AN INVESTIGATION OF THE CROSS-MODE COMPARABIITY OF A PAPER AND COMPUTER-BASED MULTIPLE-CHOICE CLOZE READING ASSESSMENT FOR ESL LEARNERS

by

Dennis Murphy Odo

B.A. (Hon.), St. Francis Xavier University, 1999
M.A., University of Leicester, 2005
M.S.Ed., D'Youville College, 2007

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

(Language and Literacy Education)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

April 2012

Abstract

This study was designed to determine whether a computer-based version of a standardized cloze reading test for second language learners is comparable to its traditional paper-based counterpart and to identify how test takers' computer familiarity and perceptions of paper and computer-based tests related to their performance across testing modes. Previous comparability research for second language speakers revealed that some studies found that the two forms are comparable while others found they are not. Findings on the connection between computer attitudes and computer test performance were also mixed.

One hundred and twenty high school ELL students were recruited for the study. The research instruments included both paper and computer-based versions of a locally-developed reading assessment. The two tests are the same in terms of content, questions, pagination and layout. The design was a Latin squares so that two groups of learners took the tests in the opposite order and their scores were compared. Participants were also asked to complete questionnaires about their familiarity with computers and their perceptions of each of the two testing modes.

Results indicate that the paper and computer-based versions of the test are comparable. A regression analysis showed that there is a relationship between computer familiarity and computer-based LOMERA performance. Mode preference survey data pointed to differences in preferences depending on each unique test feature. These results help validate the cross-mode comparability of assessments outside of the traditional discrete-point multiple choice tests which tends to predominate in current research.

Preface

This study was reviewed by the UBC Behavioral Research Ethics Board (Full Behavioral REB) and an ethics certificate (ref: H10-03325) was obtained to carry out the research.

Table of Contents

List of Tables

List of Figures

Acknowledgements

My thanks go to the many people at UBC who helped and encouraged me during my time here. In particular, I would like to acknowledge Drs. Margaret Early and Steven Talmy, my committee members, who provided an insightful perspective and invaluable feedback on my work.

I am also greatly indebted to the department for providing funding support and opportunities to continue to hone my teaching craft as I progressed through the program.

I am grateful to the many students and teachers who participated in the study. Without them the research would not have been possible, and their forbearance, good will, and insightful comments made all the difference.

I am especially appreciative to my wife and most ardent supporter Mina Kim. Thanks for being my sail and my ballast.

Lastly, I owe particular thanks to Dr. Lee Gunderson who continually demonstrated through actions rather than words the incalculable value of the wisdom and patience of a brilliant mentor.

For Mina Kim and Lee Gunderson.

Chapter 1 Introduction

The number of speakers of English as a second language (ESL) in Canada continues to grow. According to Statistics Canada (2010), by 2031, one in four Canadians will have been born in another country. Approximately 37% of learners in the Greater Vancouver school districts speak a first language other than English. This is a significant proportion of English language learners (ELL) and this trend is predicted to continue (Garnett & Aman cited in Canadian Counsel on Learning, 2008). Since Canadians have accepted these learners into Canadian society, we are responsible for helping them to succeed (Oikonomidoy, 2007). However, the alarming disappearance rates of English-as-a-second-language (ESL) learners from schools across Canada – Toronto, 53% (Radwanksi, 1987), Calgary, 73% (Watt & Roessingh, 2001) and the BC Lower Mainland, 60% (Gunderson, 2009) tells us that more needs to be done to support them in their learning. Menken (2008) argued that a significant factor in ensuring that immigrant students meet the learning targets set for them is putting appropriate and informative assessments in place.

Personnel in school districts across the lower mainland region of British Columbia employ a variety of diverse measures to assess reading ability. This practice presents a problem when students move between districts because there is very little meaningful assessment information that can be shared among schools across districts (Gunderson, Murphy Odo, D'Silva, 2010). An ESL Assessment Consortium (www.eslassess.ca) was formed to address this and a variety of related assessment issues. Members of the Consortium developed a standardized reading assessment measure for ESL and native speaking students that provides an indicator of ESL reading proficiency (Gunderson, Murphy Odo, & D'Silva, 2010). The main advantage of this measure is that it provides all of the districts participating in its development with a common,

locally-normed reading assessment that allows them to share information about reading proficiency when students move between districts. The consortium identified a measure of reading ability as a priority based on their experience of the strong connection between English language literacy and academic achievement (Gunderson, 2007). The assessment has been administered to thousands of students since 2009 in 12 different school districts with great success.

The assessment, known as the Lower Mainland English Reading Assessment (LOMERA), has served districts, teachers, and students well. It is a multiple-choice cloze that has also been validated by extensive research as an accurate indicator of reading proficiency (e.g., Espin & Foegen, 1996; Hale et al., 1989; Oller & Jonz, 1994). Consortium members agreed that the development of an online version of LOMERA was needed and that the issue of comparability between the paper-and-pencil version and the online version should be explored. In essence, the question was whether scores from the computer-based LOMERA would be comparable to those obtained from the traditional paper-based version. A careful review of relevant research literature did not provide a satisfactory answer to this question for an assessment like the LOMERA so further research seemed necessary. The present study was designed to explore the issue of comparability in addition to a number of other related questions.

1.1 Background

Second language assessments often wield considerable power over the educational trajectories of students' lives because they often determine eligibility to enter and exit programs or institutions (Shohamy, 2000). This is troubling because these tests can systematically discriminate against particular groups as a result of test bias in methods and materials (Menken, 2008). The use of computers in L2 literacy assessment, despite their apparent advantages, has not

been broadly adopted by applied linguists because of undeveloped hardware and software and apathy in the large testing programs (Chapelle, 2008). In fact, it was not until relatively recently (i.e. the mid 1990s) that computer-assisted language tests begin to be used for both large and small scale language assessment (Dooey, 2008). However, the use of computerized forms of assessments are rapidly expanding to the point where we may need to see appropriate use of technology as a facet of language ability. Chapelle and Douglas (2006) suggest that "communicative language ability needs to be conceived in view of the joint role that language and technology play in the process of communication," (p. 108) and we need to expand our view of language ability to see it as "the ability to select and deploy appropriate language through the technologies that are appropriate for a situation" (p. 107). Given these accelerating developments in language assessment technology, there is an urgent need for research that ensures computerized assessment tools being developed yield results that are valid, reliable and fair for all test takers.

1.2 Theoretical Model

It appears that most cross-mode comparability research has not used any kind of underlying explanatory framework. Trauth (2006) argues that implicit-theoretical research that neglects explicit discussion of a theory "makes it difficult for others to discuss, challenge or extend the research" (p. 1151). In this regard, schema theory provides a framework to explain findings related to any mode differences in cloze research because it is based on the notion that:

> [S]chemata, or knowledge already stored in memory, function in the process of interpreting new information and allowing it to enter and become part of the knowledge store…it is the interaction of new information with old knowledge that we mean when we use the term comprehension (Anderson & Pearson 1984, p. 37).

Schemata are thought to be integral to the reading comprehension process. The LOMERA is a test of reading ability so it seems logical to investigate whether test takers draw on the same cognitive schemas as they complete the LOMERA in each mode. Discrepancies in schema activation across modes could help explain differences in mode effect. In essence, theoretically, it is probable that there are mode or format schemata and that readers process text using these various schemata, including mode schemata.

1.3 Study Focus

This study was designed to achieve three main purposes. The first was to determine whether a computer-based version of a standardized multiple-choice cloze reading test for second language learners was comparable to its paper-based counterpart. The cross-mode comparability aspect of the study also addresses the suitability of schema theory for conceptualizing test takers' reading processes as they complete both the paper and the online versions of the LOMERA. The second aim was to establish whether there is a relationship between test takers' reported computer familiarity and their performance on the computer-based LOMERA. The third objective was to identify learners' preferences for the paper and computer-based test based on several test features such as readability, ability to choose and change answers, and their reasons for those preferences.

Results will contribute to the research literature by investigating potential differences in mode effect for test items other than the traditional discrete-point multiple-choice items that predominate in reading research. This study also addressed the relationship between assessment mode and perceptions of paper and computer-based testing with a type of second language assessment that has been overlooked in comparability research. A third reason for conducting

this research was to evaluate the judiciousness of large-scale test developers' decision to move language assessments online without enough research support. According to Chapelle and Douglas (2006), test developers often move tests to a computerized format to save money rather than because computerized tests are necessarily a more valid or authentic form of assessment. They contend that not enough evidence has been collected to make a determination either way. Lastly, most of the investigations of comparability have been done with university students taking large-scale tests like TOEFL. This study will investigate cross-mode comparability with secondary school ESL students who have historically not received as much research attention in cross-mode comparability. Results may also impact ESL assessment practices in the lower mainland because the validation of the comparability of both versions of the LOMERA will justify local district specialists' use of the computer-based version thus offering them more flexibility in their use of assessments.

1.4 Research Questions

- Is there a mode effect in ESL students' performance on paper- and computer-based versions of a multiple-choice cloze reading assessment?

- Do L2 learners who are more familiar with computers achieve higher scores on a computer-based multiple-choice cloze reading assessment than those who are less familiar with computers?

- Which test features are favored and what reasons do test takers provide for preferring certain test features in a particular mode?

1.5 Definition of Terms

1.5.1 Comparability

Choi, Kim and Boo (2003) define comparability research as an "investigation into the comparability of test methods or test tasks represented in different testing modes [i.e., paper and computer-based]" (p. 297). That is, the primary objective of comparability research is to ascertain whether test scores obtained for computer-based tests are comparable to those gotten from their traditional paper-based counterpart. Paek (2005) adds that

> Comparability studies explore the possibility of differential effects due to the use of computer-based tests instead of paper-and-pencils tests. Comparability studies help ensure that test score interpretations remain valid and that students are not disadvantaged in any way by taking a computerized test instead of the typical paper test (p 1-2).

This is the second aspect of comparability that ensures changes in mode do not disadvantage particular groups of test takers (e.g., those who are less familiar with computers).

1.5.2 Mode Effect

When a difference is identified between the paper-based and computer-based testing modes this is known as mode effect. Clariana and Wallace (2002) define mode effect as "empirical evidence that identical paper-based and computer-based tests will not obtain the same results" (p. 593). Typically, this takes the form of statistically significant discrepancies in test scores and associated deviation in rank and distribution of test takers.

1.5.3 Maze Procedure

The originators of the maze (otherwise known as multiple-choice cloze) procedure, Guthrie, Seifert, Burnham, and Caplan (1974), explain that, like the cloze, maze consists of a stand-alone piece of text that can be taken from a larger stretch of discourse such as a story or a nonfiction book. However, instead of deleting a word at a regular interval throughout the text, the test designer modifies the text by substituting three word choices where the blank would be. The test taker then has to select one of these three options. For example:

	some	
The farmer and his truck	swam	fast
	went	

Figure 1. Example maze-type test item

The administration procedure for the maze is that "The child reads the material silently and circles the alternatives which he [sic] believes are correct. The number or percentage that the child circles correctly indicates the level of his comprehension of that passage" (Guthrie, et al. 1974, p. 163). The maze is scored the same way as the cloze with one point being given for the test taker selecting the answer that the test designer considers to be correct. It was devised principally to overcome the cloze's perceived ambiguity with respect to the accuracy of some test taker supplied answers (Guthrie, et al. 1974). Therefore, unlike the traditional cloze, maze uses only exact word scoring.

1.6 Outline of Dissertation

This dissertation consists of six chapters. Chapter 1 provides a brief introduction and overview of the study presenting the development of the LOMERA, the design and focus of the

study including the research questions and defining several key concepts. Chapter 2 reviews the research literature that has explored key phenomena addressed in the study. Literature related to cloze testing with second-language learners, cross-mode comparability of language tests, perceptions of computerized test and the relationship between computer familiarity and computer test performance is examined. Chapter 3 contains an explanation of the methodology of the study including the research design, provides a rationale for the choice of mixed-methods, and describes the participants, the instruments, as well as the data collection and analysis procedures. Chapter 4 consists of a presentation of the results and discussion for each aspect of the cross-mode comparability component of the study. Chapter 5 is a review and discussion of the results for computer familiarity and test takers' perceptions of the computerized LOMERA. Chapter 6 includes conclusions and implications following from the results as well as possible limitations and suggestions for future work.

Chapter 2 Review of Relevant Research Literature

2.1 Introduction

This chapter begins with a discussion of schema theory and its application to test takers' reading processes during cloze testing. This is followed by a review of the research literature related to cloze-type assessments for second-language learners. Issues such as research support for and criticism of the validity of cloze tests as a measure of text-level reading comprehension are touched upon. This discussion includes commentary on other limitations and strengths of cloze as well as adaptations that have been made to the procedure (e.g. maze tests) based on the criticisms that have surfaced. This is followed by a review of the research into computers and cloze assessment. Studies of cross-mode comparability conducted with first and second language test takers are then discussed. Research into computer familiarity is also examined to provide some insight into the relationship between familiarity and computer test performance. Lastly, investigations that explored test taker perceptions of positive and negative features of paper and computerized tests are reviewed.

2.2 Schema Theory

2.2.1 An Overview of the Development of Schema Theory

The use of schema theory to explain cognition and comprehension processes has a long and productive history in psychological and educational research. According to Tracey and Morrow (2006), schema theory grows out of the constructivist theory of learning. In essence, constructivism emphasizes the active construction of knowledge by learners. For constructivists "learning occurs when individuals integrate new knowledge with existing knowledge" (p. 47). Two features of constructivism that are particularly germane to schema theory are the claim that learning is largely hypothesis testing (i.e. making and confirming predictions) and their emphasis

on the necessity of learners making inferences to "[fill] in the meaning gaps" (Ruddel & Ruddel, 1995, p. 54). Working in the tradition of Gestalt psychology, Bartlet (1932) was the first to use the term schema. He created the concept of schema to explain findings from his studies of memory of stories where he noticed that subjects supplied missing information from stories based on their own background and cultural knowledge.

The involvement of schemata in the first-language reading process was first acknowledged by Anderson and Pearson (1984). In their view, the role of schemata in the reading comprehension process was such that

> [S]chemata, or knowledge already stored in memory, function in the process of interpreting new information and allowing it to enter and become part of the knowledge store…it is the interaction of new information with old knowledge that we mean when we use the term comprehension (p. 37).

Essentially, they viewed schema as an "abstract knowledge structure" (p. 42) that was comprised on interconnected nodes of information with which new text information interacted.

Carrell and Eisterhold (1983) were among the first to reflect on the specific implications of schema theory for second language readers. Unique challenges for second-language readers they identified include limited cultural knowledge, incomplete content schema or lack of knowledge of the second language. All of these shortcomings often impede second language learners' comprehension of the text.

A comprehensive summary of the main criticisms that have been levelled against schema theory over the years has been provided by Grabe (2009). He reviewed several critiques presented by others (e.g. Carver 1992) such as how much of the evidence used to support schema

theory involves using unusually challenging texts and problem-solving tasks that do not represent the kind of reading that most people usually do.

Despite these critiques, prominent scholars continue to see great value in schema theory as an explanation for how readers use background knowledge to facilitate reading comprehension (e.g. Alderson, 2000; Eskey, 2005; McVee et al, 2005). For instance, Eskey (2005) mentions that although socio-cultural issues have gained increased prominence in the field of second language literacy, there is still much value in psycholinguistic conceptions of reading. He reminds us that "the reader's brain is not an empty container to be filled with meaning from the text" (p. 569). That is, readers have knowledge and understandings that they bring to a reading event that shape how they understand a given text. He maintains that a substantial aspect of the reading process is psycholinguistic and "reading as a psycholinguistic process, when performed successfully, entails both rapid and accurate decoding and construction of meaning based on prior knowledge [i.e. schema]" (p. 570).

Its robust research support and longevity in the field make schema theory a viable framework to explain the reading comprehension process from a psychological perspective. Schema theory is also a convincing explanatory framework for the mental processes at work as test takers complete a cloze passage for several reasons that will be outlined below.

2.2.2 The Relationship between Schema Theory and Cloze Testing

Klein-Braley (1997) provides insight into the underlying mental processes that might be at work as test takers complete a cloze test. She uses the Chomskyan distinction between competence and performance to argue that cloze (or any other) test performance is only a sample of test takers' actual linguistic competence. She explains her conception of competence as a "developing rule system or [level] of proficiency" (p. 53).

She refers to tests as “a controlled situation where a specific, objectively defined performance is required to order to obtain a sample which is a representative sample of the examinee’s performance” (p. 53). She then goes on to point out that test scores reflect test taker competence only indirectly because test results are always affected by such things as fatigue, concentration, motivation, testing conditions or any other factor that could be identified as a source of test error.

She explains that when test takers complete a cloze test they draw on a variety of types of linguistic cues such as phonology, morphology, syntax, collocations, pragmatic and logical clues (among others) in the text so that “In order to restore the original text the learners must make use of all the clues available in the remaining portions of the text and in their heads” (p. 52). In this way, test takers use textual cues in combination with their background knowledge to help them complete the cloze task.

Bailey (1998) complements this theory with the elaboration of two concepts initially drawn from Saussureian theory to explain the kinds of mental competence test takers use to complete a cloze task. The terms for these two postulated types of mental competence are syntagmatic and paradigmatic competence. Syntagmatic competence tells the reader the part of speech of the word being read. For instance, in the sentence “the cat ate the _____” readers would know that they have to put a noun in the blank though they may not yet know exactly what word. Paradigmatic competence makes readers aware of the necessary semantic features for the missing word. For example, in the same example sentence above, readers’ knowledge of the world tells them that they would probably not put “dog” into the blank (Bailey, 1998). These are the hypothesized types of knowledge that comprise readers' expectancies.

According to Cohen (1994), the main purpose of cloze is to measure reading skills in an interactive way. Test takers combine information from the text with their schematic knowledge to successfully fill the missing information in the cloze assessment. This is based on the logic that "in written language, a sentence with a word left out should have enough context that a reader can close that gap with a calculated guess, using linguistic expectancies (formal schemata), background experience (content schemata), and some strategic competence" (Brown & Abeywickrama, 2010, p. 241).

As discussed above, schema theory is a potentially valuable theoretical lens through which test takers' interaction with each version of the LOMERA might be viewed. It provides a credible and coherent explanation for how test takers integrate their background knowledge with the content of the maze passages to help them complete the test.

Current conceptions of schema theory typically distinguish between content and formal schema. However, it is proposed here that schema theory be expanded to include a "mode" dimension. This postulated "mode schema" could help account for other multi-modal aspects of the computerized LOMERA that are absent from the paper test experience that current theoretical explanations do not include. This enhanced model will also help integrate more up-to-date understandings of literacy such as multi-modal literacy (e.g. Kalantzis, Cope, & Harvey, 2010) with questions around assessment practices (e.g. see Asselin, Early & Filipenko, 2005). There are noteworthy differences between the paper and computer testing modes that a mode schema will help to address. For instance, the computer test is presented on a monitor that may require different kinds of visual processing than printed text. In addition, students have to coordinate their movement of the mouse with the corresponding pointer on the screen and they have to navigate the online test by interacting with various icons. These online features are

different from those encountered in a paper test experience, requiring different types of schemata to be drawn upon. These online features are probably uniquely associated with reading in the computer testing mode. Theoretically, there are particular expectations about how reading is efficiently accomplished in each mode that students bring with them that might mediate their test taking experience differently across modes. This comparability research is a first step toward identifying whether mode effect exists and whether there is compelling evidence to warrant future investigation of mode schema as a viable concept.

2.3 Research on Cloze Procedure

2.3.1 Research Supporting Cloze as a Measure of Text Level Comprehension

Cloze has been defined as "encompassing any procedure that omits portions of a text or discourse and asks readers or listeners to resupply the missing elements" (Oller & Jonz, 1994, p. 3). Several scholars have praised cloze tests as a relatively easily developed and administered assessment for ESL learners. Cohen (1994) discusses its versatility noting that it has been used to measure readability, global reading skills, and grammar. Hurley and Tinajero (2001) recommend cloze because

> … it is flexible enough to allow teachers to assess large groups of students at the same time or assess students individually [and]…information gathered in the assessment procedure can be used to select reading material for students that is challenging but not frustrating (p. 19).

Similarly, Law and Eckes (1995) contend that "close tests provide a window into the strategies the student is using to gain meaning, as well as insight into how sophisticated his or her skill level in English is." (p. 68-69)

A substantial amount of the research into the use of cloze indicates that it is useful as a measure of text comprehension beyond the sentence level for first language students (Hale, Stansfield, Rock, Hicks, Butler, & Oller, 1989; McKenna & Layton, 1990) and second language readers (Chavez-Oller, Chihara, Weaver, & Oller, 1994; Hanania & Shikhani, 1986; Yamashita, 2003). The research with second language learners is persuasive. Chihara, Oller, Kelley, Weaver and Chavez-Oller (1994) used two prose passages that were written for non-native speakers as cloze assessments. One cloze used the regular text and the other had the sentences scrambled. Two hundred and one Japanese EFL learners took the two assessments in one sitting using a Latin squares design for presenting the tests. They concluded that "cloze procedure is sensitive to discourse constraints ranging across sentences" (p. 143). Chavez-Oller et al. (1994) reviewed research on both sides of the debate and reexamined Chihara et al.'s (1994) data to discern what types of cloze items are sensitive to discourse constraints across sentence boundaries. Their conclusion was that while all cloze items were not equally able to measure discourse level knowledge many do. They also expressed misgivings regarding the methods of researchers who have not found cloze to be an effective measure of discourse-level comprehension. They contend that skeptics of cloze have reached their conclusions due to their use of rudimentary research designs such as using inappropriate texts that were unable to demonstrate the inter-sentential knowledge measurement capabilities of cloze. In general, their findings support cloze as a measure of integrative language proficiency operating at a discourse (i.e. textual) level, although, as noted previously, the nature of the words deleted determines whether inter-sentential processing is required (Fotos, 1991, Hale et al., 1989). For instance, content words tend to be considerably more difficult to replace than function words (Alderson et al., 1995).

Hanania and Shikhani (1986) correlated cloze scores with scores on writing tests to determine whether cloze predicts ESL students' writing ability. They found that the commonality between cloze and writing may "... be related to the testing of higher-order language abilities, which include the discourse-level factors of cohesion and organization" (p. 107). More recent research had Japanese EFL students provide verbal reports of what they were thinking as they completed rational cloze tests. Findings confirmed the value of cloze as an assessment of reading comprehension beyond the sentence level. The author notes "it can be said that this gap-filling [rational cloze] test's scores largely reflect the students' ability to use text-level constraints…" (Yamashita, 2003, p. 286). This research demonstrates that cloze procedure has a high degree of sensitivity to discourse-level knowledge and supports the argument that the cloze procedure is sensitive to constraints beyond the level of local syntax.

2.3.2 Criticism of Cloze as a Measure of Text Level Comprehension

Not all commentary on the capability of cloze to assess discourse-level knowledge has been in agreement. Several scholars have questioned the suitability of cloze to evaluate discourse-level reading comprehension ability. A criticism from Alderson et al. (1995) was that the choice of the first word in the close passage can dramatically influence the validity and reliability of the test. Consequently, the resulting tests can be very different depending on which word is deleted at the beginning because the subsequent deletions can be mostly function words which are relatively easy to restore or content words which are significantly more difficult to retrieve.

Some researchers have claimed that cloze is best used to measure sentence level writing ability and syntactic knowledge rather than comprehension of connected discourse (McKamey, 2006; Markham, 1987; Shanahan et al., 1982). For instance, in Markham's (1987) study the

participants included 14 ESL college students who completed a 65 item rational deletion cloze task. They were subsequently interviewed in order to ascertain their use of intra-sentential and inter-sentential cues with respect to each successfully completed item. Based on the findings from these think aloud interviews, he concluded that the cloze procedure appears to yield a valid estimate of syntactic and lexical awareness on the level of individual sentences but not beyond the sentence level.

In another influential study, Shanahan, Kamil, and Tobin (1982) carried out a series of three experiments where they administered three types of cloze texts to three different groups of university undergraduates. The tests were a standard cloze passage, scrambled cloze passages, and texts with sentences from the original passage that were decontextualized by placing them into other unrelated texts. They observed no significant differences in the mean cloze scores of undergraduates when original, scrambled and decontextualized versions were compared. Therefore, they deduced that there were no differences in students' performance regardless of whether or not the sentence was in order or more or less contextualized. Based on these results, they concluded that "cloze, as now recommended for classroom use, does not usually measure inter-sentential information integration" (p. 250). These findings were supported in other independent research studies conducted by Alderson (1980) and Porter (1983).

McKamey (2006) provides evidence that cloze may actually measure both grammatical knowledge and reading ability. She performed a factor analysis and a multiple regression analysis to establish the relationship between results of cloze tests and a battery of other language tests that were administered to the ESL students in the study. Her findings about grammar knowledge corresponded to Saito's (2003) in that grammar knowledge was found to be the highest predictor of cloze test performance. However, the results of another reading

comprehension test also greatly contributed to cloze test scores. These findings led her to conclude "I would anticipate that the respective contributions of grammar and reading abilities to cloze test scores are at least relatively equal" (McKamey, 2006, p. 148). This result suggests that cloze readers process text at the sentence level and the textual level to an equal degree. Therefore, texts may not be processed at strictly the sentence or text level but rather the level of processing may depend on the reader, the text or both.

The findings that cloze measures only sentence level syntactic knowledge are certainly not unanimous. Indeed, they are strongly contested by other prominent scholars. For instance, Oller and Jonz (1994) argue that most of the researchers who had negative findings for the effectiveness of cloze erroneously presumed they could prove that cloze never measures text-level reading comprehension. However, Oller and Jonz argue that researchers such as Shanahan et al. (1982) are mistaken because demonstrating cloze to be effective once is enough to validate it as an effective measure of reading comprehension. Oller and Jonz dismiss critics' approach to research as a misguided attempt to do science by tallying conflicting findings and declaring the side with the most number of studies victorious.

Chavez-Oller et al. (1985) concurred with this critique and add several other more specific criticisms of detractors' methods. They point out that, in the case of Alderson (1980), he compared cloze test scores for two different groups of subjects in his experiment, but he did not try to account for the likely influence of subject characteristics on variability between groups. Porter's (1983) research was flawed by too many extraneous variables. For instance, he chose texts that were not stand alone which is an essential criteria for cloze. As well, he selected a wide variety of genres of texts and then tried to compare their results to each other. Chavez-Oller et al. (1985) agreed that Shanahan et al's (1982) fatal flaw was that they attempted to prove a null

hypothesis that there is no difference in readers' scores on contextualized and de-contextualized cloze passages. Chavez-Oller et al. point out that this is simply not how science is done. As Chavez-Oller et al. (1985) mention, "the most important consideration is that null outcomes in any number of studies…are never a satisfactory basis for disregarding positive results." (p. 232)

2.3.3 Commentary on Other Limitations and Strengths of Cloze

Brown (2000) points out that the majority of items on a cloze test do not actually appear to discriminate successfully between high and low scorers (i.e. low item facility). Thus, their validity as a measure of language or reading proficiency is questionable. Furthermore, the test may be causing students unnecessary amounts of frustration answering items that are unnecessarily challenging and yet do not appear to provide much useful information. Lastly, cloze tests do not give any definitive information about the reasons why the examinee answered a particular item correctly or incorrectly. This is important when test users want to use the test for diagnostic purposes.

Brown (2000) acknowledges that cloze has several favorable aspects. For instance, cloze tests are based on contextualized written language that forces test takers to make predictions like they do when they read. As well, a properly designed cloze passage with enough items should spread students along a continuum in terms of ability. Consequently, he concludes that "it is not yet necessary to throw the baby out with the bathwater..." (p. 110) and recommends that rational cloze test be used to avoid including a lot of items that do not effectively discriminate between more and less proficient test takers.

Recently, Watanabe and Koyama (2008) conducted a meta-analysis of 33 studies that used cloze tests with second and foreign language learners and commented on several

shortcomings with research using these assessments. One limitation was that many of the studies lacked detailed descriptions of test designs (i.e., passage length or number of deletions), scoring information, descriptive statistics, and reliability estimates for the tests. An example of one crucial piece of information that many researchers neglected to mention was test takers' proficiency levels. Even those that did were not consistent in terms of how they distinguished between levels of learner proficiency. As a result, it was difficult to compare results across studies.

Watanabe and Koyama (2008) also discuss issues related to the reliability of cloze tests. They question researchers' reliance on exact word scoring (rather than acceptable word) which is convenient for them but it is of dubious benefit for L2 speakers to expect them to use the exact word that a native speaker might use. Their meta-analysis showed that the seventh word deletion and rational deletion cloze tests are most reliable. They also make several convincing points for why K-R 20 is the best reliability estimate formula when calculating the reliability estimate for a given cloze test. However, they are careful to suggest further research into how item interdependence may affect test reliability.

2.3.4 Cloze Adaptations

Cloze has been adapted several times over the years in response to many of the criticisms noted above. Two adaptations that are specifically germane to this discussion are the rational deletion cloze and the multiple-choice cloze. These are discussed here because they are both features of the LOMERA. The unique aspects of each of these adaptations as well as the research into their use with second language learners will be reviewed below.

One revision to the traditional cloze is the rational cloze or gap-fill. The main difference between a rational and the traditional cloze relates to the procedures for altering the text.

Traditional cloze deletes words at a pre-determined interval so if the interval is five then every fifth word will be deleted regardless of where it is in the text. In contrast, the rational cloze does not have a predetermined interval for deleting words from the text. Instead, words are deleted based on the test designer's own intuitions and not some pre-set criteria. Standards for choice of deletions in the rational cloze have alternatively been based on grammatical class and knowledge theorized to be necessary to fill the gap (e.g., syntactic, inter-sentential, genre, schematic) (Oller & Jonz, 1994).

Using rational cloze offers at least two advantages over traditional fixed-ratio cloze. Alderson et al. (1995) point out that an advantage of the rational cloze is that the test designer has more input into which words will be deleted for the test. Consequently, the test designer can choose words that will more realistically represent the test takers' reading comprehension ability. A second benefit of rational deletion is that it "allows the designer to avoid deleting words that would be difficult to predict from the context" (Brown, 2005, p. 202). That way the test can validly assess comprehension ability rather than ability to guess obscure words.

Empirical research conducted on rational cloze generally supports its use for L2 literacy assessment. Bachman (1982) used confirmatory factor analysis to determine whether rational deletion cloze was an effective measure of text level comprehension for college-level ESL students from various L1 backgrounds. His conclusion was that rational deletion cloze was a better measure of syntactic and discourse knowledge than traditional random deletion cloze. He asserted that "it would appear that cloze passages using a rational deletion procedure can be used to measure textual relationships beyond clause boundaries" (p. 66). Later research by Bachman (1985) with EFL university students also found that rational cloze assessed knowledge of inter-sentence connections better than the fixed-ratio cloze. Abraham and Chapelle (1992) investigated

international EFL students' performance on various types of cloze tests and had comparable results. Their findings were that rational cloze is a valid form of assessment for second language learners and it appeared to be more practical for classroom teachers as well.

Recent research continues to support the use of rational cloze for L2 learner assessment. Yamashita (2003) used qualitative verbal think aloud protocols to probe whether Japanese EFL students used sentence or text level information to complete a rational deletion cloze test. He observed that, in general, the rational cloze did access discourse-level second language reading ability. Takanashi's review (2008) concurred that rational cloze is an effective way to measure discourse comprehension for second language learners, particularly EFL students.

A second type of adapted cloze test relevant to this study is known as the maze procedure or multiple-choice cloze. It was originally created to address a concern that test takers' unexpected or ambiguous answers could interfere with the validity of the test. Maze addresses this shortcoming by limiting the number of answers that can be put into the blank because the learner is given options from which to choose rather than blanks to fill. Another advantage of maze is that it allows for rapid scoring and it requires minimal expertise to judge the acceptability of answers (Brown, 2004).

Several studies support the use of the maze procedure with first- and second-language learners. The results of the maze test with first language learners "showed trends similar to those of the cloze, both for mean scores and for accuracy of prediction [of test takers' ability]" (Pikulski & Pikulski, 1977, p. 769). Additionally, one study of middle school native English speaking students found the maze procedure to be valid and reliable measure of content area reading comprehension ability (Espin & Foegen, 1996).

These findings were comparable to those with second language test takers. A study of over 400 TOEFL test takers revealed a "relatively high correlation between total MC [multiple-choice] cloze score and total TOEFL score [which] attests to the concurrent validity of the method... Also, the [multiple-choice] cloze procedure appears to provide assessment that is as reliable as the TOEFL" (Hale et al., 1989, p. 65). Another study compared maze and traditional cloze tests with postsecondary second language learners in Israel. The results showed that maze was equally as effective as traditional cloze and multiple-choice questions for reading comprehension tests. Researchers also pointed out that maze tests allow examiners to ask more questions than the traditional cloze because test takers do not have to take as much time to construct their responses (Bensoussan & Ramraz, 1984).

Support for the effectiveness of the maze procedure is not unanimous. At least two scholars expressed reservations about the use of this form of reading assessment with second language learners. One criticism is that maze is an inauthentic reading comprehension measure. The contention is that readers of maze are forced to interrupt the natural flow of their reading when they stop to choose their answers as they read the passage (Steinman, 2002). This seems to be a specious criticism because the same comment could be made about the authenticity of any standardized reading assessment task. A second perceived limitation is that "multiple-choice cloze test construction is not an easy, mechanical task which all teachers can do, but requires sophisticated knowledge about option selection" (Propst & Baldauf, 1979, p. 685) so maze assessments need to be carefully developed and require extensive piloting. These authors recommended using "matching cloze" tasks as a more easily-developed alternative to the maze task. Few would disagree that this is probably a good idea for classroom-based assessments. However, more standardized and widely-used assessments may be worth the additional effort

required to develop and pilot a maze because they have more research support. This point is noteworthy because, at present, there does not appear to be a great deal of research verifying the validity of matching cloze besides Propst and Baldauf's (1979) research with elementary ESL students. This study showed that matching cloze tests correlated highly with learners' scores on the Gates-McGinitie vocabulary and reading subtests.

2.4 Background –Computer-Based Assessment

According to Ockey (2009), computer-based testing (CBT) is "The use of computers to deliver, score, select items, and report scores of assessments…" (p. 836). Chapelle (2008) argues that the rationale for the development of computer-based testing was to "improve the efficiency of current testing practice" (p. 127). Efficiency is achieved through computers' effectiveness in recording, collecting and analyzing test data.

Despite the apparent advantages of computer-based assessment, it was not immediately adopted by applied linguists. Chapelle (2008) reports that the late adoption of computers in language assessment was largely due to undeveloped hardware and software and apathy in the large testing programs that had the resources to develop these kinds of tests. It was not until recently (mid 1990s) that computer-assisted language tests begin to be used for both large and small scale language assessment (Dooey, 2008).

Regardless of this slow start, the use of computerized forms of assessment has become commonplace. Ockey (2009) notes that "The increasing use of computer-based testing has occurred on numerous assessment fronts, from high-stakes, large-scale commercialized tests to low stakes, small-scale assessments available for free on the Internet" (p. 836). Given the growing prominence of information technology in contemporary society, Chapelle and Douglas (2006) suggest that "communicative language ability needs to be conceived in view of the joint

role that language and technology play in the process of communication..." (p. 108) and we need to expand our view of language ability to see it as "the ability to select and deploy appropriate language through the technologies that are appropriate for a situation." (p. 107) That is, linguistic expression and technological skill are becoming increasingly intertwined so our assessments should follow suit.

At a relatively early stage in the development of computer-based language tests, Blanchard, Mason and Daniel (1989) compiled a list of their advantages and disadvantages. Although the list is over twenty years old it is interesting to note that many of the issues it addressed persist to this day. Benefits of computers that are still relevant include: unlimited patience; unlimited storage and recall of results; uniformity of test presentation; and it saves time and resources collecting, recording and storing data. Some disadvantages still applicable include technology's tendency to dehumanize and depersonalize; computer literacy demands; mechanical failures; confidentiality; and the comparability of computerized test scores with scores from traditional tests.

Questions about comparability have been addressed in first and second language testing research), but many still remain. Results from a number of studies have found computerized tests to be comparable to their traditional multiple choice counterparts with English language learners (e.g., Choi et al., 2003; Sawaki, 2001) but it is not yet clear if this is the case with other types of test items such as cloze. This study will address questions surrounding the cross-mode comparability of a previously overlooked assessment in ESL literacy assessment research (i.e., maze) with a learner subpopulation (i.e. secondary students) that has been traditionally disregarded in this area of research (Snow, 2008).

2.5 Computers and Cloze Assessment

Technology is increasingly being adopted for various forms of language assessment for first and second language learners and cloze appears to be beginning to join this trend. In one somewhat dated investigation of computerized cloze tests, Miller, Burnett, and Upitis (1983) employed a linguistic prediction task to investigate grade three and six students' use of orthographic constraints, syntactic redundancy, and syntactic and semantic knowledge using a cloze task that was developed for a desktop computer. The main problem they reported with using the computer to administer the cloze test was that it took too long to record the children's answers. Consequently, some children became bored or frustrated with the task. The authors acknowledged that better programming and a more powerful computer could probably have resolved the situation. They also insightfully noted that any changes to the program would also likely change the nature of the task. This observation highlights the need for researchers to ensure that software is adequately designed and calibrated to elicit and record the particular assessment data sought by the researchers within a suitable timeframe. That is, test takers must be able to able to complete the computerized test at a pace that is suitable to them while still providing test administrators the data they need.

Research has been conducted on computerized cloze-elide (i.e. words are omitted from the text without being replaced with corresponding blanks) techniques with Hispanic and Laotian ESL learners in elementary, middle and high school. This research revealed significant correlations between the results of the computerized cloze-elide test and traditional multiple-choice reading comprehension, maze and traditional cloze tasks (Manning, 1987). These correlations indicate that the three test forms may be measuring the same psychological construct despite the fact that one is online and the others are not. In another study, Cameron, Hinton and Hunt (1987) used a word processing program to administer cloze tests to measure the reading

ability of native speaking primary students over the course of a three-year longitudinal literacy investigation. They found the word processors simplified the administration of these assessments and test takers reported enjoying doing the cloze assessments more when they did them on the word processor than with pencil and paper. More importantly, they found no significant difference in learners' performance on the computer-based cloze as compared to the pencil and paper cloze. These interesting findings provide some preliminary data on cross-mode cloze test performance for first-language speakers. However, despite the information these studies provide, there are still many questions that remain. For instance, they do not yet reveal if there is mode effect for second language users taking multiple-choice rational cloze or the potential effect that background variables like computer familiarity may have on this particular type of test.

Taira and Oller (1994) investigated the use of computerized cloze tests with Japanese university EFL students. They reported that "cloze exercises presented through CALI [computer assisted language instruction] work well" (p. 364). They also noted that those who used cloze exercises and tests had more gains in English proficiency than the control group. Chapelle and Douglas (2006) briefly touched upon the issue of how the use of computers may allow for a more nuanced means of scoring cloze responses. They note that the powerful computational abilities of computers can potentially allow for a range of scores to be assigned to answers according to gradations in degree of appropriateness of the various answers. They contend that this unique feature of computer-based cloze tests could provide a more realistic representation of the range in understanding that test takers actually have.

Another important observation is that different types of cloze tasks such as multiple choice and elide draw on different sorts of knowledge on the part of the learner. Additionally, the amount of contextualization provided will affect the learners' answer. As a consequence, these

different types of cloze have different requirements in terms of computational resources necessary to score them and determine whether or not they can be graded on a scale. Bailey (2008) remarks

> It is not enough to identify the type of exercise, such as cloze, to be able to determine the processing required for content assessment. Many different exercises fall under the heading `cloze'. Not all cloze exercises can be processed by listing the target response(s) and then matching against the learner response (p. 48).

Thus, test developers need to be clear about the type of cloze they are developing to establish whether the language processing demands can be met by the computer scoring program to accurately and fairly score the test.

2.5.1 Problems with Computer Based Assessment

Although computer-based assessment has numerous strengths several potential limitations may have bearing on their comparability with paper-based assessments. One reservation about computer-based assessment is the potential threats to validity that it faces. Chapelle and Douglas (2006) have summarized two relevant threats to the validity of computer-based tests which include: (1) Performance on a computer-delivered test may fail to reflect the same ability as what would be measured by other forms of assessment. (2) The types of items that can be developed in computer-assisted formats are different from those that can be developed with other media so current conceptions of validity may not account for these unique computer features.

A second challenge presented by computer assessment is the ever-changing technology which necessitates scalable hardware and software as well as ongoing technical skill

development for assessment administrators (Chapelle, 2008). This makes computer-based testing potentially quite resource intensive. The significant costs of adopting these new technologies for assessment are often passed on to the test taker through increased testing fees. These increased fees can prevent learners from writing the test which, in some cases, may systematically preclude certain types of learners from gaining access to further education (Chapelle, 2008).

The problem is that access to the resources required for CBT is often determined by the test administrator or test taker's nationality. As Dooey (2008) observes:

> In general, access to computers is not equally divided among different nationality groups. This means that while access to computers worldwide is increasing, this trend is not as rapid in certain parts of the world. Furthermore, certain language groups seem to be less familiar with computers than others. This highlights the widening gap between the computer haves and have-nots (p. 28).

Limited access to computers facilities required to do well on these tests is not an uncommon problem either. Taylor et al. (2000) estimate that "one quarter to one half of the students from most regions of the world will likely need help learning how to use English word-processing programs and the Internet once they arrive at North American colleges and universities" (p. 584). Although the issue may be have been somewhat ameliorated over the past decade, it is still likely worth some consideration for students coming from developing countries.

Dooey (2008) emphasizes the importance of computer based tests measuring test takers' language ability and not their computer skills. However, at this point, "we still do not know with any certainty how computer technology in language tests affects individual test takers who may be advantaged or disadvantaged with respect to computer skills" (Douglas & Hegelheimer, 2008,

p. 116). Furthermore, paper-based assessments may require different strategies than computer-based assessments and this would affect the validity of the test (Chapelle, 2008). To answer these questions, investigations such as the present study must be conducted on the comparability between paper-based and computer-based forms of a variety of types of assessments.

2.5.2 Guidelines for Computer-Based Language Testing

Several professional organizations have published guidelines that offer recommendations for developing computer-based assessments that are comparable to their traditional paper-based counterparts. Some examples of these enumerated by Wang et al. (2008) are: Guidelines for Computerized Adaptive Test Development and Use in Education (American Council on Education, 1995); Guidelines for Computer-Based Testing (Association of Test Publishers, 2000); Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 1999); Guidelines for Computer-Based Tests and Interpretations (APA, 1986), and International Guidelines on Computer-Based and Internet-Delivered Testing (International Test Commission [ITC], 2004) (p. 6).

The manuals produced by the American Psychological Association (1986) and the International Test Commission (2004) address guidelines for developing computer-based assessments exclusively. These guides offer quite specific advice on how test developers, publishers, and users can ensure maximum comparability between assessment modes. These sources concur on several essential considerations that any developer of a computer-based version of a paper based test should give serious attention.

Both guides clearly agree that test developers must provide test users with data that confirms the comparability of their traditional and computer-based assessments. For example,

the American Psychological Association (APA) states "The test developer should report comparison studies of computerized and conventional testing to establish the relative reliability of computerized administration." (APA, 1986, p. 20) Likewise, the International Test Commission (2006) specifies that test developers need to provide

> ...clear documented evidence of the equivalence between the CBT/Internet test and non-computer versions…to show that the two versions have comparable reliabilities, correlate comparably with other tests and external criteria and produce comparable means and standard deviations or render comparable scores (p. 156-157).

The APA manual (1986) adds that, in addition to providing comparability data, test developers should provide "detailed information on the method of achieving equivalence" (p. 19).

Not only do these professional organizations both share the same view that evidence of comparability must be given, they also detail the kind of evidence required to demonstrate comparability between paper and computer-based testing modes. The APA (1986) states that

> Scores from conventional and computer administrations may be considered equivalent when (a) the rank orders of scores of individuals tested in alternative modes closely approximate each other, and (b) the means, dispersions, and shapes of the score distributions are approximately the same...(p. 18).

The ITC (2006) adds that norms being compared across modes must be based on the same versions of the test that has been administered to the same subpopulation of test takers.

Both guides also caution developers to be aware of potential effects of computer accessory devices that are being used to present items and record responses across modes. For instance, the ITC (2006) advises that designers of computer-based versions of paper-based tests

must ensure that test items are presented in the same way across modes and test takers have the same level of control over the test (i.e., can easily move between pages or revisit previous questions). They also insist that the method of responding must be comparable across modes. The APA (1986) notes that "special responding devices such as a light pen could in some circumstances affect test norms. This is especially an issue with timed responses, which are known to vary in speed for different types of required responses" (p. 19-20). Therefore, developers are encouraged to be vigilant to ensure that computer equipment use does not impair examinees' ability to respond to questions in a timely manner.

2.6 Cross Mode Comparability Research

The central objective of comparability research is to determine whether test results from computer-based tests are comparable to those obtained from their paper-based counterparts. Numerous studies have been conducted to investigate whether or not this has been the case with a variety of types of tests. The results of many of these studies are presented and discussed below. This review includes research with both first and second language test takers because the literature that exclusively reviews second language speakers is relatively sparse.

2.6.1 Comparability of First Language Reading Comprehension Tests

The research findings on comparability of first-language reading assessments are divided. Some of the research comparing paper and computer-based reading assessments has found that both forms of the test are comparable. For instance, Higgins, Russell, and Hoffmann (2005) investigated a sample of 219 fourth-graders from eight schools in Vermont. This sample included both native and non-native speaking children. They found that there were no significant differences in students' reading comprehension scores across the two modes. Another study by Evans, Tannehill, and Martin (1995) explored 51 native speaker students' performance on letter-

word identification and word attack skills on paper and computer tests in a clinical setting. These researchers also reported that the two forms of the test were comparable and that computer assessments can be used to assess the same skills as traditional assessment measures. Paek's (2005) review of comparability research for first language tests led her to conclude that evidence has accumulated to the point that it appears that computers may be used to administer tests in many traditional multiple choice test settings without any significant effect on student performance. However, she also cautions that there does not yet appear to be enough research with other types of test items. Cloze is an example of such an item type. A large scale meta-analysis was also undertaken to examine the research into testing mode effects for reading assessments of native learners from K-12. After an extensive review of the research literature, the authors narrowed the sample down to 11 studies with 42 independent data sets. Their meta-analysis of this aggregated data revealed that administration mode did not have a statistically significant effect on K–12 student reading achievement scores (Wang, Jiao, Young, Brooks, & Olson, 2008).

Other researchers report conflicting findings about native speakers' performance across modes. For instance, Pommerich (2004) studied grade eleven and twelve English first-language students doing separate multiple-choice English and reading tests. The English test asked them to make judgments about the register of texts that they were shown. The results revealed that "while some items showed no significant performance differences across administration modes, there were other items for which examinees clearly did not respond in the same way across modes or interface variations" (Pommerich, 2004, p. 40). Her analysis of individual English and reading test items led her to conclude that several factors might exert a unique and unpredictable influence on mode effects depending on each individual testing item. These unpredictable factors

were primarily differences in some of the computer interfaces and fatigue from having to colour in bubbles for the paper test. Ultimately, she did find some difference in mode effects for the reading comprehension test, and speculated that "the findings suggest that for the test forms studied, the observed performance differences might have a fairly small effect in practice" (Pommerich, 2004, p. 41). These findings suggest potential differences across the modes but they do not specify which mode is more challenging and they suggest that these differences may not have a substantial impact in practice.

Other investigators have concluded that the results from computer and paper tests cannot be used interchangeably, and that one mode is less challenging than the other (Keng et al., 2008; Kim & Hyunh, 2008). At least two studies have found that participants performed better on a paper test. One investigation of grade 8 students taking state-wide reading tests revealed that they demonstrated significant differences in performance with the paper-based group performing better (Keng et al., 2008). This result was supported by other research on middle and high school students taking an English test which found "CBT [computer] to be significantly more difficult than PBT [paper]; however, the magnitude of the difference was small" (Kim & Hyunh, 2008, p. 567). Here, as well, questions remain about the practical significance of these findings.

Although the results that test takers perform better on the paper tests are compelling, they are not unanimous. At least one study found that native speaking college students achieved better on a computer-based test than on a paper test. This discrepancy was even more pronounced for test takers with the highest achievement overall (Clarania & Wallace, 2002).

Pomplun, Frey, and Becker (2002) investigated the comparability of scores from a computerized and paper-and-pencil test for secondary and post-secondary level students. Their results revealed that the paper format and one of the computerized versions produced higher

comprehension and total scores. This indicates that differences in scores may depend not only on the mode but also the type of test being administered through that mode. Clearly, more research is required with a wider variety of test types such as the LOMERA.

Reading comprehension tests intended for first-language speakers may also be more susceptible to test administration mode effect than other forms of assessment. In their investigation of middle and high school students, Kim and Hyunh's (2008) analysis revealed that the reading comprehension test showed a comparatively large difference in Rasch ability estimate between the paper and computer-based tests. A Rasch analysis is a statistical test that indicates whether both tests have consistent scores at the item level (i.e. differential item functioning). This disparity reveals that reading comprehension tests may be more seriously affected by the test administration mode. Pommerich (2004) agreed that there may in fact be greater mode effects for reading tests than for tests of other skills.

2.6.2 Second Language Comparability Studies

Several investigators of paper and computer-based tests of second language reading comprehension concluded that forms were comparable and that there were no mode effects. Sawaki (2001) conducted a review of the literature in educational and psychological measurement as well as in ergonomics, education, psychology, and L1 reading research. Her main conclusion was that "comprehension of computer-presented texts is, at best, as good as that of printed texts, and that reading speed may or may not be affected by mode of presentation" (Sawaki, 2001, p. 49). That is, for L2 reading tests, both paper and computer-based modes are comparable. Nevertheless, she also cautioned that computer-based assessments cannot simply be viewed as digitized versions of PBTs. Instead, she urged readers to bear in mind that computerizing a test to replace an existing test "frequently means revisions of test specifications,

test format and layout with the hope that the new test form will bring about enhanced authenticity, construct validity, and measurement accuracy." (Sawaki, 2001, p. 50)

Two studies with locally developed English language assessments also found the paper and computer-based versions to be comparable. One empirical study measured 167 Saudi medical EFL students' performance on paper and computer versions of a reading comprehension test. Although the investigator found a significant difference between the scores on the two modes, this difference was not a result of the testing mode effect. He argued it was actually caused by the small number of items that were used on the test. He based these conclusions on a more in-depth analysis of the data which revealed that the reliability and validity of the tests was not affected by the testing mode (Al-Amri, 2008). An additional study of Malaysian postsecondary EFL learners yielded similar results. Test takers in that study were taking paper and computer-based forms of a locally-developed English reading test. These researchers similarly reported that there were no significant differences in students' performance across the two modes though test takers did perform slightly better on the online version (Norazah, Arshad, Razak, & Jusoff, 2010).

Analogous results were reported for larger scale assessments as well. Choi, Kim and Boo (2003) compared paper and computer-based versions of a postsecondary-level standardized English language test developed by Seoul National University in South Korea. They reported that the two modes were comparable across all subtests (listening comprehension, grammar, vocabulary, and reading comprehension). They also conducted a confirmatory factor analysis and determined that, to a certain degree, paper and computer-based subtests measure the same constructs. A more detailed analysis of the subtests also revealed that "the grammar test showed the strongest comparability, and the reading comprehension test the weakest comparability" (p.

316). This result appears to support other findings noted above that reading assessments may have greater potential to exhibit mode effects.

A comparison was made of paper and computer-based versions of the International English Language Testing System (IELTS) using a sample of 400 participants who represented the most common language groups that took the IELTS. The researchers reported that both forms were comparable and could be used interchangeably if candidates had enough computer training (Green & Maycock, 2004). However, it is noteworthy that this study was commissioned by University of Cambridge Local Examinations Syndicate (UCLES) to validate a test that they had already invested a lot of resources into developing. In essence, they are publishing the research that validates their own test. One could not be faulted for being slightly wary of the results. The same criticism can be made of Choi et al's (2003) research discussed above.

In contrast to the findings claiming cross-mode test comparability, there were at least two studies that found lack of cross-mode comparability in their results. One study of a university entrance placement test for ESL learners in the UK found that there was a significant difference in test scores between the mean of the paper and computer-based test (Fulcher, 1999). Though he does acknowledge that order effect probably accounts for some of the better computer test performance (all test takers wrote the paper test followed by the computer test) Fulcher (1999) contends that the cross-mode correlation of .82 is not high enough to justify the use of the computer-based tests as a replacement for the paper test. Coniam (2006) did not find paper and computer-based tests to be comparable for all second language students either. He investigated secondary students who took an English listening comprehension test in Hong Kong and he concluded that test takers generally performed better on the computer-based test than on the paper-based test. He argued that correlations between scores on the two test types were high

enough to justify the computer-based test's use as a low-stakes test (i.e., school-based testing), but not as a high-stakes test (i.e., territory-wide test).

Yu (2010) is one of the first to explore the comparability of paper and computer assessments beyond traditional multiple-choice item types that previously received the bulk of research attention. He studied the cross-mode differences in written summaries of various types of texts of 157 undergraduate students at a Chinese university. He found that Chinese first language computer-based summaries tended to be longer but they were not judged to be better. Intriguingly, he also observed that the type of text being summarized may have amplified the mode effect with texts that were more difficult to summarize being even more challenging in the computer mode.

2.7 Computer Familiarity

A background variable that has long been hypothesized to affect test takers' performance across modes is computer familiarity. Kirsch, Jamison, Taylor and Eignor (1998) conducted a study for Educational Testing Service (ETS) as part of the development of the computer-based version of the Test of English as a Foreign Language (TOEFL). They surveyed a sample of almost 90,000 TOEFL examinees from a wide variety of first languages. Their findings revealed small differences in computer familiarity based on age, gender and reason for taking the test. They reported larger discrepancies in computer familiarity for examinees from Japan and African countries. They also observed a small but significant difference in the performance on the paper-based test scores but, surprisingly, their study did not investigate the relationship between familiarity and computer-based test performance.

Results of later studies were that computer familiarity does not affect performance on a computer-based language test. An investigation of 1,200 TOEFL examinees from a wide variety

of first language groups determined that when English language ability was taken into account there was "no meaningful relationship between level of computer familiarity and level of performance on the CBT [i.e. computer based test] language tasks" (Taylor, Kirsch, Eignor, & Jamieson, 1999, p. 265). These researchers concluded that there was no "adverse relationship between computer familiarity and computer-based TOEFL test performance due to lack of prior computer experience" (Taylor et al., 1999, p. 219). Similar results were reported with Saudi medical students who were taking an EFL reading comprehension exam (Al-Amri, 2008). An extensive review of research into the cross-mode effects for second language reading examinees similarly determined that "computer familiarity…does not seem to manifest itself in test scores" (Sawaki, 2001, p. 44).

Other research with elementary school students that included both native speakers and second language learners concluded that reading comprehension test scores were not affected by computer literacy, but the authors did note that students with weaker computer skills were particularly disadvantaged when they had to scroll through text (Higgins et al., 2005). Some researchers that did find cross-mode differences in performance noted that computer familiarity was not related to this performance difference (Clarania & Wallace, 2002). The only possible negative factor related to testing mode familiarity had to do with the affordances of paper-based tests that are not readily available in the computerized mode. For instance, test takers may be more comfortable with being able to highlight or take notes on the printed page, but not on the computer screen. Nevertheless, some authors conjecture that computer programs can probably be designed that will allow for more interactivity with the text (Choi et al., 2003). With the advent of devices such as tablet PCs, this is very likely the case.

Examinees' ability to manipulate the accessories of the computer was a specific area of inquiry related to test interactivity and computer familiarity that was mentioned in at least two studies. Pomplun, Frey and Becker (2002) compared paper and computer versions of a reading placement test for secondary and post-secondary students. Their data revealed that, in a number of cases, differences in respondents' performance across the two modes seemed to be caused by the differences in their response speed associated with the use of a mouse compared to a pencil. That is, those who performed better on the computerized test tended to be more adept at using the mouse. Other research into secondary students' ability to answer open ended (i.e., short answer and essay) questions indicated that learners who have more experience with keyboarding performed better on the computer based test (Russell, 1999). These findings imply that facility with using computer devices may have some bearing on CBT performance.

2.8 Examinee Perceptions of Paper and Computer-based Tests

Test taker perceptions of computerized assessments are also thought to be linked to their performance on paper versus computer-based tests. Research has explored several potential influences on learners' perceptions of computer-based language assessment. For instance, gender is one variable that is believed to impact learners' perceptions of computers in general. There was also some discussion in the literature of male and female second language learners' varied perceptions of assessment mode. Interviews with English learners who took computer and paper based versions of the same listening test revealed that the males preferred the computer-based test while females favoured the paper-based test (Coniam, 2006). Another large scale meta-analysis of studies of North American native English speakers found that males saw themselves as being more competent with computers and enjoyed working with computers more. This was

particularly the case for high school learners. However, the author also warned that effect sizes were small so “any 'gender gap' that exists in computer-related attitudes and behaviour is extremely small” (Whitley, 1997, p. 15).

Other research explored the connection between attitudes towards computers and performance on tests. Findings were divided. One investigation with Saudi EFL students reported that there was no significant correlation between computer attitudes and the participants’ performance on computer-based tests (Al-Amri, 2008). In contrast, research with learners of French and Spanish found that students who had positive motivation and attitudes toward language study tended to do well on the computerized module tests that they were assigned to measure their progress through a course. The authors of the study also surmised that this was because these smaller tests required ongoing attention to ensure optimal performance. Interestingly, while there was a correlation between performance on the smaller tests and students’ attitudes, this was not the case for the final exam or the course grades. The authors speculated that “the lack of correlation may have been due to students having acquired “test-taking skills” by the time of the final exam” (Ushida, 2005, p. 67). These ambiguous findings indicate that more investigation is required.

A number of researchers have investigated learner perceptions of tests of language skills. In one study of reading skills with both first language and ESL students, Higgins et al. (2005) found that those who took a reading test with items from the National Assessment of Educational Progress (NAEP), Progress in International Reading Literacy Study (PIRLS), and New Hampshire state assessments indicated that they would prefer to take it on computer. Another study looked at post-secondary examinees’ perceptions of traditional versus computer based modes of speaking assessment (Kenyon & Malabonga, 2001). They reported that attitudes

toward the computer mode seemed to be varied. On the one hand, participants commented that they appreciated that computers could record their speech (i.e., pronunciation, vocabulary, fluency, grammar). However, they also pointed out that the natural interactivity of genuine communication was missing. Fulcher (1999) investigated the impact of second language learners' attitudes towards taking Internet-based tests and their computer test scores. He asked several questions on a Likert scale such as whether they liked the computer and the paper-based test, which test they thought they would score higher on, and, which test would they choose to take again. His results revealed that attitudes had no bearing on computer test performance.

Much of the computer test perceptions and attitudes research has also focused on respondents' views of the relative advantages and disadvantages of using computers for language assessment. Gorsuch, (2004) conducted a retroactive interview study of ESL graduate students. Participants revealed that they believed the computer version of the listening comprehension test they took actually improved their performance on a paper-and-pencil version of the same test they took four weeks later. Other positive features of the test respondents felt improved their test performance were a preparation tutorial, large computer screens, and a comfortable testing environment. Pino-Silva (2008) researched Venezuelan EFL students' attitudes toward the use of computer-based tests. Test takers reported these tests were "dynamic and motivating" (p. 151) and that the scores from these tests were "accurate and reliable" (p. 152). Participants stated that there were substantially more advantages of computer-based tests than disadvantages. The most frequently reported advantage was the ability to receive their grade report immediately. Some of the other benefits mentioned included the convenience of the test (plausible given their putative high level of computer literacy), and not having to construct written answers (and thereby avoid surface errors) (Pino-Silva, 2008). Another study of 423 examinees who took paper and

computer-based versions of the IELTS exam reported that they were satisfied with the computer-based version of the test and they considered it to be comparable to the paper-based version (Green & Maycock, 2004). Yu (2010) interviewed Chinese EFL students to ascertain their perceptions of the two testing modes. In general, they commented that they slightly favoured the paper test because of tangibility, security and visibility. Students also revealed that there were small physical and psychological effects of the text presentation mode (i.e., eye strain from reading the screen) on their performance; the test performance data did not demonstrate this to be the case.

Research into L2 learners' perceptions of computer-based assessment has also revealed reservations that learners had about the process. A study of Saudi EFL learners reported that "About 51% of the subjects think that PBT [paper-based test] measures their comprehension skills more accurately while only 23% think that CBT [computer-based test] is better in this respect" (Al-Amri, 2008, p. 40). Thus it appears some learners may not have faith in computer-based assessment of reading comprehension. This intriguing finding warrants further enquiry. Several studies have also given voice to test takers' more specific criticisms. For instance, Coniam (1999) indicated that test takers had more favourable views of computer-based tests when they were relatively undemanding (e.g., multiple-choice). However, tests with somewhat challenging tasks such as gap filling (similar to cloze) were viewed much less positively. In those instances, test takers generally preferred the paper-based version. Although the majority of test takers in Pino-Silva's (2008) study praised the use of computer-based assessments, they also expressed several concerns about them. Hardware-related misgivings had to do with anxiety over power outage and machine failure affecting their tests. A few respondents also complained about visual fatigue caused by reading from a screen though "the majority of comments pointed to

visual fatigue not being a detrimental factor in subjects' perceptions" (Pino-Silva, 2008, p. 152). Some respondents were also uncomfortable with not being able to see the entire test at will and several complained that they had to interact entirely with a machine. They explained that this was a problem because "there is no room for negotiation and/or objection" (Pino-Silva, 2008, p. 154). A few respondents expressed anxiety about being able to successfully adapt to the procedures for these new forms of assessment as well (Pino-Silva, 2008).

2.9 Summary of the Research Literature

The empirical research into the construct validity of cloze to evaluate the reading ability of second language learners is divided. The controversy centers on whether or not cloze effectively measures comprehension of the text beyond the sentence level. The research compiled for this review identified several studies that concluded cloze is a valid measure of comprehension beyond the sentence level (Chavez-Oller et al., 1994; Chihara et al., 1994; Hanania & Shikhani, 1986; McKamey, 2006; McKenna & Layton, 1990; Yamashita, 2003). However, other scholars have claimed that cloze actually measures syntactic knowledge rather than text comprehension (Markham, 1987; Saito, 2003; Shanahan et al, 1982). Rational deletion cloze was reported to be useful for L2 literacy assessment (Abraham & Chapelle, 1992; Bachman, 1982, 1985; Takanashi, 2008; Yamashita, 2003) as was multiple choice cloze (Bensoussan & Ramraz, 1984; Hale et al., 1989) though these results were not unanimous (Propst & Baldauf, 1979; Steinman, 2002). Oller and Jonz (1994) present a strong argument that cloze supporters have met their burden of proof by providing several studies where cloze has been shown to be an effective measure of text-level comprehension. Studies of computer-generated cloze tests have revealed advantages such as streamlined administration, and increased test taker enjoyment. The main problem with computerized cloze tests is that some children

became bored or frustrated with parts of the task (Cameron, 1987). An investigation of computerized cloze tests with Japanese university EFL students revealed that those who used cloze exercises and tests had more gains in English proficiency than the control group (Taira & Oller, 1994).

The advantages of computer testing discussed in the literature include unlimited storage of results and disadvantages include comparability with traditional tests. A challenge presented by computer assessment is its ever-changing technology upgrade requirements that make it quite resource intensive (Blanchard et al., 1989). Even more troubling, it appears that most experts in second-language assessment do not have a strong understanding of computer assessment in language learning (Chapelle, 2008).

Investigation into cross-mode comparability research aims to determine whether test results from computer-based tests are comparable to those gotten from their paper-based counterparts. At this stage, this area of research is still emerging. Test development guidelines promote computer assessments that are comparable with traditional paper-based forms as long as they are transparently developed (APA, 1986; ITC, 2006). Factors found to shape cross-mode comparability are computer familiarity (Al-Amri, 2008; Taylor et al., 1999), and ability to manipulate the computer accessories (Pomplun et al., 2002; Russell, 1999). Research results for cross-mode comparability of reading comprehension assessments are somewhat ambiguous. L1 research reports comparable (Evans et al., 1995; Higgins et al., 2005; Wang et al., 2008) and non-equivalent findings (Clarania & Wallace, 2002; Keng et al., 2008; Kim & Hyunh, 2008; Pommerich, 2004). These results parallel the second language comparability literature which also claims cross- mode comparability (Al-Amri, 2008; Green & Maycock, 2004; Sawaki, 2001) that is disputed (Coniam, 2006; Fulcher, 1999). Some second language research appears to

suggest that reading comprehension tests may be uniquely more susceptible to mode effect than other forms of assessment (Choi et al., 2003).

Although prior research has begun to provide some understanding of comparability, many questions remain. There are three areas in particular where further research is clearly needed. Al-Amri (2008) points out that there is not a great deal of research on the interaction between assessment modes and examinee variables. For instance, the effect of background variables such as computer familiarity, perceptions of computer-based tests, age and first language has not been fully explored. Secondly, there does not appear to be any cloze comparability studies with second language readers. Much of the comparability research with first and second language test takers appears to be based on the typical standardized multiple-choice test items (Chapelle & Douglas, 2006). It would be beneficial to investigate other types of items and tasks such as cloze. This would provide a clearer idea of whether mode affects are more predominant in particular types of test items. Lastly, many of the studies reviewed above reported that the paper and computer versions of the test were quite dissimilar in terms of their layout and functionality. These differences introduce the confounding variable of layout and functionality effect in addition to the mode effect. More research is needed with paper-based and computer-based assessments which are absolutely as similar as possible to eliminate this confounding variable. The current research study will address these issues and others.

Another variable that was hypothesized to affect test takers' performance across modes was computer familiarity. One large-scale study with a sample of almost 90,000 TOEFL examinees from a wide variety of first languages reported negligible differences in computer familiarity based on examinees' country of origin (Taylor et al., 1998). Results of more recent studies were that computer familiarity does not affect performance on a computer-based

language test (Clarania & Wallace, 2002; Higgins et al., 2005 Sawaki, 2001). Even research that did find cross-mode differences in performance noted that computer familiarity was not related to this performance difference (Coniam, 2006; Fulcher, 1999).

There are some questions about the link between test taker perceptions of computerized assessments and student performance on paper versus computer-based tests. Research has explored several potential influences on learners' perceptions of computer-based language assessment. At least one study showed that boys preferred the computer-based test while girls favored the paper-based test. Boys also saw themselves as being more competent with computers and enjoyed working with computers more (Whitley, 1997). In one study of reading skills, respondents indicated that they would prefer to take the test on computer (Higgins et al., 2005). Perceived advantages of computer-based tests were these tests were "dynamic and motivating" and that the scores from these tests were "accurate and reliable." Reservations expressed about computer-based tests were anxiety over machine failure, visual fatigue and difficulty navigating the test (Pino-Silva, 2008).

The next chapter contains a description of the design of the study. This is followed by an explanation of how participants in the study were selected and an account of the procedures followed for data collection. Materials (i.e. tests and surveys used) used during the data collection process are described and evidence of instrument validity and reliability is presented and discussed. This discussion also includes a rationale for the choice of mixed-methods used for this research. Potential threats to internal validity of the study are also addressed.

Chapter 3 Methodology

3.1 Introduction

The procedures followed in collecting the data for this project are outlined in this chapter. An explanation is given of the selection and recruitment of study participants followed by a description of the instruments that were used for data collection and the procedures for scoring those instruments. A explanation of the research site is provided and the procedures that were followed to collect the data are explained. The timeline that was followed to carry out the research is outlined. An overview of the design of the study is given as well as an account of the steps that were taken to ensure the internal validity of the study.

3.2 Selection and Recruitment of Participants

The research site was a large secondary school in a major city on the west coast of Canada. The school serves learners from grade 8 through 12. A sizable proportion of the student body is comprised of recent immigrants to Canada. The school has an ESL program that serves hundreds of students of all ages, grades and ESL levels. Participants in this study came from a variety of different countries including Brazil, China, Iran, Korea, Mexico, Philippines, Russia, and Vietnam. They were recruited by this author with the assistance of their ESL teachers.

This author approached the head teacher in the ESL department first and asked her if she thought any of her students would be interested in participating in the project. She then conferred with her colleagues and they agreed to allow this author to enter all of the ESL classes in the school and explain the project to the students to recruit volunteers to participate. This author then entered eight ESL classes at the school to solicit volunteers. Those who agreed to take part were given assent forms to sign and consent forms for their parents to sign.

The number of participants that was recruited for the study was 60 for each testing group which totalled 120. This decision follows the advice of Fraenkel and Wallen (2005) who recommend 30 as the minimum number for statistical power in an experiment.

The primary purpose for selecting these students is that secondary-school-aged participants have often been overlooked in the second-language assessment research literature (Snow, 2008). These participants were also chosen because they are the most suitable group to help establish whether the paper-based and computer-based versions of the LOMERA are comparable in their ability to discriminate ESL learners' reading proficiency. This is because the school is from the lower mainland region of British Columbia where the test is normed and the study participants are in the group for whom the test was designed – secondary school ESL students.

3.3 Instruments Used in the Research

The Lower Mainland English Reading Assessment (LOMERA) was designed and normed by ESL assessment specialists from each of the school districts represented in the Lower Mainland ESL Assessment Consortium of British Columbia. The test manual explains that the LOMERA was created to be "an easily administered measure of English reading achievement. It was designed to provide an estimation of reading levels to provide some information to help place ESL students in appropriate instructional groups." (Gunderson, D'Silva & Murphy Odo, 2010, p. 4) In essence, it is straightforward to administer and serves as a general indicator of reading proficiency. Evidence for the validity and reliability of the LOMERA will be presented in chapter 4.

3.3.1 The Paper-based LOMERA

The version of the LOMERA that is currently in use in Lower Mainland schools is a paper-based multiple-choice rational cloze style test (i.e., maze). The test consists of a series of eight passages on a variety of topics in various text genres that have been taken from textbooks in several different subject areas that are used in schools in the lower mainland. Each passage is 250 words. The first and last sentence of each passage has been kept intact to provide the reader with some context as was suggested by Guthrie et al, (1974). Each passage has been chosen based on its readability and internal coherence. The passages are arranged according to difficulty so that the first one in the text booklet is the easiest. As the examinee progresses through the test, the passages become progressively more challenging. Appendix B contains a chart identifying the part of speech from each of the deletions in each passage. In addition to being organized according to difficulty, scores on the passages are also locally normed with students from the Lower Mainland. Local test norms provide those administering the assessment with information about average passage scores by grade level and percentiles for local ESL and native speaker students. This information can then be used to compare a particular test taker's performance to other local students.

3.3.2 The Computer-based LOMERA

The computer-based version of the LOMERA was designed to be as similar to the original as possible. The ESL Assessment Consortium members stressed the importance of the comparability of both tests in terms of layout and functionality to ensure optimal comparability. Indeed, the paper and computer-based versions of the LOMERA use the same passages with the same deletions and are presented in the same order.

3.4 Instrument Administration and Data Collection Procedures

Data collection began after this author contacted teachers and district administrators and explained the study to obtain permission to conduct research at the school. An ethics review (BREB) was completed and authorization was gained to study human subjects. Participants were assured that their identities would remain confidential during and after the research. The instruments used in the study including the LOMERA and the questionnaires were explained. Informed consent for students to participate in this research was then obtained from students and their parents (see Appendix E). When all permissions were granted, schools were visited participants were randomly assigned into either a computer-first or a paper-first group and data collection began.

One hundred and twenty participants were administered the paper-based and computer-based versions of the test. Half of the sample was given the paper version and the other half was given the computer version. Approximately four weeks later, the group that has previously taken the paper-based test took the computerized version and vice versa. Although there is no prescribed minimum period to wait between test administrations (Kauffman, 2003), the interval chosen for the second administration of the test for this study was four weeks. A four-week interval should be an adequate minimum period to wait before administering the second test for two reasons discussed by Kauffman (2003). First, if the interval is too short (e.g. a few days), learners will be more likely to remember answers to specific questions and test taking strategies learned through completing the first test. These advantages will then inordinately enhance their subsequent test performance (i.e. practice effect). Equally, if the period between test administrations is too long, it becomes increasingly likely that other extraneous variables like maturation or instruction will affect performance (Kauffman, 2003). A four week interval was

thought to allow enough time for learners not to recall the test but not too much time to introduce other potential threats to validity.

In addition to taking both versions of the test, participants also completed two written questionnaires that asked them to report on their familiarity with computers and their perceptions of the comparability of the paper-based and computer-based versions of the LOMERA (see appendices C and D). Results from the familiarity surveys were then compared with computer test performance to determine the relationship between computer familiarity and computer test performance. Results from the perceptions survey were compiled and categorized into themes that arose from follow-up interviews with ten volunteer test takers.

The classroom teacher granted access to ten volunteers who gave written consent to speak individually with the interviewer (this researcher) for approximately ten minutes each. Interviewees were asked about which mode they preferred for each of the test comparison criteria (e.g. navigability, readability etc.) and then to explain why they preferred that mode. Interviews were semi-structured to allow all interviewees to speak to common themes while remaining flexible enough to pursue potentially informative responses. All interviews were audio recorded and all contents were transcribed. All interviewee answers to the same questions were reviewed and common responses were coded and compiled. These common responses were then interpreted and reported in the results and discussion chapter.

Test administration procedures are standardized and clearly outlined in an accompanying test manual. At this stage, the LOMERA is only administered by assessment consortium members. Regarding administration procedures, test takers are given 35 minutes to take the test by reading each of the passages and selecting as many correct answers as they can to fill in the blanks. The assessment includes passages from grades two through ten. There is no passage for

grade three and the eighth passage is suitable for grades eleven and twelve. Texts also represent a variety of topics from history to science and include both narrative and expository styles. The LOMERA does not provide traditional "grade level" reading level designations. However, it does provide an indication of a test taker's ESL reading level given his or her grade level and test score. It also gives some information on the test taker's potential instructional and frustration levels.

3.5 Questionnaires

The questionnaires used in this study were paper-based to avoid any possible confounds caused by presenting the questionnaires through different media. Information contained in these questionnaires was taken from past research studies on computer familiarity (Taylor et al., 1999) and perceptions of computer tests (Al-Amri, 2008). One questionnaire included eighteen questions about test takers' computer familiarity and the other asked twelve questions that had them compare their experience of taking the paper and computer tests. The questions concerning computer familiarity asked about how frequently participants have access to computers in a month and how often they actually use computers on a weekly basis. The familiarity questionnaire also enquired about how comfortable students feel about using computers. Both questionnaires were informally piloted with several second-language speakers beforehand to ensure that the language of the questions was comprehensible to participants. Questions contained in the first questionnaire asked about how often computers are available to participants, how comfortable they felt using the computer and taking tests on the computer and how often they use the computer for various tasks.

The second questionnaire was given after the second test administration because by then all participants had had the opportunity to take both tests. The second questionnaire asked

participants to compare their experiences taking the paper and computerized versions of the LOMERA on a variety of criteria such as readability and navigability. They were asked to check a box to indicate their mode preference based on a several test features that previous research identified as being potential causes of mode effect. Examples of features that were enquired about include: passage and question readability, ease of selecting and changing answers et cetera (see appendix D for the entire list of test attributes). Follow up interview questions were then asked to ten volunteers that encouraged them to expand on their answers to this questionnaire by offering reasons for their reported mode preferences for each of the test attributes. These interviews were beneficial because they provided nuanced data that post-test questionnaires would probably not elicit. Bachman and Palmer (1996) advised that reliance on only questionnaires may restrict test takers' responses so that researchers only get answers to the questions they ask. In contrast, open-ended interview questions may produce unexpected responses from test takers which would hopefully lead to new insights into the phenomenon under investigation (Bachman & Palmer, 1996).

3.6 Scoring Procedures

The LOMERA was scored by this researcher who was trained by its designers. As was mentioned above, there are a total of eight passages with twelve deletions (1 point each) per passage. This produces 96 items for the entire test. Because the items are multiple-choice, there is only one possible correct answer among three alternatives so the scoring method is exact word. One point is given for each correct answer so a perfect raw score on either version of the LOMERA is 96 points.

To determine a test taker's ESL level, the examiner must first find the sum of scores for all passages. Then, the examiner checks the test taker's final LOMERA score on the appropriate

chart on page nine of the test manual (see Table 1 for districts with four levels below). One can see that a score of 57, for example, would place a grade nine student in level two in a district with four levels.

Table 1

Test Taker ESL Level Based on Overall LOMERA Score

ESL Level	Grade 8	Grade 9	Grade 10	Grade 11	Grade 12
1	0-47	0-47	0-48	0-46	0-43
2	48-62	44-61	49-63	47-60	44-61
3	63-70	62-73	64-74	61-75	62-75
4	71-96	74-96	75-96	76-96	76-96

For this analysis, the total raw scores were used to compare test takers' performance across the two modes. A statistically significant difference in these total scores across the two modes was taken as evidence of mode effect. Raw scores on individual test passages were also correlated to establish whether or not there were statistically significant differences in the scores on individual passages across the two modes. Additionally, raw scores were used in the calculation of a delta-plot differential item functioning (DIF) analysis to ascertain whether there was mode effect for individual test items.

Scoring procedures for the questionnaires are slightly different from those of the LOMERA. The familiarity questionnaire asked participants to indicate how often computers are available to them in various locations such as home and school. They are presented with four categorical options from which to choose labeled “once a week or more often,” “between once a week and once a month,” “less than once a month,” and “never.” They are then asked about their comfort level with various aspects of computer use such as taking tests on a computer. Their

response is divided into four categories which are “very comfortable,” “comfortable,” “somewhat comfortable,” and “not comfortable.” Finally, they are asked about how often they use a computer for various tasks and they must choose among four categories labeled “more than once a day to once a week,” “less than once a week to once a month,” “less than once a month,” and “never.” The questions were scored by indicating which category was selected and the results were entered into SPSS.

The second questionnaire asked participants to indicate which test they found easier on a variety of criteria such as ability to read, record answers and so forth. They must choose from “paper”, “no difference” or “computer.” The data from the second questionnaire were used in a categorical analysis that will be explained below.

3.7 Rationale for Choice of Mixed-Methods Approach

The methodological approach used in this research was a mixed-method approach that incorporated multiple-choice cloze tests, surveys and follow up interviews. Both quantitative and qualitative data were collected to answer the research questions in this study. One reason for the integration of the quantitative data and descriptive data was that it allowed for “methodological triangulation that can help to reduce the inherent weaknesses of individual methods by offsetting them by the strength of another, thereby maximizing both the internal and the external validity of research” (Dörnyei, 2007, p. 43). For instance, a limitation in the quantitative data was that survey data could not easily access unexpected reasons why test takers had the particular perceptions of the various aspects of the test that they did. Follow up qualitative interview data allowed for test takers to elaborate on their reasons in their own voice. A unique strength of a mixed-method approach is that it emphasizes that

> Words can be used to add meaning to numbers and numbers can be used to add precision to words. It is easy to think of situations in applied linguistics when we are interested at the same time in both the exact nature and the distribution of a phenomenon (Dörnyei, 2007, p. 45).

The situation that called for greater numerical precision from this research was the desire to learn the proportion of test takers who favoured each mode for a particular test feature. Likewise, the qualitative data provided greater insight into the wide variety of reasons why a particular mode might be better preferred with respect to a given test feature.

A third reason for using a mixed methods design is the hope of reaching a variety of different audiences. Individuals who are involved in educational assessment often come from measurement (i.e. quantitative) backgrounds and therefore they tend to feel more comfortable with statistical analyses of tests. Additionally, many administrators and policy makers who may be interested in the findings reported here may also be more oriented to statistical representations of these phenomena. Therefore, it is important to address these audiences in a manner that they expect using evidence that they accept as legitimate. Equally, there are a growing number of applied linguists and assessment researchers who see great value in qualitative data and believe it is crucial for an honest and accurate portrayal of educational phenomena.

Considerable debate still exists around the commensurability of quantitative and qualitative research methodologies. On the one hand, purists contend that the widely divergent worldviews influencing the ontological, epistemological and axiological assumptions underlying quantitative and qualitative methods preclude their coherent combination (Onwuegbuzie & Leech, 2005). However, Johnson and Onwuegbuzie (2004) argue for the value of a classical

pragmatic philosophical orientation to mixed method research. In essence, the pragmatist view is that the ultimate value of an idea derives from its empirical and practical consequences. They acknowledge that at present a pragmatic view of research is not completely able to solve the tensions exists within the conflicting worldviews of qualitative and quantitative research perspectives. However, they suggest that "Mixed methods research should, instead (at this time), use a method and philosophy that attempt to fit together the insights provided by qualitative and quantitative research into a workable solution (p. 16)." In their view, the pragmatic approach must be drawn upon precisely because it allows for important research questions to be investigated rather than ignored due to uncertainty regarding how to appropriately underpin a study philosophically. Furthermore, as Rossman and Wilson (1985 cited in Dörnyei, 2007) point out, a "pragmatist" approach to second-language research offers the additional benefit that "some sort of integration of the two research methodologies can be beneficial to 'corroborate (provide convergence in findings), elaborate (provide richness in detail), or initiate (offer new interpretations) findings from the other method (p. 627)."'

3.8 Description of the Study Design

This study took place in the participants' school setting so all aspects of the environment could not be completely controlled. A modified Latin squares design (see Chihara et al, 1992) was selected for this investigation for two reasons. First, this design allowed for comparison of participants' performance on paper-based cloze tests with their performance on computer-based versions of the same test. This is essential because the central goal of this investigation was to determine whether there is a statistically significant difference in test performance between these two testing modes. Second, there is no control group in the traditional sense because each participant will be serving as his or her own control. Nevertheless,

the paper version of the test could be considered to be a type of control while the computer could be thought of as the treatment condition because the paper test has already been validated and its reliability has been confirmed. Gribbons and Herman (1997) explain that a Latin squares design allows for all groups to take part in more than one randomly-ordered treatment and control conditions. As is illustrated in figure 2 below, at the outset of this study, half of the participants were randomly assigned to take the paper-based version of the LOMERA and the other half took the computer-based version. One month later, test takers switched roles and took the test in the mode that they had not initially taken it.

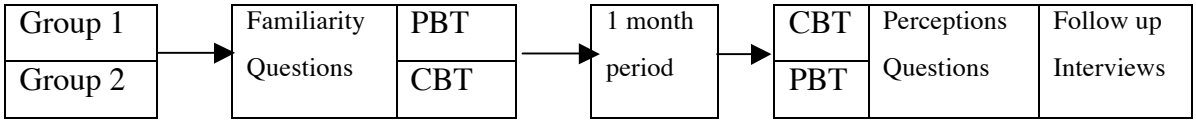


Figure 2. Diagram representing the procedures in the study

3.9 Research Timeline

The timeline for this research was as follows: Recruiting teachers and participants began in November of 2010. The school district approved the project on March 9, 2011. The UBC Review Board approved the study proposal on March 14, 2011. Permissions from parents and participants were secured in March 2011. Data was collected in March and April of 2011. These data were analyzed from May 2011 through September 2011.

3.10 Treatment of Potential Threats to Internal Validity

During the study design process, several potential threats to internal validity became apparent. To preserve the validity and strengthen the design, these threats were clarified and steps were taken to eliminate or minimize them as much as possible. These threats were: subject

characteristics, mortality, instrumentation, testing, history, maturation, and implementation. The subject characteristics that are most relevant to this investigation are: age, gender, first language, L2 proficiency, intelligence, attitude, and socioeconomic status. These variables must be taken into consideration because disproportionate differences between study participants on any of these variables may unduly influence the results.

The variable of age was accounted for by selecting study participants from within a relatively narrow age range. This group ranged in ages from 13 to 18 with a couple of 19 and one 20 year old. Exact numbers in each age category are presented in the results chapter. This group is the most appropriate choice for this research for two main reasons. First, as was mentioned above, there has not been enough research with them. Secondly, this was the group of volunteers the school personnel provided access to and parents gave permission to study. Being able to collect data from ESL learners in the higher grades was fortunate because ESL students often disappear from school over their high school career (Gunderson, 2007). Clearly, this sample will limit generalizability to only the age group selected for the study, but it will also mitigate the inordinate influence of age as an intervening variable.

Controlling for the other subject characteristic threats to internal validity of the study was accomplished by randomly assigning the subjects to the paper-first or computer-first groups. This helped distribute any distinctive background variables more evenly among the participating groups. Furthermore, study participants also served as their own control. That is, participants' performance on each of the testing modes was compared with their own performance on the other testing mode.

Participants' proficiency level was another key background variable that was specifically taken into account here as well. At the outset, their proficiency level was based on the ESL level

that they were assigned at their school. Level is an important consideration here and it was imperative that a balanced mix of levels be recruited. If not, the concern is that learners with lower proficiency might have presented too much variability in terms of their language learning between test administrations. This variability could inordinately affect the outcomes of the study. As well, they may have had difficulty understanding the questionnaire which could affect the validity of their answers. Conversely, learners with higher proficiency might have relatively smaller numbers and recruitment and mortality might have presented challenges during the course of the study. Data presented in chapter 4 show that there was a suitable balance of levels recruited.

Confounds from data collector characteristics were negligible because the researcher was the sole collector of all data used in the study. Having only one test administrator and using standardized test procedures helped avoid possible threats to validity posed by idiosyncratic implementation of the data collection procedures (Wallen & Fraenkel, 2005). Including participants from both groups one and two in each testing session helped reduce potential confounding variables such as researcher bias and possible divergent testing conditions (Chihara et al., 1992). Scorer fatigue was avoided by not scoring too many tests at one time. Lastly, a wide discrepancy in scores (i.e. intra-rater reliability) is not a serious concern because the assessment instruments are not open-ended and therefore rely less on subjective scorer judgment.

Threats from testing were controlled by waiting for a suitable amount of time between the administrations of the first and the second tests. This wait time also lessened practice effect. Threats from diverse test-taker learning experiences between administrations of the test were also addressed through minimizing the interval between test administrations and by confirming with school staff that there are no scheduled activities that could appreciably impact the results

of the study. Instrument decay was minimized by not changing any of the data collection instruments after the study had begun.

Subject mortality was reduced by maintaining a brief interval between the administration of the tests and by requesting subjects who agreed to do the first test to commit to doing a follow-up test. Significant mortality was not expected because during the initial norming of the LOMERA there was a lot of positive feedback from test takers which indicated that they somewhat enjoyed taking the test. This turned out to be the case in this research as well.

3.11 Summary

This chapter detailed the data collection and analysis procedures that were used in this study. This account included an explanation of study participants selection and recruitment. Instruments used for data collection were explained and issues related to the internal validity of this research were also discussed. Data collection procedures were outlined and the research site was described. The timeline of the research was delineated.

Chapter 4 will present the results of the data analysis. This will include a discussion of LOMERA validation and reliability checks. Participant demographic data will be presented as will results of the t-test and DIF comparability analyses. Results from the regression analysis using familiarity variables will be reported and discussed.

Chapter 4 Comparability and Familiarity Results and Discussion

4.1 Introduction

This chapter contains the results for the cross-mode comparability phase of this study. First, the procedures for the analysis of the data are explained. This is followed by the presentation of the descriptive statistics for the study sample. Next, results from the t-test, reliability and correlation analyses that were performed to establish the cross-mode comparability of the entire test are reported and discussed. Findings from the differential item functioning analysis are reviewed and considered along with p-values comparisons that were conducted to ascertain the degree to which individual test items might have functioned differently across modes. Lastly, results from the regression analysis comparing the familiarity questionnaire with performance on the computerized LOMERA are presented and discussed.

4.2 Data Analysis

Analyses of data from several sources were conducted to answer the research questions asked in this study. First, psychometric characteristics such as the distribution, rank, and correlation of scores on the two tests were examined as a preliminary sign of cross-mode comparability (Choi et al., 2003). These indicators of comparability also meet the criteria set forth by the testing organizations such as the American Psychological Association (APA) and the International Test Commission (ITC).

Upon establishing the acceptability of psychometric characteristics, several other statistical analyses were performed to evaluate the comparability of these two tests. The first of these was a series of independent-sample t-tests of mean scores for group one and group two on the paper and computer tests for the first and second test administrations. These tests ensured that

there was no statistically significant difference in the scores between the two groups which provided evidence for their comparability. Because there were no statistically significant differences in test performance between the two groups, a paired-samples t-test was done to compute the average difference between scores on the paper and computer-based test for each test taker to establish whether the average variance was significantly different from zero. A significant difference between the two scores indicates mode effect.

Inter-passage and inter-test correlations across modes were then calculated to determine how test takers' scores on one mode of the test correlates with their scores on the alternate mode. The purpose of computing these correlations was to discover whether raw test scores were similar across two modes. Results of these investigations further corroborate cross-mode comparability from the psychometric perspective (APA, 1986; Choi et al., 2003; ITC, 2006). Lastly, p-values were calculated to determine the proportion of test takers that answered each test item correctly in each mode of the test. These p-values were then plotted on a scatter plot chart to allow for a Delta-plot differential item functioning (DIF) analysis to be conducted. This DIF analysis indicates whether individual test items perform differently across modes. The results of all of these analyses are presented and discussed below.

Upon determining the degree of comparability between the two measures, participants' results from the computer familiarity questionnaire were compared with their performance on the computer-based testing mode using a multiple regression analysis. First, the categorical variables elicited by the questionnaire were recoded into dummy variables to allow for meaningful interpretation of the results. The categories in these new recoded variables were then compared with a constant to establish whether there were statistically significant differences in test takers' computer scores depending on their level of self-reported comfort with computers, comfort with

computer tests and number of tests taken on computer. Each of these variables was taken as an indicator of computer familiarity. The purpose of this analysis was to establish the degree to which computer familiarity predicted their performance on the computer version of the LOMERA. Therefore, the variables' ability to predict performance on the computer version of the LOMERA is viewed as an indicator of the effect of computer familiarity on computer test performance. Results of these analyses are reported in the following section below.

4.3 Demographic and Descriptive Findings

The sample size used for this study consists of 120 individuals. There were 61 females and 59 males in the sample. The mean age was 15.73 (SD = 1.67). The youngest participant was thirteen and the oldest was twenty. The following tables (2, 3, and 4) contain breakdowns for number and percentage of students in each grade, each ESL level and each first-language group designation.

Table 2
Frequencies of School Grade

	Grade	Frequency	Percent	Valid Percent
Valid	8	17	14.2	14.3
	9	17	14.2	14.3
	10	27	22.5	22.7
	11	33	27.5	27.7
	12	25	20.8	21.0
	Total	119	99.2	100.0
Missing	99	1	.8	
Total		120	100.0	

Table 2 above shows that there are a slightly larger proportion of participants in the higher grades than in the lower grades.

Table 3
Frequencies of ESL Level

	ESL Level	Frequency	Percent	Valid Percent
Valid	1	29	24.2	24.2
	2	31	25.8	25.8
	3	38	31.7	31.7
	4	22	18.3	18.3
	Total	120	100.0	100.0

Table 3 demonstrates that there is generally comparable proportion of levels although there are marginally fewer test takers in level four.

Table 4
Frequencies of First Language Group

		Frequency	Percent	Valid Percent
Valid	Mandarin	12	10.0	11.4
	Vietnamese	8	6.7	7.6
	Tagalog	22	18.3	21.0
	Kinyarwanda	1	.8	1.0
	Ilocano	8	6.7	7.6
	Ilongo	2	1.7	1.9
	Bisaya	5	4.2	4.8
	Cantonese	5	4.2	4.8
	Chinese	24	20.0	22.9
	Spanish	2	1.7	1.9
	Korean	1	.8	1.0
	Russian	2	1.7	1.9
	Arabic	1	.8	1.0
	Portuguese	1	.8	1.0
	Bahnar	2	1.7	1.9
	Tamil	1	.8	1.0
	Tigrinya	2	1.7	1.9
	Burmese	1	.8	1.0
	Jarai	4	3.3	3.8
	Karen	1	.8	1.0
	Total	105	87.5	100.0
Missing	99	15	12.5	
Total		120	100.0	

A number of test takers (12.5%, N = 15) did not report their first language and among those who did a number identified their first language as "Chinese" (20%, N = 20), but did not specify whether or not it was Mandarin, Cantonese or some other dialect of the language. It should also be noted that a sizable proportion of the study participants were native speakers of Asian languages. Although Immigrants from Asia make up the largest group moving to urban centers around North America (Statistics Canada, 2010) this sample may not be entirely representative of immigrant demographics in Canada.

4.4 Comparability Results

The typical methods for examining comparability are psychometric characteristics such as the distribution, rank, and correlation of scores on the two tests (Choi et al., 2003). These indicators of comparability also meet the criteria set forth by the testing organizations such as the American Psychological Association (APA) and the International Test Commission (ITC). The ITC points out that developers of computerized tests need to "…produce comparable means and standard deviations or render comparable scores (International Test Commission, 2006, p. 156-157). The mean for the first administration of the paper-based test was 65.5 and the standard deviation was 18.3. The mean for the computer-based test was 68 and the standard deviation was 21.2. The mean for the second administration of the paper-based test was 61.6 and the standard deviation was 18.7. The mean for the second administration of the computer-based test was 66.4 and the standardization was 18.7. These descriptive statistics can be found in table 5 below.

Table 5

Descriptive Statistics

	Test 1		Test 2	
	N = 120		N = 120	
	Mean	SD	Mean	SD
Paper based	65.5	18.3	61.6	18.7
Computer based	68.0	21.2	66.4	18.7

4.4.1 t-test Analyses

LOMERA test takers' scores across testing mode both within and between administrations were compared using t-tests. These tests were performed bearing in mind Siegel's (1992) advice that "there are some clear-cut cases of multiple t tests that do not pose any problem. One case is when none of the t-tests is significant" (p. 774).The purpose of this series of t-tests was to identify whether there was a statistically significant difference in test takers' performance across test modes which might call into question the comparability of the groups that were being compared.

Table 6

Results of t-test of Mean Test Scores on First and Second LOMERA Administration

	Paper			Computer					
	Mean	SD	N	Mean	SD	N	t	df	Sig.
Test 1	65.5	18.3	60	68.0	21.2	60	.70	118	.486
Test 2	61.6	18.7	60	66.4	18.7	60	-1.38	118	.169

Table 6 describes the means and standard deviations of raw scores for the first and second administration of the LOMERA by mode of test administration. The independent sample

t-tests were performed only to provide some indication that the two groups were comparable. The results of an independent-samples t-test comparison scores from the paper and computer version of the LOMERA for the first and second administrations are also reported. Results of the t-test for the first test administration revealed that there was no statistically significant difference in mean test scores between those who took the paper test (M = 65.5, SD = 18.3) and those who took the computer test (M = 65.5, SD = 18.3), t(118) = .7, p = .48 across modes. Findings from the second administration were that those who took the paper test (M = 61.6, SD = 18.7) and those who took the computer test (M = 66.4, SD = 18.7), did not perform significantly differently across modes either t(118) = .-1.38, p = .16. No statistically significant difference in the mean LOMERA test scores across modes in either test administration provides evidence of the comparability of the two groups between the paper and online versions of the test.

Table 7

Results of t-test for Order of Test Taken and Mean Test Score

	Paper first			Computer first					
	Mean	SD	N	Mean	SD	N	t	df	Sig.
Paper score	65.1	17.8	59	62.5	18.9	54	-.73	111	.464
Computer score	66	17.9	60	60.4	18.8	60	-1.65	118	.1

Results of t-tests comparing raw score means within test mode by order in which test mode was taken are shown in table 7 above. The purpose of this test was to collect further evidence that both groups displayed comparable results when they took the test in the same mode. Analysis of scores on the paper test for those who took it first (M = 65.1, SD = 17.8) compared to those who took the computer test first (M = 62.5, SD = 18.9) revealed no significant

difference in scores on the paper mode for either of these two groups t(111) = -.73, p = .46. A comparison of scores of the same two groups on the computer-based LOMERA demonstrated that the paper-first group (M = 66, SD = 17.9) did not have a statistically significant difference in their computer-based test score than the computer-first group (M = 60.4, SD = 18.8) t(118) = -1.6, p = .1. These results indicate no statistically significant difference in test scores within modes across test administrations. This finding supports the claim that the two groups of test takers demonstrate similar English ability as it is measured by each separate mode of the LOMERA.

Table 8

Results of Paired Sample t-test Comparing First and Second Test Administration

	Paper			Computer					
	Mean	SD	N	Mean	SD	N	t	df	Sig
Test 1 and test 2	63.8	18.3	113	63.3	18.7	113	.933	112	.353

The final t-test analysis completed for this study was a paired-sample t-test. Its purpose was to identify any discrepancies in each test taker's scores between one mode of the test and the other by comparing his or her mean final scores on the paper and computer tests. The t-test revealed that the mean paper test score for all test takers (M = 63.8, SD = 18.3) was not statistically different than their mean computer test score (M = 63.3, SD = 18.7) t(112) = .93, p = .35. This test yields a convincing piece of support for the absence of mode effect because, in this instance, each individual test taker's scores across both modes of the test are compared. No statistically significant difference in the means shows that when the same test taker's scores are averaged and compared across modes no mode effect is present.

As a follow up to the t test, a Cohen's d effect size of 0.166 was calculated and based on the effect size a post-hoc power analysis was performed. The results of the power analysis for a 2-tailed paired-sample t test were 0.144. Fourteen percent power indicates that that there is insufficient power in the study to detect a true difference among the means. The power of a statistical test is the probability that the test will not make a false negative decision (i.e. fail to detect a difference in means that actually existed or Type II error). As Gamst, Myers and Guarino (2008) point out "increasing the power of a statistical test allows you to increasingly discern differences that were not apparent under a lower power test" (p. 45). Consequently, the somewhat low power of this paired-sample t test reveals that it cannot be relied upon alone to ensure the cross-mode comparability of the LOMERA. Therefore, other indicators of comparability are required to support the findings of the t-test.

Overall, the results of the paired-sample t-test show no statistically significant differences in mean test taker scores across modes. These findings indicate an absence of mode effect between the paper and computer versions of the LOMERA though caution must be taken when interpreting these results due to the low power of the test.

4.4.2 Reliability Analysis

Using SPSS 16, a Chronbach's α reliability analysis was performed on the data as a measure of internal consistency for both the paper and computer LOMERA test administrations. The analysis for the paper test produced a very high reliability coefficient (96 items; $\alpha = .95$). The analysis of the internal consistency of the computer test also resulted in a very high reliability coefficient (96 items; $\alpha = .95$).

4.4.3 Correlation Analyses

A series of Pearson product-moment correlations was conducted to distinguish how test takers' scores on one version of the test correlated with their scores on the alternate version to ascertain whether examinees achieved similar scores across the two modes. A close correspondence between these two scores supports cross-mode comparability because it demonstrated that examinees ranked the same way on one mode of the test as they did on the other. A significant correlation of $r = .96$ ($p < .001$) was found when the scores from the first and second administration were correlated. This is a statistically significant and very high correlation showing that test takers' scores are closely related across modes; thus it provides additional evidence of cross-mode comparability.

Additional correlations were computed with individual passage scores across modes to discern the degree of relationship between examinees' passage scores across the paper and computer modes. The correlations that are of most interest here are those from the same passage taken across different testing modes (see table 9). Correlations between passage one on the paper and computer mode were moderately high $r = .74$ ($p < .01$). Passage two had a cross-mode correlation of $r = .76$ ($p < .01$). The paper and computer passage scores correlated at $r = .77$ ($p = .01$) for passage three. The fourth paper and computer passages had a correlation of $r = .77$, $p = .01$. The cross mode correlations for passages five and six were moderately high. The correlation for the paper and computer scores in text five was $r = .81$, $p = .01$. Passage six produced a cross-mode correlation of $r = .85$, $p = .01$. The paper version of passage seven had a lower correlation with its computer counterpart at $r = .60$, $p = .01$. The same was true of passage eight at $r = .65$, $p = .01$. All of the correlations were $r = .60$ or higher. Most were above $r = .70$ and some were as high as $r = .85$. All of these relationships were statistically significant at the .01 level as well which indicates a genuine association between the passage scores across modes.

Table 9

Correlations for Individual Passages from the LOMERA across Modes

	PBT1	PBT2	PBT3	PBT4	PBT5	PBT6	PBT7	PBT8
CBT1	.743**	.574	.654	.600	.590	.661	.495	.548
CBT2	.702	.766**	.747	.713	.642	.738	.578	.679
CBT3	.729	.712	.771**	.699	.752	.789	.579	.728
CBT4	.688	.689	.724	.776**	.768	.719	.611	.758
CBT5	.604	.684	.674	.759	.819**	.767	.550	.690
CBT6	.659	.671	.684	.731	.789	.857**	.673	.760
CBT7	.619	.594	.611	.629	.710	.714	.607**	.625
CBT8	.507	.585	.562	.625	.656	.640	.512	.653**

** $p < .01$.

High and statistically significant ($p < .01$) correlations can also been observed (see Table 9) among passages within each mode. That is, there are significant correlations (in the .5 to .78 range) between each passage in the paper test and all of the other passages in the computer version. This pattern is evidence that the test passages are generally measuring the same thing.

4.4.4 Detection of DIF

The next stage of the analysis involved the creation of a scatter plot diagram to visually represent the relationship of examinees' performance on each individual test item across modes. There were several steps involved in the process of creating the Delta plot chart. First, scores from each of the eight test passages were transformed from a per-passage score to a binary per-item score of 1 for correct and 0 for incorrect and entered into a database. After that, p-values were calculated for both the paper and computer versions of each test item. Ordinarily the p-

values are then transformed into Delta values to standardize the scores and allow for easier comparison. However, in this instance, the p-values were not transformed because they came from the same population of examinees taking the same test. Using SPSS 16, the obtained p-values were then used to create a scatter plot that plotted the intersection of test takers' scores for each individual test item on both modes. Subsequent to plotting the paper and computer score for each item on the scatterplot graph, a regression line was added to clarify the general shape and direction of the plots. Two lines demarcating the 95% confidence interval were also added in relation to the least-squares regression line. This enabled the DIF analysis by distinguishing the area within which item plots had to be located to be considered not functioning differently. Following the advice of Muniz, Hambleton and Xing (2001), items outside 95% confidence interval band around the regression line were deemed to be DIF because discrepancy in examinee performance on one mode differs significantly from their performance on the other mode.

The first noteworthy observation from the Delta-plot DIF analysis was that the regression line passes through the Y axis rather than the origin of the chart. This finding reveals a slight difference in overall performance across modes in favor of the computer version of the test. However, the t-test and correlation analysis findings determined that there are no statistically significant differences between the two modes of the test so the placement of the regression line is not a cause for concern.

Results for whether there are dissimilarities in subjects' scores across modes on individual test items were that four items demonstrated cross-mode DIF according to the Delta plot method. That is, these items were outside of the 95% confidence interval band around the regression line. As is illustrated in Figure 3, all of the DIF items were found to be biased in favor

of the computer test. Besides these four items, the others are all within the 95% confidence interval and thus do not reveal any DIF. The numbers of the DIF items were 20, 30, 44, and 92. These items were inspected more closely in an attempt to ascertain what might be causing the cross-mode DIF for them. The actual sentences and their deletions for those that are outside the confidence interval are presented in Appendix F. Some possible causes will be discussed below.

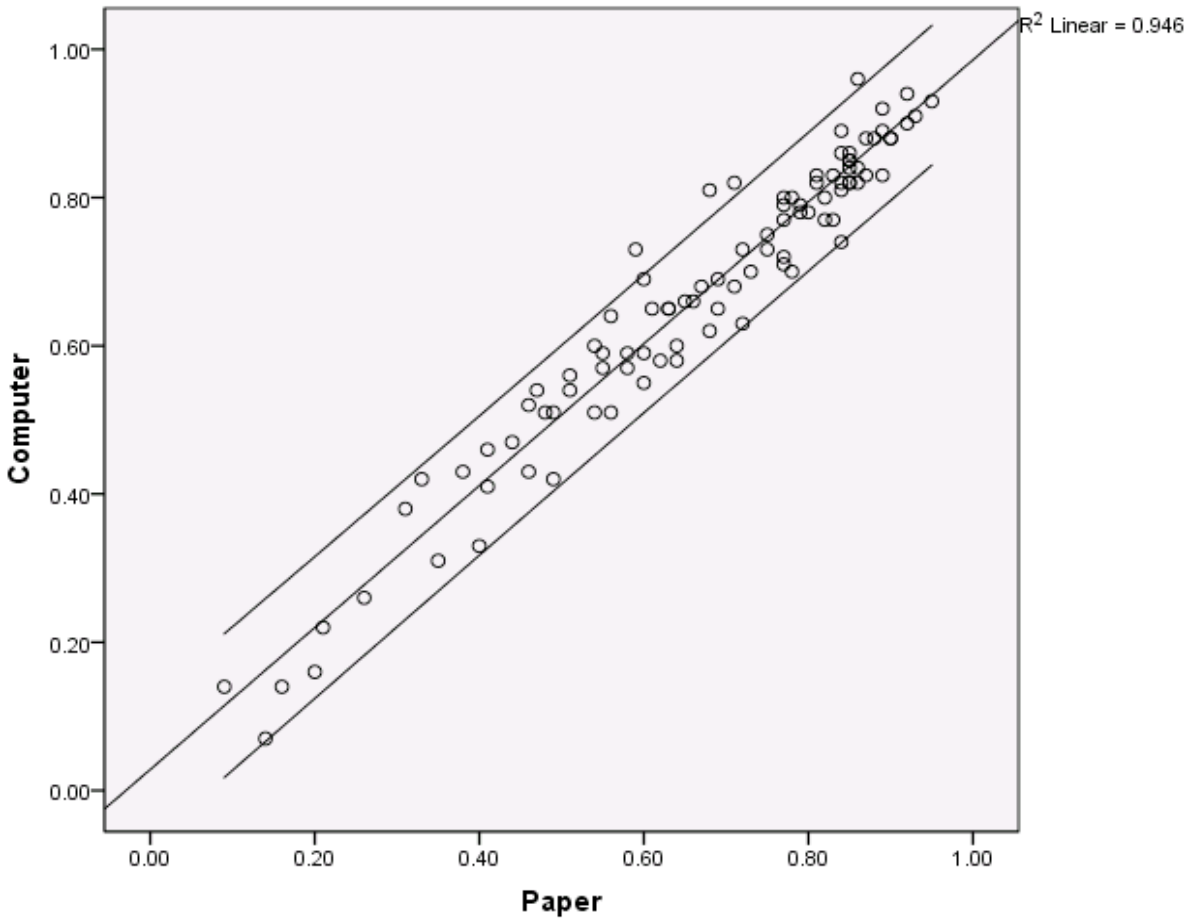


Figure 3 Scatter Plot of Item p-values on Paper and Computer Versions of LOMERA

4.4.5 P-values Comparisons

In addition to t-test statistics, correlations, and DIF analysis, item facilities (p-values) were also calculated and compared for both modes for each test item. Item facility is a statistic

that is used to determine the percentage of students who correctly answer a particular test item. The purpose of this stage was to determine whether there was any systematic discrepancy in test takers' performance on individual test items across modes. The results were that the paper test items had higher p-values for 43 items, the computer had a higher value for 40 items and 9 items had the same value. This indicates that test takers performed slightly better on a few more of the paper items. However, given the position of the majority of test items around the regression line in the scatter plot graph and the relative insignificance of the differences as determined by the DELTA plot analysis, there did not appear to be any real discrepancy in test takers' scores at the item level overall.

4.5 Results of Computer Familiarity Questionnaire Analysis

Table 10
Descriptive Statistics for Mean Scores and Standard Deviations for Familiarity Dummy Variables

	Mean	SD	N
comptotal	63.23	18.566	120
somewhat comfort	.050	.218	120
comfortable	.366	.483	120
very comfort	.550	.499	120
somewhat comfort	.266	.444	120
comfortable	.516	.501	120
very comfort	.158	.366	120
1-2	.383	.488	120
3-4	.250	.434	120
5 or more	.141	.350	120

Descriptive statistics for the criterion variable, computer test total score, and each predictor variable used are reported in Table 10. Figures included in the table are the mean score for the variable, standard deviation and number of participants.

Table 11

Correlations between LOMERA Computer Test Score and Dummy Variables Representing Self-reported Comfort with Computers, Taking Tests on Computers and Number of Tests Taken on Computers

		Somewhat Comfort (computer)	Comfortable	Very Comfort	Somewhat Comfort (Test)	Comfortable	Very Comfort	1-2 Tests	3-4	5 or More
Pearson Correlation	Comptotal[1]	-.019	-.096	.235*	-.092	.253*	-.103	.168*	.052	-.261*
	Somewhat Comfort (comp)[2]	1.000	-.175	-.254*	.294**	-.237*	.005	-.024	-.044	-.093
	Comfortable		1.000	-.841**	.128	.009	-.235*	.076	.000	-.111
	Very Comfort			1.000	-.326**	.164*	.209*	-.045	.019	.079
	Somewhat Comfort (test)[3]				1.000	-.623**	-.262*	.145	-.131	-.083
	Comfortable					1.000	-.448**	.077	.096	-.133
	Very Comfort						1.000	-.201*	.013	.217
	1-2 Tests[4]							1.000	-.455**	-.320*
	3-4								1.000	-.235**
	5 or More									1.000

* Correlation significant at the .05 level

** Correlation significant at the .001 level

[1] Final score on the computer-based LOMERA

[2] Recoded dummy variables indicating comfort with computers

[3] Recoded dummy variables indicating comfort with taking tests on computer

[4] Recoded dummy variables indicating number of tests taken previously on computer

Correlations between the predictor variable "comptotal" or total LOMERA score on the computer and each of the dummy independent variables are found in table 11. Test takers' self-reported data about how comfortable they feel using computers and how comfortable they feel taking tests on a computer were correlated with their overall scores on the computer version of the LOMERA. The correlations are all less than .3 which shows there is minimal correlation among any of these variables which indicates the lack of a strong relationship among any of them. These low correlations also demonstrate that there is not a great deal of colinearity among the variables which is an assumption that must be met to justify their use in the regression analysis.

Table 12
Summary of Multiple Linear Regression Analysis for Variables Predicting Computer-based LOMERA Score

Source	Sum of Squares	df	Mean Square	Adjusted R^2	F Value	Sig.
Regression	10905.20	9	1211.68	.206	4.42	.001
Residual	30115.72	110	273.77			
Total	41020.92	119				
Predictor Variable	B	SE B	β	t	p	
(Constant)	15.02	11.56		1.299	.197	
somewhat comfort (comp)	31.48	11.01	.371	2.859	.005	
comfortable	30.25	9.08	.789	3.329	.001	
very comfort	37.15	9.03	1.000	4.115	.000	
somewhat comfort (test)	13.79	7.20	.330	1.913	.058	
comfortable	16.06	6.73	.434	2.386	.019	
very comfort	10.22	7.47	.202	1.367	.174	
1-2 tests	4.35	4.10	.115	1.063	.290	
3-4	3.03	4.45	.071	.682	.496	
5 or more	-6.54	5.35	-.123	-1.221	.225	

a Predictors: (Constant), 5 or more, very comfort, comfortable, 3-4, somewhat comfort, very comfort, 1-2, somewhat comfort, comfortable
b Dependent Variable: comptotal

As displayed in table 12 above, adjusted R square = .20; F (9, 110) = 4.42, $p < 0.01$. (Several dummy variables were not significant predictors in this model). The standard error of estimate is 11.56. A multiple regression analysis was conducted with the dummy variables created for each category from the questions that asked test takers about their "comfort with computers," "comfort with taking tests on computers" and "number of tests taken on computer." This type of analysis was chosen to investigate the degree to which variables associated with computer familiarity influence computer test performance. The null hypothesis was rejected based on evidence that these variables associated with computer familiarity were statistically significant predictors of computerized LOMERA test performance (F (9, 110) = 4.42, $p < 0.01$). That is, there is some relationship between computer familiarity as it is measured in this study and computerized LOMERA test performance. As can be seen in Table 12 above, the combination of these predictor variables accounts for just over 20% of the variance in online LOMERA scores. Therefore, one reasonable conclusion is that while there is a statistically significant relationship between computer familiarity and computer test performance, familiarity did not exert an inordinate influence on the variability of computer test performance in this study. That is, while "comfort with computers" and "comfort with taking tests on computers" play some role in LOMERA computer test performance, it is not a predominant role.

The results of the present study establish that while variables that relate to computer familiarity do have some impact on online LOMERA performance, they are not the dominant factors in determining computer test scores. These findings indicate that further research should be conducted to ascertain more specifically if there are any other as-yet unidentified

factors related to computer familiarity that may be impacting computer test performance and what those factors might be. Additionally, other variables that are less obviously related to computer familiarity should be explored in future research as well.

4.6 Discussion

4.6.1 Descriptive Statistics and t-test Analyses

Based on the descriptive statistics and results from various t-tests reported above, there is no noteworthy disagreement in scores between the two versions of the LOMERA either within or across the two test administrations. This result provides evidence for the comparability of the computer version of LOMERA with its paper counterpart. In the case of Choi, Kim and Boo (2003), the Seoul University listening, grammar, and vocabulary subtests they studied had cross-mode discrepancies in means. The reading comprehension subtest had the largest cross-mode difference. However, they did not interpret these results as indicators of incomparability. They explained that there were significant mode effects for the listening comprehension, and reading comprehension subtest scores, but not for the grammar test. They contended that the mode effects for the listening and reading subtests were caused by the fact that most subjects found the graphic layout of the two modes of the listening and reading subtests to be quite different from each other. They also conjectured that the "negligible mode effects for the grammar subtest could be accounted for by the fact that the way in which the CBLT [computer-based language test] of grammar was presented was not very different from that of the PBLT [paper-based language test] counterpart" (p. 310). That is to say, it was the discrepancy in layout across modes rather than the content of the test itself that caused the observed mode effect.

Maycock and Green (2004) explored agreement rates between paper and computer versions of the IELTS. They found that both tests placed 50% of test takers within the same band and 95% placed them within a half band on a nine-band scale. They took this to be convincing substantiation of cross-mode comparability. In her review of cross-mode comparability research into reading tests across a wide variety of disciplines, Sawaki (2001) stated that she could only locate one study that dealt specifically with the comparability of second-language reading tests. She reported on the results of this study conducted by Yessis (2000). In the study, Yessis (2000) explored post-secondary advanced ESL students' cross-mode performance on a series of timed weekly reading tests. His design was counterbalanced so that the order of testing mode presentation was accounted for and test takers' language ability was taken into consideration. His mixed-model regression analysis revealed that there was no significant difference in test performance across modes.

Three studies reported statistically significant differences in test performance across modes as measured by t-test analyses. In his comparability study of a post-secondary ESL placement test, Fulcher (1999) found that there was a statistically significant difference between the paired-samples t-test that was used to compare subjects' performance on the two forms of the test. He used this finding and a correlation result discussed below to contend that the two forms of the test were not entirely comparable in contrast to the results of the present study. Coniam (2006) reported similar findings for the independent t-tests in his study. Four groups from two different schools taking a listening test were found to have statistically significant differences in their mean test scores. Al-Amri (2008) stated that the three paired-sample t-tests in his study showed statistically-significant differences in cross-mode scores but he pointed out that the small number of test items and large sample was

likely the cause of the discrepancy in scores. He contended that descriptive statistics for the three tests were more revealing about how similar the results were across modes. He highlighted that there was only a slight divergence in means and standard deviations across modes and that was better evidence for lack of mode effect.

4.6.2 Correlations

Correlations between both administrations of the LOMERA are .96 which indicates considerable agreement in scores across modes. This finding provides additional evidence for the cross-mode comparability of both versions of the LOMERA. The correlations on the individual passages across modes are generally between .70 and .85. These statistics are moderate to high and statistically significant. The lowest are .6 and .65 for passages seven and eight. These lower correlations among the more challenging passages could be caused by examinees' performance being variously affected by fatigue, possibly depending on the testing mode, toward the end of the assessment.

This result is in general accordance with the findings of previous research. Al-Amri (2008) also performed a cross-mode correlation analysis of the tests he studied and reported a correlation of .74 which he identified as being moderate. Choi, Kim and Boo's (2003) comparison of each of the subtests they studied with its cross-mode counterpart revealed that reading comprehension subtest had a correlation of .62 which was the lowest among all of the subtests. This contrasts with the overall correlation for the whole test which was .88 which appeared to satisfy these researchers. Correlations among the subtests ranged from .62 to .75. The relatively high correlations for six of the eight passages in the present study are generally in accordance with the moderate to high correlations reported in other research.

Fulcher (1999) found a correlation of .82 between his two versions of an English test to be an insufficient correlation to judge the two versions of the test as being comparable. Overall, it appears that the correlations reported in the present study are higher than those in the research literature. However, this may be due to factors such as greater test similarity or the type of test task. Only additional research can better illuminate the causes of the discrepancies between findings reported here and those of some previous researchers.

4.6.3 DIF Analyses

The results in the present study were that only four items demonstrated DIF were at odds with previous research. However, a review of relevant literature has only provided two studies that have used DIF methods to investigate mode effect. Schwarz, Rich, and Podrabsky (2003) used the Linn-Harnish and nonparametric Standard Mean Difference methods to analyze adult students' scores on the "In View" adult aptitude test. They found that eight items out of twenty demonstrated mode effect. That means almost 40% of test items revealed mode effect for the In View test. In contrast, the proportion of test items that showed cross-mode DIF in the present study was substantially less at only four percent. This considerable incongruity between these two studies in the number of items that showed DIF across modes may be due to the fact that Schwarz's et al. (2003) study was with adult learners who may have had less familiarity with computers than the secondary students in the present study. Another study of results for the Texas statewide standardized achievement test used a Mantel-Haenszel type Rasch Item functioning analysis (Keng, McClarty, & Davis, 2008). Their findings were that "Reading/ELA items that were longer in passage length…or involved scrolling in the online administration tended to favor the paper group (p. 221). They

did not discuss the proportion of items that were differentially functioning but only noted that there were discernable differences in cross-mode performance on particular items. Of relevance to the present investigation is their observation that one of the potentially problematic item types is related to reading comprehension. Clearly, further investigation is warranted with other types of reading test items such as those used in the present investigation.

There are several possible causes of the DIF identified in this study. A review of the DIF items revealed that all of the items were not of the same grammatical class; they were not spelled in a similar way nor were they the same length so these features of the key and distracter words do not appear to be the cause of the DIF. The only obvious similarity that all of the DIF items shared was that they were in sentences with at least three blanks in the same sentence. It could be speculated that examinees' possible increased level of enjoyment from using a computer for the test might have allowed them to persist in completing the item despite the increased cognitive load of having to complete sentences with multiple blanks which they may otherwise find to be excessively challenging in the paper format. Alternatively, measurement error or simply random chance could explain the DIF exhibited by only these four items. Further research will better clarify possible causes.

4.6.4 Familiarity

The findings of this study are generally consistent with other research literature on computer familiarity (see Clarania & Wallace, 2002; Higgins et al., 2005). Taylor, Kirsch, Eignor, and Jamieson's (1999) study of 1,200 TOEFL examinees reported that computer familiarity does not affect performance on a computer-based language test. They concluded

that there was “no meaningful relationship between level of computer familiarity and level of performance on the CBT [i.e. computer based test] language tasks” (p. 265). Another investigation with Saudi learners taking a reading test reached similar conclusions (Al-Amri, 2008). Sawaki (2001) offered similar observations in an exhaustive review of computer assessment-related research literature from a wide variety of academic disciplines.

The only cautions mentioned in other studies was the difficulty some students had with scrolling down the screen (Higgins et al., 2005) or having to use other interactive computer hardware such as the mouse (Pomplun et al., 2002) and the keyboard (Russell, 1999). In the present study, students only had to use the mouse for the test and they did not have to type anything using the keyboard. Nevertheless, the possibility exists that some test takers may have struggled slightly with the mouse. It might be advisable to investigate this aspect of the LOMERA test-taking experience more closely in the future. Another noteworthy limitation of the online test taking experience mentioned by Choi et al. (2003) was examinees’ inability to interact with the computer screen by “marking it up” as they might do with a paper test. During the testing sessions, a colleague observed that this interactivity was not as constrained in the computer mode as it first appears. In fact, there were several examinee-improvised methods of interacting with the text of the online LOMERA that appeared to be advantageous for those who were able to discover them. Two examples of such inventiveness were the manipulation of font sizes to make the test easier to read and using the cursor to “highlight” portions of text upon which the test taker was focusing. Future research might explore how test takers invent alternative means of making the online test taking experience more interactive. Designers might then attempt to incorporate these adaptations into future test iterations.

4.7 Summary

This chapter reviewed the findings for cross-mode comparability. These involved the descriptive statistics for the study sample, results from the t-test, reliability and correlation analyses and results from the differential item functioning and p-value comparison analyses. Additionally, results from the regression analysis comparing the familiarity questionnaire with performance on the computerized LOMERA were reported and examined. Chapter five will present the results of the mode perceptions survey and follow up interviews that asked participants to elaborate on their reasons for why they preferred one mode or the other on various key criteria such as readability and navigability among others.

Chapter 5 Mode Perceptions Results and Discussion

5.1 Introduction

A review of the findings of the perceptions survey and the interviews are presented in this chapter. The methods by which the data were analyzed are first described. LOMERA test takers' preferences for test mode are presented as well as their reasons for selecting their most preferred mode. This chapter is organized into three main sections: 1) the results and discussion for the test features that were generally more preferred on the paper test; 2) the results for features that were more preferred on the computer test, and; 3) the results for test features for which no mode preference was indicated. Each section includes the results of the survey followed by a presentation of common themes that arose in the interviews. Lastly, relevant research literature is discussed.

5.2 Data Analysis

Test takers' perceptions of paper and computer-based testing were investigated with a post-test survey. They were asked to report their mode preference on a variety of test features (e.g. readability or stressfulness) that past research has identified as potential sources of cross-mode discrepancy in performance. The percentages of respondents who favoured a particular mode on a given test feature were then calculated to provide an indication of which mode was preferred by the largest proportion of test takers for a particular test feature.

Ten separate one-on-one audio-recorded interviews with a subset of volunteer participants were conducted. The interviews allowed participants to elaborate on their responses to the perceptions survey. In addition to identifying the general mode preferences for particular test features, the interviews allowed respondents to elaborate on the reasons why they preferred a particular test mode with regards to a given test feature. A content

analysis of the interview data was conducted because it allows for convenient collection of perspective information with declarative data; in addition, it provides interviewees a voice in the research (Codó, 2009). The content analysis consisted of a thematic analysis of the data. This included coding and categorizing similar responses to each question based on common themes (Seidman, 1998).

The following sections present the survey results and the results of the analysis of the interview data. The analyses of these data address the final research question for the study. That is: which test features are favored and what reasons do LOMERA test takers provide for preferring certain test features in a particular mode?

5.3 LOMERA Test Taker Perceptions of the Paper and Computerized LOMERA

Table 13
Total Mode Preference Count by Mode

	No Difference	On paper	On Computer	Total
Total Mode Selection across All Criteria	27.65% (391)	38.33% (542)	34.01% (481)	1414

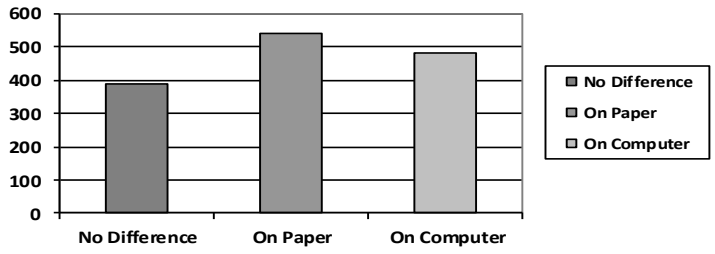


Figure 4. Total mode preference count by mode

Table 12 above summarizes the tally of LOMERA test takers' mode preferences based on test features commonly identified in the research literature to be potential causes of mode effect. The table and chart reveal that, according to the overall tally, LOMERA test takers appear to prefer the paper LOMERA (38.33%) over the computer test (34.01%). In the sections below, results of interviews are also presented with summary statistics indicating LOMERA test takers' overall preference for each specific LOMERA feature such as navigability, readability et cetera. Combining the questionnaire and interview results in this way provides a snapshot of LOMERA test takers' preferences in general and specifies some of the reasons why those volunteers who were interviewed favored one mode over the other. Both strengths and limitations of each mode are discussed with respect to each test feature in order to offer a broader perspective on LOMERA test takers' views regarding the constraints and affordances of each mode.

Table 14
Summary of Mode Preference by Test Feature Preferred on Paper Test

	Mode Preference			Sample size
	Paper	No difference	Computer	
Passage Navigability	45.8% (N = 55)	26.7% (N = 32)	25.8% (N = 31)	118
Passage Readability	55.8% (N = 67)	21.7% (N = 26)	20.8% (N = 25)	118
Answer Readability	41.7% (N = 50)	30% (N = 36)	24.2% (N = 29)	115
Less Tiring Test	42.5% (N = 51)	26.7% (N = 32)	29.2% (N = 35)	118
More comfortable test to take	44.2% (N = 53)	20% (N = 24)	34.2% (N = 41)	118
More accurately measured your reading comprehension skills	45% (N = 54)	33.3% (N = 40)	20.8% (N = 25)	119

5.4 Passage Navigability

Table 14 shows the results for test features that were preferred in the paper mode. The survey question asked LOMERA test takers to disclose which test was easier to navigate. The greatest number of respondents (45.8%, N = 55) stated that they thought the paper version was the easiest to navigate. By contrast, 27.7% (N = 32) claimed there was no difference and 25.8% (N = 31) contended that the computer test was easier to navigate. In the follow-up interviews that asked them to provide reasons for their mode preferences, several claimed to find the paper test to be more easily navigated while just as many preferred the computer test. One interviewee reported that there was no difference.

Those who said the paper test was easier to navigate provided three main reasons for their choice. One was that it was easier to turn the pages on the paper test than to select and

click on the page buttons on the computer screen. Betty[5] claimed that she preferred the paper version of the test because "it's easy to flip [pages] because you have [it] in your hands." John agreed that it was easier just to turn the page than to use the mouse saying with paper "it's easier like you can just flip it. And computer you need to point and hold the mouse." Innis was of the same opinion saying "I think that the paper is more easier because you can flip it back and then you can erase the answer the same as a computer…" These comments were somewhat counterintuitive because designers believed that being able to click a button to switch easily between the test passages would be one of the more appreciated features of the online version. Consequently, it was unexpected that some interviewees reported they found it more challenging to point and click the mouse on the page icon to turn to the next page than to physically turn the page. A second reason given for easier paper test navigability was the distraction caused by the clock. Betty remarked that the clock on the computer test distracted her because she felt that "…you're like getting hurry, oh the time's going, the time's going and you get nervous and stuff…" Designers believed that the clock would reassure LOMERA test takers by allowing them to keep track of and efficiently allocating their time during the test. However, as evidenced by Betty's response, this feature caused undue anxiety for some LOMERA test takers. Innis offered a third reason why he favored the navigability of the paper test saying "when you do the test in a computer it's like the monitor is going to hurt your eyes so I think it's more better doing the paper test." This indicates he believed that eyestrain from reading the computer test made it more difficult to navigate. In summary, it appears that being able to physically turn the pages of the test, distraction caused

[5] The names given to all respondents are pseudonyms

by the clock and eye strain were the main reasons given for why LOMERA test takers found the paper version of the test easier to navigate.

Those who preferred the computer test said that they liked being able to simply click the button to switch between pages. They reported finding it to be more bothersome to have to actually flip back and forth between pages. Donna expressed her preference for the computer mode explaining "paper you have to use [your] hands to flip around a computer you can just click it and it's easy." Jennie agreed saying "computer because you just click around, that's the main thing. Paper you have to flip back always." Test designers expected that the majority of LOMERA test takers would have this reaction, but, at least with this group, it appears that a few prefer the familiarity of holding the paper and turning the pages over the convenience of simply pointing and clicking. This result seems to run counter to the common stereotype that today's youth are more comfortable doing most tasks online. A second explanation given for the easier navigability of the computer test was its layout. Alice liked the navigability of the computer test "because I think it looks clear and the font is not that big and you don't have to switch pages around so it's easier in one page you can see everything and in the paper the font is bigger and I get confused." She appeared to prefer the fact that everything is located in the same "place" online in that she can go wherever she needed to with only one click. She also complained about the font in the paper version of the test saying it confuses her. Although care was taken to ensure that the fonts were as similar as possible across modes, a perceived discrepancy caused by mode of delivery may have resulted in discomfort for some LOMERA test takers.

One informant replied that the he found no difference between the two modes with respect to their navigability. Rodney commented that it was just as easy to turn the pages as it was to click a bar at the top of the screen that contained numbers of the test pages.

In summary, it appears that the ability to click on buttons to move from page to page within the test was viewed favorably by a number of LOMERA test takers. However, there were at least a few who did not prefer this feature though it seems their reasons for disliking it may relate to a disinterest in using computers in general. Reasons for perceived mode effect described in the research literature generally overlapped with those identified in this study. Pino-Silva's (2008) research with Venezuelan post-secondary EFL students revealed that lack of eyestrain and ease of moving among the pages were provided as reasons for the superior navigability of the paper test. In particular, he mentioned that a "few reports [were] received about reading on screen being bothersome" (p. 153). Participants in at least two studies also mentioned that the built-in clock indicating time remaining caused distraction while they were taking the computer test (Al-Amri 2008; Pino-Silva 2008). At least two computer mode criticisms were found in the research literature were not mentioned by any respondents here. One was that some participants were uncomfortable with not being able to see the entire test at will (LOMERA test takers could). Another source of distraction mentioned by a participant in Gorsuch's (2004) research was the sound of clicking computer keys from other participants who were doing the test. Findings from previous research are also consistent with the results of this study that participants find the paper test to be easier to navigate overall (Al-Amri 2008; Gorsuch 2004; Pino-Silva 2008).

5.5 Passage and Answer Readability

The second survey question had respondents identify in which mode they found the reading passages to be easier to read. A majority (55.8%, N = 67) favored the paper test. Of the remaining respondents, 21.7% (N = 26) reported that there was no difference and 20.8% (N = 25) favored the computer in this regard. Several reasons emerged for why the paper test was claimed to be easier to read. One explanation was the discomfort some felt reading from the monitor. This was somewhat surprising because the monitors that were used in the study were flat screen and state of the art. Nevertheless, Sally mentioned that "I can read it [the paper] easier because when I look at the screen, I think it's kind of dizzy [the screen makes you dizzy?] Yeah." Innis hinted at a similar problem saying "it [paper] is more easier because you can read more easily but then with the computer it's kind of confusing and yeah kind of when you see the monitor it's like kind of moving or something." Both of these comments reveal a potential problem with eye strain caused by reading the computer monitor.

A related readability feature mentioned by two respondents was the font on the computer test. Betty articulated how the paper was easier to read because "in the paper it [the font] was a little bigger, I think, and it's more easier…it makes you concentrate on only that paper and stuff…" John seemed to agree and commented that it was easier to look at the paper and how that allowed him to "understand each [and] every word." Jennie discussed difficulties with having to scroll through some parts of the LOMERA. She indicated that she preferred reading the paper test because "you don't really have to scroll down but in the computer you have to scroll down like it makes me kind of confused with the questions." This response was quite unexpected because the test was designed to require no scrolling. Each passage was carefully laid out on the page so that both the reading text and the

questions could be viewed at the same time. However, inevitably the browser settings on a few computers did not allow the entire passage to be shown on one screen and the LOMERA test taker had to scroll down approximately 3 or 4 lines. Interestingly, this minimal amount of scrolling was still noticed and viewed unfavorably.

Jennie reiterated a reason that was touched on in relation to navigating the test. That was how it made her feel more comfortable to actually be able to physically hold onto the paper. This comment seems to relate to another point that was raised in the interviews. Two other respondents remarked that they felt the paper test was easier to read because they had more experience with doing tests on paper. Bobby said "I feel comfortable with paper because it is used to, like, we have been using paper for the tests always so it's more comfortable." Donna expressed how she thought "Paper [was more readable] because computer you have to think about the answer and for me it's just hard to read. I come from China and we did a lot of tests on paper we never use computer to do a test so I just find papers easier." This point may be relevant to the issue of computer familiarity and test performance. That is, although LOMERA test takers may have the requisite computer skills, they prefer the paper test because that is the mode with which they are more familiar for taking tests. In other words, the issue may not be lack of computer familiarity per se but rather a preference for paper tests based on greater experience with taking them (i.e. more developed paper mode schemata).

Only two people who were interviewed said that they found the computer test to be more readable. When asked why she thought so, Alice said it was because "it looks really nice, it looks clear. Yeah, it looks great." This seems to indicate that at least some students may actually find computer screens to be easier to read and they do not experience eye strain.

Another male interviewee explained that he found the computer easier to read because "I think maybe I like computers better." When asked why he liked them better he answered that "yeah, I use computer a lot so…" (Rodney) These comments show that he believes his familiarity and experience increases his preference for the computer. Both answers serve as a reminder that not all learners experience eye strain.

Respondents were also asked about the readability of the test answers. As with the reading passages, the largest proportion indicated that they found the answers easier to read on the paper test (41.7%, N = 50). A follow up question was not asked about this survey response during the interviews because this researcher felt that the answer would overlap unnecessarily with the question about how easy it was to read the passage. That is, it appeared as though they were essentially the same question. Due to time constraints for interviews, questions were chosen that were thought to provide the most insights and useful information. In retrospect, it might have been worthwhile to ask this question in case it did yield unexpected answers.

Interview responses supporting the view that paper tests are more readable than computer versions are similar to published research findings. Eye strain was recognized as a major impediment to online test readability in the present study and in previous studies. Sawaki (2001) argued that the improvement of monitor technology would likely make this less of an obstacle in the future. However, 10 years later, the problem may not have been resolved. Participants in more recent research still complained about eye fatigue caused by the computer monitors (Al-Amri 2008; Gorsuch 2004; Pino-Silva 2008). Pino-Silva's (2008) research revealed that visual fatigue was one of the most frequently reported disadvantages of the computerized test. In another study, three of the participants felt that taking the test on

computer made their eyes overly tired (Gorsuch 2004). Likewise, Al-Amri (2008) reported that eye fatigue caused at least one participant great distress due to his overuse of computers as part of his daily routine. The question remains whether many secondary students are in the same position. Certainly, the risk exists for postsecondary students.

Comfort or familiarity with the testing mode (rather than with computers in general) was also mentioned as another determinant of readability. Positive or negative associations also had some bearing on readability in the research literature. Prior unpleasant experience with a specific testing mode (i.e. traditional paper based) made several Saudi respondents feel anxious and uncomfortable taking a computer-based test and provided them with a justification to resist efforts by school staff to introduce computer tests (Al-Amri 2008). A study of Canadian students' attitudes to the use of technology in foreign language teaching classrooms in several universities across the country revealed that reading texts (not necessarily for assessment purposes) on the computer was the only activity that students disliked and judged to be useless across all institutions surveyed (Peters, Weinberg & Sharma, 2009). The authors speculated that the cause of these negative perceptions may be the "intensive" focus on word-by-word reading as a teaching method that is often employed in their teaching contexts. In this way, unpopular teaching methods may taint students' perceptions about using computers in the classroom both for learning and assessment.

5.6 Less Tiring Test

In response to the question about which mode they found less tiring, the largest proportion of LOMERA test takers (42.5%, N = 51) stated that the paper test made them less tired. This compares to 26.7% (N = 32) who said there was no difference and 29.2% (N = 35)

who said the computer test was less tiring. The reasons provided also related to discomfort from reading the screen. Sally claimed that the paper test made her less tired because the computer test made her dizzy. Jennie mentioned that she found the paper made her less tired because **"**maybe it's the light in the computer [the light from the computer? How does it make you feel?] My eyes will feel stingy, more tired." Once again, eye strain causing fatigue appears to contribute to partiality for the paper over the computer mode.

Interviewees provided a wider variety of reasons for why they thought the computer test was less tiring. One commonly mentioned reason was ease in recording and changing answers. For instance, John replied that the computer test was less tiring because "you just need to click the answer." Betty agreed saying "computer because it was easier, like, if you want the answer you just click on it and stuff even though you're not getting it, it's easier to answer it." These remarks suggest that the convenience of clicking rather than filling in a Scantron card may alleviate some of the mental exertion caused by struggling with comprehending the passage or choosing the right answer while doing the test. Innis pointed out that having to turn through the pages of the paper test was more tiring. He said "yeah, I think still paper [is more tiring] because if you have some mistake you have to flip it back again and review it again." Rodney touched on the greater effort required to complete the paper test as well. He mentioned that "with paper I will have to fill. I will have to sit and use my hands but computer it's just a mouse, just move it around and, Sometimes paper you make a mistake and then you have to look for something to white out something, but computer it's not a big deal. You make a mistake you just change it right there. Only takes a second." He praised the computer test because it requires much less page turning, marking and erasing answers. This statement also speaks to the relative ease of changing answers on

the computer which will be discussed in more depth below. A final explanation for why the computer test was less tiring came from Donna. She said "…paper you have to very focus on it. You don't want an answer to be wrong and computer [is] just much easier to flip to next page and do really fast. To flip the page you can just click it and jump to the next page." This comment seems to relate to the computer test requiring less concentration to change pages or choose and change answers thus making it less tiring for some.

These results conflict with Al-Amri (2008) who reported that 55% of his participants answered that the computer test was less tiring. The participants in his study were given follow-up interviews but he did not elaborate on why interviewees found the computer test less tiring beyond mentioning that it was easier to read. One speculation is that the difference in test format between Al-Amri's test and the LOMERA might help explain why his participants found the computer test less tiring while LOMERA test takers reported the paper test was less tiring.

5.7 More Comfortable Test

The largest percentage of LOMERA test takers viewed the paper test as being more comfortable to take than the computer test (44.2%, N = 53). This was compared to 34.2% (N = 41) who preferred the computer version and 20% (N = 24) who said there was no difference. One common answer given for why paper was viewed as being more comfortable to take was familiarity with the mode. For instance, Donna said "definitely paper because I'm used to it." John appeared to agree claiming "For me, paper because I'm the kind of person that I always want to take everything on paper I'm not really good at computer…I'm not comfortable taking tests or something like that." This remark suggests that some students

may lack computer familiarity and do not enjoy using the computer. Jennie made a cultural connection when she mentioned “Paper, because I don't really feel any stress, and because, in my own country, I use paper that's why [Do you think that if you had more practice using a computer you might feel better about it or no?] No, I still like paper.” This echoes comments made by others who noted that tests done in their own country were all on paper. Therefore, that mode may be viewed as the most legitimate form of test and other testing modes may have questionable face validity for them. Accordingly, LOMERA test takers’ view of the “seriousness” of the tests may affect the effort they are willing to expend on doing the computer version.

Three justifications emerged for why the computer was thought to be more the comfortable test. The first was simply that it was more enjoyable. Bobby responded “it's like playing a game. It's fun.” As noted above, this theme ran through responses to several questions. Some of those who have experience using computers for gaming and other enjoyable tasks seem to transfer that positive experience to the online version of the LOMERA. The second reason was touched on by Alice. She said “Yeah, everything is easier because you just move the mouse around and just choose.” This comment echoes the sentiment that, for some, the computer was easier to navigate, and choose and change answers with only a mouse click which makes it more comfortable for some. Rodney raised an insightful point when he referenced his own physical discomfort with taking the paper test. He shared that the computer test was better because “…I don't have to bend my back. Like me, for example, I can't bend my back because I have a problem and with the computer I can just sit however I want and use my hand, but with the paper I had to do like this [slouches over].” This statement may be worth further consideration. Providing students with

the alternative of taking the computer test may provide an ergonomically comfortable option for some which may allow for optimal performance.

Conclusions from other studies are generally in accordance with the results presented here. In one study, interview and questionnaire data suggested that the majority of participants were not completely comfortable with taking foreign language listening comprehension tests on computers (Gorsuch 2004). Al-Amri (2008) also noted that participants who indicated paper as their preferred testing mode provided ease and comfort as their primary reasons why. Anxiety was also touched on as variable that has received some attention as a potential influence on computerized achievement scores. While significant differences in performance between the computer and paper participants were not found by Ward, Hooper, and Hannafin (1989), they did report that students testing on computers tended to have significantly higher anxiety levels. It is worthwhile to note that although students in the computer test group reported feeling significantly more anxious and agreed that testing by computer was more difficult, they still achieved comparable performance scores. Wise and Plake (1990) also found no evidence that computer anxiety had any connection to computer test scores for first-language college students. The issue of examinee physical comfort was addressed as well. Second language respondents in Pino-Silva (2008) mentioned the need to sit in comfortable chairs and maintain appropriate posture as important for preventing undue fatigue for those who took his computer test. This observation contradicted Rodney's comments about the relative physical discomfort of taking the paper test. In Rodney's view, ergonomic considerations are more of a concern for the paper test.

5.8 More Accurate Test

A large proportion of examinees also indicated that they thought the paper mode was a more accurate measure of reading ability. Findings were that 45% (N = 54) of respondents favored the paper mode compared to 33.3% (N = 40) no difference and 20.8% (N = 25) computer.

One reason provided for why the paper LOMERA was thought to more accurately measure reading skills was that some felt they had more difficulty reading the text on the computer screen. For instance, Sally said "[paper] is not going to make you, like, dizzy and something confused when you read the word." This statement hints at eye strain possibly making her feel as though she was not performing as well on the computer test. John mentioned a similar reason saying "You can look at [paper] easier and if you want, for example you have trouble in your eyes like not that good and you can see it clearly [the paper test?] Yeah [what do you mean about the computer? you can't see it clearly?] Because sometimes there's a computer that's like blurry, sometimes, and in the computer if your eyes get tired, like, it's hard to see the words [do you find as you're doing the computer test that your eyes get more tired?] Yeah, if the test has too much questions." This comment suggests that eye strain is a source of John's self-perceived discrepancy in performance across modes. He is also pointing out that problems with eye strain can be exacerbated by the quality of the monitor and the length of the test. Betty made a comment that seemed to allude to similar concerns about the accuracy of the computer test. She said that "I was like in that paper I was really concentrating and stuff and I didn't even worried about the time and stuff so I was just trying my best to, like, just answer the right question." This remark was a bit more complicated to interpret. She mentioned having worries about the time. This appears to relate

to the online test feature of being able to have a clock at the top of the screen which may have made her more anxious. A third reason given for perceived superior accuracy of the paper test relates to a theme that was also touched on earlier. This was the remark that the computer test may not be perceived to be as serious (i.e., lacking face validity) as the paper test. Donna noted that "paper is serious [paper is serious? Okay, can you tell me a little bit more about that?] Well... I'm not a computer person I don't play computer games. I don't, like, stay on the computer forever. I'm not the study person I like using just actually thing to finish the serious stuff." She seems to be saying that, for her, the paper was the more serious (i.e., legitimate or valid) test.

Not all interviewees said that the paper was more accurate. Two had more faith in the computer version. For instance, Jennie's answer that she thought the computer was more accurate was somewhat unexpected because she seemed to express a general preference for the paper test in her interview. She said "I think it is computer... it just gives me the feeling that, I don't know, of understanding more [the computer did?] Yeah, but I still like paper [so you like paper but you think that the computer does a better job?] Yeah, maybe, but I don't really care about flipping over again. I just like paper." Unlike some of the others who felt they were better able to read the paper test, she said she was better able to read the computer test. Rodney began his answer by saying that there was no difference in the accuracy with which he felt the two modes measured his reading ability. He said "I don't know I think both [Both?] yeah, both. I written the same so both." However, about half way through his answer, he changed his mind saying "But probably the computer will be more, like I said, sometimes I can't really see very far small thing I can't see it very clear so maybe paper was too small. I won't see something right but I can make the computer [font] bigger." This was a point that

he made earlier. He liked the fact that he was able to adjust the font size on the computer test to see it better. This was something that most LOMERA test takers could not do. This comment seems to indicate that he thought the computer test more accurately measured his ability because he had more control over manipulating the size of the font.

Pino-Silva (2008) surveyed EFL students' attitudes toward the use of computer-based tests and reported that respondents thought they were "accurate and reliable" (p. 152). He explained that "…a computerized test, as subjects clearly appreciate, guarantees that the scores obtained faithfully reflect student performance and are not the product of human error in the correction process" (Pino-Silva 2008, p. 152). Another study of 423 who took paper and computer-based versions of the IELTS exam reported that they were satisfied with the computer-based version of the test and they considered it to be comparable to the paper-based version (Green & Maycock, 2004). In contrast, Al-Amri's (2008) study of Saudi EFL learners showed that "About 51% of the subjects think that PBT measures their comprehension skills more accurately while only 23% think that CBT is better in this respect" (p. 40). This finding indicates that at least some learners may not have complete faith in computer-based assessments of reading comprehension. Al-Amri's results appear to coincide with the findings of the present study. That is, the majority reported that they thought the paper test was a more accurate test of their reading ability.

Table 15
Summary of Test Preferences by Mode by Test Feature Preferred on Computer Test

	Mode Preference		
	Paper	No difference	Computer
Less stressful test	36.7% (44)	25% (30)	37.5% (45)
Easier test to choose answers	35.8% (43)	20% (24)	43.3% (52)
Easier test to change answers	18.3% (22)	12.5% (15)	67.5% (81)
More enjoyable test to take	27.5% (33)	31.7% (38)	39.2% (47)

5.9 Less Stressful Test

Survey respondents indicated partiality for four aspects of the computer test. The largest proportion – albeit only marginally – (37.5%, N = 45) reported that the computer test created less stress for them compared to 25% (N = 30) who said there was no difference and 36.7% (N = 44) who said the paper LOMERA was less stressful (see Table 15). Among examinees who said the computer test was less stressful, two mentioned that, comparatively, the paper test was more stressful because of its layout. Alice said that the spaces the paper test used to demarcate the multiple choice answers confused her. She also complained that the different sizes of the font between that used in the passages and that used in the answers made her confused. Rodney found some fault with the layout of the paper version of the test as well. His point was that with the paper test he was forced to read the font size that was on the page. However, with the computer test, he had sufficient technical skills to adjust the size of the font (a feature that was unknown to the developers at the time). As a result, he explained “I don't know, like, when I had the paper, I have to hold it and look. With the

computer, I can make it bigger when the writing is small I can make it bigger. I have to use my brain more. With the computer I can make it bigger or smaller it's faster so it was easier for me." This was a useful discovery that the computer offers the affordance of being able to adjust the font size for those who know how. This feature allows passages to be more readable to some who may otherwise have difficulty seeing the text in either mode. Bobby provided another intriguing reason for why the computer test caused him less stress. He remarked "you know students, like, they always play games on the computer so we might be doing the test, like, we might think it's less stressful, I think, like you are playing a game." In this instance, he described how he found the test to be less stressful because he enjoyed playing games on the computer and this test felt like a game in some respects. His point raises the issue of the positive associations that many have with computers. It may be that online tests actually help those who enjoy using computers compensate for potential test anxiety by doing the test online. Therefore, it may be helpful to permit students to have the option to do the LOMERA online if it lowers their anxiety and allows for optimal test performance.

Some informants reported the paper test was less stressful because they had difficulty comprehending what they were reading on the computer. For instance, Sally remarked that the computer test was stressful because "it's dizzy and sometimes I can read words but it's just like I don't get it. [Is it the same way with the paper one? Do you ever read it and not get it?] No it's different." Innis agreed claiming that "you can understand more in the paper better than a computer." Eye strain seemed to be a related aspect of the experience of reading the computer test that caused some stress for at least one examinee. Jennie mentioned how she found the paper less stressful "because maybe it's the light in the

computer the light [the light from the computer? How does it make you feel?] My eyes will feel stingy ... more tired." The issue of struggling to read the test due to eyestrain or monitor-induced fatigue may have been the cause of at least some stress. Another reason given for why the paper test was less stressful was limited experience of time pressure. Betty found the paper less stressful "because you get to see, like, the time, when you are telling us the time it was like if there is, like, five minutes left at the last time there's only five minutes you have to tell or 10 minutes left you tell us but before that you don't get to know what time it is so you just concentrate on your paper and then you do it, so yeah..." This comment indicates that she felt pressured by having access to the clock on the computer screen while she took the online test. It appears that Betty desired to be periodically reminded of the time rather than have continuous access to a clock. John made a similar point when he mentioned feeling less stress from the paper test because "in every page there is only like one paragraph you won't be surprised like how many pages left." Because the test has one passage per page, test takers can know how many passages remain simply by counting the pages. This observation is actually mistaken because those taking the computer test can also know where they are simply by looking at the buttons along the top of the screen and counting how many passages remain after the highlighted button that indicates where they are in the test. The way of navigating among pages was explained, but it might have to be more carefully demonstrated for those who have less familiarity with computers.

One source of stress that was identified both during interviews and by at least one research study was the use of the computer clock. In the report, some interviewees stated the presence of the clock during the test caused them stress (Al-Amri, 2008). One participant in the present study raised the same criticism. Secondly, there is the game-like perception that

students often have of computers and it was associated with perceived stress. Several respondents stated that the computer test was easier to do and that made it less stressful. Al-Amri (2008) reported similar findings noting that the computerized test was convenient and quick to do. Pino-Silva (2008) reached similar conclusions and speculated that the socio-economically advantaged subjects in his study felt less stress with the computer test because they belong to the "post-Nintendo generation, and computers are second nature to them" (p. 152). Of course, not all immigrant students have easy access to computers so this statement does not apply to everyone. One stress-alleviating computer test feature mentioned in the research literature but not by respondents in the present study was the advantage of immediately receiving a grade report. As Pino-Silva (2008) noted, the period between test administration and reporting of results can cause students stress. However, "the instantaneous report of results allowed by the computerized test would seem to spare students this distress and thus the high frequency of perceived advantage of this category" (p. 151). None of those interviewed for this research mentioned it probably because only their teacher was provided with access to their score reports.

5.10 Easier Test to Choose Answers

Students were also asked about which test they believed was easier to choose answers for and the largest number (43.3%, N = 52) selected the computer. The largest proportion of those interviewed said they found the computer to be easier to choose answers. In contrast, 35.8% (N = 43) claimed that the paper was easier and 20% (N = 24) said that there was no difference. The common reason for choosing the computer version was the convenience of pointing and clicking. All of those in favor of the computer appeared to agree with Bobby's

comment that it was easier “because it's just one click.” Rodney elaborated saying “the paper, the letter I will have to highlight it all so that I will make it like easy to see, but computer it's just a button just press on it and it is highlight by itself.” He appears to believe he has to spend valuable test time clearly identifying his chosen answer in the paper test (by coloring in the Scantron bubble darkly) so that those marking the test can be sure about which answer he chose. However, with the computer, he simply has to click the right answer and he can be sure that it will be unambiguously identified and graded by the computer.

Only one said she found the paper easier to choose the answer. Sally explained “I think paper because when I use computer and I click it I think I'm going to miss it. Like miss the page by accident [like you'll miss a question or something is that what you mean?] Yeah.” Her concern seems to be that she may click on the answer to the wrong question and not be able to realize that she has made a mistake. She also mentioned that the screen made her dizzy. It may be that she was concerned that her disorientation could cause her to miss questions and lose points.

5.11 Easier Test to Change Answers

A high percentage (67.5%, N = 81) felt that the computer mode allowed them to change their answers more easily while only 18.3% (N = 22) reported that the paper test did and 12.5% (N = 15) said there was no difference. Interviewees unanimously stated that the computer was easier to change the answers for several different reasons. They remarked that it was easier to change answers on the computer test because wrong answers did not have to be erased; the new choice simply had to be clicked. As Donna explained, “…you can circle it [paper] and it's hard to change. Computer, if you choose “A” and think about “B” and you

can just click “B” you don't have to erase anything.” Sally noted a related advantage of the computer saying “you just click and then you change it and if you use paper you need to erase it and then maybe you forgot what you circled again…when I erase it maybe I say ‘I'm going to read it again’ and then I thought I circled it so I just go back.” She appears to be saying that it is more difficult for LOMERA test takers to come back to a question that they are unsure about later because they may become confused. That is, they may have trouble remembering whether or not they wanted to change a particular answer because it is more difficult to completely erase the answer they already had. Innis seemed to echo this feeling when he said “I think the computer because you can like uncheck the answer and then go to the next page and then review the answer again unlike the paper you have to erase it.” He also seems to be saying that it is easier to select an answer and they de-select it and come back to it later because it has clearly been left blank.

The findings in the present study regarding respondents’ preference for choosing and changing answers on the computer were corroborated by Al-Amri (2008). He noted that those who selected computers as the superior testing mode cited straightforwardness of recording and changing answers as one of the primary reasons for their choice. Among the respondents to his questionnaire, 57.5% said it was easier to choose answers on the computer compared to 23.4% on paper. When he asked his respondents which test it was easier to change answers, 90.4% said the computer test. He did not elaborate on why his participants thought one mode was easier than the other. It is unclear how much not having to circle or manually erase answers on a paper test sheet factored into their decision.

The research reported in the present study showed that LOMERA test takers tended to prefer the computer test for choosing and changing answers. However, several

interviewees and much of the relevant research literature illuminated potential pitfalls to be addressed to ensure valid and reliable test answer selection. For instance, some participants said it was sometimes difficult for them to remember to come back to a question that they were unsure about. Likewise, some reported having difficulty remembering whether or not they wanted to change a particular answer. Early computer-based tests often exacerbated this problem by having the limitation of not allowing those taking the test to move ahead or back in the test. However, Spray et al. (1989) conducted a relatively early study that allowed participants to skip items and to review and change answers after completing the test. They did not have significant cross-mode differences in test performance. The authors contended that both versions of the test should be as similar as possible in terms of functionality to ensure comparability across modes. Higgins et al. (2005) allowed participants to use a highlight feature and found that participants used it because they planned to return to the highlighted test items. Participants stated that they used the "mark for review" feature because they were uncertain of an answer so they intended to attempt an answer later. These observations reveal that while there is relative convenience of choosing and changing responses on the computer there are still particular pitfalls that test designers need to take into consideration.

5.12 More Enjoyable Test

Respondents also identified the computer test as being more enjoyable mode. The largest proportion (39.2%, N = 47) found the computer version to be more enjoyable to take. Nevertheless, there were still sizable proportions of respondents who stated that they found the paper version more enjoyable (27.5%, N = 33) and that they found no difference (31.7%,

N = 38). Their reasons were varied. Alice said that she thought it was because it was more comfortable. She elaborated saying “I don't have to flip pages, I don't have to write. I just have to choose and everything was really clear...” She seems to be saying that it was enjoyable because it required less effort to navigate the test and choose the answers than with a pencil. Novelty was also identified as an issue. John mentioned “It's new in your years, like test in a computer. It like kind of sounds fun especially if teachers usually give a test in a paper so for me it's a new, like, idea.” The novelty of the computer test makes it more enjoyable. Bobby elaborated saying “it is new to do the test on the computer because we never do the test on the computer before.” He also approves of the novelty of the test. Donna said “computer is kind of new technology and the people play a computer game a videogame on computer on the screen and they can just think about fun and not taking seriously something I don't know everybody, but I'm just feeling by myself.” She also touches on the newness of the technology, but then goes on to mention a theme that has been raised by several other respondents. This is the notion that while the computer-based test may be more enjoyable there is still the risk that it may be too game-like and test takers may not take it seriously. Rodney’s perspective was a bit different than the others. He talked about how he thought “computer is better because it makes me feel like I'm doing something important. It is all my life I was doing paper and now it's something changed so it's probably important for me.” This serves as an interesting counterpoint to Donna’s comments. Donna said she felt that using the computer was like a game while Rodney emphasized that the shift to the computer mode made him feel as though the test was more important.

Respondents provided reasons for why the paper was more enjoyable that were related to other factors mentioned above, namely, its navigability and readability. When she

was asked about why she thought the paper version was more enjoyable, Sally said that it was because it was easier to read and move around. This reply indicates a belief that the paper mode is more enjoyable which may be a result of her competence with reading and navigating the test. The second reason cited related to perceived convenience of the mode. Innis remarked the paper test was more enjoyable to take "because you can answer the paper in whatever place unlike the computer you can go to the library or something just to answer the paper test [so computer you have to go to the library or to the lab. You don't like that?] No [why not?] Just for me, I don't really like seeing a computer for like an hour. It's better doing the paper. You can answer [in] whatever place." This response suggests there is a perceived portability and convenience of the paper test that causes it to be more enjoyable. Innis said he liked not having to go to the lab and look at the computer screen for a long time. For him, being able to take the paper LOMERA without having to move to a computer lab was another positive advantage of the paper mode.

Relevant research literature that explores English-as-an-additional-language test takers' levels of enjoyment while taking computer tests is surprisingly sparse. Al-Amri (2008) asked which mode his respondents thought was more enjoyable. The majority (76%) said the computer was more enjoyable compared to 12% who said the paper test was. Similarly, in follow up interviews conducted by Pino-Silva (2008) students commented that the computerized test was more dynamic and motivating which in the investigator's view "implies not only an acceptance of the computerized test, but also a good disposition to be evaluated through this instrument" (p. 151). None of the reasons provided during the interviews in the present study were mentioned in any of the studies in the research literature. No other studies identified the novelty of their computer assessment or participants' possible

feelings of doing something important as being possible causes for their enjoyment of the computer test. Coniam (1999) observed that in his study participants had more favourable views of computer-based tests when they were relatively undemanding (e.g., multiple-choice). However, tests with somewhat challenging tasks such as gap filling were viewed much less positively. In those instances, students generally preferred the paper-based version. It may be the computer tests are more enjoyable as long as they are perceived as being easier. The results reported here seem to contradict Coniam's finding in that the computer mode was found to be more enjoyable even though the test format was what he identified as "challenging" (i.e. gap fill).

Table 16
Summary of Test Preferences by Mode by Test Feature with No Mode Preference

	Mode Preference		
	Paper	No difference	Computer
More likely to guess the answer	25% (30)	39.2% (47)	33.3% (40)
Likely to receive the same score a second time	33.3% (40)	39.2% (47)	25% (30)

5.13 Likelihood of Guessing the Answer

Respondents stated that there was no difference across modes for two features. The greatest proportion (albeit not by a large margin) (39.2%, N = 47) felt that the testing mode had no impact on whether or not they would guess the answer to a question (see Table 16). They felt no more or less likely to guess the answer to an item in either mode. However, many also said they were more likely to guess their answer in the paper version (25%, N =

30) and the computer version (33.3%, N = 40). These findings are similar to Al-Amri's (2008) that 52.7% of his respondents said that there was no difference across modes in terms of guessing the answer. No research studies were located that addressed the topic of participants' tendency to guess more in one mode or the other. Nevertheless, the interviews conducted for the present study provided several interesting comments. Several interviewees stated that they felt there was no difference between the two testing modes, but none elaborated on why they felt this to be a case beyond Bobby's comment "I think both because it doesn't really matter between the computer and the paper."

Those who claimed they were more likely to guess the answers on the paper test provided one main reason for their choice. Jennie mentioned that she was more willing to guess answers on the paper test because "it makes me comfortable. I can think more [you said 'I can think?'] Yeah, I can think more [what do you mean by that?] It means that, uhm, I don't know, it just gives me this kind of feeling that it suits me with paper." This may indicate that because Jennie felt more comfortable with the paper format of the test she could think more clearly and thus she may have been more willing to take the risk of guessing the answers. Sally related that she was more likely to guess answers in the paper version "because it's easier to read." This too appears to signal a level of comfort with the presentation of the paper mode that allowed her to feel more at ease about guessing.

One interviewee thought it was easier to guess answers in the computer test. John believed the computer was easier because "computer like it's easier if you're running out of time you just need to click all, better than circle it." If the convenience of clicking rather than circling answers causes test takers to be more careless about their selections that may have an impact on the validity of the computerized test.

5.14 Perceived Test Reliability

The largest percentage of survey respondents (39.2%, N = 47) also reported that they thought they would receive the same score if they took the test a second time in either mode. This finding compares with 33.3%, (N = 40) who stated that the paper test was more reliable and 25%, (N = 30) who said the computer test was. Al-Amri (2008) asked which test examinees thought they would be more likely to receive the same score if they took it a second time and the largest proportion (41.9%) said there would be no difference. This is in accordance with the findings of the present study. Pino-Silva (2008) also noted that several participants in his research commented that an advantage of computer tests is that they are more accurate and reliable. He contended that computers were more reliable because the element of human error was removed from the recording, correction and reporting process. He remarked "For designers of computerized tests, this finding is encouraging, since it shows that subjects trust the computer in the scores it reports" (p 152). Table 17 below summarizes interviewees' reasons for why they preferred each mode of the LOMERA on a particular test feature.

Table 17

Summary of Interviewee Reasons for Preference of Each Mode

	Interviewee Stated Reasons for Preference
Paper	
Navigability	• Distraction caused by the clock
	• Being able to physically turn the pages of the test
Passage readability	• Eye strain caused by the computer monitor
Answers readability	• Smaller font size of the computer test
	• Having to scroll on some computers due to browser settings
Less tiring test	• Computer screen caused "dizziness," eye strain and caused fatigue
More comfortable test	• Familiar with the mode
More accurate	• Difficulty reading the text on the computer screen
	• Possible distraction from the computer clock
	• Reservations about the "seriousness" of the computer test
No Difference	
Likely to receive same score	• No reasons given
More likely to guess	• No reasons given
Computer	
Less stressful test	• Less confusing layout
	• Game-like experience of using a computer
Easier to choose answers	• No need to erase their initial answers
	• No need to worry about returning to questions later
Easier to change answers	• Less effort to navigate the test and choose the answers
More enjoyable test	• Its novelty
	• Its game-like character
	• Felt like doing something important

5.15 Summary

This chapter has provided a summary of students' general mode preferences based on various important features of the test. The purpose of this analysis was to answer the research question of which test features are favored and what reasons students provide for preferring certain test features in a particular mode. Based on these data, secondary school ESL learners favored the paper LOMERA largely in terms of its navigability and readability, ease of use and measurement accuracy of the paper test. They also reported that the paper test made them less tired and more comfortable. They remarked that they believed the paper test more accurately measured their reading ability. Results showed that respondents preferred several features of the computer test as well. The features that were favored were ability to choose and change answers. They also found the computer test to be more enjoyable to take. They expressed a strong preference for the computer mode on several crucial aspects such as reduced stressfulness, greater enjoyment and ease of selecting and changing answers. The two test features that most respondents felt no major differences were their likelihood of guessing an answer and perceived score reliability within the same mode. On each point of comparison between modes, LOMERA test takers typically favored one mode or the other. In fact, there were only two aspects (i.e. reliability and likelihood of guessing the answer) where respondents indicated that there were no differences across modes.

Reasons why the paper test was thought to be more navigable were: the distraction caused by the clock, being able to physically turn the pages of the test, and eye strain caused by the computer monitor. The paper test was preferred for passage and answer readability because of perceived smaller font size and not having to worry about scrolling. Several interviewees found the paper test less tiring because it presented no problems with eye strain

and associated fatigue. The paper test was considered more comfortable because some interviewees were familiar with it. The paper test was thought to be more accurate because of trouble reading from the computer screen, the distraction of the computer clock, and the face validity of the computer test.

The computer mode was favored for several reasons. The computer mode was said to be less stressful because some thought it had a less confusing layout. Interviewees stated it was less stressful because it was in many respects like a game. LOMERA test takers also said that the mouse made choosing and changing answers easier on the computer test simply by pointing and clicking. Finally, the computer test was reported to be more enjoyable because it was more comfortable, it was new and different, it was like a game, and one interviewee mentioned the computer test felt important. Interestingly, however, though students generally preferred one mode over the other for each of the test features, there were no mode effects displayed in the actual tests. Some of these issues will be discussed more fully in chapter 6.

Chapter 6 will outline several conclusions that were reached based on analyses of the data collected for this project. Conclusions based on the results for the cross-mode comparability aspect of the study are presented. Issues around the relationship between LOMERA test taker computer familiarity and computer test performance are also addressed. A discussion is provided of the conclusions arrived at regarding LOMERA test takers' perceptions of particular features of the LOMERA based on results from survey data as it is compared with answers to the follow-up interview questions. Possible strengths and limitations of the study are also enumerated. Research contributions as well as potential

contributions to assessment practice are discussed. Lastly, prospective directions for future research are mentioned.

Chapter 6 Conclusion

6.1 Introduction

School districts in the Lower Mainland of British Columbia have collaborated in recent years to explore ways to share assessment information. Addressing this issue becomes increasingly urgent as the proportion of immigrant and ESL learners rises. The formation of the ESL Assessment Consortium has been a positive development in the move toward closer cooperation among districts. One particularly valuable outcome of the consortium has been the development of the Lower Mainland English Reading Assessment (LOMERA). The successful integration of this test into district assessment schemes motivated consortium members to develop a computer version of the LOMERA.

6.2 Background

The eclectic assortment of reading measures used by school district personnel impeded information sharing when students moved between districts. To overcome this dilemma, Consortium members developed a standardized secondary ESL reading assessment that provides an estimate of reading proficiency (Gunderson, Murphy Odo, D'Silva, 2010). This measure is a common locally-normed reading assessment that allows districts to use and to share information about learners' reading proficiency.

Consortium members were satisfied with the assessment tool known as the LOMERA and they sought to develop a computer-based version of the test. The question was soon raised about whether scores from the computer-based LOMERA would be comparable to those obtained from the traditional paper-based version. A review of research literature yielded no answer so further research was deemed necessary.

6.2.1 Review of Research

The research for comparability of the paper- and computer-based reading assessments for first language speakers is somewhat divided. On one hand, there are those who contend that the two forms are comparable (Higgins et al., 2005; Wang et al., 2008). However, others claim they are not (Clarania & Wallace, 2002; Kim & Hyunh, 2008; Pommerich, 2004). Somewhat similar results are reported in the second language comparability literature. A number of scholars insist that assessments are comparable across modes (Al-Amri, 2008; Green & Maycock, 2004; Sawaki, 2001) while others disagree (Coniam, 2006; Fulcher, 1999). Additionally, several sources have mentioned that reading comprehension may be the skill that shows the greatest discrepancy in cross-mode performance (Choi et al., 2003; Kim & Hyunh, 2008).

Other research has explored the connection between computer familiarity and performance on tests. Early research found a relationship between computer familiarity and computer test scores (Kirsch et al., 1998), but recent findings suggest there is no significant connection (Al-Amri, 2008; Taylor et al., 1999; Sawaki, 2001). Examinees' perceptions of computerized assessments revealed that respondents preferred to take a reading test on computer (Higgins et al., 2005) though it depended on the type of test being taken (Coniam, 1999). Some advantages of computer-based tests were that they were motivating while producing scores that were considered to be accurate and reliable (Pino-Siva, 2008). Misgivings regarding computer-based tests included system failure, eye fatigue and anxiety about the inability to properly operate the test software (Pino-Siva, 2008).

6.2.2 Study Objectives

The three main objectives of this study were: (1) determine whether paper and computer-based versions of a standardized m-c cloze reading test for second language learners are comparable (2) identify whether learners' computer familiarity predicted their performance on the computerized LOMERA (3) learn about participants' perceptions of features of the paper and computer-based LOMERA.

6.2.3 Procedures

Upon receiving IRB approval, I recruited participants from a secondary school in a major western Canadian city with a high proportion of ESL learners. The research instruments included both paper- and computer-based versions of the LOMERA. The two tests were the same in terms of content, questions, pagination and layout. They differed only with respect to method of recording answers (i.e., pencil vs. mouse) and the fact that test takers had limited ability to make notes or highlight particular questions on the computer as they could with a paper-based test. In addition to these tests, there were also questionnaires that asked participants closed, self-report questions about their computer familiarity and perceptions of computer-based tests.

The study design was counterbalanced to avoid order effects so that two groups of learners took the tests in the opposite order and their scores were compared. The tests were administered to two different randomly-assigned groups. To minimize practice effect, group one took the paper-based test and four weeks later they took the computer-based test. Group two did the opposite. Just prior to the first administration of the first test, participants were asked to complete a questionnaire that asked about their computer familiarity. Upon finishing

the second test administration, they completed a second questionnaire that asked about the perceptions of each of the two testing modes and ten volunteers were interviewed to obtain reasons why the preferred certain test features in a particular mode.

6.3 Research Findings and Conclusions for Comparability

The present study was designed to address three research questions. Each of the following sections provides conclusions for each of the questions based on the findings of the study.

Research question 1: Is there a mode effect in ESL students' performance on paper-based and computer-based versions of a multiple-choice cloze reading assessment?

The objective with this question was to determine whether the paper and computer versions of the LOMERA were comparable based on a variety of criteria established in previous research as being indicators of cross-mode comparability. Evidence from a paired-sample t-test, correlation analyses and delta-plot (DIF) analyses was assembled to answer this question. The paired-sample t-test comparing scores from the first and second administration of the LOMERA was used to test the null hypothesis that there would be no statistically-significant difference between modes in test scores for either test administration. Results showed that there was no statistically significant difference in the test scores produced by two randomly-assigned groups of students who were each taking the LOMERA in a different mode. The results of the paired-sample t-test confirmed that test takers' scores on the paper version of the test were comparable with their score on the computer version.

These findings corresponded with much of the research literature into cross-mode comparability (Choi et al., 2003; Maycock & Green, 2004; Sawaki, 2001; Yessis, 2000).

Other cross-mode comparability research that has employed t-test analysis has tended to find mode effect (Al-Amri, 2008; Coniam, 2006; Fulcher, 1999). However, there are some issues with the previous use of t-tests for this research. For instance, Al-Amri (2008) provides two compelling reasons to doubt his own findings. He acknowledges that significant differences in his participants' scores were due to the low number of test items on each of his tests and small differences in a large sample size such as the one in his study can result in inaccurate significant results. Fulcher (1999) also acknowledged that "the increase in mean score on the CBT is due in large part to an order effect..." (p. 294) (he did not counterbalance the mode of administration to account for order effect) but he defends his finding of mode effect by insisting that "this in itself is not enough to account for the possible variation in scores as indicated by the standard deviation of the difference of means" (p. 294). Nevertheless, this "possible variation in scores" allows for some skepticism about the conclusiveness of his findings. In light of these acknowledged limitations, it is not unreasonable to seek further confirmation of the results reported above or to accept that the significance tests used in this study could reveal absence of mode effect.

The t-tests were followed with a series of cross-mode inter-passage correlations that were used to assemble further evidence for the comparability of the LOMERA tests. The inter-passage correlations ranging from .6 to .85 were satisfactory. The cross mode correlation of .96 for the entire test was quite remarkable. These correlations were higher but generally in accordance with those reported in other research though some were more

impressive (r = .82) (Fulcher, 1999) (r = .88) (Choi et al., 2003) than others (r = .74)(Al-Amri, 2008).

A Delta-plot differential item functioning (DIF) analysis was the final piece of research that explored comparability of the LOMERA tests at the item level. Several useful insights were gained. First, this study demonstrates that the Delta-plot DIF analysis method is a useful tool for identifying particular items that are causing mode effect in dual-mode tests. The benefit of using this tool in addition to traditional methods of comparing tests across modes is that it can provide information about which specific test items are causing the mode effect. Although at least some previous studies have attempted to apply DIF methods to cross mode comparability questions, the present study appears to be the first time that the Delta-plot method has been used with second-language test takers. A second observation based on this research is that there appear to be a few particular test items that demonstrate greater mode effect than others. These potentially problematic items may have to be modified or replaced. However, surprisingly, there is currently no guidance regarding the proportion of DIF test items above which it would be advisable to consider a test incomparable. This apparent oversight may deserve further consideration.

The results of this DIF analysis were that only four test items out of 96 showed cross-mode discrepancies in item-facility p-values. Potential causes of the divergence in these four items might be examinees' possible increased enjoyment of the computer test strengthening their patience to complete sentences with multiple blanks which test takers might consider too bothersome in the paper test. Alternatively, it could simply be measurement error or random chance. Locating relevant research to inform this analysis was somewhat challenging

primarily because there has not been a great deal of research that has used this technique to evaluate cross-mode comparability. In one study, Schwarz et al. (2003) analyzed their adult basic education students' scores and found that approximately 40% of test items demonstrated mode effect compared to approximately four percent in the present study. This difference in the results of the present study may relate to the present sample being ESL secondary school learners while Schwarz's et al. was with adult basic education students who might have been less familiar with and more anxious about using computers. Keng, McClarty, and Davis' (2008) study of the Texas statewide standardized achievement test did not discuss the proportion of items that were differentially functioning but they did tentatively speculate that some reading test items might be vulnerable to mode effect. The present study did not confirm this finding.

The results of the comparability component of this study have several key implications for the research literature. First, this study adds to the relatively scarce cross-mode assessment comparability literature with second language learners. Second, it expands on the types of assessments being investigated by evaluating a multiple-choice cloze (maze) test. It explored cross-mode comparability with forms of assessment that go beyond traditional multiple-choice discrete-point item types that most comparability research has tended to focus on. Indeed, this appears to be the first cross-mode comparability study in the literature that explores the phenomenon with an integrative type of assessment. Third, this research incorporates a method of cross-mode analysis that has not been used with second-language learners on these types of assessments. The delta-plot DIF analysis technique goes beyond many traditional comparability research methods to enable the cross-mode

comparison of both versions of the LOMERA at the level of individual test items. The combination of these analysis techniques allows for evaluation of cross-mode comparability at both the test and item level. This combined "top-down" and "bottom-up" approach should provide a more nuanced and complete description of how both versions of the LOMERA relate to each other.

Results also have relevance for schema theory. The existence of mode effect could be an indication that the test presentation mode is causing test takers to draw on different schemata as they do the test (i.e. some type of "mode schema"). For instance, it might be that minor layout alterations or prior associations with the mode itself cause examinees to activate schema structures that are unique to a particular mode which may consequently impact their test performance. However, the minimal mode effect evident in this study indicates that, with respect to the maze-type test tasks used here, both tests seem to be activating similar schema structures. Therefore, at least for this particular form of cross-mode assessment, mode schema may exist but they do not have a powerful influence on cross-mode test performance. Thus, either the difference in schematic structures being activated is negligible or mode schemata do not exist. Only future research can determine which is true.

The evidence for cross-mode comparability presented here will give assessment consortium members confidence to use the online version of the LOMERA to substitute for the paper test. Replacing the paper LOMERA with the online version will allow members to administer the test without having to actually take the paper test into the schools and worry about potential breaches of security if a test form were to go missing. Furthermore, administering the test via computer will also save their respective school boards valuable resources that would otherwise be spent on performing necessary clerical duties associated

with administering the paper test such as organizing, scoring, and record keeping. All of this is accomplished automatically with the computer test.

Some might ask whether it would be more prudent to replace the paper with the computer version entirely and re-norm the test on the computer so as to avoid having to conduct comparability research altogether. Additionally, migrating the tests exclusively online would exploit their many time and labour saving affordances while reducing the threat of human error in scoring. Besides, many would argue that schools, like the rest of society, are moving in the direction of increased integration of technology rather than away from it. Moving the test entirely online would appear to be an ideal solution. However, the problem in many school contexts (as was learned through discussions with several consortium members) is that lack of resources prevents purchase of the most up-to-date technology. In many instances, the technology currently in place is outdated or unreliable. Therefore, when these problems inevitably arise, local ESL assessors need to be prepared with a hard-copy of the test that will produce comparable results.

There is also a move to incorporate other types of handheld wireless devices into many educational contexts. The advantage of these devices is that many students are familiar with them and there is much less of an expensive infrastructure required for their use. Having comparable online assessments in place will allow teachers much more flexibility to integrate these powerful technological tools into their classroom for an even wider variety of purposes. In this way, tools such as IPads can become portable assessment devices in addition to being e-readers, word processors and educational activity platforms.

6.4 Research Findings and Conclusions for Computer Familiarity

Research question 2: Do L2 learners who are more familiar with computers achieve higher scores on a computer-based multiple-choice cloze reading assessment than those who are less familiar with computers?

Findings are consistent with other research literature that computer familiarity does not have an inordinate impact on a computer-based language test performance. Although there is a statistically significant relationship between indicators of computer familiarity and online LOMERA performance, these variables actually do not explain a great deal of the variance in computer-based LOMERA test performance. A corollary of this conclusion is that computer familiarity may not be an important consideration in deciding whether to adopt the online LOMERA in local districts. Nevertheless, future studies might explore whether there may be a threshold of familiarity after which it becomes less of an issue. It could be that most of those who took the test in the present study were beyond that threshold.

The findings for the computer familiarity component of the study have several implications for assessment practices with the LOMERA in the lower mainland. As mentioned above, local test administrators do not have to be unduly concerned about the computer familiarity of those taking the test. In fact, one anecdote serves as a reminder that learners can often surpass our expectations and teachers may sometimes underestimate the computer familiarity of their ESL students. At one point during the test administration the test was given to a group of level-one students. The teacher was initially doubtful about whether several students in the group would be able to take the computer test because they were later-to-literacy learners who had not had a great deal of experience with computers. The teacher was surprised during the administration of the test when all of the students were

able to navigate the test with much less difficulty than had been anticipated. This example illustrates that it may be worthwhile to allow the student to attempt the test before deciding that he or she would probably be better served by doing it on paper. Nevertheless, if teachers suspect that a test taker has had insufficient previous exposure to computers (e.g. using the test at a reception centre) they might consider developing some type of brief pre-screening instrument or protocol to ensure that they have the requisite computer skills to guarantee a valid and comparable test administration.

6.5 Research Findings and Conclusions for Mode Perceptions

Research Question 3: Which test features are favored and what reasons do test takers provide for preferring certain test features in a particular mode?

Analysis of descriptive statistics for test takers' responses to the follow-up test perceptions survey revealed that the largest proportion of test takers preferred the paper LOMERA across all of the comparability criteria. The specific paper test features that test takers expressed preference for were passage navigability, passage and answer readability, being less tiring, being more comfortable and more accurately measuring reading comprehension skills. Test takers who were interviewed after both test administrations provided several intriguing reasons for their preference.

Three justifications provided for favoring the paper test were: the distraction caused by the clock, being able to physically turn the pages of the test, and eye strain caused by the computer monitor. Reasons articulated for the preference of the paper test for passage and answer readability had to do with perceived smaller font size of the computer test (by one interviewee) and having to scroll on some computers due to browser settings. A number of

interviewees found the paper test to be less tiring because the computer screen made them experience "dizziness" and eye strain which caused fatigue. The paper was identified as being more comfortable because some interviewees were familiar with the mode. Additionally, the paper version was viewed as the most accurate form of testing mode by at least a few interviewees. The paper test was thought to be more accurate because of the difficulty reading the text on the computer screen, possible distraction from the computer clock, and reservations about the "seriousness" of the computer test.

Mode preference in favour of the computer mode for several criteria was identified as well. Some interviewees said the computer mode was less stressful because they thought it had a less confusing layout. That is, some reported that there was a closer consistency between the font in the passages and that in the questions. Interviewees also reported the game-like experience of using a computer made it less stressful. They said that the computer was an easier test to choose answers and change answers because they did not need to erase their initial answers as well. One interviewee mentioned that the computer test allowed her to not have to worry about returning to questions later because it was clear whether or not a final answer had been chosen for each question. The computer test was reported to be more enjoyable than the paper version because it was more comfortable (i.e. required less effort to navigate the test and choose the answers), its novelty and its game-like character. One interviewee also remarked that the LOMERA on the computer made him feel like he was doing something important.

Several research and practice implications arise from the conclusions regarding mode perceptions presented above. First, approximately two-thirds of test takers said they thought one mode of the test was more accurate than the other and on most of the test criteria test

takers tended to prefer one mode over the other. However, there was no significant difference in actual test scores. This indicates that there may be some kind of discrepancy in perceived performance on a particular (possibly less favored) mode and actual test performance. This appears to be a type of "mode perception performance paradox" or M3P. This M3P risks causing test takers unnecessary test anxiety and possibly affecting their test performance. The existence and potential effect of M3P should be explored in greater depth in future research.

Results of the present study also suggest that preference for a particular mode depends on distinct test features. The paper test was generally viewed as being more navigable, readable and accurate while the computer test was seen as less stressful and more enjoyable. Researchers should investigate whether preferred test features vary across modes depending on the type of test items being used. Results from this research could then enable test designers to specifically address particularly problematic features when they develop computer versions of a paper-based test to minimize negative perceptions of the test and ensure optimal test performance.

Longstanding and more recent concerns related to the computer mode continue to require further exploration. For instance, eye strain continues to be identified as a common cause of discomfort for some test takers. Ten years ago, researchers were confident that monitor technology would eliminate this problem by now and yet it persists. As well, the entertaining or game-like quality of computer tests mentioned by several interviewees should be explored in greater depth. Interview data indicate that it may be a double-edged sword. On the one hand, the "fun factor" makes the test more enjoyable but questions persist concerning whether some test takers will take the test less seriously as a result. There may also be a

cultural component to this as Chinese interviewees were more likely to express misgivings about the "seriousness" of the computer test.

Results suggest that a certain proportion of members of the so-called "digital generation" or "gen-e" actually prefer not to take tests on computers. Therefore, if the resources are available, test administrators should allow LOMERA test takers a choice to do the test in the mode that they favor. However, test administrators should also recognize that mode preference may not necessarily dramatically affect LOMERA test takers' actual test performance so if administrators are unable to provide a choice they can be reassured about the validity and reliability of the test results.

A final implication for local testing practice coming from the results of the perceptions survey and interview is related to the impact of the technology used during the test on the quality of some test takers' test experience. Those administering the LOMERA need to be aware that the older monitors in schools may cause test takers excessive eye strain that, while probably not seriously affecting their test scores, may nevertheless make them unnecessarily uncomfortable. Accessing up-to-date minimum-glare monitors would help reduce problems with eye strain. Additionally, checking browser settings beforehand to enable optimal display of the test will avoid students having to needlessly scroll between passages and questions.

6.6 LOMERA Administration Issues

Several other noteworthy observations were made throughout the administration of the LOMERA. For instance, there are technological obstacles in schools that must be satisfactorily addressed to ensure smooth integration of the online LOMERA into local ESL

literacy assessment schemes. Some technological challenges that need to be dealt with are the widespread use of out-dated hardware and software in some districts, restricted access to computer labs to administer tests and limited availability of a technology person who can troubleshoot hardware or software problems. All of these issues can be incredibly daunting for a marginally technologically-inclined ESL examiner who is attempting to administer a test under conditions that allow for optimal student performance.

A second issue has to do with test security. In particular, it was noted that there was the danger of test takers looking at each other's screens as they completed the test. It was soon discovered that this problem could be avoided by providing sufficient space for students during the test and ensuring that the computers in the lab were arranged so that all were readily visible by test proctors. Alternatively, cardboard screens set up between computers have proven to be quite effective. Discussions with potential users of the online LOMERA raised another security-related concern which is that teachers or test proctors still need to be present and observant during the test administration. Some teachers seem to believe that the use of technology will allow them to leave their students to do the test without supervision and the online test will somehow monitor test takers' behaviour. Teachers and proctors may need to be reminded that this new technology still requires the time-honoured practice of careful test invigilation.

A third point relates to the fact that this test was born of a need for local school districts to share information about the ESL students in their districts. This goal is important because of the high mobility of local ESL students (Gunderson & Murphy Odo, 2009). Several district representatives noted that due to the wide variety of assessments used in different districts, it was virtually impossible to share meaningful assessment information

among the districts. The LOMERA allows for the sharing of meaningful assessment results across districts for the first time. Another welcome but unanticipated by-product of the development of the LOMERA was that many ESL specialists reported that it was also one of the more objective (i.e. less dependent on individual assessor subjective judgment) assessments they had in their repertoire. Certainly, establishing the online version in local districts stands to provide all of the previously enumerated benefits of its paper counterpart in addition to the convenience, reduced clerical labour and diminished paper consumption that comes with using an online tool.

The ESL consortium also deserves enthusiastic praise for being a group that has dedicated their precious time and expertise to collaborate and share information and resources that will benefit all ESL students. Hopefully, the consortium will have the opportunity to continue to efficiently and effectively produce and refine these valuable assessments for their students. This group is truly a model for inter-district resource sharing for the benefit of all that others would be wise to emulate.

6.7 Possible Strengths and Limitations

There are several unique strengths of this study that add to the second language assessment research literature. First, this research was conducted with secondary ESL students who have been identified as being underrepresented in the research literature (Snow, 2008). Second, this study explores the cross-mode comparability of an integrative test task type which has not received a great deal of previous research attention. Third, the test designers took great care to ensure that both versions of the LOMERA were as similar as possible in layout and functionality across modes. This is a feature of research design that has

often been overlooked in previous studies but it is fundamental to the valid design of any cross-mode comparability study. Fourth, the present study investigates comparability at the item level and test level. The majority of previous studies focussed on comparability at the test level only.

One limitation of this research is the generalizability of the results. The sample used for this study was a convenience sample so it may not be completely representative of ESL learners from other contexts. In addition, this study took place in the complex environment of a school. Although strenuous efforts were made to control for intervening variables such as learning or practice effect (see discussion in chapter 3) one can probably never entirely eliminate contaminating influences in such an environment.

Limitations of the perceptions interview methodology must also be acknowledged. One frequent criticism is that respondents' actual opinions may differ from their answers and this may lead to false conclusions on the part of the researcher. However, this potential threat would appear to be more serious in interviews about sensitive topics such as interviewee participation in deviant behavior. It seems that this would be less of an issue when interviewees are being asked to report on why they preferred one testing mode over another, especially when they have been assured that the test will have no bearing on their course grade.

6.8 Research Contributions

In addition to establishing the cross-mode comparability of the paper and computer-based LOMERA, this research will make several other important contributions to the research literature. One particular area where this project can enhance comparability research

is its investigation of differences in mode effect for test items other than the traditional multiple-choice discrete-point items used in a great deal of comparability research (Yu, 2010). This project attempts to address this gap in existing research by specifically investigating possible mode effects for multiple-choice cloze type integrative test items. Secondly, this study addresses the relationship between assessment mode and background variables such as computer familiarity and perceptions of paper and computer-based testing with another form of assessment. In this instance as well, all of the comparability studies that investigate background variables use tests that are comprised of traditional multiple-choice test items. Therefore, researchers need to further explore whether these background variables are more or less influential with other types of test items.

A second contribution of this research relates to the striking number of studies in the research literature that admitted to having the design flaw that the paper-based (PBT) and computer-based (CBT) versions of their tests often had quite different layout and functionality. This dissimilarity between instruments seriously compromises the integrity of any investigation of mode effect. To overcome these limitations, more research is needed with assessments that are virtually indistinguishable except for the mode of presentation such as those used in the present study. This would help verify that mode effects are not confounded with the other differences in the paper and computer-based tests.

A third research contribution of this study addresses the fact that there is not a lot of research that has used the Delta-plot DIF method to probe mode effect at the level of individual test items. This study incorporates the Delta-plot differential item functioning (DIF) procedure to focus on individual test items that may exhibit differing levels of mode effect across modes. This analytical technique has not been used for this purpose in most

previous comparability research and thus it appears to be a useful tool that has been overlooked.

Lastly, the findings of this study contribute at least two potentially useful insights to the research literature. First, this research identified the potential positive and negative impact that the "fun-factor" of the computer test may have on the way that students perceive and interact with computerized tests in comparison to their paper-based counterparts. Although there is a substantial amount of overlap in test takers' comments about why they preferred certain features on a given mode with comments that were reported in previous studies, the enjoyment of the game-like quality of the test and its potential threat to their perception of the face validity of the test was a theme discovered in this research that was not mentioned in previous studies. Second, the notion of M3P identifies the apparent discrepancy between many test takers' stated feelings about taking a test in a particular mode and their actual test performance. It may be informative to explore this concept further.

6.9 Contribution to Practice

A crucial reason for conducting this research relates to the large-scale migration of many language assessments online frequently for financial rather than pedagogical reasons. As a consequence, test developers often do not fully consider the possible negative effects of potential mode effect differences in paper and computer based tests (Chapelle & Douglas, 2006). To remedy this situation, disinterested researchers need to engage in independent investigation of cross-mode comparability to ensure that test developers' claims about cross-mode comparability are supported by convincing evidence. This investigation exemplifies

the type of independent research that can lend greater credibility to the encouraging cross-mode findings previously reported by test developers.

The results of this study will also impact ESL assessment practices in the lower mainland because its support for the comparability of paper and computer versions of the LOMERA will justify examiners' use of the computer-based version for local assessment. The ability to use the computer test has the potential to considerably streamline current LOMERA administration procedures by automatizing clerical tasks such as scoring and organizing test data. The online test also eliminates the danger of scorer error and thus increases test fairness.

6.10 Future Research

This study revealed several possible avenues for future research. One potentially interesting investigation could explore the cross-mode comparability of other types of test items (elide, matching cloze, summary/paraphrase, short answer etc.) to see if there are greater mode effects than with traditional multiple choice and maze test items. Additionally, the cross-mode comparability of cloze assessments used for other purposes such as listening could be investigated to determine if there are differences depending on the language skill being assessed. Future projects could also explore the relationship between test performance across modes and other types of background variables such as gender and language proficiency. It might also be informative to compare younger and older learners' cross mode performance to determine if there are discrepancies between "digital natives" (i.e., younger learners) and "digital immigrants" (older learners) (Prensky, 2001).

Other DIF identification tools might also be useful for identifying items exhibiting mode effect. Two methods used by Schwarz et al., (2003) (i.e., Linn-Harnish and nonparametric Standard Mean Difference) may also be useful with second-language learners to support the results of the Delta plots. One caveat is that other DIF methods must be chosen carefully because they often assume that the two test taking groups will be different whereas this is not the case with cross-mode investigations. Methods that make this assumption are not useful for cross mode investigations. Additionally, other statistical (e.g., multiple regression) and qualitative tools (e.g., think-aloud protocols) may be helpful for identifying sources of DIF across modes.

Another possible area of future exploration would be to investigate the effects of test takers' inability to interact with the online test by writing on it as they can with the paper version. One speculation is that the ability to interact with or "mark up" the computer screen may not be as limited in the computer mode as was originally presumed. Future research might also study how test takers improvise alternative means of making the online test taking experience more interactive. Designers could then attempt to incorporate these adaptations into future test iterations.

The concept of mode schema may also be worthy of future exploration. The findings presented here did not identify any significant mode effect across both versions of the LOMERA so a strong case for the existence of mode schema cannot be made. However, it may prove to be a useful theoretical concept in future studies of cross-mode comparability with other types of assessments. Researchers might attempt to identify whether learners' understandings about reading on computers improve or hinder their computer test performance. Questions might also be asked about whether there are specific aspects of the

computer or paper test taking experience that might have an inordinate influence on mode schema and mode effect. Features such as the ability to add recorded test instructions that are unique affordances of the computer mode could also be systematically manipulated to determine whether that has a relationship with mode effect. Alternatively, an experiment could be devised such that both modes of the test would be as similar as possible except for the ability of test takers to control the size of the font on the computer version. Researchers could investigate whether this unique computer-related affordance causes mode effect when both forms of the test are taken.

The association between mode preference and actual test performance (i.e., M3P) should also be explored in greater depth in forthcoming studies. For example, it may be worthwhile to investigate the perception that the computer test is more enjoyable and game-like. Additionally, future studies may assess whether there is any need to be concerned about potential risk that some test takers from non-western cultures will take the test less seriously because of its higher entertainment value and if that in turn greatly affects their test performance.

These promising comparability results reported here encourage further research on how to move other types of non-traditional forms of assessment online. This kind of research is necessary particularly in K-12 ESL contexts that have traditionally been underexplored (Snow, 2008). Even large-scale test publishers (e.g. University of Cambridge ESOL Examinations) are moving away from traditional discrete-point forms of testing toward more integrative and authentic (i.e. performance-based) assessments. This kind of research is clearly a step in the right direction.

References

Abraham, R. G. & Chapelle, C. A. (1992). The meaning of cloze test scores: An item difficulty perspective. *The Modern Language Journal,* 76, 468-479.

Al-Amri, S. (2008). Computer-based testing vs. paper-based testing: A comprehensive approach to examining the comparability of testing modes. *Essex Graduate Student Papers in Language & Linguistics,* 10, 22-44.

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

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

American Psychological Association Committee on Professional Standards and Committee on Psychological Tests and Assessment. (1986). *Guidelines for computer-based tests and interpretations*. Washington, DC: Author.

Asselin, M. Early, M. & Filipenko, M. (2005). Accountability, assessment, and the literacies of information and communication technologies, *Canadian Journal of Education,* 28, 802-826.

Bachman, L. F. (1982). The trait structure of cloze test scores. *TESOL Quarterly,* 16, 61–70.

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

Bailey, K. M. (1998). *Learning about language assessment: Dilemmas, decisions, and directions*. Toronto: Heinle & Heinle.

Bailey, S. M. (2008). *Content assessment in intelligent computer-aided language learning: Meaning error diagnosis for English as a second language*. Unpublished doctoral dissertation, Ohio State University, Columbus, OH. Retrieved June 10, 2010, from http://www.ling.ohio-state.edu/~s.bailey/papers/bailey_thesis.pdf

Bensoussan, M. & Ramraz, R. (1984). Testing ESL reading comprehension using a multiple-choice rational cloze. *The Modern Language Journal,* 68, 230-239.

Blanchard, J. S., Mason, G. E., & Daniel, D. (1989). *Computer applications in reading* (3rd ed.). Newark: International Reading Association.

Brown, H. D. & Abeywickrama, P. (2010). *Language assessment: Principles and classroom practices* (2nd ed.). New York: Pearson Longman.

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment*. New York: McGraw-Hill.

Brown, J. D. (2002). Do cloze tests work? Or, is it just an illusion? *Second Language Studies,* 21(1), 79-125.

Cameron, C. A. Hinton, M. J. & Hunt, A. K. (1987, June). *Automated cloze procedures as research and teaching tools*. Paper presented at the Annual Meeting of the Canadian Psychological Association. Retrieved June 11, 2010, from ERIC Academic Database.

Canadian Counsel on Learning (2008). *Lessons in learning: Understanding the academic trajectories of ESL students.* Accessed March 24, 2012. http://www.ccl-cca.ca/pdfs/LessonsInLearning/Oct-02-08-Understanding-the-acedemic.pdf

Carrell, P. L. & Eisterhold, J. C. (1988). Schema theory and ESL reading pedagogy. In P. L. Carrell, J. Devine, & D. E. Eskey (Eds.), Interactive approaches to second language reading (pp. 73-92). Cambridge: Cambridge University Press.

Chapelle, C. A. (2008). Utilizing technology in language assessment. In E. Shohamy & N. H. Hornberger (Eds.). *Encyclopaedia of language and education*, 2nd Edition, Volume 7: language testing and assessment.

Chapelle, C. A. & Douglas, D. (2006). *Assessing language to computer technology*. Cambridge: Cambridge University press.

Chavez-Oller, M.A., Chihara, T., Weaver, K.A. & Oller, J.W. (1994). When are cloze items sensitive to constraints across sentences? In Oller, J.W. Jr. & Jonz, J., (Eds.), *Cloze and coherence* (pp. 229–245). London: Associated University Press.

Chen, Z. & Henning, G. (1989). Linguistic and cultural bias in language proficiency tests. *Language Testing,* 2, 155-163.

Chihara, T., Oller, J. W., Weaver, K. A.., & Chavez-Oller, M. A. (1992). Are cloze items sensitive to constraints across sentences? *Language Learning, 27*, 63-73. In J. W. Oller, & J. Jonz (Eds.), *Cloze and coherence* (pp. 135-147). London: Associated University Press.

Choi, I. C., Kim, K. S., & Boo, J. (2003). Comparability of a paper-based language test and a computer-based language test. *Language Testing,* 20, 295–320.

Clariana, R. & Wallace, P. (2002). Paper-based versus computer-based assessment: Key factors associated with the test mode effect. *British Journal of Educational Technology,* 33, 593-602.

Codó, E. (2009). Interviews and questionnaires. In L. Wei, & M. G. Moyer (Eds.), *The Blackwell guide to research methods in bilingualism and multilingualism* (pp.158-176). Malden, MA: Blackwell.

Cohen, A. (1994). *Assessing language ability in the classroom* (2nd ed.). Boston: Heinle & Heinle.

Coniam, D. (1999) Subjects' reactions to computer-based tests. *Journal of Educational Technology Systems,* 23, 195–206.

Coniam, D. (2006). Evaluating computer-based and paper-based versions of an English-language listening test. *ReCALL,* 18, 193-211.

Dooey, P. (2008). Language testing and technology: Problems of transition to a new era, *ReCALL*, 20, 21-34.

Dörnyei, Z. (2007). Research methods in applied linguistics. Oxford: Oxford University Press.

Douglas, D., & Hegelheimer, V. (2007). Assessing language using computer technology. *Annual Review of Applied Linguistics,* 27, 115–132.

Espin, C. A. & Froegen, A. (1996). Validity of general outcome measures for predicting secondary students' performance on content-area tasks. *Exceptional Children,* 62, 497-514.

Evans, L. D., Tannehill, R. & Martin, S. (1995). Children's reading skills: A comparison of traditional and computerized assessment. *Behavior Research Methods, Instruments, & Computers,* 27, 162-165.

Fotos, S. (1991). The cloze test as an integrative measure of EFL proficiency: A substitute for essays on college entrance examinations? *Language Learning,* 41, 313-336.

Fraenkel, J. R. & Wallen, N. E. (2005). *How to design and evaluate research in education*. New York: Mcgraw-Hill.

Fulcher, G. (1999). Computerizing an English language placement test. *ELT Journal,* 53(4), 289-299.

Gamst, G., Meyers, L. S., & Guarino, A. J. (2008). *Analysis of variance designs: A conceptual and computational approach with SPSS and SAS*. New York: Cambridge University Press.

Gorsuch, G. (2004). Test takers' experiences with computer- administered listening comprehension tests: Interviewing for qualitative explorations of test validity. *CALICO Journal,* 21, 339-371.

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. New York: Cambridge University Press.

Green, T. & Maycock, L. (2004). Computer-based IELTS and paper-based versions of IELTS. *Research Notes,* 18, 3-6.

Gribbons, B. & Herman, J. (1997). *True and quasi-experimental designs*. *Practical Assessment, Research & Evaluation*, 5(14). Retrieved August 9, 2010 from http://PAREonline.net/getvn.asp?v=5&n=14.

Gunderson, L. (2007). *English-only instruction and immigrant students in secondary schools: A critical examination*. Mahwah, NJ: Lawrence Erlbaum Associates.

Gunderson, L. (2009). *ESL (ELL) literacy instruction: A guidebook to theory and practice* (2nd ed.). New York: Routledge.

Gunderson, L. (2007). *English-only instruction and immigrant students in secondary schools: A critical examination.* Mahwah, NJ: Lawrence Erlbaum Associates.

Gunderson, L., D'Silva, R. & Murphy Odo, D. (2010). *The Lower Mainland English Reading Assessment (LOMERA) Manual*. Vancouver: The Lower Mainland ESL Assessment Consortium. Retrieved August 9, 2010 from http://www.eslassess.ca/esl/

Gunderson, L., D'Silva, R. & Murphy Odo, D. (2010). Assessing English language learners. In P. Afflerbach (Ed.), *Handbook of Language Arts*. New York: Routledge.

Gunderson, L. & Murphy Odo, D. (2009). Predicting Young Immigrant Students' Academic Achievement. Paper presented at the annual meeting of the National Reading Conference, Albuquerque, New Mexico, Dec 5, 2009.

Guthrie, J. T., Seifert, M., Burnham, N. A. & Caplan, R I. (1974). The maze technique to assess, Monitor reading comprehension. *The Reading Teacher,* 28, 161-168.

Hale, G. A., Stansfield, C. W., Rock, D. A., Hicks, M. M., Butler, F. A. & Oller, J. W. (1989). The relation of multiple-choice cloze items to the Test of English as a Foreign Language. *Language Testing,* 6, 47-75.

Hanania, E. & Shikhani, M. (1986). Interrelationships among three tests of language proficiency: Standardized ESL, cloze, and writing. *TESOL Quarterly,* 20, 97-109.

Hartas, D. (2010). The epistemological context of quantitative and qualitative research. In D. Hartas (Eds.), Educational Research and Inquiry: Qualitative and Quantitative Approaches. New York: Continuum.

Higgins, J., Russell, M., & Hoffmann, T. (2005). Examining the effect of computer-based passage presentation on reading test performance. *Journal of Technology, Learning, and Assessment, 3*. Retrieved June 9, 2010, from http://www.jtla.org

Hurley, S. R. & Tinajero, J. V. (Eds.). *Literacy assessment of second language learners* (pp. 64-83). Needham Heights, MS: Allyn & Bacon.

International Test Commission. (2006). International guidelines on computer-based and Internet-delivered testing. *International Journal of Testing,* 6, 143–171.

Johnson, R.B. & Onwuegbuzie, A.J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher, 33*(7), 14-26.

Kalantzis, M., Cope, B. & Harvey, A. (2010). Assessing multiliteracies and the new basics. *Assessment in Education: Principles, Policy & Practice,* 10, 15-26.

Kauffman, A. S. (2003). Practice effects. *Speech and Language Forum*. http://www.speechandlanguage.com/cafe/13.asp

Keng, L., Larsen McClarty, K. & Davis, L. L. (2008). Item-level comparative analysis of online and paper administrations of the Texas Assessment of Knowledge and Skills. Applied *Measurement in Education,* 21, 207–226.

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

Kim, D. H. & Hyunh, H. (2008) Computer-based and paper-and-pencil administration mode effects on a statewide end-of-course English test. *Educational and Psychological Measurement,* 68, 554-570.

Kirsch, I. Jamison, J. Taylor, C. & Eignor, D. (1998). *Computer familiarity among TOEFL examinees.* TOEFL Research report 59, March, 1998, Princeton, NJ: Educational Testing Service.

Klein-Braley, C. (1997). C-tests in the context of reduced redundancy: an appraisal. *Language Testing,* 14(1), 47–84.

Law, B. & Eckes, M. (1995). *Assessment and ESL: On the yellow big road to the withered of Oz*. Manitoba: Peguis.

Manning, W. H. (1987). Development of cloze-elide tests of English as a second language. TOEFL Research report 23, April, 1987, Princeton, NJ: Educational Testing Service.

Markham, P. L. (1987). Rational deletion cloze processing strategies: ESL and native English. *System,* 15, 303-311.

McKamey, T. (2006). Getting closure on cloze: A validation study of the "rational deletion" method. *Second Language Studies,* 24, 114-164.

McKenna, M.C. & Layton, K. (1990). Concurrent validity of cloze as a measure of intersentential comprehension. *Journal of Educational Psychology,* 82, 372–77

McVee, M. Dunsmore, K. & J. R. Gavelek. (2005). Schema theory revisited. *Review of Educational Research,* 75, 531–566.

Menken, K. (2008). *English learners left behind: Standardized testing as language policy*. Toronto: Multilingual Matters.

Miller, L. Burnett, D. & Upitis, R. (1983, October 7-9). *Reading as an interactive process.* Paper presented at the 8th Annual Meeting of the Boston University Conference on Language Development, Boston, Massachusetts. Retrieved June 6, 2010, from ERIC Academic Database.

Muniz, J., Hambleton, R. K. & Xing, D. (2001). Small sample studies to detect flaws in item translations. *International Journal of Testing,* 1(2), 115-135.

Norazah Mohd Nordin, N. M., Arshad, S. R., Razak, N. A., & Jusoff, K. (2010). The Validation and Development of Electronic Language Test. *Studies in Literature and Language,* 1, 1-7.

Ockey, G. J. (2009). Developments and challenges in the use of computer-based testing for assessing second language ability, *The Modern Language Journal,* 93, 836-847.

Oikonomidoy, E. (2007). I see myself as a different person who [has] acquired a lot …': Somali female students' journeys to belonging. *Intercultural Education,* 18, 15–27.

Oller, J. W. & Jonz, J. (Eds.).(1994). *Cloze and coherence*. Cranbury, NJ: Bucknes University Press.

Onwuegbuzie, A.J. & Leech, N.L. (2005). On becoming a pragmatic researcher: The importance of combining quantitative and qualitative research methodologies. *International Journal of Social Research Methodology, 8*(5), 375–387.

Paek, P. (2005). *Recent trends in comparability studies*. Pearson Educational Measurement Research Reports. Research Report 05-05. Pearson Educational Measurement. USA.

Peters, M., Weinberg, A. & Sharma, N. (2009). To like or not to like! Student perceptions of technological activities for learning French as a second language at five Canadian universities. *The Canadian Modern Language Review,* 65, 869–896.

Pikulski, J. J. & Pikulski, E. C. (1977). Cloze, maze, and teacher judgment. *The Reading Teacher,* 30, 766-770.

Pino-Silva, J. (2008). Student perceptions of computerized tests. *ELT Journal,* 62, 148-156.

Pommerich, M. (2004). Developing computerized versions of paper-and-pencil tests: Mode effects for passage-based tests. *Journal of Technology, Learning, and Assessment, 2*. Available from http://www.jtla.org

Pomplun, M., Frey, S. & Becker, D. (2002). The score equivalence of paper-and-pencil and computerized versions of a speeded test of reading comprehension. *Educational and Psychological Measurement,* 62, 337-354.

Porter, D. (1983). The effect of quantity of context on the ability to make linguistic predictions: A flaw in a measure of general proficiency. In A. Hughes & D. Porter (Eds.), *Current developments in language testing* (pp. 63-74). London: Academic Press.

Prensky, M. (2001). Digital natives, digital immigrants: Part 1. *On the Horizon,* 9, 1 – 6.

Propst, I. & Baldauf, R. B. (1979). Use matching cloze tests for elementary ESL students. *The Reading Teacher,* 32, 683-690.

Radwanski, G. (1987). *Ontario study of the relevance of education and the issue of dropouts*. Toronto: Ontario Ministry of Education.

Russell, M. (1999). Testing on computers: A follow-up study comparing performance on computer and on paper. Retrieved from http://epaa.asu.edu/ojs/issue/view/7

Saito, Y. (2003). Investigating the construct validity of the cloze section in the examination for the certificate of proficiency in English. In Johnson, J. S. (Ed.). *Spaan Fellow Working Papers in Second or Foreign Language Assessment* (pp. 39-82). English Language Institute University of Michigan: Ann Arbor, MI.

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

Schwarz, R. D. Rich, C., & Podrabsky, T. (2003, 22-24 April). *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.

Seidman, I. (1998). *Interviewing as qualitative research: A guide for researchers in education and the social sciences* (2nd ed.). New York: Teachers College Press.

Shanahan, T., Kamil, M.L. & Tobin, A.W. (1982). Cloze as a measure of intersentential comprehension. *Reading Research Quarterly,* 17, 229–55.

Shohamy, E. (2000). *The power of tests: A critical perspective on the uses of language tests*. Essex: Pearson.

Siegel, A. (1992). Multiple t tests: Some practical considerations. *TESOL Quarterly,* 26, 773-775.

Snow, C. (2008). Crosscutting themes and future research directions. In D. August & Shanahan, T. (Eds.), *developing reading and writing in second language learners*. New York: Routledge.

Spray, J. A., Ackerman, T. A., Reckase, M. D. & Carlson, J. E. (1989). Effect of medium of item presentation on examinee performance and item characteristics. *Journal of Educational Measurement,* 26(3), 261-271.

Statistics Canada (2010). *The Daily Statistics Canada*. Retrieved July 16, 2011, from: http://www.statcan.gc.ca/daily-quotidien/100309/dq100309a-eng.htm

Steinman, L. (2002). A touch of...class! *The Canadian Modern Language Review,* 59, 921-301.

Taira, T. & Oller, J. W. (1994). Cloze and episodic organization. In J. W. Oller & J. Jonz (Eds.), *Cloze and coherence* (pp. 345-369). Toronto: Bucknell.

Takanashi, Y. (2008). Can cloze tests measure discourse competence in ESL/EFL appropriately? *Bulletin of Fukuoka University of Education,* 57, 47-57.

Taylor, C., Jamieson, J. and Eignor, D. (2000) Trends in computer use among international students. *TESOL Quarterly*, 34, 575–85.

Taylor, C., Kirsch, I., Eignor, D. & Jamieson, J. (1999). Examining the relationship between computer familiarity and performance on computer-based language tasks. *Language Learning,* 49, 219–274.

Trauth, E. M. (2006). Theorizing gender and information technology research. In E. M. Trauth (Ed.), *Encyclopedia of Gender and Information Technology*, (Vol 2, 1154-1159). Hershey, PA: Idea Group Publishing.

Ushida, E. (2005). The role of students' attitudes and motivation in second language learning in online language courses. *CALICO Journal,* 23, 49-78.

Wang, S. Jiao, H. Young, M. J. Brooks, T. & Olson, J. (2008). Comparability of computer-based and paper-and-pencil testing in K-12 reading assessments: A meta-analysis of testing mode effects. *Educational and Psychological Measurement,* 68, 5-24.

Ward, T. J. Hooper, S. R. & K. M. Hannafin, (1989). The effect of computerized tests on the performance and attitudes of college students, *Journal of Educational Computing Research,* 5, 327–333.

Watanabe, Y. & Koyama, D. (2008). A meta-analysis of second language cloze testing research. *Second Language Studies,* 26(2), 103-133.

Watt, D. & Roessingh, H. (2001). The dynamics of ESL dropout: Plus ca change... *Canadian modern language review*, 58, 203-222.

Whitley, B. E. (1997). Gender differences in computer-related attitudes and behavior: A meta-analysis. *Computers in Human Behavior,* 13, 1-22.

Wise, S. L. & Plake, B. S. (1990). Computer-based testing in higher education, *Measurement and Evaluation in Counselling and Development,* 23, 3–10.

Yamashita, J. (2003). Processes of taking a gap-filling test: comparison of skilled and less skilled EFL readers. *Language Testing*, 20, 267-293.

Yu, G. (2010). Effects of presentation mode and computer familiarity on summarization of extended texts. *Language Assessment Quarterly,* 7(2), 119-136.

Appendices

Appendix A Sample Passages from the Paper and Computer LOMERA

A.1 Sample LOMERA Passage

The Trip to the Store

Every Saturday morning Alice and Sam go with their father to the store to buy food. They ___1___ a list of things ___2___ buy. Sam writes the ___3___, while Alice tries to ___4___ what they need to ___5___. They need to have ___6___ shopping carts to hold ___7___ food. Sam reads the ___8___ and their dad looks ___9___ the best prices. Alice ___10___ Sam always get to __11___ on one favourite food ___12___ buy like ice cream. They both like Saturday morning.

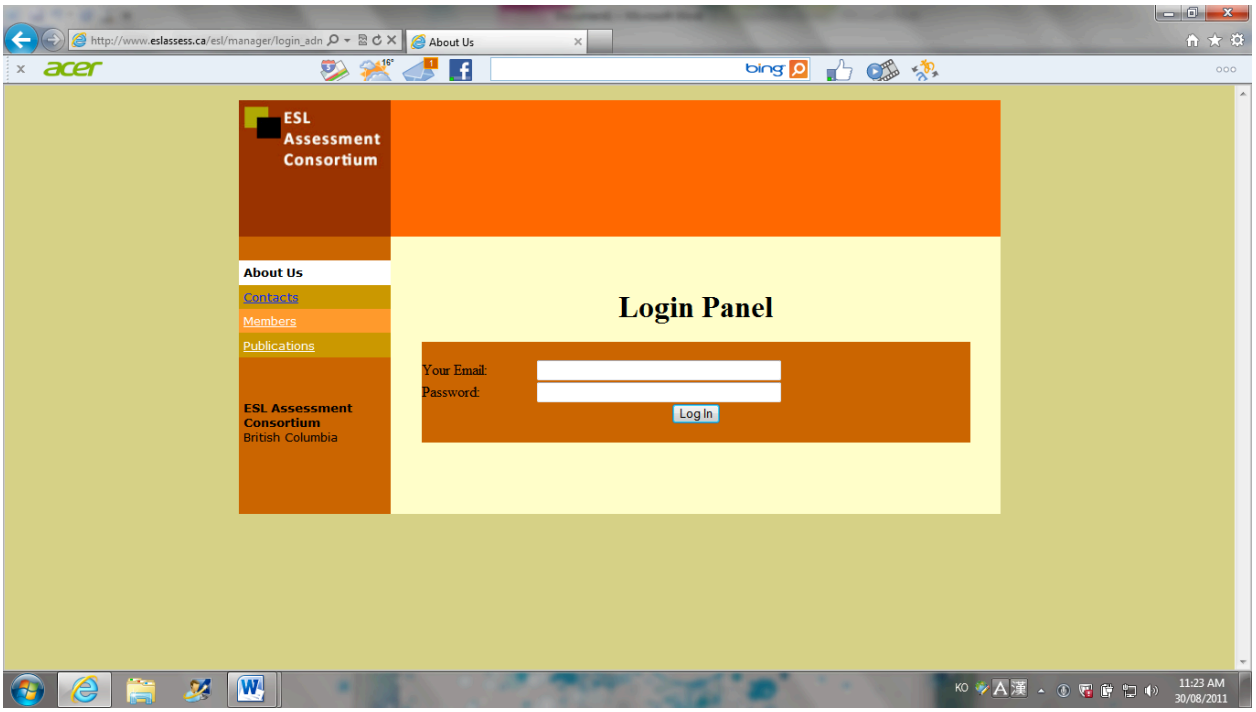
1:	2:	3:	4:
a: wrote	a: and	a: list	a: argue
b: work	b: which	b: paper	b: remember
c: have	c: for	c: book	c: have
d: write	d: to	d: note	d: want

5:	6:	7:	8:
a: by	a: two	a: them	a: book
b: bye	b: to	b: those	b: list
c: buy	c: too	c: their	c: note
d: bought	d: a	d: there	d: paper

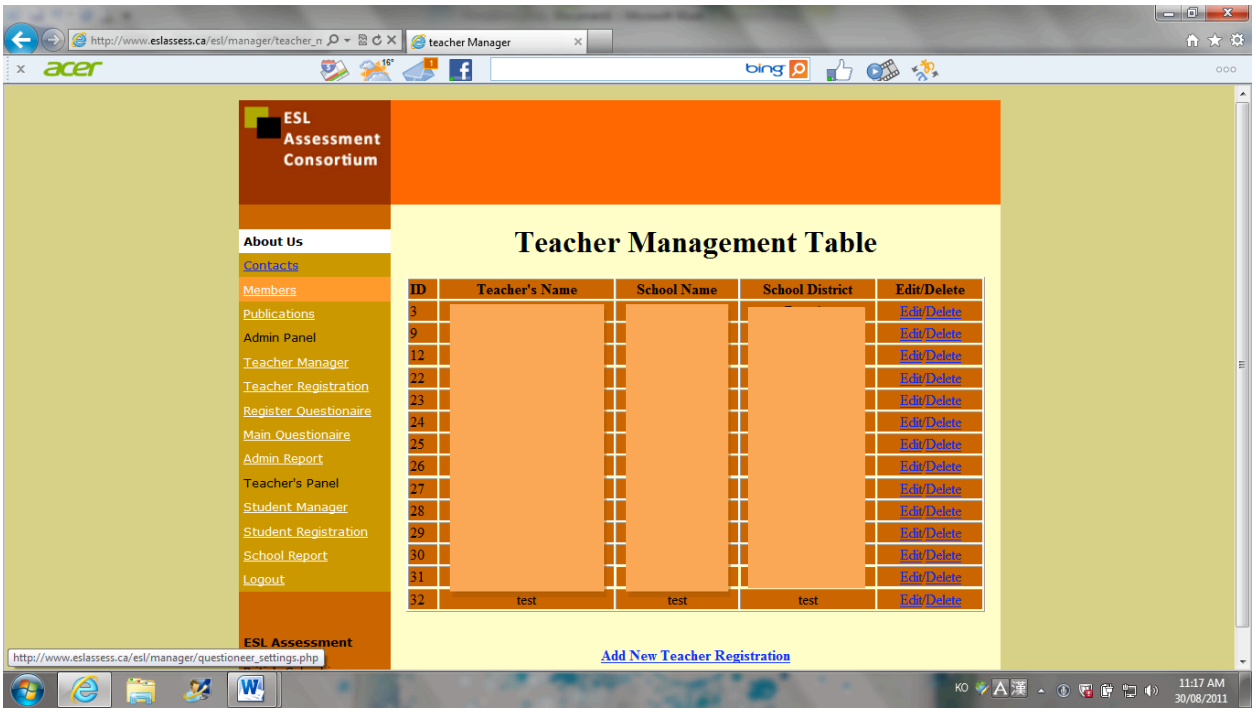
9:	10:	11:	12:
a: from	a: an	a: gobble	a: too
b: for	b: with	b: grab	b: two
c: form	c: but	c: decide	c: toot
d: with	d: and	d: chew	d: to

A.2 LOMERA Screenshots

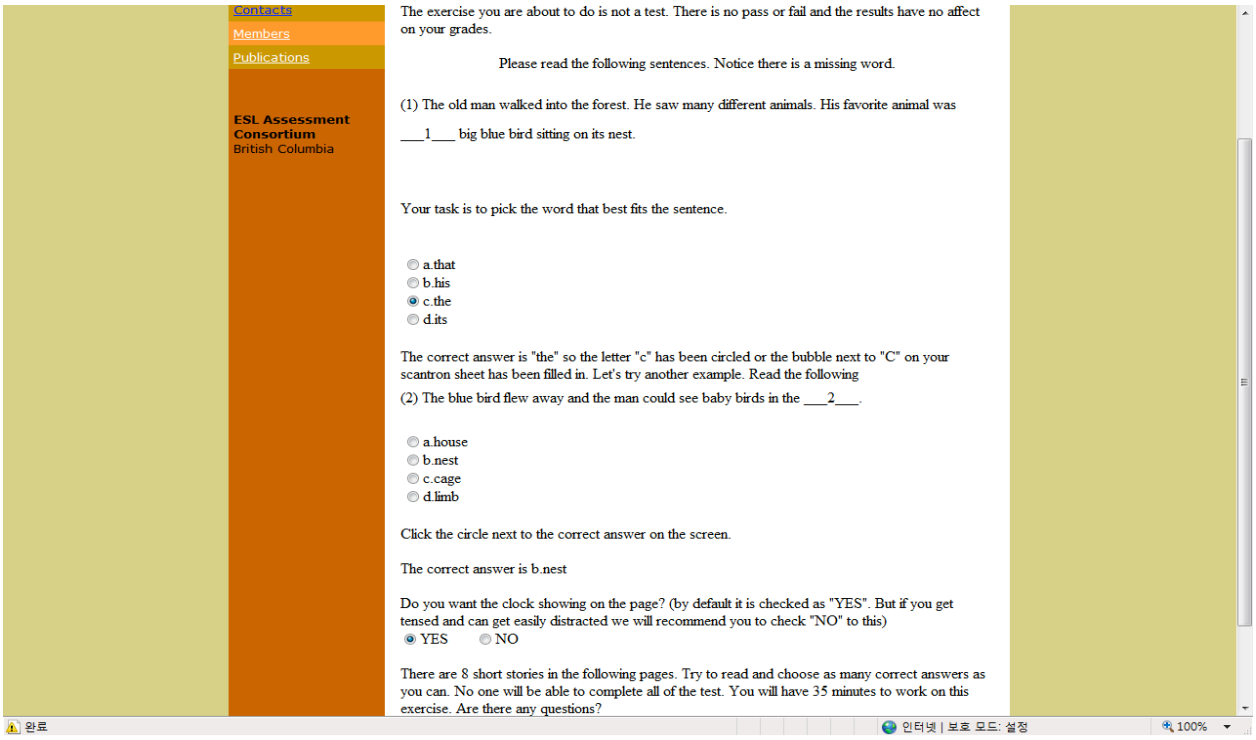
Screenshot of the online administrator login panel



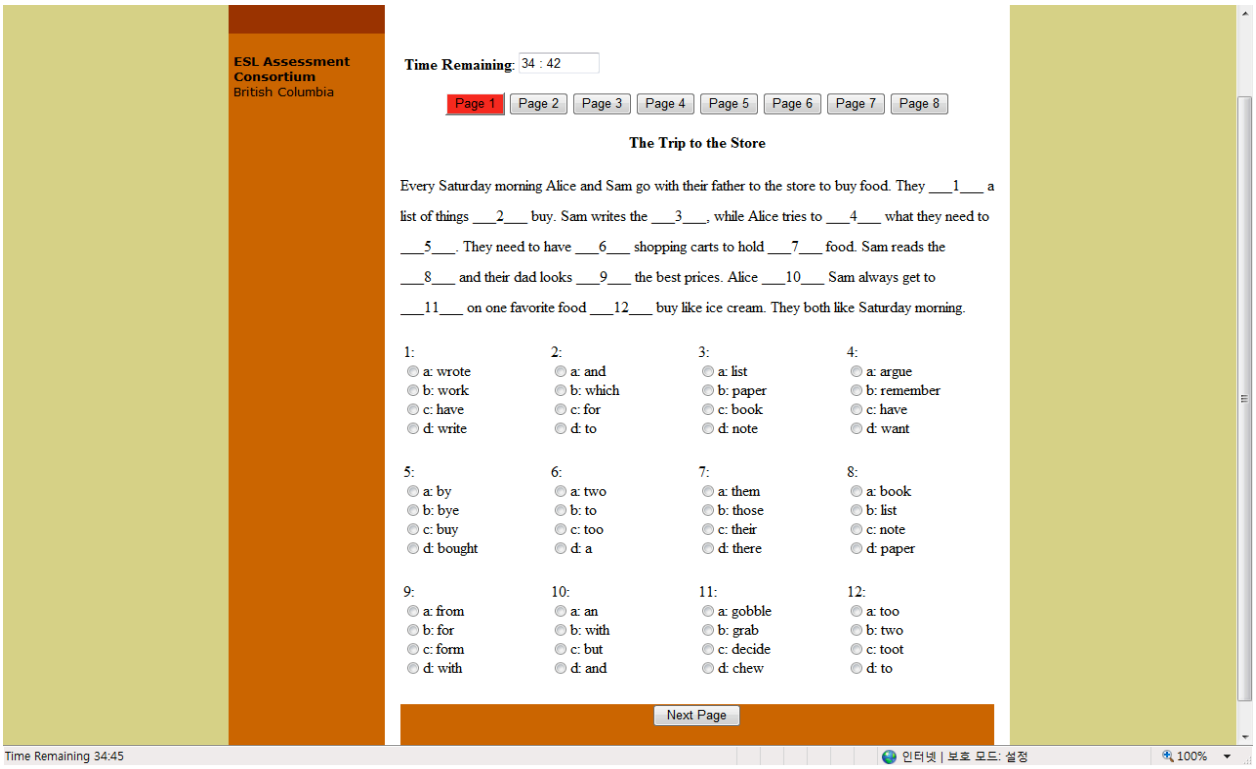
Screenshot of the online administrator report of test results for each teacher



Screenshot of the online test instructions



Screenshot of an online LOMERA test passage



Appendix B Word Class for Each Mutilation per Passage

	Readability[6]	Noun	Article	Preposition	Adjective	Conjunction	Verb	Pronoun	Infinitive	Adverb
1	2	2			1	2	4	1	2	
2	4	3	1	2	1	1	2	2		
3	5	3	2	3	1			2		
4	6/7	5	1	1			1	3		1
5	8	3		2	3		4			
6	9	4	1	1	2	1	1	2		
7	10	1	3	1	3	1	1	1		1
8	11/12	6	1	1			1	2		1

[6] Readability levels were determined using a computer program that used the Fry, Flesch-Kincaid and Raygor different readability formulas

Appendix C Computer Familiarity Questionnaire[7]

How often is there a computer available for you to use at these places?		Once a week or more often	Between once a week and once a month	Less than once a month	Never
home					
school					
friend's					
Library that you use					
		Very comfortable	comfortable	somewhat comfortable	Not comfortable
How comfortable are you with using a computer?					
How comfortable are you with using a mouse?					
How comfortable are you with using a computer to write a paper?					
How comfortable would you be taking an English test on a computer?					
How many tests or examinations have you taken on a computer?		5 or more	3-4	1-2	0
How would you rate your ability to use a computer?		Excellent	Good	Fair	Poor
How often do you use a computer?		More than once a day to once a week	Less than once a week to once a month	Less than once a month	never
How often do you use the Internet?					
How often do you use a computer to send or receive e-mail?					
How often do you use each of these types of computer programs?					
	Games				
	Word processing				
	Spreadsheets				
	Graphics				

[7] Adapted from Taylor et al. 1999

Appendix D Perceptions of PBT and CBT Questionnaire[8]

In which test were reading passages easier to navigate through?	On paper	No difference	On computer
In which test were reading passages easier to read?			
In which test was the text in the items easier to read?			
Which test created less stress?			
Which test made you less tired?			
In which test was it easier to record answers?			
In which test was it easier to change answers?			
Were you more likely to guess the answer in			
Which test was more comfortable to take?			
On which test would you be more likely to receive the same score if you took it a second time?			
Which test was more enjoyable to take?			
Which test more accurately measured your reading comprehension skills?			

[8] Adapted from Al-Amri (2008)

Appendix E Consent Forms

E.1 Consent Form for Parents

The University of British Columbia
Department of Language and Literacy Education
2125 Main Mall
Vancouver, B.C. V6T 1Z4
Phone: (604) 822-5788, Fax: (604) 822-3154

Consent Form for Parents
Comparability Study of a Paper and Computer-based Multiple-choice Cloze Reading Assessment for ESL Learners

Principal Investigator: Lee Gunderson, Department of Language and Literacy Education, The University of British Columbia, Phone: xxx-xxx-xxxx

Co-Investigator(s): Dennis Murphy Odo, PhD Candidate, Department of Language and Literacy Education, The University of British Columbia, Phone: xxx-xxx-xxxx

This data generated from this research will be used in a dissertation for a doctor of philosophy degree in education. The results of this dissertation study will be made available to other scholars and the general public via the University of British Columbia library.

Purpose:

This research will study whether secondary school ESL students perform differently on paper-based and computer-based versions of the same multiple-choice reading test. Another aim is to investigate the relationship between learners' computer familiarity or perceptions of computer-based tests and their test performance.

Study Procedures:

- Groups will be assigned for the study randomly, but both groups will be taking the same assessments except in a different order. The results of these assessments will have no bearing on the student's standing in their class.

- The research will take place over two separate sessions. Each session will take about 35 minutes to complete the assessment and about 15 each session to do each of the surveys. Some students will also be invited to explain some of their answers to the survey questions in a third interview session.

- If the student chooses not to participate in the study, he or she will work on another class assignment chosen by his or her teacher during the time that the other students are involved with the study.

- This study involves analysis of tests that are a part of the regular class routine. However, the results of those who do not participate will not be included in the research. Instead, those results will be given to the classroom teacher for her records.

Potential Risks:
The risks to participants are possible anxiety from taking the tests. They may also be slightly inconvenienced by answering two 10 minute questionnaires.

Potential Benefits:
The participants may benefit from this study because their English reading ability will be assessed using a test that has been developed specifically for students in the Lower Mainland.

Confidentiality:

The participant's identity will be kept confidential by only identifying documents with code numbers (rather than student names) and keeping them in a locked filing cabinet. Participants will not be identified by name in any reports of the completed study. Records will be kept on a computer hard disk that is password protected.

Compensation:

Participants will be *entered into a draw for prizes* in the amount of - $100.00. Participation in this draw is not dependent on completion of the project, and anyone who decides to withdraw from the study after it has begun will still be eligible for the draw.

Contact for information about the study:

If you have any questions or desire further information about this study, you may contact Lee Gunderson at xxx-xxx-xxxx or gunderso@interchange.ubc.ca. Alternately, Dennis Murphy Odo can be contacted at xxx-xxx-xxxx or dmodo@interchange.ubc.ca.

Contact for concerns about the rights of research subjects:

If you have any concerns about your treatment or rights as a research subject, you may contact the Research Subject Information Line in the UBC Office of Research Services at 604-822-8598 or if long distance e-mail to RSIL@ors.ubc.ca or toll free 1-877-822-8598.

PLEASE KEEP THIS PART OF THE FORM FOR YOUR RECORDS

Consent:

Please read the information below, circle your consent, sign and print your name. Then return the last page of this consent form signed and dated to **Ms. A. M.** at xxx Secondary School.

Your child's participation in this study is voluntary and you or your child may refuse to participate or withdraw from the study at any time without jeopardy to his or her class standing.

Your child's data can be withdrawn from the study at any point prior to the data analysis. If it is withdrawn, it will be destroyed immediately.

Your signature below shows that you were given a copy of this consent form for your records.

Parent or guardian, please indicate your choice below.
'I consent/ **I do not consent** (circle one) to my child's participation in this study.'

_____

Parent or Guardian Signature Date

_____

Printed Name of Parent or Guardian signing above

PLEASE RETURN THIS SECTION OF THE FORM

E.2 Consent Form for Students

The University of British Columbia
Department of Language and Literacy Education
2125 Main Mall
Vancouver, B.C. V6T 1Z4
Phone: (604) 822-5788, Fax: (604) 822-3154

Assent Form for Students
Comparability Study of a Paper and Computer-based Multiple-choice Cloze Reading Assessment for ESL Learners

Principal Investigator: Lee Gunderson, Department of Language and Literacy Education, The University of British Columbia, Phone: xxx-xxx-xxxx

Co-Investigator(s): Dennis Murphy Odo, PhD Candidate, Department of Language and Literacy Education, The University of British Columbia, Phone: xxx-xxx-xxxx

The results of this research will be used for a doctor of philosophy (Ph.D) degree in education. When this study is finished, it will be at the University of British Columbia library for anyone who wants to read.

Purpose:
We will study whether secondary school ESL students get different scores on the same reading test if it is on paper or on a computer. We also want to learn if students who use computers more or like the computer test more will have better scores on the computer test.

Study Procedures:

- Students will be put into two groups for the study randomly. Both groups will be taking the same tests except in a different order. The results of these tests will not affect their class grade.

- The study will have two different meetings. Each meeting will take about 35 minutes to complete the test and about 15 minutes each meeting to do each of the surveys. Some students will be asked to explain some of their answers to the survey questions in an interview.

- If the student chooses not to join the study, he or she will do other class work chosen by his or her teacher during that time.

- This study involves analysis of tests that are a part of your regular class work. However, the scores from students who are not in the study will not be used in the study.

Risks:
You might feel a bit nervous about taking the two tests. You may also be a little bothered by answering two 10 minute questionnaires.

Benefits:
You may benefit from this study because your English reading will be tested using a test that was made for ESL students in the Lower Mainland.

Confidentiality and Privacy:
Your name will not be on any documents in the study. We will only use code numbers. We will keep the documents a locked filing cupboard. Records will be on a computer that has a secret password.

Compensation and Rewards:

You will be *entered into a draw for prizes* totaling $100.00. You do not have to join the study to be in the draw, and anyone who wants to leave the study after it starts will still be in the draw.

Contact for information about the study:

If you have any questions about this study, you can call Lee Gunderson at xxx-xxx-xxxx or email at gunderso@interchange.ubc.ca. You can also call Dennis Murphy Odo at xxx-xxx-xxxx or email at dmodo@interchange.ubc.ca.

Contact for concerns about the rights of research subjects:

If you are worried about your treatment or rights, you may contact the Research Subject Information Line in the UBC Office of Research Services at 604-822-8598 or if long distance (free) 1-877-822-8598 or e-mail to RSIL@ors.ubc.ca.

PLEASE KEEP THIS PART OF THE FORM FOR YOUR RECORDS

Assent:

Please read the information below, circle your assent, sign and print your name. Then return the last page of this consent form signed and dated to **Ms. A. M.** at xxx Secondary School.

You do not have to join this study and you can leave the study any time without your changing your class grade.

Your information can be taken out from the study anytime before we analyze it. If it is taken out, it will be destroyed at that time.

Your signature below shows that you were given a copy of this consent form.

Assent

Participant please indicate your choice below.
'I assent/ **I do not assent** (circle one) to participating in this study.'

_____

Participant Signature Date

_____

Printed Name of Participant signing above

PLEASE RETURN THIS SECTION OF THE FORM

Appendix F LOMERA Items Showing DIF

20

__17___ are the same resources __18___ people used in the __19___, and they may be __20___ for the same purposes.

a: tried
b: held
c: used
d: washed

30
The ___25___swim up stream, and __26____ even travel far into __27___ interior of British Columbia. __28___ the end of their __29___, they spawn and then __30___ die.

a: they
b: those
c: them
d: she

44
__43___ tribe of people made __44___ with materials they could __45___ easily.

a: hats
b: houses
c: hearts
d: tents

92
The __92__ of how DNA carries __93___ blueprint was one of ___94__ greatest achievements of 20th-century ____95__.

a. mystery
b: discovery
c: carries
d: caught