CHAPTER ONE

INTRODUCTION

1.1 Context of the problem

Testing speaking is complex and challenging. For decades, language testers have been searching for better methods of testing second language speaking. The use of oral interviews has been a widespread practice for many years. However, in addition to long-standing concerns over validity (e.g., Young & He, 1998; Johnson, 2000) and reliability (e.g., Lazaraton, 1996) of this method, its practical problems, such as being labor-intensive in terms of administration, have led to the development of speaking tests using various forms of technology (e.g., a tape-recorder).

Recent improvements in computer technology have encouraged the delivery of speaking tests through computers. The computer-delivered speaking test has gained popularity, particularly in large-scale tests, for its cost-efficiency, ease for scoring, and flexibility in presenting tasks with various sources of multi-media input. However, these advantages do not provide sufficient grounds for its use in preference to other conventional tests. Empirical evidence is clearly needed to demonstrate its validity from various aspects to justify its use.

Unlike other types of speaking test, a computer-delivered test requires the examinee to talk to a computer without an interlocutor to interact with. This feature, however, has raised several issues regarding the test’s validity. Of primary concern is the extent to which examinees perform differently compared with when they take a face-to-face test, which has been considered to have superior authenticity over other
types of speaking test (Brown, 2005). Exploration of this issue is of importance, because it concerns a fundamental question of test validation, namely, to ensure the score interpretation.

In test validation, language testers have long held an interest in specifying and minimizing the factors that confound score interpretation. In recent years, the effect of computer delivery mode has attracted much attention from researchers as a major potential source of construct-irrelevant variance for computer-based tests. The most "ubiquitous" concern raised about technology for language assessment, according to Chapelle and Douglas (2006), is that examinees' performance on a computer test may not reflect the same ability as that measured by other forms of assessment. Some researchers have called for more research to be carried out on the effect of computer delivery mode (e.g., Chalhoub-Deville & Deville, 1999; Alderson, 2004), particularly by comparing the performance on the computer test with that on the conventional types of test (Chapelle, 2003).

A substantial number of studies have investigated the effect of computer delivery mode on testing reading (e.g., Yessis, 2000; Higgins, Russell, & Hoffmann, 2005), listening (e.g., Kasten, 1995; Coniam, 2001; Coniam, 2006; Ockey, 2007), and writing (e.g., Lee, 2002; Lee, 2004; Wolfe & Manalo, 2005; Li, 2006) by comparing test performance on computer-delivered tests with that on paper-and-pencil tests. Due partly to the fact that the computer-delivered speaking test is a new phenomenon, studies that focus on the effect of computer delivery mode on testing speaking are few in number and, to the author's knowledge, none have made comparisons to face-to-face tests.

In the context of testing speaking, there could be variability in examinees' performance due to the change of delivery mode in both test scores and examinees'
speech samples.

Turning first to test scores, the issue is whether an examinee could obtain a score from the computer mode that was comparable to the score from the face-to-face mode. Also, psychometrically, do the speaking tests delivered by the two test modes measure a similar speaking ability? The need to provide evidence of score equivalence for any new computer-delivered test with conventional versions has been underscored by professional standards, such as those of the American Psychological Association (1986). Test end users also need the information to make a direct comparison of examinees who have taken a speaking test in different modes. In addition to comparing test scores directly, examining how the underlying constructs differ across modes will provide us with valuable insights into the interpretation of the scores produced with the computer mode.

The second aspect of performance difference is associated with examinees' speech sample, particularly in terms of linguistic performance. The question arises of the extent to which the language produced by examinees varies across modes. Several researchers have cast doubts on the validity of technology-based speaking tasks as instruments for eliciting linguistic performance (Iwashita, McNamara, & Elder, 2001; Elder & Iwashita, 2005). Their concerns could be justified with evidence from the second language acquisition (SLA) research (Skehan, 1998; Ortega, 2005), according to which language learners may give utterances that are more accurate and complex but less fluent when a listener is absent. However, whether this proposal is applicable to a test situation still remains uninvestigated.

One characteristic of examinees that needs to be considered to understand the comparability of speech samples across modes is examinees' proficiency. The concern
is whether examinees of a certain proficiency level would be disadvantaged by the change of delivery mode. It is important to explore the possible interaction effect of examinees’ proficiency with delivery mode on test performance, because it affects the fairness of the test (Messick, 1989; Bachman, 1990). However, how the examinee’s proficiency affects the potential variability in speech samples across modes is another unanswered question.

In addition to concerns over equivalence of examinees’ speaking performance across the two test modes, another issue that emerges with the introduction of the computer in testing speaking is the face validity of the test. Does a speaking test delivered by computer seem acceptable to examinees? A test that does not appear to be valid to examinees may not be taken seriously, which in turn would affect the response validity of the test (Alderson, Clapham, & Wall, 1995). Since oral communication is mainly a human phenomenon, how examinees feel about talking to a computer is an important question that needs to be addressed as an integral part of test validation.

In contrast to the paucity of studies investigating the effects of computer delivery mode on testing speaking, there has been substantial research that has examined the comparability of tape-based speaking tests and face-to-face tests. Evidence from these studies has verified that the change of delivery mode may have an impact on underlying constructs (O’Loughlin, 2001), speech samples (e.g., Shohamy, 1994; Luoma, 1997; Koike, 1998), and examinee attitudes (e.g., Clark, 1988; Hill, 1998). However, as the computer-delivered test involves multi-media prompts, talking to a computer may be somewhat different from responding to a tape-recorder. Thus, whether the findings of these studies are generalizable to the computer-delivered speaking test is debatable.

Methodologically, these studies have suffered from several limitations that could
endanger their generalizability as well. One of the problems concerns the high sample dependency due to the relatively small sample sizes involved in most studies. Moreover, most research has been based on a comparison of tests including both monologic and interactive tasks, which makes the comparison of the two types of tests problematic and the interpretation of the results difficult.

The issue of the confounding factor of tasks also needs to be addressed when comparing the computer-delivered speaking test and the face-to-face test. Although computer-delivered speaking tests can replicate a realistic situation through a combination of audio and video prompts and can offer preset verbal feedback, doubts remain as to the feasibility of assessing interactive ability on the computer (Norris, 2001). In fact, the current technology of automatic speech recognition has not made it possible to deliver an interactive task through a computer. The computer is still not able to act as a fully-fledged conversational partner, who can give adequate verbal feedback to examinees' responses. A solution to this practical constraint is to focus on monologic tasks across modes rather than making a direct comparison of the two types of tests that include different task types.

In sum, despite the increasing use of the computer in the delivery of speaking tests and growing concerns over the effects of computer delivery mode on speaking test performance and examinee attitudes, research comparing computer-delivered speaking tests with face-to-face tests is scant. Concern about methodological drawbacks in the previous studies is another motivating factor for the proposed speaking assessment research.
1.2 Purpose of the study

The purpose of the present study was to explore the effects of computer delivery mode on testing second language speaking by comparing monologic tasks delivered in the computer and the face-to-face modes. This general issue was investigated from three aspects: (1) score comparability, (2) speech sample comparability, and (3) examinee attitudes toward testing speaking in the two modes.

More specifically, score equivalence was explored by comparing test scores and underlying constructs across modes. This study addressed the comparability of speech samples by focusing on linguistic features in terms of fluency, accuracy, and complexity, and additionally by examining the interactions of test modes with examinees’ proficiency. Examinee attitudes were investigated through their responses to questionnaire items regarding various aspects of testing speaking in the two modes and also through a direct comparison of the two modes.

1.3 Significance of the study

The findings of this study are expected to have both theoretical and practical significance for the fields of language assessment and second language acquisition (SLA).

The benefits of the findings on any observed performance differences between the computer mode and the face-to-face mode will be numerous. From a theoretical perspective, the study sheds light on examinees’ speaking test performance when an interlocutor is absent. As such, it enriches our understanding of the nature of speaking performance and accordingly the construct of speaking ability measured in the context
of computer-delivered assessment.

This research also provides information on the validity and score interpretation of the computer-delivered speaking test. It provides empirical evidence on the effects of computer delivery mode as a construct-irrelevant factor and thus contributes to the establishment of an argument for the validity of the computer-delivered speaking test. A better understanding of the construct of speaking ability helps to interpret test results.

Moreover, this research is unique in its attempt to explain possible differences in examinees’ speech samples within a theoretical framework in SLA research. The findings in turn assist researchers looking for validity evidence of computer-delivered speaking tasks as elicitation instruments in SLA research. In this regard, this study represents a response to calls from researchers (e.g., Bachman & Cohen, 1998; Shohamy, 1998) for interface studies linking language testing and SLA research.

With respect to test development, this study could contribute in two ways. First, it helps the development of a scoring rubric for computer-delivered speaking tests. If any discrepancy between the results on differences of speech sample and test scores on any rating elements emerges across modes, test developers could add those features into the scoring rubric. Second, feedback from examinees on taking the speaking test delivered by computer provides test developers with useful information on converting face-to-face tests to the computer mode.

Furthermore, the findings of this study are beneficial to examinees seeking information on the comparability of test scores across modes and to test end users wishing to make valid comparisons of speaking ability of examinees who have taken different modes of a speaking test.

Finally, this study employed a much larger sample than previous studies in an
attempt to improve the generalizability of the results. Given that the collection, rating, and transcription of oral data is rather time-consuming and labor-intensive, it is quite a common practice for comparative studies in speaking assessment research to use a sample of 10 to 20 subjects from one location. In the present study, efforts were made to collect the data from nearly 100 participants from several sources.