more complex tasks requiring advanced computer skills (*e.g.*, writing code in HTML) (see Appendix I).

A maximum score of 50 indicated that participants reported using a computer for all listed tasks on a daily basis. The median score of 31 was used as the cutoff in this study. The score of 31 indicated that participants used a computer for all listed tasks monthly or more frequently. Specifically, individuals who received 31 or less as a total score were categorized as less computer use group. Individuals who received a score of more than 31 were categorized as more computer use group.

First of all, one-, two-, and three-way descriptive statistics were undertaken in order to answer these research questions. According to one-way descriptive statistics, presented in Tables 4.4, 4.5, and 4.6, self-reported years of English study seemed to affect test results of the spoken English test significantly. However, neither test delivery format or self-reported computer use seemed to affect the spoken English test results significantly.

According to two-way descriptive statistics, presented in Table 4.7, years of English study and test delivery format seemed to interact significantly based on the observation that several confidence intervals did not overlap each other. On the other hand, years of English study and computer use, in Figures 4.7 and 4.8, seemed not to interact significantly. Also, test delivery format and computer use, shown in Table 4.8, seemed not to interact significantly.

According to three-way descriptive statistics, given significant mean differences presented in Table 4.11 and Figure 4.12, treatment effects seemed to be present. Thus, an

omnibus F test analysis was undertaken to investigate any possible significant interaction and/or main effects simultaneously.

The results of ANOVA statistics in Table 4.13 suggested that all three factors, $F(1, 202) = .431$, $p > .05$, did not jointly produce a significant effect on the spoken English test scores. Given non-significant three-way interaction, three two-way interactions were reviewed. First, as shown in Table 4.13, $F(1, 202) = 6.5$, $p < .05$, self-reported years of English language study and test delivery format significantly affected the speaking test results cooperatively. Second, self-reported years of English language study and computer use, $F(1, 202) = .007$, $p > .05$, did not jointly produce any significant effect on the speaking test results. Finally, the interaction between test delivery format and self-reported computer use, $F(1, 202) = 3.824$, $p > .05$, seemed not to be significant.

Since self-reported years of English language study significantly interacted with test delivery format, it was not meaningful to use solely the significant main effect of years of English study to explain the variance in the test scores. Instead, further analyses were conducted to investigate the simple effects of years of English language study on the computerized speaking test and on the taped speaking test. The results of this analysis, in Table 4.14, revealed that self-reported years of English language study produced a significant effect on the test results of the computerized test, but not those of the taped test.

In addition, the simple effects of test delivery format were analyzed at the level of less English study and at the level of greater English study. The results of this analysis, in Table 4.15, suggested that less English study group significantly interacted with test

delivery format, while greater English study group did not significantly interact with test delivery format.

Since simple effects were significant, the Scheffe post hoc test was undertaken to further analyze the pattern of interaction between self-reported years of English language study and test delivery format. The results of the post hoc comparisons, in Table 4.16, detected two significant mean differences: 1) between less English study group and greater English study group that took the computerized test; and 2) between less English study group that took a computerized test and greater English study group that took the taped test.

Specifically, greater English study group performed significantly better on the computerized version of the speaking test than less English study group. In addition, greater English study group that took the taped speaking test performed significantly better than less English study group that took the computerized test. In other words, test delivery format had an impact on the test scores for those that had self-reported less English language study but not for those that reported greater English language study.

Consequently, the study found that the variance in test scores on the speaking test could not be accounted for solely by a single factor. Self-reported years of English language study, however, significantly interacted with test delivery format. These two factors jointly produced a significant effect on the speaking test scores. In particular, the computerized speaking test, not the audio-taped SPEAK test, seemed to affect test results more for less English study group than for greater English study group. Less English study group performed better on the taped version of the spoken English test than on the computerized version of the spoken English test. However, contrary results were

observed with self-reported greater English study group. Although the mean difference was not significant, self-reported greater English study group performed better on the computerized spoken English test than on a taped version of the test. Further study is needed to identify both why an individual reacts differently to taking different test delivery formats and if the results are sustained with non-self-reported data.

On the other hand, self-reported computer use did not significantly affect test results during oral proficiency assessment. Specifically, computer use alone did not account for the variance in test scores on the English speaking test. Furthermore, computer use neither significantly interacted with years of English language study nor with test delivery format. It also seemed that computer use did not affect performance during oral proficiency assessment. The results of this study were consistent with the findings of Taylor *et al*., (1998) that no significant correlation was found between test takers' computer familiarity and their test scores on a computer-based TOEFL test measuring language ability in listening and reading. In other words, computer familiarity did not affect the test results on the computerized TOEFL test.

The results of the present study, however, contrasted with the findings of Jeong (2003) that showed a positive relationship between the electronic literacy and the English oral proficiency of the examinees who took DVOCI in an English as a foreign language context. That is to say, the better an individual carried out the computer tasks, the better the individual performed on the computerized English speaking test, DVOCI.

There might be several reasons for these mixed findings on the relationship between test takers' computer skills and the results of a computerized language test. First, the use of different definitions of computer skill might result in different findings. For

example, the present study used self-reported computer use that was defined in terms of the frequency of computer use for certain computer tasks. Taylor *et al*. (1998) used computer familiarity in terms of accessibility to, attitude toward, and experience with a computer. Jeong (2003) used the concept of electronic literacy defined in terms of electronic communication, cyberspace construction, and academic research. Given these varying treatments, conflicting results in the studies are not surprising.

Second, the use of different research instruments might be another possible reason. Specifically, the present study used the SPEAK test developed by the ETS. Taylor *et al*. (1998) used the sections of listening and reading of the TOEFL test, which did not measure spoken language abilities. Jeong (2003) used the spoken English test developed by the instructors at the research site.

Third, the different backgrounds of the research participants might make a difference. For instance, the participants for the present study were diverse international students in an ESL context, specifically the United Sates. Taylor *et al*. (1998), on the other hand, collected the data in both ESL and EFL settings. The participants for Joeng's study were Korean students, particularly cadets in military school in an EFL context, specifically Korea.

In sum, differences in terms of the definition of factors, research instruments, and research participants might explain these mixed findings on the relationship between test takers' computer skills and performance on a computerized language test. Thus, the different findings of these studies should be interpreted with caution.

Discussion has thus far focused on the physical test delivery format, such as a computer. At the same time, attention should be paid to the test delivery mechanism of

the computerized test and the tutorial. Specifically, the computerized version of the English speaking test in this study adopted a linear mechanism in that test items were delivered automatically in a linear manner from the beginning to the end of the test.

This linear automatic delivery mechanism did not require a test taker to select test items manually. The logic behind this linear automatic mechanism was to minimize possible distraction elements that might be caused by the test delivery mechanism itself, such as using a computer mouse. One of the main weaknesses of this linear mechanism, however, was that it did not allow the test taker to have a second chance to correct previous test items. This study leaves the pros and cons of test delivery mechanism for further study. Instead, the researcher had to use the testing mechanism used by the Educational Testing Service for the original SPEAK test.

In addition to the automatic test delivery mechanism, all the participants took a tutorial immediately before taking the actual speaking test. The tutorial was designed to help the test takers become familiar with the testing equipment, the test administration procedures, and the testing environment. For example, test takers who were not familiar with operating testing equipment, particularly the computer, had an opportunity to practice the computer uses needed for the computerized version of the test. If the test delivery mechanism and tutorial were not controlled, the study might have produced different results.

5.3 Implications

5.3.1 Language testers and relevant stakeholders

A constant effort has been made to develop valid and reliable testing tools for measuring language proficiency. Aligned with the current common agreement that a test result should be explained by the ability in question (Lord, 1980; Messick, 1989; Bachman, 1990; Angoff, 1993; Raju & Ellis, 2002), the concept of validity has been refined. Particularly, in the late 1980s, as mentioned earlier, the AERA/APA/NCME and Messick extended the concept of validity by adding the social significance of the interpretation and use of test results (Angoff, 1988; Messick, 1989; Bachman, 1990).

This new movement made language testers aware of their social responsibility. Further, awareness of the social significance of the interpretation and use of test results has called for more attention to test taker characteristics. This concern has called for research on potential effects of test taker characteristics on test results. Examples are studies on test validation with a focus on fairness, test bias, and differential item function.

The present study investigated whether test taker characteristics (*i.e.*, self-reported years of English language study and self-reported computer use) influenced performance on an English speaking test delivered in two different test formats (*i.e.*, conventional audio-tape player and computer). The results of this study suggested that certain test taker characteristics may significantly interact with test delivery format. Specifically, less English study group seemed to interact more with test delivery format than test takers who studied English language relatively longer, specifically 12 years or longer in this study. Therefore, throughout the entire process of developing a test, more caution should be paid to the format of test delivery, particularly for beginners.

In addition, this study revealed that the computer use of test takers did not affect performance during an oral English speaking test. However, it should be remembered that the participants in the present study completed a mandatory tutorial designed to help them become familiar with testing equipment and procedures. Corroborating the findings of Taylor *et al*. (1998), this research has suggested that a tutorial about the testing equipment and testing procedures should be provided to test takers. Accordingly, appropriate use of a tutorial may reduce the possible confounding effect of testing equipment (*i.e.*, test delivery format), testing administration procedures, and the testing environment.

5.3.2 Language educators and policy makers

Foreign/second language education programs vary in different countries. For example, some language learners are first exposed to target language study at the pre-elementary school level, while others are introduced at the elementary, middle, high school, or university level. The present study revealed that self-reported years of English language study significantly influenced test scores on an English speaking test by interacting with test delivery format. Specifically, self-reported greater English study group, particularly individuals exposed to English at the pre-elementary or early elementary level, seemed to perform better on the spoken English test than self-reported less English study group.

In other words, self-reported years of English language study strongly interacted with test delivery format. This interaction produced a significant impact on the results of the spoken English test. Specifically, the group that self-reported greater English study

seemed less influenced by the test delivery format. Thus, a test delivery format should be selected according to test taker characteristics along with the purpose of a language test.

5.4 Limitations

This study was limited in terms of a relatively small sample size, a single test, and mainly self-reported data. First, this research was conducted at a large US public university with a significant population of students from diverse countries. A total of 210 volunteers were recruited for this study. Considering the relatively small sample size of 210, further study with a larger sample size is necessary to generalize the findings beyond the participants in this study.

Second, a dependent variable for this study was test scores on a single spoken English test to assess oral proficiency. The SPEAK test was used to measure the oral performance of participants in English because of its established reasonable reliability and validity as published by the Educational Testing Service. Considering that various language tests are available, further research with several different language tests will provide a deeper understanding of the interaction between certain test taker characteristics and the test results.

Third, of the three independent variables for this study, years of English language study and computer use were based on self-reported information. It would be useful to collect data from various information resources including students' official documents, as well as other data from the participants themselves. In this study, however, the data on years of English language study and computer use was limited because the information was self-reported via a participant questionnaire. Of course, it was assumed that the

volunteer study participants responded to the questionnaire accurately and honestly, but they may have inaccurately reported their language study experience and their computer abilities.

5.5 Recommendations for further research

As validity is the essential element of testing, an ongoing effort is needed to ensure language test validity. First of all, further research is necessary to explore the relationship between test taker characteristics and test items. A communicative approach to language testing encouraged language test makers to utilize various task-based test items. Certain types of task-based test items might be biased for or against individuals in a particular group.

Under the item response theory (IRT), differential item functioning (DIF) analysis is useful for identifying the relationship between test item characteristics and test taker characteristics between different groups (Clauser & Mazor, 1998). The IRT approach provides rich information about test takers and test items because the IRT "estimates the value of the trait using the inferred relationships between the item responses and the trait being measured" (Thissen, 2003, p. 592). Moreover, the IRT is a useful theoretical framework for DIF statistics by providing "between-group differences in the item parameters for the specific model" (Clauser & Mazor, 1998, p. 32). Currently, the computer programs BILOG and MULTILOG are available for IRT-based DIF analysis.

In addition, generalizability theory (G-theory) is efficient for analyzing the relationships among multi-facets (*e.g.*, item parameter, person parameter) (Cronbach *et al*,

1972; Lee, 2006). The computer program GENOVA is available for G-theory analysis. The findings would theorize the effect of test taker characteristics and more importantly enhance fairness of testing.

Second, further study is needed to standardize the definition of oral proficiency and subsequently to enhance accountability of spoken language test results. For example, the ILR and ACTFL defined oral proficiency in terms of function, context/content, accuracy and text type (see Appendices A & C). On the other hand, for the use of the SPEAK test first published in 1982, the ETS described oral proficiency in terms of pronunciation, grammar, fluency, and comprehensibility (see Appendix D). Currently, the ETS defines oral proficiency in terms of general description, delivery, language use, and topic development for the use of the TOEFL iBT. Thus, the findings of key elements of oral proficiency would possibly be useful in standardizing a definition of oral proficiency in the profession. Further, a standardized definition of oral proficiency might be useful in promoting validity and accountability of language tests by providing essential constructs for oral proficiency.

A structural equation model (SEM) might be appropriate to identify the multiple latent variables that determine oral proficiency. The SEM effectively analyzes a number of variables simultaneously. For the SEM analysis, the computer programs AMOS and LISREL are commonly used.

Finally, further study is needed about scoring reliability because it is one of the important components to promote test validity. Considering that a speaking test adopts a polytomous scoring model, attention should be paid to ensure scoring reliability. Measures of scoring reliability include inter-rater reliability and intra-rater reliability.

Thus, in-depth research on rater reliability is necessary to enhance test validity. Many-facet Rasch measurement is useful to investigate rater reliability. Many-facet Rasch measurement calibrates multiple facets (*e.g.*, item parameter, judge parameter *etc.*) on a single scale to allow estimation of the effect of each facet at a glance (Hambleton *et al.*, 1991; McNamara, 1996). The computer program FACETS is widely used for many-facet Rasch measurement analysis.

5.6 Conclusions

Various language tests have been developed corresponding to myriad purposes. Despite this diversity in language testing, all of the tests pursue a common goal: measuring the ability of interest appropriately and accurately. Therefore, validation is an essential process for the development of a test.

Until the late 1980s, instead of a unitary concept of validity, there were several sub-categories for validity. For example, according to the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME) (1985), construct validity was defined as the extent to which a test measured the ability in question. Content validity focused on the extent to which a test dealt with the subject content of interest. Criterion validity was the extent to which a test score fit certain criteria for the ability of interest.

The late 1980s, however, witnessed a new approach to validity with a change in perspective about the validity of a test. The AERA/APA/NCME (1985), and Messick(1989) introduced a more comprehensive concept of validity. Particularly, Messick(1989) proposed the unitary concept of validity with multi-facets.

According to Messick(1989), validity is “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 or other modes of assessment” (p. 13). This unified validity framework includes two facets: “the source of justification of the testing” and “the function or outcome of the testing” (p.20). The source of justification of the testing means “either an evidential basis or a consequential basis,” whereas the function or outcome of the testing means “either test interpretation or test use” (p.20). This new movement also emphasized the social responsibility of testing by including the facets of test use and test interpretation.

With respect to test use, the awareness of the social responsibility of testing brought a change in perspective about tests. For instance, a test has been used mostly for the purpose of screening and selection against a benchmark set by language testers, educational institution administrators, or future employers. With increased social awareness, however, consideration was given to not only the request of language testers, educational institution administrators, or future employers, but also to the needs of test takers. In other words, all interested parties should benefit from testing.

With respect to test interpretation, the awareness of the social responsibility of testing called more attention to test takers, specifically possible influences of test taker characteristics on test results (Messickk, 1989; Bachman, 1990). For example, test takers at the same level of ability in question are expected to earn the same test results (Lord, 1980; Hulin, Drasgow, & Parsons, 1983; Hambleton & Swaminathan, 1985; Messickk, 1989; Bachman, 1990; Hambleton, Swaminathan, & Rogers, 1991; Angoff, 1993; Raju & Ellis, 2002). In other words, test results should not be interpreted primarily based on

particular test taker characteristics but, rather, by the ability in question, even though the test taker characteristics may be involved in some way.

This study found that certain characteristics of test takers might produce a significant influence on test results during an oral English test. Therefore, this research recommends that a multifaceted effort be made to ensure that a test functions fairly across various test takers, regardless of their individual backgrounds.

This study also suggests sharing ownership of testing among test makers, test takers, and test users, which allows all of interested parties to have an opportunity to benefit appropriately from a test. For example, test users collect psychometric information in accordance with the purposes of test use such as selecting or placing candidates. Testers use the psychometric information to develop better tests. Test takers also use psychometric information to improve their ability of interest. Thus, as testing is an interrectual property of various stakeholders, profit motives for testing should not be conflict each other.

Finally, the study has corroborated previous research and studies that call for a focus on the appropriate use of different speaking test formats according to the purposes of the tests and the characteristics of the individuals who take certain speaking tests in English.

REFERENCES

Alderson, J. C. (2000a). *Assessing reading*. Cambridge: Cambridge University Press.

Alderson, J. C. (2000b). Technology in testing: the present and the future. *System*, 28, 593-603.

Allen, M. & Yen, W. (1979). *Introduction to measurement theory*. IL: Waveland Press.

American Council on the Teaching of Foreign Languages. (1999). *ACTFL proficiency guidelines -- speaking: Revised 1999*. New York: Author.

American Council on the Teaching of Foreign Languages & Interagency Language Roundtable. (1999). *OPI 2000 tester certification workshop training manual*.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1985). *Standards for Educational and Psychological testing.* Washington, DC: Authors.

Angiolillo, P. (1947). *Armed forces foreign language teaching*. New York: Vanni.

Arnett, K. & Haglund, J. (2001). American Council on the Teaching of Foreign Languages Oral Proficiency Interview. *The Canadian Modern Language Review,* 58 (2), 312-318.

Angoff, W. (1988). Validity: An evolving concept. In H. Wainer & H. Braun (Eds.), *Test Validity* (pp. 19-32). Hillsdale, NJ: Erlbaum.

Angoff, W. (1993). Perspectives on differential item functioning methodology. In P. Holland, & H. Wainer, (Eds.), *Differential item functioning* (pp.3- 23). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. (2000). Modern language testing at the turn of the century. *Language testing*, 17(1), 1-42.

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

Bachman, L. & Savignon, S. (1986). The evaluation of communicative language proficiency: a critique of the ACTFL oral interview. *Modern Language Journal,* 70, 380-390.

Bachman, L. F. (1988). Problems in examining the validity of the ACTFL oral proficiency interview. *Studies in Second Language Acquisition*, 10, 149-164.

Bachman, L. F. & Savignon, S. (1986). The evaluation of communicative language proficiency: a critical of the ACTFL oral interview. *Modern Language Journal*, 70, 380-390.

Barnwell, D. (1996). *A history of foreign language testing in the United States*. Tempe, Arizona: Bilingual Press.

Bateman, A. & Griffin, P. (2003). The appropriateness of professional judgment to determine performance rubrics in a graded competency based assessment framework. Paper presented at New Zealand Association for Research in Education /Australian Association for Research in Education. Auckland, New Zealand.

Becker, L. (1999). Measures of effect size. Retrieved June 7, 2006, from http://web.uccs.edu/lbecker/SPSS/glm_effectsize.htm#Eta%20squared%20(h2)

Berk, R. (1982). Introduction. In R. Berk (Ed.), *Handbook of methods for detecting test bias*. Baltimore: Johns Hopkins University Press.

Bernhardt, E. (2000). If reading is reader-based, can there be a computer-adaptive test of reading? In M. Chalhoub-Deville (Ed.), *Issues in computer-adaptive testing of reading proficiency* (pp. 1-10). Cambridge: University of Cambridge Local Examinations Syndicate and Cambridge University Press.

Brown, J.D. (2004). For computerized language tests, potential benefits outweigh problems. *Essential Teacher*, 1 (4), 37-40.

Brwon, A. & Iwashita, N. (1998). The role of language background in the validation of a computer-adaptive test. In A.J. Kunnan (Eds.), *Validation in language assessment* (pp.195-207). Mahwah, NJ: Lawrence Erlbaum Associates.

Brown, J. D. (1997). Computer in language testing: present research and some future directions. *Language Learning and Technology*, 1 (1), 44-59.

Brown, J. D. & Hudson, T. (1998). The alternatives in language assessment. *TESOL Quarterly*, 32, 653-675.

Byrnes, H. (1987). Features of pragmatic and sociolinguistic competence in the oral proficiency interview. In A. Valdman (Ed.), *Proceedings of the symposium on the evaluation of foreign language proficiency* (pp. 167-77). Bloomington, IN: Indiana University.

Burstein, J., Frase, L.T., Ginther, A., & Grant, L. (1996). Technologies for language assessment. *Annual Review of Applied Linguistics*, 16, 240-260.

Byrnes, H. (1987). Second language acquisition: insights from a proficiency orientation. In H. Byrnes & M. Canale (Eds.), *Defining and developing proficiency: Guidelines, implementation and concepts* (pp. 107-131). Lincolnwood, IL: National Textbook Company.

Canale, M. (1984). Testing in a communicative approach. In G. A. Jarvis (Eds.), *The challenge for excellence in foreign language education* (pp. 79-92). Middlebury, Vt.: The Northeast Conference Organization.

Canale, M., & Swain M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1, 1-47.

Center for Applied Linguistics. (2003). Retrieved July 17, 2003, from http://www.cal.org

Center for Applied Linguistics. (2005). Retrieved June 7, 2005, from http://www.cal.org/projects/copi.html

Chapelle, C. (1999). Validity in language assessment. *Annual Review of Applied Linguistics*, 19, 254-272.

Clark, J. L., & Swinton, S. S. (1979). *An exploration of speaking proficiency measures in the TOEFL context*. (TOEFL Research Report 4). Princeton, NJ: Educational Testing Service.

Clark, J. L., & Swinton, S. S. (1980). *The Test of Spoken English as a measure of communicative ability in English-medium instructional settings*. (TOEFL Research Report 7). Princeton, NJ: Educational Testing Service.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *ITEMS* (Spring).

Chalhoub-Deville, M. (1997). The Minnesota articulation project and its proficiency-based assessments. *Foreign Language Annals*, 30, 492-502.

Chalhoub-Deville, M. (2001a). Language testing and technology: past and future. *Language Learning & Technology*, 5 (2), 95-98.

Chalhoub-Deville, M. (2001b). Task-based assessments: Characteristics and validity evidence. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks second language leaning, teaching and testing* (pp. 210-221). Essex, England: Pearson Education Limited.

Chalhoub-Deville, M. & Deville, C. (1999). Computer adaptive testing in second language contexts. *Annual Review of Applied Linguistics*, 19, 273-299.

Chapelle, C. (1999). Validity in language assessment. *Annual Review of Applied Linguistics*, 19, 254-272.

Chapelle, C. (2001). *Computer applications in second language acquisition*. Combridge: Combridge University Press.

Choi, S. (2000). *Teaching English as a foreign language in Korean middle schools: exploration of communicative language teaching through teachers' beliefs and self-reported classroom teaching practices*. Unpublished doctoral dissertation. Ohio State University, Columbus.

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

Clinch, J.J., & Keselman, H.J. (1982). Parametric alternatives to the analysis of variance. *Journal of Educational Statistics*, 7, 207-214.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. N.J. : L. Erlbaum Associates.

Cole, N. S. (1981). Bias in testing. *American Psychologist*, 36, 1067-1077.

Cole, N. S. (1993). History and development of DIF. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp.25- 33). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 292-334.

Cronbach, L.J, Gleser, G.C., Nanda, H. & Rajaratnam, N. (1972). *The dependability of behavioral measurements: theory of generalizability*. New York: John Wiley.

Di Pietro, R. (1989). *Strategic interaction: Learning languages through scenarios*. Cambridge: Cambridge University Press.

Dunkel, P. (1999). Research and development of a computer-adaptive test of listening comprehension in the less-commonly taught language Hausa. In M. Chalhoub-Deville (Ed.), *Issues in computer-adaptive testing of reading proficiency* (pp. 91-121). Cambridge: University of Cambridge Local Examinations Syndicate and Cambridge University Press.

Douglas, D. (1997). Language for specific purposes testing. In *Encyclopedia of language in education* (Vol. 7, pp. 111-20). Dordrecht: Kluwer Academic.

Educational Testing Service. (1982a). *Guide to SPEAK*. Princeton, NJ: Educational Testing Service.

Educational Testing Service. (1982b). *Test of Spoken English: manual for score users*. Princeton, NJ: Educational Testing Service.

Educational Testing Service. (1996). TOEFL: Announcing computer-based testing. Princeton, NJ: Educational Testing Service.

Eignor, D., Taylor, C., Kirsch, I., & Jamieson, J. (1998). *Development of a scale for assessing the level of computer familiarity of TOEFL examinees.* Princeton, NJ: Educational Testing Service.

Farhady, H. (1982). Measures of language proficiency from the learners' perspective. *TESOL Quarterly*, 16(1), 43-59.

Fulcher, G. (1996). Testing tasks: issues in task design and the group oral. *Language Testing*, 13(1), 23-51.

Fulcher, G.. (1997). The testing of speaking in a second language. In C. Clapham & D. Corson, (Eds.), *Language testing and assessment* (Vol. 7, pp. 75-85). Dordrecht: Kluwer Academic.

Fulcher, G. (2000). The communicative legacy in language testing. *System*, 28, 483-497.

Fulcher, G. (2003). *Testing second language speaking*. London: Longman/Pearson Education.

Grabe, W. (1999). Developments in reading research and their implications for computer-adaptive reading assessment. In M. Chalhoub-Deville (Ed.), *Issues in computer-adaptive testing of reading proficiency* (pp. 11-48). Cambridge: University of Cambridge Local Examinations Syndicate and Cambridge University Press.

Gruba, P. & Corbel, C. (1997). Computer-based testing. In C. Clapham & D. Corson, (Eds.), *Language testing and assessment* (Vol. 7, pp. 141-149). Dordrecht: Kluwer Academic.

Hair, J., Anderson, R., Tatham, R, & Black, W. (1998). *Multivariate data analysis*. (5th ed.). New Jersey: Prentice-Hall, Inc.

Hambleton, R.K. (1994). Guidelines for adapting educational and psychological tests: a progress report. *European Journal of Psychological Assessment*, 10, 229-244.

Hambleton, R.K. (2001). The next generation of the ITC test translation and adaption guidelines. *European Journal of Psychological Assesessment*, 17, 64-172.

Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory: principles and applications*. Norwell, MA: Kluwer.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory*. Thousand Oaks, CA: Sage.

Hancock, C. R. (1994). Glossary of selected terms. In C. R. Hancock (Ed.), *Teaching, testing, and assessment: making the connection* (pp.235-240). Lincolnwood, IL: National Textbook Company.

Hancock, C. R. (1996). Alternative assessment in foreign/second language: What do we in foreign language know? In Z. Moore (Eds), *Foreign language teacher education* (pp. 75-88). Lanham, Maryland: University Press of America.

Hawisher, G. E., & Self, C.L. (Eds.) (2000). *Global literacies and the World-Wide Web*. London: Routledge.

He, A., & Young, R. (1998). Language proficiency interviews: a discourse approach. In R. Young & A. He (Ed.), *Talking and Testing: Discourse approaches to the assessment of oral proficiency* (pp. 1-26). Philadelphia: John Benjamins.

Henning, G. (1991). Validating an item bank in a computer-assisted or computer-adaptive test. In P. Dunkel (Ed.), *Computer-assisted language learning and testing: Research issues and practice* (pp. 209-222). New York: Newbury House.

Hill, K. (1998). The effect of test-taker characteristics on reactions to and performance on an oral English proficiency test. In A. Kunnan (Ed.), *Validation in language assessment* (pp. 209-229). New Jersey: Lawrence Erlbaum.

Hicks, M. (1989). *The TOEFL computerized placement test: Adaptive conventional measurement* (TOEFL Research Report No. 31). Princeton, NJ: Educational Testing Service.

Holland, P.W., & Thayer, D.T. (1986). *Differential item performance and the Mantel-Haenszel Procedure* (Research Report No. 86-31). Princeton, NJ: Educational Testing Service.

Hopkins, K. (1998). *Educational and psychological measurement and evaluation*. (8th ed.). MA: Allyn & Bacon.

Hulin,C.L., Drasgow, F., & Parsons, C.K. (1983). *Item response theory: applications of psychological measurement*. Homewood, IL: Dow Jones-Irwin.

Hymes, D. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics* (pp. 269-293). Harmondsworth, England: Penguin Books.

Jeong, T. (2003). *Assessing and interpreting students' English oral proficiency using d-VOCI in an EFL context*. Unpublished doctoral dissertation. Ohio State University, Columbus.

Johnson, M. (2001). *The art of non-conversation: a re-examination of the validity of the Oral Proficiency Interview*. New Haven, CT: Yale University Press.

Jonassen, D., Howland, J., Moore, J., & Marra, R. (2003). *Learning to solve problems with technology: constructive perspective*. New Jersey; Pearson Education.

Jones, R. (1975). Testing language proficiency in the United States Government. In R. Jones & B. Spolsky (Eds.), *Testing language proficiency* (pp.1-9). Arlington, Virginia: The Center for Applied Linguistics.

Kachru, B.B. (1992). Models for Non-Native Englishes. In B. B. Kachru (Ed.), *The Other tongue: English across cultures* (pp. 48 -74). (2nd ed.). Urbana: University of Illinois Press.

Kaulfers, W. (1944). War-time developments in modern language achievement tests. *Modern Language Journal*, 28, 136-150.

Kenyon, D.M. & Malabonga, V. (2001). Comparing examinee attitudes toward computer-assisted and other oral proficiency assessments, *Language Learning & Technology*, 5 (2), 60-83.

Keppel, G. (1991). *Design and analysis*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Khattri, N. & Sweet, D. (1996). Assessment reform: Promises and challenges. In M. Kane, & R. Mitchell (Ed.), *Implementing performance assessment* (pp. 1-21). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Kim, M. (2001). Detecting DIF across the different language groups in a speaking test. *Language Testing*, 18 (1), 89-114.

Kim, S.H., Cohen, A.S., & Park, T.H. (1995). Detection of differential item functioning in multiple groups. *Journal of Educational Measurement*, 32, 261-276.

Kim, W. (2002). *The test of English for international communication (TOEIC) as measure of Korean adult English language oral proficiency*. Unpublished doctoral dissertation. University of Kansas.

Kirk, R.E. (1995). Experimental design: procedures for the behavioral sciences (3rd ed.). CA: Brooks/Cole Publishing Company.

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

Kitao, S. & Kitao, K. (1996). *Testing communicative competence*. (ERIC Document Reproduction Service No. ED 398 260).

Language Acquisition Resource Center. (2003). Retrieved July 17, 2003, from http://larcnet.sdsu.edu/testing.php?page=dvoci

Lazaraton, A. (1992). The structural organization of a language interview: a conversation analytic perspective. *System*, 20, 373-386.

Lazaraton, A. (1996). Interlocutor support in oral proficiency interviews: the case of the CASE. *Language Testing*, 13 (2), 151-172.

Lazaraton, A. (1997). Performance organization in oral proficiency interviews: the case of language ability assessments. *Research on Language and Social Interaction*, 30 (1), 53-72.

Lee, Y. (2006). Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks. *Language Testing*, 23 (2), 131-166.

Lewkowicz, A. (2000). Authenticity in language testing: some outstanding questions. *Language Testing,* 17 (1), 43-64.

Liskin-Gasparro, J. (1984a). The ACTFL proficiency guidelines: a historical perspective. In T. Higgs (Ed.), *Teaching for proficiency, the organizing principle* (pp. 11-42). Lincolnwood, IL: National Textbook Company.

Liskin-Gasparro, J. (1984b). The ACTFL proficiency guidelines: gateway to testing and curriculum. *Foreign Language Annals*, 17 (5), 475-489.

Liskin-Gasparro, J. (1987). The ACTFL proficiency guidelines: an update. In A. Valdman (Ed.), *Proceedings of the Symposium on the Evaluation of Foreign Language Proficiency* (pp. 19-27). Bloomington: Indiana U.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Lowe, P. (1988). The unassimilated history. In P. Lowe & C.W. Stansfield (Eds.), *Second language proficiency assessment: current issues* (pp. 11-51). Englewood Cliffs: Prentice Hall Regents.

Lynch, A. J. (1982). Authenticity in language teaching: some implications for the design of listening materials. *British Journal of Language Teaching*, 20, 9-16.

Masters, G..N. & Wright, B.D. (1997). The partial credit model. In W.J. Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 101-122). New York: Springer.

Malone, M. (2000). Simulated Oral Proficiency Interviews: recent developments. *Digest*, December, EDO-FL-00-14.

Mead, A.D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: a meta-analysis. *Psychological Bulletin*, 114 (3), 449-458.

Messick, S. (1989). *Validity*. In R.L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13-103). New York: Macmillan.

McNamara, T. (1996). Measuring second language performance. London, New York: Longman.

McNamara, T. (1997). Performance testing. In Encyclopedia of language and education (Vol. 7, pp. 131-139).

Miles, J. (2004). *Test of spoken English*. Paper presented at the Annual Meeting of the TESOL (Long Beach, CA, March 31-April 3, 2004).

Morrow, K. (1979). Communicative language testing: revolution or evolution? In C.J. Brumfit & K. Johnson (Eds.), The communicative approach to language teaching (pp. 143-157). Oxford: Oxford University Press.

Morrow, K. (1991). Evaluating communicative tests. In S. Anivan (Ed.), *Current developments in language testing* (pp.111-18). Singapore: SEAMEO Regional Language Center.

National Institute of Standards and Technology. (2003). *NIST/SEMATECH e-Handbook of Statistical Methods*. Retrieved August 7, 2006, from http://www.itl.nist.gov/div898/handbook

Norris, J.M. (2001). Concerns with computerized adaptive oral proficiency assessment, *Language Learning & Technology*, 5 (2), 99-105.

Nunan, D. (1999). *Second language teaching and learning*. MA: Heinle & Heinle Publishers.

Ohio State University, (2005). English as a Second Language Programs. Retrieved September 20, 2005 from http://www.esl.ohio-state.edu/Index.html

O'Loughlin, K. (2001). *The equivalence of direct and semi-direct speaking tests*. Cambridge: Cambridge University Press.

Omaggio, A. C., (1983). Methodology in transition: the new focus on proficiency. *The Modern Language Journal*, 67 (4), 330-341.

Omaggio, A. C., (1986). *Teaching language in context*: *Proficiency-oriented instruction*. Boston, Mass.: Heinle & Heinle.

Purdue University. (2005). Oral English proficiency program. Retrieved June 7, 2005 from http://www.purdue.edu/OEPP/

Raffaldini, T. (1988). The use of situation tests as measures of communicative ability. *Studies in Second Language Acquisition*, 10, 197-216.

Raju, N.S. & Ellis, B.B. (2002). Differential item and test functioning. In F. Drasgow & N. Schmitt (Eds.), Measuring and analyzing behavior in organizations (pp. 156-188). San Francisco: Jossey-Bass.

Riggenbach, H. (1998). Evaluating learner interactional skills: Conversation at the micro level. In R. Young & A. W. He (Eds.), *Talking and testing: Discourse approaches to the assessment of oral proficiency* (pp.53-67). Amsterdam: John Benjamins.

Roever, C. (2001a). *A web-based test of interlanguage pragmatic knowledge: Implicatures, speech acts, and routines*. Unpublished manuscript, University of Hawai'i at Manoa.

Roever, C. (2001b). Web-based language testing. *Language Learning & Technology*, 5 (2), 84-94.

Salaberry, R. (2000). Revising the revised format of the ACTFL Oral Proficiency Interview. *Language Testing*, 17 (3), 289-310.

Sato, K. & Kleinsasser, R. (1999). Communicative language teaching: practical understandings. *The Modern Language Journal*, 83 (IV), 494-517.

Sawaki, Y. (2001). Comparability of conventional and computerized tests of reading in a second language. Language Learning & Technology, 5 (2), 38-59.

Sarwark, S., Smith, J., MacCallum, R. & Cascallar, E. (1995). *A study of characteristics of the SPEAK test* (TOEFL Research Report No. 49). Princeton, New Jersey: Educational Testing Service.

Sarwark, S. (2006). From conversation during the SPEAK test rater training workshop at The Ohio State University.

Savignon, S. J. (1985). Evaluation of communicative competence: the ACTFL provisional proficiency guidelines. *Modern Language Journal*, 69, 129-34.

Scheuneman, J.D. (1981). A new look at bias in aptitude tests. In P. Merrifield (Ed.), *New directions for testing and measurement: measuring human abilities* (pp. 3-33). San Francisco: Jossey-Bass.

Schwarz, R.D., Rich, C., & Podrabsky, T. (2003). A DIF analysis of item-level mode effects for computerized and paper-and-pencil tests. Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, April 22-24, 2003).

Shohamy, E. (1988). A proposed framework for testing the oral language of second/foreign language learners. *Studies in Second Language Acquisition*, 10, 165-79.

Sollenberger, H. E. (1978). Development and current use of the FSI oral interview test. In J. L. D. Clark (Ed.), *Direct testing of Speaking Proficiency: Theory and Application* (pp. 1-12). Princeton, NJ: Educational Testing Service.

Spolsky, B. (1975). *Testing language proficiency*. Arlington, Virginia: Center of Applied Linguistics.

Spolsky, B. (1985). The limits of authenticity in language testing. *Language Testing,* 2 (1), 31-40.

Spolsky, B. (1995). *Measured words*. Oxford: Oxford University Press.

Spurling, S., & Ilyin, D. (1985). The impact of learner variables on language test performance. *TESOL Quarterly*, 19(2), 283-301.

Stark, S., Chernyshenko, S., Chuah, D., Lee, W. & Wadlington, P. (2001). *IRT tutorial*. IRT Modeling Lab at University of Illinois at Urbana-Champaign. Retrieved August 12, 2004, from http://work.psych.uiuc.edu/irt/dim_main.asp

Stansfield, C. (1996). *Test development handbook: Simulated Oral Proficiency Testing (SOPI)*. Washington, DC: Center for Applied Linguistics.

Stansfield, C. W. & Kenyon, D., (1996). Comparing the Scaling of Speaking Tasks by Language Teachers and by the ACTFL Guidelines. In A. Cumming & R. Berwick (Eds), *Validation in Language Testing* (pp. 124-153). England: Multilingual Matters Ltd.

Subkoviak, M. J. (1985). Review of Test of Spoken English. In Mitchell, J. V. (ed.), *The ninth mental measurements yearbook* (pp. 1592-1593). Lincoln, NE: The University of Nebraska Press.

Swender, E. (Ed.). (1999). *ACTFL oral proficiency interview tester training manual.* Yonkers, NY: American Council on the Teaching of Foreign Languages.

Tan, W. Y. (1982). Sampling distributions and robustness of t, F and variance-ratio in two samples and ANOVA models with respect to departure from normality. *Communications in Statistics – Theory and Methods*, 11, 2485-2511.

Tatsuoka, K. K. (1985). Review of Test of Spoken English. In Mitchell, J. V. (ed.), *The ninth mental measurements yearbook* (pp. 1593-1594). Lincoln, NE: The University of Nebraska Press.

Taylor, C., Jamieson, J., Eignor, D., & Kirsch, I., (1998). *The relationship computer familiarity and performance on computer-based TOEFL test tasks*. Princeton, NJ: Educational Testing Service.

Thissen, D., Steinberg, L. & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. Holland, & H. Wainer (Eds.), *Differential item functioning* (pp. 67- 113). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Thissen, D. (2003). Estimation. In M.Toit (Ed.), *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT* (pp. 592-617). Lincolnwood, IL: Scientific Software International.

Thompson, G. (1996). Some misconceptions about communicative language teaching. *ELT Journal*, 50, 9-15.

Toit, M. (2003). (Ed.). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT*. Lincolnwood, IL: Scientific Software International.

Underhill, N. (1987). *Testing Spoken Language*. Cambridge: Cambridge University Press.

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

Warschauer, M. (1999). *Electronic literacies: language, culture, and power in online education*. Mahwah, NJ: Lawrence Erlbaum Associates.

Warschauer, M. & Healey, D. (1998). Computers and language learning: an overview. *Language Teaching*, 31, 57-71.

Wood, R. (1993). *Assessment and testing*. Cambridge: Cambridge University Press.

Yoffe, L. (1997). An overview of ACTFL Proficiency Interviews: a test of speaking ability. *JALT Testing & Evaluation SIG Newsletter*, 1 (2), 3-9.

Zickar, M. (2002). Modeling data with polytomous item response theory. In F. Drasgow & N. Schmitt (Eds.), *Measuring and analyzing behavior in organizations* (pp. 123-155). San Francisco: Jossey-Bass.

Appendix A: The Interagency Language Roundtable Proficiency Scale – Speaking

	Function: Tasks accomplished, attitudes expressed, tone conveyed; what a person can do	**Context/Content:** Topics, subject areas, activities and jobs addressed; settings	**Accuracy:** Acceptability, quality and correctness of message conveyed	**Text Produced:** Length and organization of utterance; kinds of discourse
0	No functional ability	None	None	Individual words and phrases
1	Can create with the language, ask and answer questions, participate in short conversations, and resolve a basic situation	Everyday survival topics and courtesy requirements	Intelligible to a native speaker used to dealing with foreigners	Discrete sentences
2	Able to fully participate in casual conversations; can give instructions, describe, report facts, narrate in present, past and future, and resolve a basic situation with a complication	Concrete topics such as own background, family, interests, work, travel, and current events	Understandable to a native speaker not used to dealing with foreigners; sometimes miscommunicates	Full paragraphs, minimally cohesive
3	Can converse in formal and informal situations, resolve problems in unfamiliar situations, deal with abstract topics, provide explanations, offer supported opinions; hypothesize	Practical, social, professional and abstract topics, particular interests, and special fields of competence	Errors virtually never interfere with understanding and rarely disturb the native speaker. Only sporadic non-patterned errors in basic structures	Extended discourse
4	Able to tailor language to fit audience, counsel, persuade, negotiate, represent a point of view, and interpret informally for dignitaries	All topics normally pertinent to professional needs	Nearly equivalent to a well-educated native speaker. Speech is extensive, precise, appropriate to every occasion with only occasional errors	Speeches, lectures, debates, conference discussions. Well organized extensive discourse
5	Functions in a manner that is equivalent to that of a well-educated native speaker	All subjects	Performance equivalent to that of a well-educated native speaker	All texts controlled by a highly articulate, well-educated native speaker

Source: American Council on the Teaching of Foreign Languages & Interagency Language Roundtable. (1999). *OPI 2000 tester certification workshop training manual*. P. 42.

Appendix B: The ACTFL Proficiency Guidelines—Speaking

SUMMARY HIGHLIGHTS

ACTFL PROFICIENCY GUIDELINES—SPEAKING (REVISED 1999)

SUPERIOR	ADVANCED	INTERMEDIATE	NOVICE
Superior-level speakers are characterized by the ability to:	**Advanced-level speakers are characterized by the ability to:**	**Intermediate-level speakers are characterized by the ability to:**	**Novice-level speakers are characterized by the ability to:**
• participate fully and effectively in conversations in formal and informal settings on topics related to practical needs and areas of professional and/or scholarly interests • provide a structured argument to explain and defend opinions and develop effective hypotheses within extended discourse • discuss topics concretely and abstractly • deal with a linguistically unfamiliar situation • maintain a high degree of linguistic accuracy • satisfy the linguistic demands of professional and/or scholarly life	• participate actively in conversations in most informal and some formal settings on topics of personal and public interest • narrate and describe in major time frames with good control of aspect • deal effectively with unanticipated complications through a variety of communicative devices • sustain communication by using, with suitable accuracy and confidence, connected discourse of paragraph length and substance • satisfy the demands of work and/or school situations	• participate in simple, direct conversations on generally predictable topics related to daily activities and personal environment • create with the language and communicate personal meaning to sympathetic interlocutors by combining language elements in discrete sentences and strings of sentences • obtain and give information by asking and answering questions • sustain and bring to a close a number of basic, uncomplicated communicative exchanges, often in a reactive mode • satisfy simple personal needs and social demands to survive in the target language culture	• respond to simple questions on the most common features of daily life • convey minimal meaning to interlocutors experienced with dealing with foreigners by using isolated words, lists of words, memorized phrases and some personalized recombinations of words and phrases • satisfy a very limited number of immediate needs

Source: American Council on the Teaching of Foreign Languages. (1999). *ACTFL proficiency guidelines -- speaking: Revised 1999*. Hastings-on-Hudson, NY: Author. p. 10.

Appendix C: The ACTFL OPI Assessment Criteria--Speaking

Proficiency Level	Global Tasks & Functions	Context / Content	Accuracy	Text type
Superior	Discuss topics extensively, support opinions and hypothesize. Deal with a linguistically unfamiliar situation.	Most formal and informal settings / wide range of general interest topics and some special fields of interest and expertise.	No pattern of errors in basic structures. Errors virtually never interfere with communication or distract the native speaker from the message.	Extended discourse
Advanced	Narrate and describe in major time frames and deal effectively with an unanticipated complication.	Most informal and some formal settings/ topics of personal and general interest.	Understood without difficulty by speakers unaccustomed to dealing with non-native speakers.	Paragraph
Intermediate	Create with language, initiate, maintain, and bring to a close simple conversations by asking and responding to simple questions.	Some informal settings and limited number of transactional situations / predictable, familiar topics related to daily activities.	Understood, with some repetition, by speakers accustomed to dealing with non-native speakers.	Discrete sentences
Novice	Communicate minimally with formulaic and rote utterances, lists and phrases.	Most common informal settings / most common aspects of daily life.	May be difficult to understand, even for speakers accustomed to dealing with non-native speakers.	Individual words and Phrases

Source: ACTFL. (1999). ACTFL oral proficiency interview tester training manual. p. 31

Appendix D: The SPEAK Scoring Key

Overall Comprehensibility

0 - 90	Overall comprehensibility too low in even the simplest type of speech.
100 - 140	Generally not comprehensible because of frequent pauses and /or rephrasing, pronunciation errors, limited grasp of vocabulary, or lack of grammatical control.
150 - 190	Generally comprehensible but with frequent errors in pronunciation, grammar, choice of vocabulary items, and with some pauses or rephrasing.
200 - 240	Generally comprehensible with some errors in pronunciation, grammar, choice of vocabulary items, or with pauses or occasional rephrasing.
250 - 300	Completely comprehensible in normal speech, with occasional grammatical or pronunciation errors in very colloquial phrases.

Subcategories:

Pronunciation

0	:	Frequent phonemic errors and foreign stress and intonation patterns that cause the speaker to be unintelligible.
1	:	Frequent phonemic errors and foreign stress and intonation patterns that cause the speaker to be occasionally unintelligible.
2	:	Some consistent phonemic errors and foreign stress and intonation patterns, but speaker is intelligible.
3	:	Occasional nonnative pronunciation errors, but speaker is always intelligible.

Grammar

0	:	Virtually no grammatical or syntactical control except in simple stock phrases.
1	:	Some control of basic grammatical construction but with major and /or repeated errors that interfere with intelligibility.
2	:	Generally good control in all constructions with grammatical errors that do not interfere with overall intelligibility.
3	:	Sporadic minor grammatical errors that could be made inadvertently by native speakers.

Fluency

0 : Speech is so halting and fragmentary or has such a nonnative flow that intelligibility is virtually impossible.

1 : Numerous nonnative pauses and/or a nonnative flow that interferes with intelligibility.

2 : Some nonnative pauses but with a more nearly native flow so that the pauses do not interfere with intelligibility.

3 : Speech is smooth and effortless, closely approximating that of a native speaker.

Comprehensibility

0 : Overall comprehensibility too low in even the simplest type of speech.

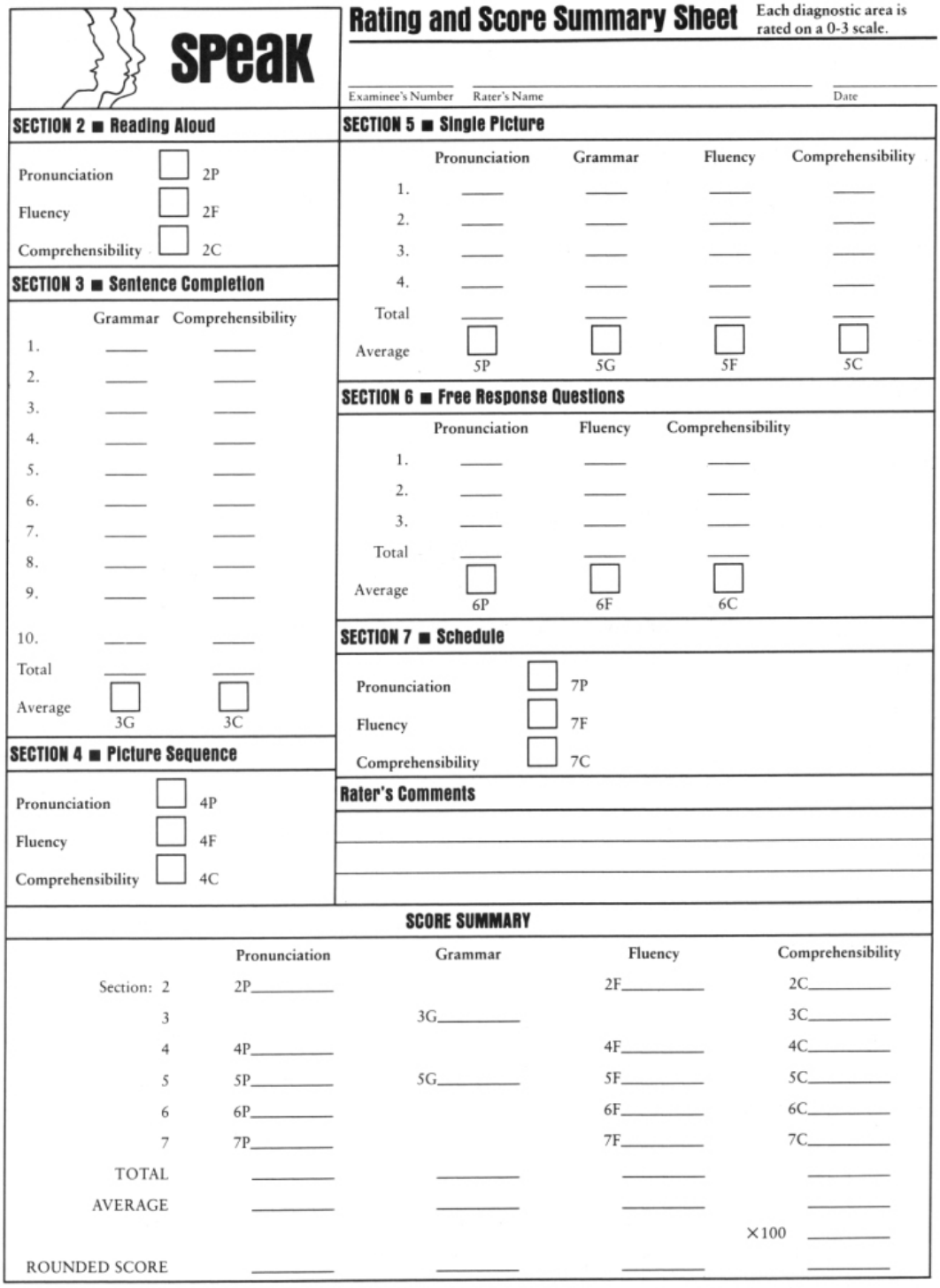
1 : Generally not comprehensible because of frequent pauses and /or rephrasing, pronunciation errors, limited grasp of vocabulary, or lack of grammatical control.

2 : Comprehensible with errors in pronunciation, grammar, or choice of vocabulary items, or infrequent pauses or rephrasing.

3 : Completely comprehensible in normal speech with occasional grammatical or pronunciation errors.

Source: Educational Testing Service. (1982a). *Guide to SPEAK*. p. 8 & 16.

Appendix E: The SPEAK Rating and Score Summary Sheet



SPEAK

Rating and Score Summary Sheet Each diagnostic area is rated on a 0-3 scale.

Examinee's Number _____ Rater's Name _____ Date _____

SECTION 2 ■ Reading Aloud

Pronunciation	☐ 2P
Fluency	☐ 2F
Comprehensibility	☐ 2C

SECTION 3 ■ Sentence Completion

	Grammar	Comprehensibility
1.	____	____
2.	____	____
3.	____	____
4.	____	____
5.	____	____
6.	____	____
7.	____	____
8.	____	____
9.	____	____
10.	____	____
Total	____	____
Average	☐ 3G	☐ 3C

SECTION 4 ■ Picture Sequence

Pronunciation	☐ 4P
Fluency	☐ 4F
Comprehensibility	☐ 4C

SECTION 5 ■ Single Picture

	Pronunciation	Grammar	Fluency	Comprehensibility
1.	____	____	____	____
2.	____	____	____	____
3.	____	____	____	____
4.	____	____	____	____
Total	____	____	____	____
Average	☐ 5P	☐ 5G	☐ 5F	☐ 5C

SECTION 6 ■ Free Response Questions

	Pronunciation	Fluency	Comprehensibility
1.	____	____	____
2.	____	____	____
3.	____	____	____
Total	____	____	____
Average	☐ 6P	☐ 6F	☐ 6C

SECTION 7 ■ Schedule

Pronunciation	☐ 7P
Fluency	☐ 7F
Comprehensibility	☐ 7C

Rater's Comments

SCORE SUMMARY

	Pronunciation	Grammar	Fluency	Comprehensibility
Section: 2	2P_____		2F_____	2C_____
3		3G_____		3C_____
4	4P_____		4F_____	4C_____
5	5P_____	5G_____	5F_____	5C_____
6	6P_____		6F_____	6C_____
7	7P_____		7F_____	7C_____
TOTAL	_____	_____	_____	_____
AVERAGE	_____	_____	_____	_____
				×100 _____
ROUNDED SCORE	_____	_____	_____	_____

Source: Educational Testing Service. (1982a). *Guide to SPEAK*. p. 17.

Appendix F: The SPEAK Format and Section Description

Format of the Test

The speaking proficiency test included in SPEAK consists of seven sections, each requiring a different speaking activity. The first section is an unscored "warm-up" in which the examinee responds orally to a few brief biographical questions provided on the test tape.

In the second section, the examinee is allowed time for preliminary silent reading of a passage of about 125 words and then is instructed to read the passage aloud. Scoring is based on pronunciation and overall clarity of speech.

In the third section, the examinee is asked to complete a series of 10 partial sentences in a way that conveys meaning and is grammatically correct.

In the fourth section of the test consists of six line drawings that tell a continuous story. After studying the drawings briefly, the examinee is asked to tell the story that is depicted, using past tense narration.

In the fifth section, the examinee looks at a single line drawing and answers several spoken questions about the picture.

In the sixth section consists of a series of spoken questions intended to elicit relatively free and somewhat more lengthy responses. Questions requiring both straightforward descriptions of common objects and fairly open-ended expressions of opinion are included. The linguistic quality and adequacy of communication, not the factual content of the responses, are at issue in scoring.

In the seventh and final section, the examinee sees a printed schedule, such as the outline for a course or a conference, and is asked to describe the schedule aloud, as though informing a group of listeners.

Scores. Each examinee receives four different scores: an overall comprehensibility score and scores for each of three diagnostic areas—pronunciation, grammar, and fluency. Overall comprehensibility scores are based on a scale ranging from 0 to 300; each of the three diagnostic area scores is based on a scale ranging from 0.0 to 3.0.

Source: Educational Testing Service. (1982a). *Guide to SPEAK*. p. 7.

Appendix G: Sample test items for the SPEAK

GENERAL INFORMATION

The purpose of the SPEAK® test is to determine the spoken English proficiency of people whose native language is not English. The test is given in one session and is divided into seven parts. For some parts of the test, you will read material that is printed in a test book; however, directions and questions for the entire test will be on an audiotape. All your answers to the test questions will be recorded on another tape. You will not have to write anything for the test.

TAKING THE TEST

When you take the SPEAK test, you will be given a test book and asked to read a set of general directions before you begin. Each section of the test has special instructions. These instructions are similar to those given for the practice questions that follow. It is a good idea to become familiar with the instructions before the day of the test.

The whole test is recorded. The voices on the test tape will ask you various questions. When you answer, you will speak into a microphone, and your answers will be recorded on another tape. To help you overcome any nervousness, you may wish to practice recording your voice before the day of the test.

The actual testing time is about 20 minutes. The test tape will announce the beginning and the end of each section. Listen carefully to each question and answer it when the voice on the tape tells you to do so. Speak in a natural tone of voice, but loudly enough for the machine to record clearly what you say. Your test scores will be based on what is recorded on the tape. **Do not stop your tape recorder at any time during the test unless you are told to do so by the test supervisor.**

PRACTICE QUESTIONS

The following practice questions and the instructions for each test section are similar to those you will find in the test. To get the most benefit from these practice questions, try to answer them just as you would during the actual test.

SECTION ONE: Directions

In this section of the test, you will be asked to answer some questions about yourself. After each question, you will have a short time to answer the question. On the actual test, you will have approximately 15 seconds to answer each question. Be sure to speak clearly after you hear each question.

Here are some sample questions:

1. What is your name?
2. How many brothers and sisters do you have?

In the actual test, questions in Section One will **NOT** be printed in the test book; you will hear them on the audiotape.

SECTION TWO: Directions

In this section, you will be asked to read a printed paragraph aloud. First, you will be given one minute to read the paragraph silently to yourself. Then, you will have one minute to read the paragraph aloud. Now, begin reading the paragraph below silently to yourself.

RADIO BROADCASTING

Radio reached its peak as a form of entertainment in the 1930's and 1940's. Every broadcast, no matter how modest, required a large number of talented people to get it on the air, from writers and directors to sound engineers and musicians. The unsung heroes and heroines were the people responsible for sound effects, those wizards who could produce any sound—from a teacup being placed in its saucer to a raging fire. In the early days, few effects were prerecorded, so the staff experimented to make the sounds just right. For example, they tried hitting hollow coconut shells on a flat surface to imitate the clip-clop sound of a horse's trot. Sound effects helped listeners to imagine that everything being portrayed was truly happening.

Now read the paragraph aloud with expression. Allow yourself exactly one minute to read the paragraph aloud.

SECTION THREE: Directions

In this section, you will see partial sentences and will be asked to make complete sentences using these parts. Look at Example X:

Example X: When the library opens. . .

There are a number of possible completions for this sentence. You could say, for example:

When the library opens, I will return the book.
OR
When the library opens, the students will go there to study.
OR
When the library opens, Mary will look for a new novel.

These are only sample completions. There are many other possibilities. Try to make the completed sentence meaningful and grammatically correct.

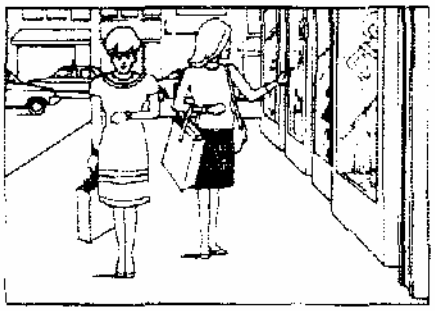
Now, complete each of the three partial sentences that follow. During the actual test you will hear only the number of each question. Speak when you hear the number, and be sure to say the complete sentence.

1. Whenever John comes home. . .
2. Before we left for class. . .
3. Because the restaurant is closed. . .

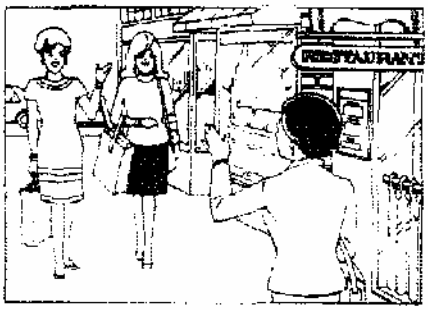
SECTION FOUR: Directions

In this section, you will see a series of pictures that tell a continuous story. You will b asked to tell the story that the pictures show. First, you will have one minute to study each of the following pictures silently, beginning with picture number 1 and going through picture number 6. Then, you will have one minute to tell the story. In the actual test, you will be told how to begin your story.

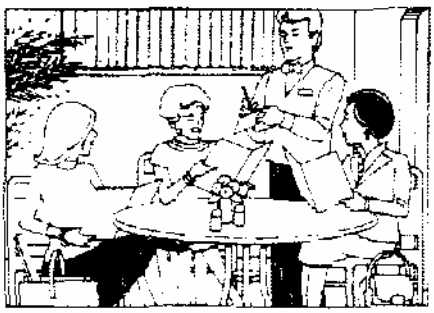
Now study the pictures on this page for exactly one minute and then tell the story in exactly one minute.



1

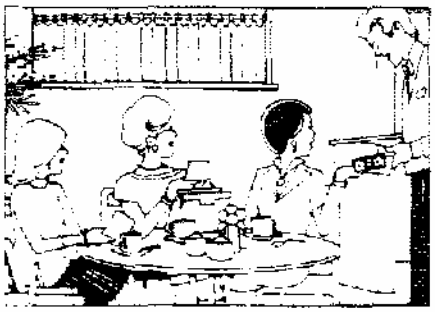


2

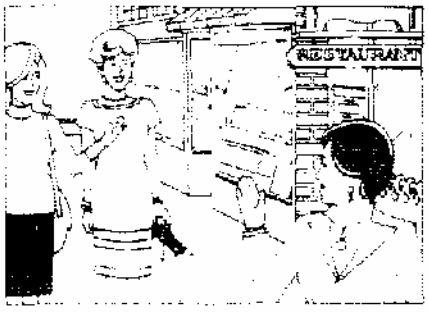


3

4



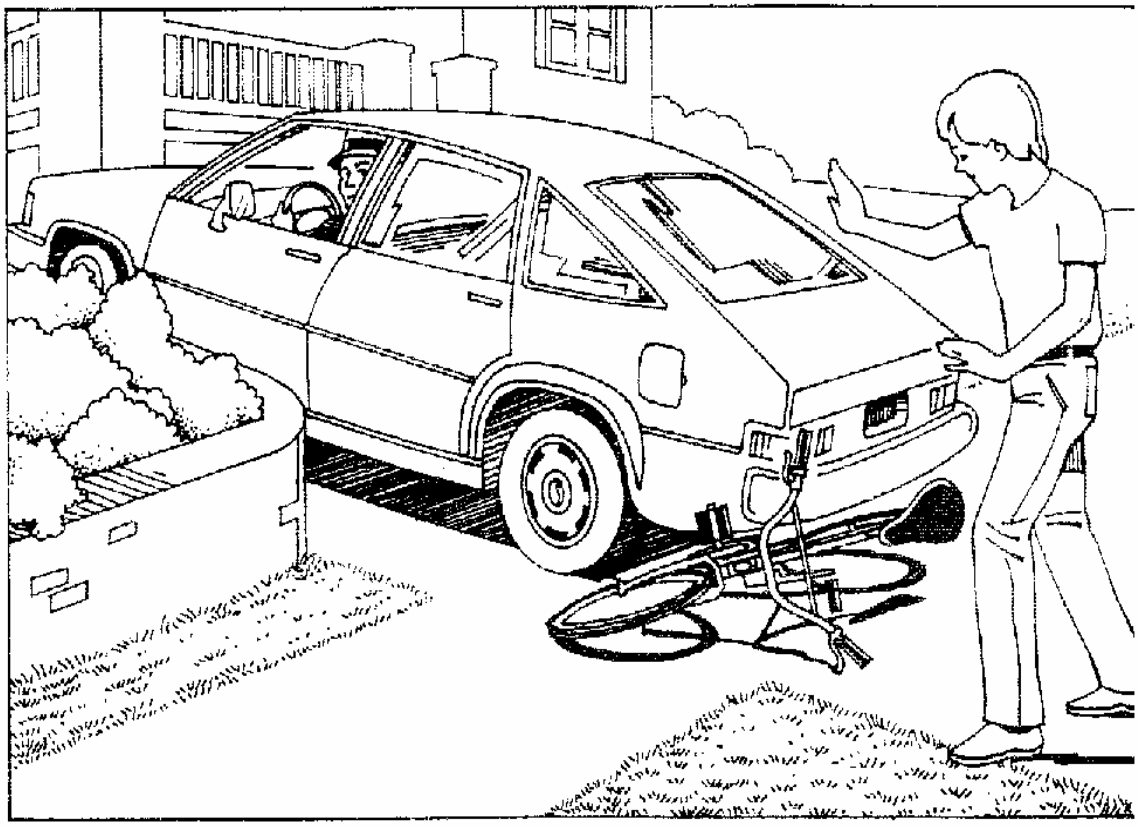
5



6

In this section, you will be asked four questions about a picture. There are many different ways each question can be answered correctly. However, you should answer each question as completely as possible. You will have 30 seconds to study the picture silently before you hear the questions.

First, study the picture below silently for 30 seconds. Then answer each of the following practice questions. You will be given approximately 12 seconds to answer each question.



1. Where is this scene taking place?
2. What has just happened?
3. What will the boy probably do after this?
4. How could this situation have been avoided?

(In the actual test, the questions in Section Five will not be printed in the test book. You will hear them on the audiotape.)

SECTION SIX: Directions

In this section, you will be asked to describe certain objects and to give your opinions on topics of general interest. Be sure to say as much as you can in responding to each question.

Remember that this is simply a test of spoken English. The graders will be interested in how you express your ideas, not the actual ideas. If you are using SPEAK Test Form 5 or 6, you will be given 15 seconds to prepare your answer to each question and approximately 45 seconds to answer the questions. If you are using SPEAK Test Form 2, 3, or 4, you will need to begin answering immediately after the question is asked.

1. Describe the things that make a perfect day.

2. Describe a telephone in detail.

3. The number of automobiles being manufactured in the world increases yearly. As a consequence, air pollution has also increased. How do you think the problem of automobile pollution should be handled?

In the actual test, the questions in Section Six will not be printed in the test book. You will hear them on the audiotape.

SECTION SEVEN: Directions

In this section of the test, you will see a schedule or a notice containing information about a club, conference, contest, etc. You will be asked to explain the schedule or notice to a group of people. For example, the schedule below describes the activities c a nature club. Imagine that you are the club president and must explain the schedule t the club members. Remember to include all important details in your description. Firs you will have one minute to silently study the information given below.

In your presentation, do not just read the information printed, but present it as if you were talking to a group of people. Now, make your presentation of the schedule. Allow yourself exactly one minute.

MOUNTAINVILLE NATURE CLUB

Quarterly Meeting:	April 15, 7:30 p.m. Mountainville Nature Center 58 Fairview Drive
Speakers:	Professor Alice Welton Biology Department—State University "The American Bald Eagle: An Endangered Species" Mr. Kenneth Shelby Author of *Focus on Nature* "Photographing Animals Close Up"
Future Travel Plans:	October 3-5 Florida Everglades Total Cost: $500 Contact: Peter Jenkins, Tour Director

Appendix H: Rater 1 vs. Rater 2 Independent Scores

Test-taker ID	Rater 1	Rater 2
t07	230	230
t11	200	200
t20	140	140
t33	230	240
t48	130	130
t51	220	220
t59	220	220
t61	300	300
t62	230	230
t77	300	300
t89	170	170
t99	280	290
t102	220	220
c05	120	120
c11	220	230
c12	260	260
c18	150	150
c19	240	240
c22	160	170
c25	240	230
c59	190	190
c85	280	280
c89	160	160
c102	190	200
c103	220	220

Appendix I: The Participant Questionnaire

Date of participation: _____ ____ ***2005*** ***Code:***
mm ***dd***

Participant Questionnaire

This study investigates spoken English tests. Your views will help non-native English-speaking students to receive fair and accurate evaluations. This questionnaire was adapted from the work of the ETS (1982b), Eignor et al., (1998), Hill (1998), Keynon & Malabonga (2001), and Jeong (2003). The questionnaire has four parts: Part I is about your experience of taking a spoken English test. Part II asks you about your computer skill. Part III asks about your English study experience. Part IV asks you for some background information.

Remember that all the information you provide will be kept completely confidential. So, your name will NOT be published. Your answers will be assigned a code number that will be used to analyze data. Your information and that of all study participants will thus be anonymous. Please take about 15 minutes to answer the questions. Please answer EVERY question as truthfully as you can.

Part I. Your experience of taking a spoken English test

Please specify below test name, month and year that each test was taken (e.g., "Test of Spoken English" by ETS, May 2003).

1. Specify test name, month, and year when you took spoken English tests on a computer before today.

	Test name	Month	Year
A.			
B.			
C.			

2. Specify test name, month, and year when you took spoken English tests on an audio-tape player before today.

	Test name	Month	Year
A.			
B.			
C.			

3. Specify test name, month, and year when you took spoken English tests in live face-to-face interviews before today.

	Test name	Month	Year
A.			
B.			
C.			

Part II. Your Level of Computer skill

Please specify the frequency of your computer use by circling the best response in the respective column (e.g., daily means at least once a day). Answer each question.

	My use of a computer	Frequency				
		Never	Yearly	Monthly	Weekly	Daily
4.	I access a computer at home…….	1	2	3	4	5
5.	I access a computer at school……	1	2	3	4	5
6.	I use a computer to send or receive email……	1	2	3	4	5
7.	I use a computer to read or write articles on website bulletin boards…….……	1	2	3	4	5
8.	I use a computer to write academic papers or assignments..	1	2	3	4	5
9.	I use a computer to prepare for presentations (e.g., PowerPoint)….	1	2	3	4	5
10.	I use a computer to make a database (e.g., Excel or Access)…….…..	1	2	3	4	5
11.	I use a computer to listen to music or to watch movies (e.g., DVD)…	1	2	3	4	5
12.	I use a computer to participate in online chats…….	1	2	3	4	5
13.	I use a computer to manage a web page…….	1	2	3	4	5
14.	I use a computer for advanced webpage authoring using either HTML source code or Java……	1	2	3	4	5
15.	I use computer programming languages (e.g., C++, Pearl)……..	1	2	3	4	5
16.	I use a computer to create multimedia projects using video/audio editing…….	1	2	3	4	5
17.	I use a computer to create interactive applications or projects similar in complexity to a computerized speaking test…….....	1	2	3	4	5
18.	I use VOIP telephony technologies such as SKYPE……..	1	2	3	4	5
19.	Other: please specify:	1	2	3	4	5

Please answer the questions by circling the response below that best describes your opinion about yourself.

		Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
20.	I am generally comfortable using a computer……	1	2	3	4	5	6
21.	I enjoy exploring the capabilities of different computer operating systems (e.g., Windows, Macintosh)……	1	2	3	4	5	6
22.	I am comfortable browsing the Internet……	1	2	3	4	5	6
23.	I am comfortable solving software problems on a computer……...	1	2	3	4	5	6
24.	I am comfortable solving serious hardware problems on a computer (e.g., replacing broken hardware)……..…	1	2	3	4	5	6
25.	Overall, I would rate my own ability as a computer user as (circle one)…….	Poor, Fair, Good, Excellent					

Part III. Your English study experience

26. When I was _____ years old, I began learning English.

27 – 54 Please check (✓) to indicate the following about yourself:

When I was pre-elementary, I:

27. had English classes at regular school
___Never ___1-2 times a week ___3 times or more a week
28. had English class at private language institution
___Never ___1-2 times a week ___3 times or more a week
29. had private tutoring
___Never ___1-2 times a week ___3 times or more a week
30. visited or stayed in English-speaking countries
___Never ___less than 6 months ___6 months to 1 year ___more than 1 year

When I was enrolled in elementary school, I:

31. had English classes at regular school
___Never ___1-2 times a week ___3 times or more a week
32. had English class at private language institution
___Never ___1-2 times a week ___3 times or more a week
33. had private tutoring
___Never ___1-2 times a week ___3 times or more a week
34. visited or stayed in English-speaking countries
___Never ___less than 6 months ___6 months to 1 year ___more than 1 year

When I was enrolled in middle and/or high school, I:

35. had English classes at regular school
___Never ___1-2 times a week ___3 times or more a week
36. had English class at private language institution
___Never ___1-2 times a week ___3 times or more a week
37. had private tutoring
___Never ___1-2 times a week ___3 times or more a week
38. visited or stayed in English-speaking countries
___Never ___less than 6 months ___6 months to 1 year ___more than 1 year

When I was enrolled in college and/or graduate school I:

39. had English classes at regular school
___Never ___1-2 times a week ___3 times or more a week
40. had English class at private language institution
___Never ___1-2 times a week ___3 times or more a week
41. had private tutoring
___Never ___1-2 times a week ___3 times or more a week
42. visited or stayed in English-speaking countries
___Never ___less than 6 months ___6 months to 1 year ___more than 1 year

43. I have been enrolled in ____elementary, ____middle, or ____high school (check ✓ for all applicable) where most subjects are taught in English for a total of _____ months _____years.

44. I have been enrolled in college or university instruction in English-speaking countries for a total of _____ months _____years.

45. I have lived in English-speaking countries for a total of _____ months _____ years.

46. I have been studying English for a total of ____ months ____ years.

IV. Your Background Information

47. Name: (optional)

48. Year of birth:

49. Academic major:

50. Gender (circle one): Female Male

51. Native language:

52. Home country:

53. Current academic status at the OSU (check only one):

MA _____1st year, _____2nd year, _____3+ year

PhD _____1st, _____2nd, _____3rd, _____4th, _____5+ year

Thank you for your participation in this study.

Appendix J: Recruitment letter

Title: Get free evaluation of your spoken English ability and enter prize drawings for $50

Dear international students,

Greetings. I am Eunjyu Yu, a doctoral candidate in Foreign/Second Language Education at the Ohio State University. My academic advisor, Dr. Charles R. Hancock, and I are conducting research to investigate the effect of different spoken English test format procedures.

I am currently seeking participants for this study. The benefits of participation in this study include a free trial of the SPEAK test or a new speaking proficiency test, free evaluation of your spoken English ability, and prize drawings for $50. Three prize drawing winners will be announced on January 20th, 2006. In return, you will be asked to take a 20-minute spoken English test and fill out a questionnaire. The whole procedure will last no longer than one and a half hours. This study will be conducted during Autumn Quarter 2005.

Your participation is very valuable to help you and future non-native English-speaking students because the results of this study will contribute to developing fair and accurate spoken English tests.

Remember that all the information you provide will be kept entirely confidential. Your name will NOT be published. A code number will be used to analyze all data obtained for this study.

Participation eligibility is restricted to the following categories. To participate in this study you MUST meet all of the following condition:

1. Your native language must not be English.
2. You should be between 20 and 35 years old.
3. You should be enrolled in full-time or part-time study in graduate degree program during Autumn Quarter 2005. The students who major or minor in foreign/second language education are excluded.
4. The Ohio State Spoken English Program has requested that you not participate in this study if you plan to take the SPEAK test right after participating this study during Autumn Quarter 2005 for an official record required for a TA position at the OSU. You, however, CAN participate in this study if you already took the SPEAK test.

If you are interested in participating please contact Eunjyu Yu via email at yu.211@osu.edu. Include the following information: your 1) native language, 2) age, 3) gender, 4) academic major, and 5) computer skills.

Your participation will be very much appreciated.
Sincerely,

Appendix K. Consent form

THE OHIO STATE UNIVERSITY

College of Education

146 Arps Hall
1945 North High Street
Columbus, OH 43210

Phone: (614) 247-7806

Protocol #: 2005E0542

CONSENT FOR PARTICIPATION IN RESEARCH

I, _____, consent to participating in research entitled:
Printed Name
"A comparative study of effects of a computerized English oral proficiency test format and a conventional SPEAK test format."

Charles R. Hancock, Principal Investigator, or his authorized representative, Eunjyu Yu, has explained the purpose of the study, the procedures to be followed, and the expected duration of my participation. Possible benefits of the study have been described, as have alternative procedures, if such procedures are applicable and available. I understand that my response to the spoken English proficiency test will be recorded on an audio-tape or CD.

I acknowledge that I have had the opportunity to obtain additional information regarding the study and that any questions I have raised have been answered to my full satisfaction. Furthermore, I understand that I am free to withdraw consent at any time and to discontinue participation in the study without prejudice to me.

Finally, I acknowledge that I have read and fully understand the consent form. I sign it freely and voluntarily. A copy has been given to me.

Signed: _____ Date: _____
Participant

Signed: _____
Principal Investigator or his authorized representative

HS-027E Consent for Participation in Exempt Research

Appendix L. Letter of support

THE OHIO STATE UNIVERSITY

Office of the University Registrar

320 Lincoln Tower
1800 Cannon Drive
Columbus, OH 43210-1233

Phone 614-292-8500
FAX 614-292-7199
E-Mail REGISTRAR@OSU.EDU

Institutional Review Board
The Office of Responsible Research Practices
The Ohio State University
1960 Kenney Rd.
Columbus, OH 43210

October 7, 2005

Dear Madam or Sir:

Ms. Eunjyu Yu, who is a doctoral candidate in Foreign/Second Language Education at the Ohio State University, has contacted the Office of the University Registrar requesting our support for her research related to the effect of different formats of spoken English test procedures.

The information she is requesting from the Registrar's Office would include e-mail addresses for international graduates students enrolled at the Ohio State University.

The Office of the University Registrar has agreed to provide Ms. Yu with this information for her project pending approval or exemption by the Institutional Review Board.

Sincerely,

Linda S. Katunich
Manager
Student Enrollment Reporting and Research Services
Office of the University Registrar
The Ohio State University

Appendix M: Letter of permission

THE OHIO STATE UNIVERSITY

College of Education

60 Arps Hall
1945 North High Street
Columbus, OH 43210
Phone: (614) 292-5005

PERMISSION

I, Susan Sarwark, Director of the Spoken English Program in the Ohio State University, give permission for Eunjyu Yu, a doctoral student in the School of Teaching and Learning at OSU, to use the SPEAK test items for her doctoral dissertation research. She has explained the study, its purpose, procedures and possible benefits. I understand that the test items will be used solely for the dissertation research purposes and that the test items used in this study will be destroyed as soon as the study is completed.

Signature

Date

Printed Name