Language Learning & Technology May 2001, Vol. 5, Num. 2
http://llt.msu.edu/vol5num2/sawaki/default.html pp. 38-59

COMPARABILITY OF CONVENTIONAL AND COMPUTERIZED TESTS OF READING IN A SECOND LANGUAGE

Yasuyo Sawaki
University of California, Los Angeles

ABSTRACT

Computerization of L2 reading tests has been of interest among language assessment researchers for the past 15 years, but few empirical studies have evaluated the equivalence of the construct being measured in computerized and conventional L2 reading tests and the generalizability of computerized reading test results to other reading conditions. In order to address various issues surrounding the effect of mode of presentation on L2 reading test performance, the present study reviews the literature in cognitive ability testing in educational and psychological measurement and the non-assessment literature in ergonomics, education, psychology, and L1 reading research. Generalization of the findings to computerized L2 assessment was found to be difficult: The nature of the abilities measured in the assessment literature does not necessarily involve language data; mode of presentation studies in the non-assessment literature involving L2 readers are scarce; and there are limitations in the research methodologies used. However, the literature raises important issues to be considered in future studies of mode of presentation in language assessment.

INTRODUCTION

Reading from computer screens is becoming more and more common in our daily lives as the amount of reading material available on line is rapidly increasing. This influence has been seen in the field of language assessment where computerized testing, such as computer-based tests (CBTs) and computer-adaptive tests (CATs, a special case of computer-based testing, where items administered to examinees are tailored to the individual examinee's ability on the construct being measured), are attracting the attention of researchers, language learners, and test users alike, as exemplified by the implementation of CATs at institutional levels in the past 15 years (Kaya-Carton, Carton, & Dandolini, 1991; Larson, 1987; Madsen, 1991; Stevenson & Gross, 1991; Young, Shermis, Brutten, & Perkins, 1996). Regardless of the rapid growth of demand in this area, development and implementation of this new mode of testing is currently in its initial stages. Therefore, sufficient empirical data, which would allow researchers to look into the soundness of computerized language tests with regard to construct validity and fairness, are yet to be available.

One issue which requires prompt investigation is the effect of mode of presentation on comparability of the information obtained from computerized and paper-and-pencil (P&P) tests. In their comprehensive summary of issues surrounding CATs in L2 contexts, Chalhoub-Deville and Deville (1999) point out the scarcity of comparability research in L2 language tests and the importance of conducting comparability studies in local settings to detect any potential test-delivery-medium effect when a conventional test is converted to a computerized test. In terms of L2 reading comprehension tests in particular, the current move toward computerized testing is proceeding without sufficient empirical evidence that reading from a computer screen is the same as reading in print for L2 readers. Since presence of a mode effect on reading comprehension test performance would seriously invalidate score interpretation of computerized reading tests, language assessment researchers have discussed the necessity of examining (a) the degree to which

computerized reading comprehension tests measure the same construct as P&P tests and (b) the extent to which results of computerized reading tests can be generalized to other contexts (Alderson, 2000; Bachman, 2000). In order to seek future directions in investigating the effect of mode of presentation on L2 reading test performance, the present study reviews two distinct areas of previous literature: (a) studies that address general construct validity issues of computerized tests in cognitive ability as well as language assessment; and (b) studies that shed light on the effects of mode of presentation on reading performance conducted mainly in ergonomics, education, psychology, and L1 reading research.

ASSESSMENT LITERATURE

In order to support construct validity of computerized tests such that the construct being measured is not being affected by the mode of presentation, the equivalence of corresponding conventional and computerized test forms must be established from various directions. In this section, potential task changes caused by a shift to the computer administration mode will be reviewed first. Then, the criteria that have been used to evaluate cross-mode equivalence of test forms and various psychometric and statistical issues, such as stability of item parameter estimates and linking tests across modes, will be summarized. This section will close with a discussion of the impact of mode of presentation on examinees, namely, the interaction of test taker characteristics with testing conditions and the comparability of decisions made across modes.

Comparability of Tasks Across Modes of Presentation

As the first step in establishing the equivalence of computerized and conventional test forms, the content covered by the two tests should be comparable. To achieve this goal, several promising algorithms to control for content coverage have been implemented in L2 CATs in the last decade (for summaries of recent developments in content balancing algorithms, see Chalhoub-Deville & Deville, 1999, and Eignor, 1999). Even when the content coverage in a given computerized test is carefully controlled to mirror the test content specification, potential "task change" may still occur across modes of presentation, as pointed out by Green (1988). A task change is the possibility that the nature of a test task may be altered when the item is presented in a different mode, which may in turn induce unexpected changes in item difficulty. Green states, "If computer presentation changes tasks, so that the correlation between scores on the computer and conventional versions is low, then validity is threatened" (p. 78).

Greaud and Green (1986) reported low cross-mode correlations in a speeded clerical skills test, which may indicate a task change caused by a shift to the CAT format. They investigated the effect of mode of presentation on the numerical operations (NO) and coding speed (CS) subtests of the Armed Services Vocational Aptitude Battery (ASVAB) administered to applicants for the U.S. military services. Fifty college students took short versions of the two subtests. The CAT versions were completed faster by the subjects, who did better on the CAT versions in general. Moreover, when the average number of correct responses per minute was used as the test score, the between-mode correlation coefficients for the coding speed subtest remained low to moderate when corrected for attenuation, while the within-mode correlations for both subtests and the between-mode correlations for the numerical operations subtest were high. Possible explanations provided by the authors were that (a) "marking a bubble" on an answer sheet in a P&P test and "pressing a button" to enter an answer on a CAT may require different motor skills (p. 33); and (b) keeping track of the location of the items presented as a group was part of the task in the highly-speeded P&P test, while it was not the case for the CAT version, where items were displayed one by one on a computer screen (pp. 31-32).

Results of Mead and Drasgow's (1993) meta-analysis concurred with Greaud and Green's (1986) findings regarding potential presentation mode effects on speeded test performance. In their meta-analysis of 159 correlations obtained in the previous mode of cognitive ability assessment presentation studies, Mead and Drasgow found that, after correcting for measurement error, the estimated cross-mode correlations were

.97 and .72 for timed power tests and speeded tests, respectively. Based on these results, the authors concluded that mode of presentation may affect speeded tests but not timed power tests. Susceptibility of speeded tests to presentation mode effects, however, was not supported by Neuman and Baydoun (1998). In their study of mode effects on a speeded clerical test, consistent high cross-mode correlations were found between the P&P and computer modes for the instrument's subtests, and a structural equation modeling suggested that the constructs being measured in the P&P and CBT versions of the tests were equivalent.

Another source of a task change may be differences in test administration conditions across modes of presentation. Spray, Ackerman, Reckase, and Carlson (1989) argued that presentation mode effects on test performance found in previous research may be partly due to differences in the flexibility of test administration conditions. In their comparative study of P&P and CBT versions of three end-of-unit tests for the Ground Radio Repair Course at a Marine Corps Communication-Electronics School, Spray et al. allowed test takers to skip items and to review and change answers after completing the test. This is not permitted on many other computerized tests. As a result, mean scores and cumulative score distributions for the raw scores across modes on this test were not significantly different between the P&P and computerized testing groups. Additionally, no item bias due to presentation mode effects was found. Based on their findings, the authors concluded that P&P and computer-based test results would be equivalent when the same test-taking condition flexibility is maintained across modes.

Psychometric Equivalence of Conventional and Computerized Tests

Criteria for Equivalence Between P&P and Computerized Tests. In response to the growth of interest in converting conventional P&P tests to computerized forms in cognitive ability assessment over the last two decades, the 1985 version of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1985) raised the concern for the parallelism of test forms when conventional and computerized tests are used interchangeably. A year later, the American Psychological Association published *Guidelines for Computer-based Tests and Interpretations* (APA, 1986), which set forth the widely used criteria for achieving psychometric equivalence of P&P and computerized tests (Bugbee, 1996; Mead & Drasgow, 1993). The *Guidelines* specifies the psychometric equivalence of P&P and computerized tests as follows:

> When interpreting scores from the computerized versions of conventional tests, the equivalence of scores from computerized versions should be established and documented before using norms or cutting scores obtained from conventional tests. 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, or have been made approximately the same by rescaling the scores from the computer mode. (p. 18)

In the *Guidelines*, criterion "a" is considered to be a prerequisite for achieving psychometric equivalence of P&P and computerized tests, while rescaling methods can be used to place the P&P and computerized test scores into the same scale when criterion "b" is not met. Conversely, if "b" but not "a" is met, then test forms cannot be equivalent, despite the fact that the two tests can still be transformed to have similar distributions.

Some variations of these criteria also exist. After reviewing previous mode of presentation studies on CBTs in particular, Bugbee (1996) claimed that the equivalence criteria could be altered, depending on how a CBT is used. For example, if a CBT is used as an alternative for a conventional form, then demonstrating high correlation and nearly equal means and variances between the modes may suffice. If a CBT is to be used as an exchangeable form, however, then satisfying the criteria for parallel tests in the Classical Test Theory, which requires equal means and standard deviations across modes and equal correlations with a criterion measure, should be pursued.

In addition to the criteria suggested by the *Guidelines*, Steinberg, Thissen, and Wainer (1990) described how structural equation modeling can be used for investigating the equivalence of the number and loading of latent factors across modes for construct validation of CATs. Quite a few studies have utilized factor analysis or structural equation modeling approaches in order to investigate factorial similarity as part of the cross-mode equality requirements (Green, 1988; Moreno, Wetzel, McBride, & Weiss, 1984; Neuman & Baydoun, 1998; Staples & Luzzo, 1999; Van de Vijver & Harsveld, 1994). For example, in a study of the ASVAB, Green (1988) gave P&P and CAT versions of the test to 1,500 Navy recruits and conducted an exploratory factor analysis to compare the underlying factor structure of the two forms of the test. Due to the similarity of the obtained underlying factor structures, Green (1988) concluded that construct validity of the CAT version of the ASVAB seemed to be supported.

Stability of Item Parameter Estimates. Green, Bock, Humphreys, Linn, and Reckase (1984) and Henning (1991) argue that there is no guarantee that item parameter estimates, such as item difficulty and discrimination, will remain constant across modes. A promising strategy would be to recalibrate item parameters when sufficient data become available from a CAT to see if the P&P estimates are invariant across modes. Then, items with unstable parameter estimates can be reconsidered. For example, Stone and Lunz (1994) examined stability of item parameter estimates for multiple-choice items in a medical technologist certification exam administered by the Board of Registry. The item parameter estimates in this study were obtained by using item response theory (IRT), which specifies by a mathematical function the relationship between the observed examinee performance and the unobservable examinee abilities considered to underlie it (Hambleton & Swaminathan, 1985). When the equivalence of item difficulty parameters obtained in the P&P and CAT forms of the certification exam was evaluated in terms of the standardized differences, Stone and Lunz found that, although the text-only items showed a strong trend of parameter estimation equivalence, items with graphics tended to be less stable than text-only items. Further investigation of the items suggested that the significantly different difficulty estimates obtained across modes seemed to be accounted for by different picture quality as well as by image and character sizes used across the CAT mode.

Linking Tests Across Modes. When a computerized test is used as an alternative or replacement for a conventional test, the score relationship between the computerized and conventional test forms must be established. This can be achieved in two steps. First, qualitative and quantitative analyses of equivalence of the construct being measured and psychometric properties between the test forms must be examined. Second, if sufficient evidence based on the analysis supports equivalence of the test forms, then the forms can be placed onto the same scale and appropriately linked using conventional test equating methods (Staples & Luzzo, 1999). A variety of test equating methods have been used in conventional tests to link separate forms of a test built to the same test specifications, but the stringent criteria required for equating are not likely to be satisfactorily met when equating is attempted across modes. A concern here is that a CAT has a different pattern of measurement accuracy from a conventional non-adaptive test. A CAT is often designed to have equal measurement accuracy across a score scale, while a conventional test is not (Green et al., 1984; Kolen & Brennan, 1995). This is one of the main reasons why researchers have questioned the feasibility of equating a CAT to a P&P test. This point was challenged, however, by Green et al. (1984), Wainer (1993), and Wainer, Dorans, Green, Mislevy, Steinberg, and Thissen (1990), who argued that, since ideal test equating, as defined by traditional testing literature, may not be achieved across modes, calibrating rather than equating of test scores should be sought.

Linn's (1993) definition and description of calibrating as a less stringent form of test linking method, as compared to test equating, could be employed for establishing a statistical relationship between conventional and computerized tests. According to Linn, calibrating, unlike equating, can be applied to link tests designed to measure the same construct but not built to the same specification and associated with different score reliability patterns. Moreover, failing to satisfy the stringent equating assumptions does not keep researchers from employing conventional statistical equating methods. The differences

between equating and calibrating are that the calibrated forms cannot be exchangeable, and the potential decrease in stability of test linking results across time and samples requires close monitoring of calibration results. By using Linn's criteria above, mathematics achievement scores obtained from statewide tests and the National Assessment of Educational Progress (NAEP) have been successfully calibrated by means of an equating method called equipercentile equating (Ercikan, 1997; Linn & Kiplinger, 1995). Other large-scale testing programs have also calibrated CATs and P&Ps, utilizing various conventional equating methods (e.g., Lunz & Bergstrom, 1995; Segall, 1997). For interpreting these test calibrating results, Pommerich, Hanson, Harris, and Sconing's (2000) guidelines for interpreting linking results between tests built to different test specifications would be useful.

Impact of Introduction of Computerized Tests to Examinees

Interaction of Examinee Characteristics and Testing Conditions. Another concern related to construct validity of CBTs is the effect of examinee backgrounds on test performance and attitudes toward new forms of language tests. Investigation of these issues is important because a test score obtained from a computerized test should reflect the construct of interest only. That is, if the test score represents *both* language ability and computer familiarity, for example, then valid generalization of test scores across modes is no longer possible. A set of studies has focused on computer familiarity and its potential effects on performance on CBTs and CATs.

Oltman (1994) investigated the effects of complexity of mouse manipulation on performance in reading and math subtests of the Computer-Based Academic Skills Assessments for the Praxis Series: Professional Assessment for Beginning Teachers. Two types of mouse manipulation were required by the tasks involved: "simple" (items which require a single click to mark an answer) and "complex" (items which require more than one click to mark an answer). The reading and math subtests were given to 333 minority (Hispanic, Native American, and Black) and 148 white university students who were not experienced computer mouse users. An ANOVA analysis showed a significant interaction effect of ethnic group and task type, suggesting that minorities, who took longer and scored lower than white students, were affected by the complexity of the task types. However, the interaction effect accounted for only 1.2% of the total variance, which led Oltman to conclude that the difference was statistically significant but not so pronounced as to be considered of practical importance.

Taylor, Jamieson, Eignor, and Kirsch (1998) conducted a large-scale study that investigated the effects of computer familiarity on examinees' performance on the CBT version of the TOEFL after providing examinees with computer familiarity training and making adjustments for the P&P TOEFL ability level. A CBT version of the TOEFL was administered at 12 worldwide sites to a sample of TOEFL examinees, which was comparable to the examinee population of the operational TOEFL. The examinees were classified into either "computer familiar" or "computer unfamiliar" groups based on their responses to a computer familiarity scale (Eignor, Taylor, Kirsch, & Jamieson, 1988; Kirsch, Jamieson, Taylor, & Eignor, 1998). Because of the extremely large sample size used, which makes even a small difference in means statistically significant, the authors evaluated Cohen's (1988) practical importance measure as well as results of statistical tests of significance. Results differed depending on how language ability was treated. Before adjustments were made for ability, differences on performance between the familiarity groups were statistically and practically significant. However, after adjustments were made for ability as measured by the P&P TOEFL, only the examinee background (number of times TOEFL taken) significantly interacted with computer familiarity on the TOEFL reading subtest, barely reaching practical significance. The effect of familiarity estimated by an alternative differential item functioning approach was an average of 1.3 point difference on the TOEFL total score. Thus, the researchers concluded that computer familiarity does not play a major role in CBT TOEFL performance.

In terms of examinees' reactions to new forms of testing, Madsen's (1991) study is one of the few studies that provides details based on a self-report questionnaire. Madsen administered an attitude questionnaire

on the CAT version of an ESL placement test of reading, structure and listening at Brigham Young University. He found that although students' reactions to the new test were generally positive, differences in attitudes were observed across language groups. Spanish speakers in his study reported that it was easier to read on the computer screen than in print and that they were interested in and willing to take the CAT in the future. On the other hand, Japanese students' reactions were rather negative. They claimed that it was more difficult to read on the screen and reported anxiety about taking the CAT, even though the Japanese subjects were "more experienced" users of computers than the Spanish-speaking students. Thus, Madsen concluded that experience with computers does not reduce test anxiety, and effects of examinee language background on affect must be investigated more closely.

Comparability of Decisions. Since mode of presentation may also affect decisions made about examinees, comparability of decisions, therefore, must be investigated as part of the effect of mode of presentation as well. Some equating studies for large-scale testing programs have addressed this issue. Segall's (1997) equating study for the ASVAB involved an investigation of differential item functioning across gender and ethnic background of U.S. military applicants, based on equated scores and a calculation of a series of conditional probabilities. The purpose of the study was to see what proportion of female applicants would be affected by selection decisions based on the concurrent use of CAT and P&P forms. Lunz and Bergstrom (1995) also investigated the relationship between cut score and standard errors of IRT ability estimates to calculate how many examinees of the Board of Registry medical technologist certification exam would change their pass/fail status depending on the mode of presentation. It was found that such a small portion of examinees was affected in each case (0.07% and 2.2% in the ASVAB and Board of Registry studies, respectively) that the effect of mode of presentation on the selection and certification decisions based on the tests was not of practical importance.

Examples of such decision-making comparisons can also be seen in language assessment placement testing. Hicks (1986) investigated the comparability of the Multi-level TOEFL (a form of CAT) and the conventional P&P TOEFL. The within-level Pearson correlations between the Multi-level and P&P TOEFL scores after correction for attenuation were high, ranging from .79 to .95, when a strict branching criterion was used. Moreover, placement of examinees into three different levels was highly similar across the two modes. Hicks therefore concluded that the examinees were assigned to their appropriate levels when the items in the Multi-level TOEFL were branched into levels, using P&P TOEFL as the criterion. This also suggests that virtually the same information was obtained by administering the Multi-level TOEFL.

Contrary to Hicks (1986), however, Stevenson and Gross (1991) found that placement decisions were considerably altered for a locally-developed standardized ESL placement test targeted at grade school pupils in the Montgomery County public school district in Maryland. The results of the study showed that the CAT version generally placed the students into higher working levels than the conventional version, while rank ordering of the students was similar across modes. Stevenson and Gross interpreted the observed difference as favorable, attributing the change to the dramatically higher CAT performance of the 6^{th} and 7^{th} graders who were previously disadvantaged by taking a common P&P test, which included items too difficult for all grade levels.

Finally, Fulcher (1999) addressed potential presentation mode effects on placement decisions made for an ESL placement test, which was intended to place candidates into upper intermediate and advanced ESL courses at a UK university. As part of his analysis of the 80-item multiple-choice grammar test given as P&P and Web-based forms, Fulcher utilized an ANCOVA analysis with the P&P score as the covariate. The purpose of the study was to investigate potential biases on CBT performance associated with candidates' computer familiarity, attitudes toward taking tests on the Internet, and background information (age, gender, L1, and field of study). As a result, a significant main effect on the candidates' CBT performance was found only for L1. Meanwhile, separate one-way ANOVA analyses of the P&P and CBT tests with the final placement groups as the independent variable revealed that mean scores of

the final placement groups were significantly different on the CBT, while this was not the case for the P&P form. The above findings indicated that the CBT provides better information for placement decisions, but also that the CBT may place certain L1 groups (East Asian students in this case) into lower levels.

Summary of Assessment Literature

Although the criteria used for assessing the equivalence of test forms across modes seem to be sufficiently standardized with the *Guidelines* (APA, 1986) as the base, the empirical findings as to comparability of conventional and computerized tests are rather mixed. On one hand, the reported stability of parameter estimates and factorial similarity of test forms across modes may suggest that the construct being measured by the tests administered in conventional and computerized forms are comparable. On the other hand, however, the effect of examinees' characteristics, such as computer familiarity, does not seem to manifest itself in test scores. Moreover, linking tests across modes seems to be feasible when sufficient care is taken with regard to the content comparability of tests across modes and interpretation of test linking results. Empirical findings regarding effects of mode of presentation on speeded tests performance, however, are mixed. Since L2 reading tests used as selection, diagnostic, achievement, and placement tests, for example, are often designed as timed power tests, findings related to speeded tests may not be an issue for L2 reading tests. Moreover, inconsistencies can be seen in the impact of computerization of tests on placement decisions based on conventional and computerized tests. The seriousness of such inconsistencies should be evaluated in terms of the stakes of a test in the local context. As Fulcher (1999) argued, potential misplacement of candidates in lower levels may not be a source of great concern in ESL contexts, because such misplacements can often be detected; and necessary arrangements can quickly be made by language instructors.

MODE OF PRESENTATION AND READING

Unfortunately, little empirical investigation on the effects of mode of presentation on reading comprehension has been done in L2 reading research. Yessis (2000) addressed cross-mode L2 reading performance differences in reading rate in an advanced ESL course at a North American university. In his study, 44 undergraduate and graduate students participated in weekly timed and paced reading exercises on paper, while another 9 students performed these exercises on computer. Toward the end of the quarter, the participants read two 1,000-word passages at the 8^{th} grade readability level, one on paper and the other on computer, and answered 10 multiple-choice reading comprehension questions after each passage. A series of mixed model regression analyses showed that when the order of presentation mode and passages were counterbalanced and language ability differences were controlled by entering the participants' ESL placement scores into the equation, the mode differences on comprehension and speed were not significant. Moreover, while the computer practice group read more slowly than the paper practice group on the second occasion, they performed significantly better. Yessis pointed out that the observed performance differences between practice groups might be due to differences in practice conditions. Specifically, the paper practice group followed the pace set by their instructors, while the computer practice group was allowed to set their own pace. This might have led the computer practice group to focus more on the content, as compared to the paper practice group. Based on the participants' responses to a computer attitude questionnaire, Yessis also found that a positive attitude was a significant predictor of better comprehension, but not of reading speed. Finally, a chi-square analysis of pausal protocols of 9 students who participated in a follow-up study showed that frequencies of various reading strategies used by the participants were not significantly different across the modes of presentation.

Although Yessis' (2000) study provides an insight into how L2 reading process may or may not be affected by presentation mode, other empirical L2 reading studies that would allow us to evaluate Yessis' findings are not available in the L2 reading literature.

Accuracy and Speed of Reading in a First Language

An extensive body of literature in ergonomics, education, psychology, and L1 reading has directly compared text information processing of computer-based and paper-based text reading in a first language. The studies to be reviewed here are deemed to have implications for construct validation of CBTs in particular because the computer-based reading items utilized in the studies were not adaptive to participants' reading ability.

Dillon (1992) extensively reviewed ergonomic studies on the effect of mode of presentation (paper vs. computer screen) on reading. He classified numerous studies according to their focus of investigation (outcome or process) as well as factors that potentially accounted for often-reported differences in reading outcome and process across modes. These are outlined in Tables 1 and 2.

Table 1. Factors Previously Investigated in Mode of Presentation Research (Dillon, 1992)

Factors	Definition / Description
Outcome measures	
Reading speed	task completion time
Accuracy of reading	accuracy of proofreading (e.g., identification of spelling mistakes)
Fatigue	visual fatigue and eye strain
Comprehension	level of reading comprehension of texts
Preference	paper vs. computer presentation of texts
Process measures	
Eye movement	frequency and duration of eye fixation
Manipulation	manipulation techniques (e.g., turning pages with fingers; placing a finger as a location aid; flipping through pages while browsing through a document)
Navigation	devices that let the reader know the present location in the document (e.g., table of contents)

Table 2. Factors That Potentially Account for the Differences in Reading Outcome and Process Across Modes (Dillon, 1992)

Factors	Definition / Description
Basic ergonomic factors	
Orientation	orientation of text/screen presentation (e.g., vertical vs. horizontal)
Visual angle	angle created by the length of lines presented on the computer screen and the distance between the screen and the reader's eyes
Aspect ratio	ratio of width to height of computer displays
Dynamics	screen filling style and duration (e.g., rate and direction of text scrolling)
Flicker	frequency of scanning phosphor surface of screen to generate a character that is apparently stable
Image polarity	positive image polarity (dark characters presented on a light background) vs. negative polarity (light characters presented on a dark background)
Display	fonts (e.g., character size, line characteristics spacing, character spacing)
Anti-aliasing	effect of adding various gray levels to individual characters in order to perceptually eliminate the jagged appearance of edges of characters to display sharp continuous characters
User characteristics	degree of user familiarity with computer systems, reading speed, reading strategy and susceptibility to external stress
Interaction of display characteristics	interaction of the above variables

Manipulation facilities	
Scrolling vs. paging	scrolling (the ability to move the text up and down on the screen smoothly by a fixed increment to reveal information currently out of view) vs. paging (the ability to move text up and down in complete screens in a manner similar to turning pages of printed texts)
Display size	number of lines that can be displayed on a computer screen at one time
Text splitting across screens	splitting of paragraphs mid-sentence across successive screens
Window format	single vs. multi-window format (whether two windows can be simultaneously presented to display different parts of a single document)
Search facilities	various means of manipulating and locating information in a document (e.g., word/term searches, checking references, locating relevant sections)
Input devices	tracker ball, mouse, function keyboard, joystick, light pen, etc.
Icon design	facilities that allow rapid and easy manipulations of the text as well as access to the document through numerous routes (e.g., boxes, arrows, circles, buttons, etc.)

The main conclusions of Dillon's literature review can be summarized as follows:

1. It is difficult to draw any firm conclusions from empirical findings based on the studies reviewed due to various concerns, such as the limited scope of the studies, the unique nature of the procedures used, the unclear participant selection criteria, insufficient control of variables of interest, and the use of unrealistic reading tasks (e.g., proofreading for misspelling). However, the literature review suggests that reading from computer screens is, in fact, different from reading in print and that reading computer-presented texts generally takes longer than reading printed materials.
2. The effects of mode of presentation on process measures listed in Table 1 are not yet clear because no adequate empirical method to measure reading processes has yet been established.
3. Differences between modes seem to be caused by interactions of individually non-significant effects, and it is, therefore, impossible to attribute differences to any single factor. Moreover, in a long text that does not fit into one screen and therefore requires scrolling or paging, factors that determine the quality of visual image presented to readers as well as availability and quality of text manipulation facilities listed in Table 2 become important.

One of the limitations of the ergonomics studies reviewed by Dillon (1992) is that many of them studied proofreading rather than reading comprehension, while reading comprehension is more relevant to language assessment. Additional empirical studies utilizing reading comprehension tasks have been primarily conducted in psychology, education, and L1 reading research, nine of which are listed in the appendix. These studies were selected for review here because they included (a) experimental conditions for paper-based and computer-based reading conditions without manipulation or navigation facilities available only on computers (which are feasible for relatively long texts and often beyond the scope of language assessment), and (b) reading comprehension and/or reading speed as dependent variables, which are widely studied as outcome measures of information processing in mode of presentation. Most of these studies were conducted in the 1980s. Due to the advancement of computer technology in the past two decades, use of currently available equipment may yield different results from the studies cited here. However, more recent empirical studies meeting the above selection criteria were not available. The studies, therefore, will be reviewed here to provide a historical perspective on mode effects in reading performance. This will also provide baseline information for considering future research designs. Factors of focus, procedures used, and main findings of the studies are summarized in the appendix.

As shown in the appendix, these studies were conducted in widely different conditions, and the following issues should also be remembered when interpreting their results:

1. *Ergonomic factors.* Various ergonomic issues raised by Dillon (1992), such as display characteristics (e.g., character fonts, size, and line spacing) and features of computer displays (e.g., display size, resolution, image polarity, upper-case only or mixed character use, flicker, and orientation) involved in these studies seem to be varied, but such features were not always reported with sufficient details.
2. *Time limit.* Heppner et al.'s (1985) research is the only study that conducted the experiment under a timed conventional testing condition. None of the other studies set a time limit.
3. *Characteristics of participant.* Four of the studies (Feldmann & Fish, 1988; Reinking, 1988; Reinking & Schreiner, 1985; Zuk, 1986) incorporated grade-school children, while the other studies consisted primarily of traditional age college students or older learners. Moreover, participants' backgrounds as to computer familiarity were mixed in these studies; and descriptions of their language background were not provided.
4. *Characteristics of reading texts and tasks.* Lengths of reading texts ranged from 90-200 words per passage at the shortest (Fish & Feldmann, 1987) to a chapter of an introductory psychology textbook at the longest (McGoldrick et al., 1992). Comprehension tasks involved information search in a textbook chapter (McGoldrick et al., 1992) as well as reading for details and general semantic content in reading passages of conventional lengths often found in reading texts. Moreover, multiple-choice was the preferred item format, although some studies utilized open-ended or short-answer reading comprehension questions (McGoldrick et al., 1992) or form-completion tasks (Feldmann & Fish, 1988; Fish & Feldmann, 1987) as well. Availability of text while answering comprehension questions also differed across studies. Some studies allowed reviewing the text while responding to the questions (Heppner et al., 1985; McGoldrick et al., 1992), while others did not (Belmore, 1985; Reinking, 1988; Reinking & Schreiner, 1985).
5. *Characteristics of experimental designs.* Only two of the studies (Reinking, 1988; Zuk, 1986) counter-balanced the order of text presentation, while the others either presented the texts in the same order or did not report whether the order effect was controlled.
6. *Definition of reading speed.* Belmore (1985) and Reinking (1988) measured time spent on reading assigned texts only. Others included time required to complete the reading comprehension tasks as well (Fish & Feldmann, 1987; McKnight et al., 1990; Zuk, 1986).
7. *Distracters.* The main focus of Zuk's (1986) study was to investigate elementary school children's attention to reading tasks. Thus, a Walt Disney cartoon was played continuously as a distracter while the 3rd and 5th graders worked on the reading tasks in his study.

The findings of these studies are as follows. In terms of the level of reading comprehension, six studies out of the nine reported that comprehension level was similar across the modes (Feldmann & Fish, 1988; Fish & Feldmann, 1987; McGoldrick et al., 1992; McKnight et al., 1990; Reinking, 1988; Zuk, 1986), while one favored paper (Heppner et al., 1985), and two showed interactions -- one with the passage (Belmore, 1985) and the other with the text difficulty and the type of text manipulation (Reinking & Schreiner, 1985). The similarity of reading comprehension level across the modes is consistent with the finding of Dillon's literature review described above. Meanwhile, it is difficult to interpret the results of the two studies that showed interaction effects between mode of presentation and other factors. In Belmore's (1985) study, comprehension of the first set of passages favored print, but the effect disappeared for the second set. As pointed out by Belmore, the fixed order of passage presentation makes it difficult to separate potential order and/or practice effects. In Reinking and Schreiner's (1985) study, 5th and 6th graders scored lower on passages designated to be easier based on standard readability formulas than the other set of passages, which had higher estimates of readability. This may suggest, as pointed out

by the authors, that text characteristics not captured by readability formulae may have affected the text difficulty.

Findings on reading speed in the studies are rather mixed. In three studies (Belmore, 1985; McGoldrick et al., 1992; Zuk, 1986), reading took longer on screen than on paper; three others reported that reading rates were not significantly different across modes (Feldmann & Fish, 1988; Fish & Feldmann, 1987; McKnight et al., 1990); and two studies (Belmore, 1985; Fish & Feldmann, 1987) reported a gain in computer-based reading speed as the experiments proceeded, indicating that after a reasonable amount of exposure to the screen-based reading tasks, the effect of mode on reading rate may diminish.

Although only three of the studies in the appendix that investigated reading speed reported that reading from computer screens was slower than reading from print, quite a few studies, including those reviewed by Dillon (1992), replicated the result favoring print in the effect on reading speed. Some studies attempted to explain why this might be the case.

First, Oborne and Holton (1988) attributed the often-reported differences in reading speed to insufficient control of extraneous variables in previous empirical studies. When they controlled orientation of text presentation, retinal distances from the computer screens, image polarity, and page layout, no significant differences were found in either reading speed or in comprehension scores, regardless of the mode and image polarity. However, the strict control of extraneous variables in this study makes it difficult to generalize the results to real-life reading contexts. For example, it is unlikely that in real life readers would use book stands to vertically present the printed text, or to keep an equal distance between their retinas and the texts across modes, as attempted in this study. It has been widely accepted by ergonomists that computer-presented texts are read at greater distances than conventional paper text (Dillon, 1992; Gould, Alfaro, Finn, Haupt, & Minuto, 1987).

Second, limitations in the research methodology used in previous studies may be another source of the observed reading rate differences across modes. Hansen, Doung, and Whitlock (1978) investigated how subjects in their study spent time while taking a computer science computer-based test. Although these results may not be directly applicable to research on reading performance, the authors' explanations on why their subjects took longer to complete the CBT deserve closer attention. In their study, 7 participants took a computer-based test on introductory computer science. Four of them were videotaped. There were two sources of differences in the time spent by the two groups: (a) computer system requirements, that is, time used to go back to the table of contents to select the next task and time taken by the computer to generate problems and display them; and (b) participants' unfamiliarity with computers. These factors may no longer be relevant, considering the powerful computers available and characteristics of computer users in the 21st century. However, it is worth noting that the 4 participants who were videotaped in Hansen's study expressed discomfort with the testing condition and took significantly longer to finish the test than those who were not videotaped. Moreover, when participants' answers were marked on the screen, they were afraid that their answers would be seen by the proctor. The authors suspected that this might have contributed to the longer work time of the videotaped participants, one of whom reported in the post-hoc questionnaire that "…with PLATO *you are 'broadcasting' your answer to the world*" (Hansen et al., 1978, p. 514). Although the use of videotapes by Hansen et al. provided valuable information as to why the CBT took longer, employment of videotaping must be reconsidered since it might be intrusive for participants; and such discomfort could seriously affect the reliability of the data.

As a possible third explanation, a series of extensive empirical studies that focused on the image quality of text presented on computer screens seem to imply that graphic quality of texts may affect early stages of visual information processing rather than later cognitive information processing; and an improved image quality may, therefore, facilitate reading rate on computers. IBM researchers investigated a wide range of variables, which could account for reading rate differences (Gould, Alfaro, Barnes, Finn, Grischkowsky, & Minuto, 1987; Gould, Alfaro, Finn, et al., 1987). Gould, Alfaro, Barnes, et al. (1987)

investigated the effects of potentially important variables, such as task variables (e.g., paper orientation and visual angle), display variables (e.g., dynamic characteristics of CRT displays, quality of CRT displays, image polarity and fonts), and reader characteristics (e.g., familiarity with computer-based reading and age), on proofreading and reading comprehension performance independently in 10 separate quasi-experimental studies utilizing ANOVA designs. They failed to find any single variable that was strong enough to account for the rather sizable reading rate differences of approximately 25%, which were found in previous research when variables were studied separately. In six experiments, Gould, Alfaro, Finn, et al. (1987) focused on independently or simultaneously manipulating image quality factors, which were selected based on the above experiment results, such as character font and size, polarity, anti-aliasing, page layout, screen resolution and flicker. These authors concluded that the combination of positive image polarity, high display resolution, and use of anti-aliasing seemed to have contributed to eliminating the reading rate differences across the modes, suggesting that the image quality may play a crucial role.

The studies by Gould and his associates share the same concern as those reviewed by Dillon, however. Most of the reading tasks used were very short proofreading tasks looking for misspelling. The extent of relevancy of the proposed combination of image quality variables for reading comprehension tasks was thus obscured. For example, Feldmann and Fish's (1988) study reviewed in the appendix provides counter evidence to those of Gould and his associates. In Feldmann and Fish's study, computer-based reading comprehension tasks were presented only in upper case with negative polarity on a then-commercially available computer display, the quality of which was undesirable, according to Gould and his associates. Even under this condition, a rate and comprehension difference across modes was not found.

Furthermore, a study conducted by Ziefle (1998) challenged the position of Gould and his associates that performance differences may diminish when screen resolution is improved. When the same computer monitor was used across experimental conditions and when variables associated with character sets (size and color of fonts and backgrounds) were strictly controlled, Ziefle found that both proofreading speed and accuracy were still superior in the paper condition. Computer monitors that display text of equal or better quality, as compared with those used in the Ziefle et al.'s experiments, are commercially available already. The mixed results of the above studies seem to suggest, however, that even state-of-the art computer technology, where the use of high resolution monitors with negative polarity and anti-aliasing has quickly become a standard, may not provide the comfort of paper-based reading.

Summary of Mode of Presentation and Reading Literature

The general trends found in these studies indicate 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. Unlike studies conducted in the 1980s, issues such as participants' familiarity with computers and then-current computer system requirements, which made computer presentations of text slow, are quickly becoming less of a concern because of rapid advancements in computer technology. Other explanations that are still pertinent in the 2000s for differential performance between paper-based and computer-based reading proposed in previous studies included insufficient control of extraneous variables, uncomfortable test-taking conditions induced by videotaping during test sessions, and the graphic qualities of text as well as their effects on visual information processing. Although the methodological concerns raised by previous researchers will facilitate the design of future studies, strict control of extraneous variables may limit the generalization of research findings to practical test-taking conditions. Moreover, the mixed results obtained regarding visual explanations suggest that discussion along this line is still inconclusive.

DISCUSSION

Several conceptual and empirical issues raised in the course of the development of mode of presentation research deserve further consideration. Belmore (1985) and Oborne and Holton (1988) explicitly questioned attempts made in previous studies to closely replicate paper-based reading conventions in computer-based conditions. For example, Belmore pointed out that computers are usually introduced in education with the expectation that they would enhance learning and instruction; and computer functions that are not available in text should, therefore, be incorporated whenever existing instructional material is computerized. For example, experiments conducted by Reinking (1988) and Reinking and Schreiner (1985) included two computer conditions with text manipulation facilities, which called up definitions of words, background information, main ideas and easier paraphrases of passages. Fifth and sixth graders who had optional (Reinking, 1988) and mandatory (Reinking, 1988; Reinking & Schreiner, 1985) menus in the experimental conditions performed significantly better on comprehension tests of short expository texts than on those in the paper and computer conditions without the menus, especially when the assigned passages were difficult.

From the perspective of typography research, Gabringer and Osman-Jouchoux (1996) argued that, although the manipulation of text presentation conditions on computers may make the direct comparison of reading processes across modes difficult, the focus of investigation should be on the effect of mode of presentation on reading comprehension when the reading text is optimally presented in paper- and computer-based reading conditions, rather than on when text display variables are strictly controlled across modes. Computer legibility research has not yet confirmed that preferred visual characteristics in printed text generalize to computer-presented text (Gabringer & Osman-Jouchoux, 1996; Muter, 1996). Seeking optimal presentation of reading comprehension tests on computers should, therefore, go hand in hand with the advancement of computer legibility research.

Presenting reading texts and items in an optimal way for the specific mode of presentation suggested above is consistent with the current development in computer-based testing. Computerizing a test as a supplement or a replacement for 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.

Based on the present literature review, issues that should be taken into consideration in designing mode of presentation studies in L2 reading assessment can be summarized as follows:

1. Future comparability studies for L2 reading tests should evaluate equivalence of P&P and computer-based tests from various perspectives. Some issues that should be covered in comparability studies include the possibility of task change, similarity of factor structures, feasibility of test linking results, equivalence of decisions based on the tests, and interactions of examinees' characteristics and testing conditions. Seriousness of discrepancies in test results across modes should be assessed in a local test context.
2. The employment of large sample sizes and the inclusion of the computer presentation mode as a within-subject independent variable in empirical studies would be beneficial. Such designs would allow not only the ANOVA-based approaches employed in many of the non-assessment studies reviewed here for group comparisons but also multivariate analyses (e.g., factor analysis and structural equation modeling) employed in the assessment literature.
3. In order to compare reading comprehension in computerized and conventional tests, data should be collected under conditions that closely resemble operational testing conditions of interest. For example, a testing session has to be timed if it is conducted under the operational testing conditions in question, and computer equipment actually used for operational testing should be employed.

Moreover, ergonomic factors such as those discussed by Dillon (1992) must be described in sufficient detail in research reports in order to facilitate replication and comparison of results across studies.

4. A variety of item types must be included for investigating the mode effect on L2 reading comprehension. Considering the growing interest in performance assessment in educational and language assessment, more research that utilizes performance-based tasks, such as the form-filling tasks used by Feldmann and Fish (1988) and Fish and Feldmann (1987), would be informative.
5. In order to address the effect of mode of presentation on the process rather than on the product of reading, adequate process measures should be devised and included in empirical studies. Methodologies such as analysis of eye movement and verbal protocol analysis as well as post hoc interviews and questionnaires may be useful for this purpose, as has been done in human factors research (Kolers, Ducknicky & Ferguson, 1981) and hypertext research (Dillon, 1996; McKnight et al., 1990) for keeping track of how readers interact with reading materials while processing visual information.

Limitations of the Present Literature Review

The scope of the present survey of literature was limited in two ways. First, this literature review did not investigate issues that are more pertinent to longer texts, which require paging or scrolling. Some previous studies suggest that when a text becomes longer, paper-based reading may facilitate construction of spatial memory about discrete pieces of information included in a reading passage (Matthew, 1997; Piolat, Roissey, & Thunin, 1997). This may be another reason for often-claimed preferences for paper-based reading. This issue must be investigated because presenting long reading passages is common in more advanced L2 reading tests. Second, the present study did not cover empirical studies that involved reading passages accompanied by visual prompts such as figures, graphics, and schematics. If such visual prompts are incorporated into a computerized test, effects of these visuals on examinees' cognitive processing and their potential effects on performance differences across modes should also be addressed. For this line of research, educational media literature may be informative (e.g., Kozma, 1991; Wetzel, Radtke, & Stern, 1994).

CONCLUSION

The present review of literature in cognitive ability as well as language assessment, ergonomics, education, psychology, and L1 reading demonstrates the complexity of the effects of mode of presentation on computerized L2 reading test performance. The literature suggests that effects of mode of presentation on test performance may be observed in a change in the nature of a test task, in a decision based on a test score, in test completion time, and in test takers' affect, for example, whereas the test score itself may not necessarily be influenced. However, the wide range of characteristics of participants, test tasks, test administration conditions, computer requirements, and the degree of control over extraneous variables observed in the studies reviewed in this article, as well as the scarcity of mode of L2 presentation research, make it difficult to draw conclusions based on these studies and to generalize the results to L2 reading assessment.

With the current state of knowledge on the effect of mode of presentation on L2 reading performance, some reading CAT development projects in ESL/EFL are proceeding with caution. For example, Kaya-Carton et al. (1991) and Young et al. (1996) employed an eclectic approach, where reading materials were presented in a printed form as test booklets, while accompanying reading comprehension questions were computer-adaptive and presented on computer screens "in order to minimize method differences and in the interests of test validity" (Young et al., 1996, p. 29). Further mode of presentation research in L2 reading assessment must be continued in order to close the gap between the limitation of empirical data on the effect of mode of presentation on L2 reading performance and the empirical findings on

presentation mode effects on L1 reading performance compiled in other disciplines. This suggests that investigation of the effect of mode of presentation should be an integral part of future construct validation of computerized tests of reading in a second/foreign language.

APPENDIX
Participant and Reading Task Characteristics and Research Designs in Previous Studies

Study	Belmore (1985)	Feldmann & Fish (1988)	Fish & Feldmann (1987)	Heppner, Anderson, Farstrup, & Weiderman (1985)
Participant characteristics				
Sample size	20	Study 1: 93 Study 2: 112 Study 3: 95	Study 1: 36 Study 2: 23	85
Age/ occupation	Undergraduates in an introductory psychology course	Study 1: 4-5th graders Study 2: 8th graders Study 3: 9-12th graders	Age range (Studies 1 & 2) 23-60 years old	University students, staff, and faculty
Computer familiarity	Most were unfamiliar	25 elementary school pupils had some experience; others were familiar	"Similar across groups"	47 regular users; 38 nonusers
Reading task characteristics				
Text length/ number	8 passages (127-221 wds)	Informational texts: 3 passages (150-250 wds) Directional texts: 1 passage in Studies 1-2, 1 set of passages in Study 3	Informational texts: Study 1: 4 passages (90-200 wds) in 2 forms Study 2: 3 passages (160-275 wds) Directional texts: No information given	8 passages (one approx. 600 wds and seven approx. 200 wds)
Text content/level	Varied (e.g., narrative, expository, and persuasive)	Level ranged from grade level to above grade level	Informational texts: Study 1, Mainly philosophy Study 2: Topic not mentioned (but easier than those in Study 1) Directional texts: Business transactions	Not mentioned
Text source	Standardized reading texts	Outdated reading achievement tests	Informational texts: Study 1: GRE materials and a philosophy review book Study 2: ITBS and Davis Reading Test Directional texts: Not mentioned	Old versions of the Nelson-Denny Reading Test (2 forms)
Question format/ number	3-5 MCQs (given on paper)	Informational text: 3-6 MCQs after each text Directional text: Form filling task (32 blanks) All answers recorded on paper answer sheets	Informational text: 13-14 MCQs in total Derivational text: Form filling task (32 blanks), all answers recorded on paper answer sheets	MCQs (8 Qs for the long passage and 4 Qs for the other passages) given on paper
Nature of task	Comprehension (general semantic content)	Comprehension (Informational text: Requires recall or inference making; Directional text: following instructions)	Comprehension	Comprehension

Experimental condition				
Time limit	No	No	No	20 min.
Random assignment	Not mentioned	Yes	Yes	Not mentioned
Availability of text while answering questions	No	Not explicitly mentioned	Not explicitly mentioned	Yes
Counter-balancing of text presentation	No	Not mentioned	Not mentioned	No
Experimental design				
Statistical test	ANOVA	ANCOVA, ANOVA (Covariate: most recent statewide reading achievement test)	Multivariate ANCOVA (Covariates: GRE-Verbal in Study 1; Stanford Test of Academic Skills reading subtest in Study 2)	Mann-Whitney U test
Dependent variables	Reading time Comprehension	Comprehension	Task completion time Comprehension	Comprehension
Independent variables	Mode	Mode Gender	Mode	Mode Age Computer familiarity Reading habit
Mode factor between or within?	Within	Between	Within	Within
Findings				
Reading speed (see Note 1)	Paper > Computer (when computer first)	N/A	Paper = Computer (practice effects observed regardless of the mode)	N/A
Reading comprehension (see Note 2)	Paper > Computer (when computer first)	Generally, Paper = Computer (1 case of Computer > Paper for the Informational text; 2 cases of Female < Male, 1 for each text type)	Paper = Computer	Paper > Computer
Miscellaneous	--	Negative polarity (White x Green); Checklist for subject; background and preference	Negative polarity (White x Black); Upper case letters only; Efficacy ratings collected	Negative polarity; Number of Qs unanswered was greater for the Computer condition; Questionnaire on subjective opinions given
Description of computer/monitor	Apple II Plus	Apple IIe	Study 1: Apple II Study 2: Apple IIe	Apollo "Domain" 15 inch, Black x White screen

Participant and Reading Task Characteristics and Research Designs in Previous Studies (continued)

Study	McGoldrick, Martin, Bergering, & Symons (1992)	McKnight, Richardson, & Dillon (1990)	Reinking (1988)	Reinking & Schreiner (1985)	Zuk (1986)
Participant characteristics					
Sample size	80	16	33	104	55
Age/occupation	Undergraduates in an introductory psychology course	Members of a research center (21-36 years old)	5-6th graders	5-6th graders	3rd & 5th graders
Computer familiarity	Not controlled	Familiar	Familiar	A few have used computers before	Generally unfamiliar
Reading task characteristics					
Text length/number	A book chapter (number of words not available)	1 passage	4 passages (140-180 wds)	6 passages (140-180 wds)	2 passages (390 & 403 wds)
Text content/level	Psychology	Introduction to Wines (a basic guide to the history, production, and appreciation of wine)	Expository passages on various topics; 2 low and 2 high mean difficulty passages based on standardized readability formulas	Expository passages on various topics; 3 low and 3 high mean difficulty passages based on standardized readability formulas	Japanese short stories
Text source	An introductory psychology textbook	A widely distributed hypertext	Reading rate builder kit	Reading rate builder kit	Stimuli extensively used for reading comprehension research
Question format/number	6 factual open-ended Qs (Presented one by one on index cards in random order)	12 partial-credit Qs (format not specified)	6 MCQs for each passage	6 MCQs for each passage	14 MCQs for each passage
Nature of task	Information search	Items that require the use of search facilities to elicit a range of information	Comprehension of textually implicit information	Comprehension of textually implicit information	Comprehension of general semantic content and main ideas
Experimental condition					
True limit	No	No	No	No	No
Random assignment	Yes	Not mentioned	Not mentioned	Yes	N/A
Availability of text while answering questions	Yes	Yes	No	No	Not mentioned
Counter-balancing of text presentation	N/A	N/A	Yes	No	Yes

Experimental design					
Statistical test	MANOVA	ANOVA	ANOVA; regression	ANOVA	MANOVA
Dependent variables	Time (5 measures including time to complete locating an answer) Comprehension Frequency measures for the use of menus	Estimated document size Completion time Time spent in contents/index Comprehension	Reading time Comprehension Passage preference Estimation of learning	Post hoc standard test of reading ability Comprehension Number of attempts made to pass the criterion	Time (3 measures including task completion time) Comprehension
Independent variables	Mode Menu formatting	Presentation format (mode; linear vs. hyper)	Mode and availability of menus Reading ability Text difficulty	Mode and availability of menus Reading ability Text difficulty	Mode Grade enrolled Reading achievement
Mode factor between or within?	Between	Between	Within	Between	Within
Findings					
Reading speed (see Note 1)	Paper > Computer (time to look for answers; time in glossary; efficiency of extracting information)	Paper = Computer	Paper = Computer	N/A	Paper > Computer (total completion time)
Reading comprehension (see Note 2)	Paper = Computer	Paper = Computer	Paper = Computer (Same result obtained also after adjusting score for speed)	Low difficulty passage: Paper > Computer High difficulty passage: Paper = Computer	Paper = Computer
Miscellaneous	Reviewing answers not allowed; Computer with no menu condition not included	Concurrent verbalization required; Subjects were videotaped	Those who were in the conditions with manipulations did better than those who were not. Preference/ efficacy Qs after each passage	Use of manipulation menus significantly increased comprehension	Students were videotaped and monitored Use of cartoons as a distracter Post hoc interview on preferences given
Description of computer/ monitor	IBM microcomputer	Macintosh II screen IBM color screen	Apple II3 Apple color monitor	Apple II Plus Black x White video display monitor	Not specified

Notes

1. Reading speed
   - Paper = Computer (mode difference not significant)
   - Paper > Computer (significantly faster on paper)
   - Paper < Computer (significantly faster on computer screens)
2. Reading comprehension
   - Paper = Computer (mode difference not significant)
   - Paper > Computer (significantly better comprehension on paper)
   - Paper < Computer (significantly better comprehension on computer screens)

ACKNOWLEDGMENTS

The author is grateful to Lyle Bachman, Leslie Winston, Brent Green, and three anonymous reviewers for their detailed comments on earlier versions of this manuscript.

ABOUT THE AUTHOR

Yasuyo Sawaki is currently a student in the Interdepartmental Ph.D. program in Applied Linguistics at the University of California, Los Angeles. Her research interests include L2 reading assessment (English and Japanese) and application of item response theory and generalizability theory for modeling performance assessment ratings.

E-mail: ysawaki@ucla.edu

REFERENCES

Alderson, C. (2000). *Assessing Reading.* Cambridge, UK: Cambridge University Press.

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

American Psychological Association. (1986). *Guidelines for computer-based tests and interpretations.* Washington, DC: Author.

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing, 17*(1), 1-42.

Belmore, S. M. (1985). Reading computer-presented text. *Bulletin of the Psychonomic Society, 23*(1), 12-14.

Bugbee, A. C. (1996). The equivalence of paper-and-pencil and computer-based testing. *Journal of Research on Computing in Education, 28,* 282-299.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Dillon, A. (1992). Reading from paper versus screens: A critical review of the empirical literature. *Ergonomics, 35*(10), 1297-1326.

Dillon, A. (1996). TIMS: A framework for the design of usable electronic text. In H. van Oostendorp & S. de Mul (Eds.), *Cognitive aspects of electronic text processing* (pp. 99-119). Norwood, NJ: Ablex.

Eignor, D. (1999). Selected technical issues in the creation of computer-adaptive tests of second language reading proficiency. In M. Chalhoub-Deville (Ed.), *Issues in computer-adaptive testing of reading proficiency* (pp. 167-181). Cambridge, UK: Cambridge University Press.

Eignor, D., Taylor, C., Kirsch, I., & Jamieson, J. (1988). *Development of a scale for assessing the level of computer familiarity of TOEFL examinees* (TOEFL Research Report 60). Princeton, NJ: Educational Testing Service.

Ercikan, K. (1997). Linking statewide tests to the National Assessment of Educational Progress: Accuracy of combining test results across states. *Applied Measurement in Education, 10,* 145-159.

Feldmann, S. C., & Fish, M. C. (1988). Reading comprehension of elementary, junior high, and high school students on print vs. microcomputer-generated text. *Journal of Educational Computing Research, 4,* 159-166.

Fish, M. C., & Feldmann, S. C. (1987). A comparison of reading comprehension using print and microcomputer presentation. *Journal of Computer-Based Instruction, 14,* 57-61.

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

Gabringer, R. S., & Osman-Jouchoux, R. (1996). Designing screens for learning. In H. van Oostendorp & S. de Mul (Eds.), *Cognitive aspects of electronic text processing* (pp. 181-212). Norwood, NJ: Ablex Publishing Corporation.

Gould, J. D., Alfaro, L., Barnes, V., Finn, R., Grischkowsky, N., & Minuto, A. (1987). Reading is slower from CRT display than from paper: Attempts to isolate a single-variable explanation. *Human Factors, 29,* 269-299.

Gould, J. D., Alfaro, L., Finn, R., Haupt. B., & Minuto, A. (1987). Reading from CRT displays can be as fast as reading from paper. *Human Factors, 26,* 497-517.

Greaud, V., & Green, B. F. (1986). Equivalence of conventional and computer presentation of speed tests. *Applied Psychological Measurement, 10,* 23-34.

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

Green, B., Bock, R. D., Humphreys, L., Linn, R., & Reckase, M. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21,* 347-60.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer-Nijnhoff.

Hansen, W. J., Doung, R. R., & Whitlock, L. R. (1978). Why an examination was slower on-line than on paper. *International Journal of Man-Machine Studies, 10,* 507-519.

Henning, G. H. (1991). Validating and item bank in a computer-assisted or computer-adaptive test: Using Item Response Theory for the process of validating CATs. In P. Dunkel (Ed.), *Computer-assisted language learning and testing: Research issues and practice* (pp. 209-222). New York: Newbury House.

Heppner, F. H., Anderson, J. G. R., Farstrup, A. E., & Weiderman, N. H. (1985). Reading performance on a standardized test is better from print than from computer display. *Journal of Reading, 28,* 321-325.

Hicks, M. (1986). Computerized ESL testing, a rapid screening methodology. In C. W. Stansfield (Ed.), *Technologies in language testing* (pp.79-90). Washington, DC: TESOL.

Kaya-Carton, E., Carton, A., & Dandolini, P. (1991). Developing a computer-adaptive test of French reading proficiency. In P. Dunkel (Ed.), *Computer-assisted language learning and testing: Research issues and practice* (pp. 259-284). New York: Newbury House.

Kirsh, I., Jamieson, J., Taylor, C., & Eignor, D. (1998). *Computer familiarity among TOEFL examinees* (TOEFL Research Report 59). Princeton, NJ: Educational Testing Service.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices.* New York: Springer.

Kolers, P. A., Ducknicky, R. L., & Ferguson, D. (1981). Eye movement measurement of readability of CRT displays. *Human Factors, 23,* 517-527.

Kozma, R. (1991). Learning with media. *Review of Educational Research, 61*(2), 179-211.

Larson, J. W. (1987). Computerized adaptive language testing: A Spanish placement exam. In K. M. Baily, T. L. Dale, & R. T. Clifford (Eds.), *Language testing research* (pp. 1-10). Monterey, CA: Defense Language Institute.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education, 6,* 83-102.

Linn, R. L., & Kiplinger, V. L. (1995). Linking statewide tests to the National Assessment of Educational Progress: Stability of results. *Applied Measurement in Education, 8,* 135-155.

Lunz, M. E., & Bergstrom, B. A. (1995). Equating computerized certification examinations: The Board of Registry series of studies. Paper presented at the annual meeting of the National Council of Measurement in Education, San Francisco, CA, April, 1995. (ERIC Document Reproduction Service No. ED 388 696)

Madsen, H. S. (1991). Computer-adaptive test of listening and reading comprehension: The Brigham Young University approach. In P. Dunkel (Ed.), *Computer-assisted language learning and testing: Research issues and practice* (pp. 237-257). New York: Newbury House.

Matthew, K. (1997). A comparison of the influence of interactive CD-ROM storybooks and traditional print storybooks on reading comprehension. *Journal of Research in Computing in Education, 29*(3), 263-275.

McGoldrick, J. A., Martin, J., Bergering, A. J., & Symons, S. (1992). Locating discrete information in text: Effects of computer presentation and menu formatting. *Journal of Reading Behavior, 14*(1), 1-20.

McKnight, C., Richardson, J., & Dillon, A. (1990). A comparison of linear and hypertext formats in information retrieval. In R. McAleese & C. Green (Eds.), *Hypertext: State of the art* (pp. 10-19). Oxford, UK: Intellect.

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

Moreno, K. E., Wetzel, C. D., McBride, J. R., & Weiss, D. J. (1984). Relationship between corresponding Armed Service Vocational Aptitude Battery (ASVAB) and Computerized Adaptive Testing (CAT) subtests. *Applied Psychological Measurement, 8*(2), 155-163.

Muter, P. (1996). Interface design and optimization of reading of continuous text. In H. van Oostendorp & S. de Mul (Eds.), *Cognitive aspects of electronic text processing* (pp.161-180). Norwood, NJ: Ablex Publishing Corporation.

Neuman, G., & Baydoun, R. (1998). Computerization of paper-and-pencil tests: When are they equivalent? *Applied Psychological Measurement, 22*(1), 71-83.

Oborne, D., & Holton, D. (1988). Reading from screens versus paper: There is no difference. *International Journal of Man-Machine Studies, 28,* 1-9.

Oltman, P. (1994). The effect of complexity of mouse manipulation on performance in computerized testing (ETS-RR-94-22). Princeton, NJ: Educational Testing Service. (ERIC Document Reproduction Service No. 395 023).

Piolat, A., Roussey, J.-Y., & Thunin, O. (1997). Effects of screen presentation on text reading and revising. *International Journal of Human-Computer Studies, 47,* 565-589.

Pommerich, M., Hanson, B. A., Harris, D. J., & Sconing, J. A. (2000). *Issues in creating and reporting concordance results based on equipercentile methods* (ACT Research Report 2000-1). Iowa City, IA: American College Testing.

Reinking, D. (1988). Computer-mediated text and comprehension differences: The role of reading time, reader preference, and estimation of learning. *Reading Research Quarterly, 23,* 484-500.

Reinking, D., & Schreiner, R. (1985). The effects of computer-mediated text on measures of reading comprehension and reading behavior. *Reading Research Quarterly, 20*, 536-552.

Segall, D. O. (1997). Equating CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 219-226). Washington, DC: American Psychological Association.

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.

Staples, J. G., & Luzzo, D. A. (1999). *Measurement comparability of paper-and-pencil and multimedia vocational assessments* (ACT Research Report 99-1). Iowa City, IA: American College Testing.

Steinberg, L., Thissen, D., & Wainer, H. (1990). Validity. In H. Wainer, N. J. Dorans, R. Flaughter, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (pp. 187-232). Hillsdale, NJ: Lawrence Erlbaum Associates.

Stevenson, J., & Gross, S. (1991). Use of a computerized adaptive testing model for ESOL/Bilingual entry/exit decision-making. In P. Dunkel (Ed.), *Computer-assisted language learning and testing: Research issues and practice* (pp. 223-236). New York: Newbury House.

Stone, G. E., & Lunz, M. E. (1994, April). *Item calibration considerations: A comparison of item calibrations on written and computerized adaptive examinations.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA. (ERIC Document Reproduction Service No. [ED] 371 016)

Taylor, C., Jamieson, J., Eignor, D., & Kirsch, I. (1998). *The relationship between computer familiarity and performance on computer-based TOEFL test tasks* (TOEFL Research Report 61). Princeton, NJ: Educational Testing Service.

Van de Vijver, F. J. R., & Harsveld, M. (1994). The incomplete equivalence of the paper-and-pencil and computerized versions of the General Aptitude Test Battery. *Journal of Applied Psychology, 79*, 852-859.

Wainer, H. (1993, Spring). Some practical considerations when converting a linearly administered test to an adaptive format. *Educational Measurement: Issues and Practice, 12*(1), 15-20.

Wainer, H., Dorans, N. J., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (1990). Future challenges. In H. Wainer, N. J. Dorans, R. Flaughter, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (pp. 233-272). Hillsdale, NJ: Lawrence Erlbaum Associates.

Wetzel, C. D., Radtke, P. H., & Stern, H. W. (1994). *Instructional effectiveness of video media.* Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Young, R., Shermis, M. D, Brutten, S. R., & Perkins, K. (1996). From conventional to computer-adaptive testing of ESL reading comprehension. *System, 24*(1), 23-40.

Zeifle, M. (1998). Effects of display resolution on visual performance. *Human Factors, 40*(4), 554-568.

Zuk, D. (1986). The effects of microcomputers on children's attention to reading. *Computers in the Schools, 3*, 39-51.

An investigation into the comparability of two tests of English as a Foreign Language: The Cambridge-TOEFL comparability study Lyle F. Bachman, F. Davidson, K. Ryan, I.-C. Choi. Test taker characteristics and performance: A structural modeling approach Antony John Kunnan.Â An empirical investigation of the componentiality of L2 reading in English for academic purposes Cyril Weir. The equivalence of direct and semi-direct speaking tests Kieran Oâ€™Loughlin. A qualitative approach to the validation of oral language tests Anne Lazaraton.Â 6 Legibility and the rating of second-language writing Annie Brown. 7 Modelling factors affecting oral language test performance: a large-scale empirical study Barry Oâ€™Sullivan. 8 Self-assessment in DIALANG. Computerization of L2 reading tests has been of interest among language assessment researchers for the past 15 years, but few empirical studies have evaluated the equivalence of the construct being measured in computerized and conventional L2 reading tests and the generalizability of computerized reading test results to other reading conditions. In order to address various issues surrounding the effect of mode of presentation on L2 reading test performance, the present study reviews the literature in cognitive ability testing in educational and psychological measurement and the non-assessment Ã˜To what extent is a computerized two-turn speaking situation test task able to elicit L2 learnersâ€™ pragmatic performance? Ã˜What are the distinguishing patterns of L2 learnersâ€™ pragmatic ability at each of the assessed levels? Conversation Management. Construct Definition for Operationalization.Â â€¢ Test-takers listen to and read a short description of a situation. Then, a conversation about the situation is simulated between an interlocutor (a recorded voice) and the test-taker. â€¢ Test-takers have a maximum of 15 seconds to respond for each turn. Two-Turn Speaking Situation: Ite. 7. m Design. A description of a situation (30~60 words, i.e., appr. 4~6 sentences). First turn Interlocutor: _ Test-taker The second paragraph gives examples of what people feel about learning and speaking a second language but it gives us NO INFORMATION about most or few New Zealandersâ€™ belief on teaching children a second language. So, the answer is:NOT GIVEN. Question 30: Chinese is the most common foreign language in New Zealand.Â Skimming will come handy and previous reading of the text can come in use. Therefore, other questions should be done first before answering this question.] Question 39: Which TWO people stopped speaking one language as a child?