An Investigation of the Cross-Mode Comparability of a Paper and Computer-Based Multiple-Choice Cloze Reading Assessment for ESL Learners

Dennis Murphy Odo (dmurphyodo@yahoo.ca)
Georgia State University, U.S.A.

Abstract

This study investigates whether a computer-based version of a multiple-choice cloze reading test for English-language learners is comparable to its traditional paper-based counterpart. 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 as similar as possible in terms of content, questions, pagination, format and layout. The design was counterbalanced so that two groups of learners took the tests in the opposite order and their scores were compared to address concerns about practice and order effect. Results indicate that the paper and computer-based versions of the test are comparable. These findings help validate the cross-mode comparability of assessments outside of the traditional discrete-point multiple choice tests which predominates in current research.

Keywords: Cross-mode comparability, mode effect, second language assessment, reading assessment

INTRODUCTION

As ESL (ELL) populations increase across North America, access to (the availability of) ESL assessments that are both valid and reliable becomes critical to the planning and implementation of instruction that meets the needs of these learners. Among the many challenges educators face when developing assessments to meet their needs is that they are minimally resource-intensive and they provide teachers and decision makers with valid and useful information.

The issue of being able to communicate assessment information across a wider constituency becomes increasingly important as the proportion of immigrant ESL learners continues to rise in North America and ESL student mobility among districts continues to increase. At present, one major impediment to wider information sharing among districts is the eclectic assortment of reading measures used by school district personnel. To address this dilemma, The ESL assessment consortium developed a standardized secondary ESL reading assessment that provides a general indicator of English reading proficiency (Gunderson, Murphy Odo, D'Silva, 2010). This measure – known as the Lower Mainland English Reading Assessment (LOMERA) – is a locally-normed reading assessment that allows districts to gather and to share information about their ESL learners' reading proficiency.

Consortium members were pleased with the usability and the valuable information provided by the LOMERA but they also realized that such a test, if administered online, could reduce the use of resources in administering, scoring and collecting data. Therefore, they began to consider the development of a computer-based version of the test. In their deliberations, the question was soon raised about whether scores from the computer-based test would be comparable to those obtained from the traditional paper-based version. This concern had to be addressed because the computer test would not be replacing the paper test. Instead, the paper test would be used in contexts where insufficient technological resources existed. A review of research literature into cross-mode comparability yielded no satisfactory answer about whether tests are equivalent across modes for cloze-type tests so further research was deemed necessary.

DESCRIPTION OF THE ASSESSMENT TOOL

The version of the LOMERA that is currently in use is a paper-based multiple-choice rational cloze style test (i.e., maze). The test is comprised 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 local districts. 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. In addition to being organized according to difficulty, scores on the passages are also normed with students from local school deistricts. 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 for comparison to a particular test taker’s performance.

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 equivalence 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.

RESEARCH EXPLORING CROSS-MODE COMPARABILITY OF LANGUAGE ASSESSMENTS

Cross Mode Comparability Research

The central objective of comparability research is to determine whether test results from computer-based tests are equivalent 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 second-language tests. The results of many of these studies are presented and discussed below.

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 research 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. A more recent review by Leeson (2006) examined research into participant and technical variables that could potentially cause mode effect. Her main conclusions were that with regard to participant variables such as ethnicity, cognitive ability, familiarity, and anxiety the findings appeared to be mixed. In terms of interface-legibility and interactivity, she pointed out several possible issues with screen size, fonts, line length and whitespace. She also touched on debates around item presentation, item review and scrolling. One of her main conclusions was that a great deal of research still needs to be done.

Two empirical studies with locally developed English language assessments found the paper and computer-based versions to be comparable. One 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 an 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 raise the issue that reading assessments may have greater potential to exhibit mode effects which is of particular interest in the present investigation because the test being studied is a reading assessment. 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 equivalent and could be used interchangeably if candidates had enough computer training (Green & Maycock, 2004).

In contrast to the findings claiming cross-mode test equivalence, there were at least two studies that did not report 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).

Review of previous research into cross-mode comparability provides several justifications for undertaking the present research study. First, this study adds to the relatively scant cross-mode second language assessment comparability literature. There is presently an insufficient amount of research specifically evaluating cross-mode comparability for second-language assessments. The necessity for further research is demonstrated by the continued debate about whether there is in fact cross-mode comparability across second language assessments. Though much current research points to comparability, at least two highly-regarded language assessment researchers have reported findings that language tests have mode effect (i.e., are not comparable) which demands further study to help resolve these contradictory findings. Thirdly, the types of assessments being investigated by evaluating a multiple-choice cloze (maze) test have thus far been limited to traditional discrete-point multiple choice item types. The present research begins to explore cross-mode comparability with forms of assessment that go beyond traditional 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. A final rationale given for the need to conduct this research is that reading comprehension tests seem to show the greatest potential to have mode effects (see Choi, Kim & Boo 2003). All of these reasons seem to provide a reasonable justification for the present research.

Some commentators might also ask whether it would be more prudent to simply 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, moving the tests exclusively online would exploit their many time and labor 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. Similar circumstances are sure to exist in other districts with high concentrations of ESL learners. Chapelle (2001) also points that language tests are designed to evaluate a particular construct such as reading proficiency. If the computer test produces vastly different results then the integrity of the test construct is called into question. Therefore, comparability research must be conducted to ensure construct validity. Until these challenges can be overcome, cross-mode comparability research must continue.

STUDY PROCEDURES

The main objective of this study was to determine whether paper and computer-based versions of a standardized m-c cloze reading test for second language learners are comparable.

Upon receiving IRB approval, participants were recruited 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, font 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.

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.

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 1 below.

Table 1
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

t-test Analysis

LOMERA test takers' scores across testing modes both between administrations were compared using a paired-sample t-test.

Table 6
Results of Paired Sample t-test Comparing First and Second Test Administration

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

The purpose of the t-test analysis 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) = .933, p = .353. 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.

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.

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; α = .95). The analysis of the internal consistency of the computer test also resulted in a very high reliability coefficient (96 items; α = .95).

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 equivalence because it demonstrated that examinees scored similarly on both modes of the test. A statistically significant correlation of r = .96 (p < .001) was found when the scores from the first and second administration were correlated. This is an impressive 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 the paper and computer modes to discern the degree of relationship between examinees' passage scores. The correlations that are of most interest here are those from the same passage taken across different testing modes (see table 7). 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 which is considered to indicate a moderate to strong relationship between the two passages (Cohen, 1988). 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 7
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 7) 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 construct (i.e., general L2 reading proficiency).

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. The DIF analysis adds a unique and informative dimension to the analysis because, unlike the correlations and t-test, it provides evidence for comparability at the item rather than test level. There were several steps involved in the process of creating the Delta plot chart. First, scores from individual test passages were transformed from a per-passage score to a binary per-item score and entered into a database. Following 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 chart that plotted the intersection of all examinees' scores for each individual test item on both testing modes. Subsequent to plotting the relationship between the paper and computer score for each item on the graph, a regression line was added to clarify the general direction of the plots. Two lines demarcating the 95% confidence interval were also added to enable the DIF analysis by distinguishing the area within which item plots had to be located not to be considered functioning differently. Following the advice of Muniz, Hambleton and Xing (2001), items outside the 95% confidence interval band around the regression line were deemed to be DIF because discrepancy in examinee performance on one mode differs significantly from the norm.

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 criteria. That is, these items were outside of the 95% confidence interval band around the regression line. As is illustrated in Figure 1, 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 appear to be functioning differentially. 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. Some possible causes will be discussed below.

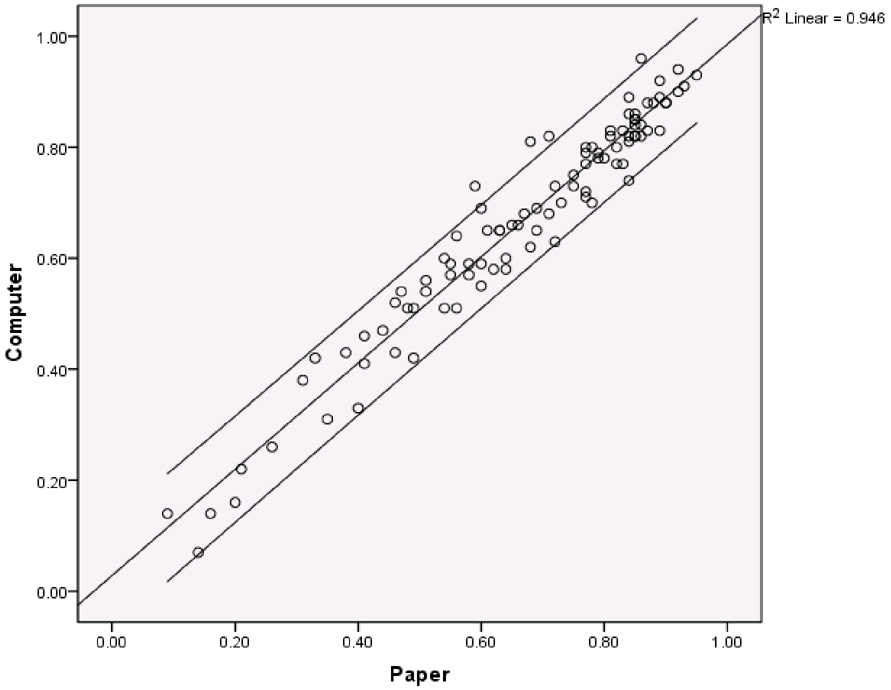

Figure 1. Scatter Plot of Item p-values on Paper and Computer Versions of LOMERA

DISCUSSION

Descriptive Statistics and t-test Analyses

Based on the descriptive statistics and results from the t-test reported above, there is no noteworthy disagreement in scores between the two versions of the LOMERA either within or across the two test administrations. The paired-sample t-test result showed that there was no statistically significant difference between test takers' scores in each mode of the test. This finding provides evidence for the comparability of the computer version of LOMERA with its paper counterpart by demonstrating that there are no sizeable differences in the scores for each test taker from one test to the other. 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 cited in Sawaki, 2001) 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 other 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.

Correlations

Correlations between both administrations of the LOMERA are .96 which indicates considerable agreement in scores across modes. That is, those who scored highly on one mode of the test tended to score highly on the other as well while those with lower scores on one mode typically had lower scores on the other. This finding provides additional evidence for the cross-mode equivalence of both versions of the LOMERA by demonstrating that test takers are likely to get scores that were similar across modes of test. 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 satisfied 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 as being equivalent. On the whole, 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.

DIF Analyses

A review of relevant literature has only provided two studies that have used DIF methods to investigate mode effect. The two studies located for this review both reported considerable differences in the performance of items across modes. However, the findings from the present study were that there were only 4 differentially-functioning items out of 96 test items across test modes which indicate minimal discrepancy in item-level performance across modes. 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 basic education 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 favour 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. This result also diverges from findings in the present study. Based on the limited amount of research currently available, it appears at this juncture it is too early to state definitively whether mode effect at the item level depends largely on the type of test that is being investigated. Clearly, further investigation is warranted with other types of reading test items such as those used in the present study.

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. One speculation is that examinees' possible increased level of reported 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.

CONCLUSION

The objective with this investigation 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 useful indicators of cross-mode equivalence. 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 test scores in each mode for either test administration. Results showed that there was no statistically significant difference in the test scores which confirmed that test takers' scores on the paper version of the test were comparable with their scores 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 the present study could reveal a genuine absence of mode effect.

The t-test procedure was followed with a series of cross-mode inter-passage correlations that were used to collect 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 applied DIF methods to cross-mode comparability questions, the present study seems 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 p-values. Potential causes of the divergence in these four items might be examinees' possible increased enjoyment of the computer test strengthening their patience, measurement error or simply 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. 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 this comparability 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 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 equivalence 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.

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.

REFERENCES

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.

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

Cohen. J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

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.

Chapelle, C. A. (2001). *Computer applications in second language acquisition: Foundations for teaching, testing and research.* Cambridge: Cambridge University Press.

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

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

ESL Assessment Consortium. (2009). *Assessment consortium mission statement.* Retrieved March 25, 2012 from http://www.eslassess.ca

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

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

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/

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.

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

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

Keng, L., 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.

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.

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.

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.

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

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.

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.

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.

Yessis, D. B. (2000). Comparing paper mode vs. computer mode in rate development reading assessments. Unpublished master's thesis, University of California, Los Angeles.