



inTASC

TECHNOLOGY AND ASSESSMENT
STUDY COLLABORATIVE

Examining the Effect of Computer- Based Passage Presentation on Reading Test Performance

**Part of the New England Compact
Enhanced Assessment Project**

Jennifer Higgins, Michael Russell, & Thomas Hoffmann
Technology and Assessment Study Collaborative
Boston College
332 Champion Hall
Chestnut Hill, MA 02467

www.intasc.org

Examining the Effect of Computer-Based Passage Presentation on Reading Test Performance

Jennifer Higgins, Michael Russell & Thomas Hoffmann
Technology and Assessment Study Collaborative
Boston College
Released October 2004

This study has been funded by the New England Compact Enhanced Assessment Project through US Department of Education Grant #S368A030014.

The New England Compact (Maine, New Hampshire, Rhode Island, Vermont) provides a forum for the states to explore ideas and state-specific strategies, build and expand a collective knowledge base, and initiate cross-state collaborative activities that benefit each state with economies of scale and cost-efficiency. A primary focus for the New England Compact is the implementation of the No Child Left Behind (NCLB) legislation. Current Compact projects include activities initiated by the Commissioners of Education, Deputy Commissioners, State Assessment Directors, and Title III and Special Education Directors and participation in an Enhanced Assessment Grant funded by the Federal Department of Education.
(www.necomact.org)

Copyright © 2004 Technology and Assessment Study Collaborative, Boston College



Examining the Effect of Computer-Based Passage Presentation on Reading Test Performance

Jennifer Higgins, Michael Russell, and Thomas Hoffmann
Technology and Assessment Study Collaborative
Boston College

Released October 2004

Background/Purpose

The No Child Left Behind legislation requires that all students in grades 3–8 be tested each year in English Language Arts and mathematics. For many states, complying with this new law requires a significant increase in the amount of student testing conducted each year. This proliferation of testing without an associated increase in funding to pay for the new tests has caused many state assessment divisions to examine new ways to make the assessment process more efficient. One alternative that several states have considered is computer-based delivery of assessments. A 2003 Education Week reported that 12 states and the District of Columbia were planning to administer some form of computer-based state exams in 2002–2003 (Editors, 2003). Although computer-based test administration has the potential to save time and money while also returning results in a more timely manner, research must be undertaken to understand how changes in testing mode affect student performance.

To date, a substantial body of research has examined the comparability of scores provided by paper-based and computer-based tests (Bunderson, Inouye, & Olsen, 1989; Bangert-Downs 1993). The majority of this research, however, has focused on young adults and adult populations. In addition, few existing studies have focused specifically on reading comprehension tests, particularly for young readers.

When reading comprehension tests that require examinees to read extended passages are presented on a computer, test developers are faced with an important design decision: how to allow students to navigate through a passage. Currently, one of two options are typically used: a) the passage is broken into subsections and

each subsection is presented as a single block of text; or b) the passage is presented as a single block of text through which the examinee must scroll. While students who are accustomed to using the Internet or reading text passages on a computer may find scrolling through a reading passage to be a natural process, those students who are less familiar with technology may find scrolling to be a relatively novel experience and in turn be negatively affected by this format. If the students' ability to scroll through a reading passage does affect their performance on a reading comprehension assessment, the computer-based version of this assessment introduces a new construct, namely computer literacy. If computer literacy is not part of the construct being measured by a reading comprehension test, then this "ancillary non-construct element" (Haertel & Wiley, 2003, 1) should be eliminated or reduced as much as possible from the reading comprehension assessment. In other words, the mode of assessment administration should not unfairly bias the scores of students based on their computer literacy skills. As Pommerich and Burden (2000) highlight, "because of subtle differences in test presentation across administration modes, care must be taken to ensure that examinees respond to item content only" (p. 1).

Research on the impact of transitioning passage-based tests from paper to computers has been conducted for almost two decades. In 1986 Haas and Hayes found that students taking tests on computer where the reading passages were more than one page in length achieved lower test scores. Mazzeo and Harvey (1988) also found that passage-based tests administered on computer may be more difficult than paper based tests. This increased difficulty may be due to students being uncomfortable with navigating through a reading passage on computer. However, as the student to computer ratio decreases in schools and students are given more access to computers outside of school, some experts predict that the modal effect will become less of an issue. As an example, Clariana and Wallace (2002, p. 599) write, "As students become as familiar with computer-based testing as they are with paper-based testing, the test mode effect should decrease or disappear." Until that time, however, it is important for test designers to develop a better understanding of the extent to which modal effects exist for different populations and to develop strategies that reduce the effect for these populations.

To this end, in 1998 and 2000 Pommerich conducted research on English, Reading, and Science passage-based tests using more than 20,000 11th and 12th graders (Pommerich, 2004). In the 2000 administration, students were randomly assigned to take the test on paper, on computer using scrolling text, or on the computer using paging text. Although students who took the English and Science test on computer scored significantly higher than students who took these tests on paper, there were no significant differences in scores across modes for the Reading test. Pommerich also notes that while significant differences were found in the English and Science tests, adjustments for multiple comparisons were not made and "the effect sizes for the mean differences observed across both studies were no larger than +/- 0.15 in any content area, which is considered small by Cohen's

(1988) standard” (Pommerich, 2004, 20). At an item level, Pommerich found some significant differences and concluded that “examinees responded differently to some items under different (presentation) variations” (p. 37).

In 2002, Choi and Tinkler presented findings from a comparability study that examined differences in performance for 3rd and 10th grade students on the Oregon state math and reading tests administered on paper and computer. Using a counterbalanced design, all students were administered half of the items in each subject area on paper and half of the items on computer. The largest difference in test scores was found on the 3rd grade reading test. The authors concluded that “scrolling reading passages on computer screens seems to interfere with students’ test-taking behavior, especially the test-taking behaviors of younger students. Providing page-up and page-down buttons in place of a vertical scroll bar may alleviate the interference that might have affected younger students” (Choi & Tinkler, 2002, 10).

The purpose of the study presented here is to examine differences in performance when two different computer-based test formats and a traditional paper-and-pencil based format were used to present reading passages to 4th grade students. Specifically, the research questions are:

1. When 4th grade reading comprehension assessments are transitioned to computers, what is the impact on students’ test scores?
2. Are there differences in test scores when reading passages are presented on the computer using scrolling text versus paging text?
3. How is student performance (modal effect) related to their computer skills, computer literacy, and prior computer use?

Results from this research will provide evidence of whether computer presentation, with scrolling or whole page text format, interferes with measuring reading comprehension skills of fourth grade students. This research is federally funded through the New England Compact Enhanced Assessment Program and conducted collaboratively with Vermont, Rhode Island, New Hampshire, Maine, the Education Development Center (EDC), and the Center for Applied Special Technology (CAST).

Design

To examine whether and how the presentation of reading passages affected the test performance of 4th grade students, 219 students from eight Vermont schools were randomly assigned to perform the same reading test in one of three modes: 1) on paper, 2) on computer with scrolling reading passages; and 3) on computer with passages divided into sections that were presented as “whole pages” of text. The participating schools were selected with the cooperation of the state Director of Assessment. When selecting schools, we aimed to acquire a mix of rural, suburban, and urban schools and a mix of students that included English Language Learners (ELL), students identified with special needs in the area of language/reading, as well as non-special education native English speaking students. Since ELL students tended to be clustered within schools, and since the location of the school could not be manipulated, random assignment occurred within rather than across schools. Through this three group randomized design, the research compares the effect of using paper and pencil versus text presented using scrolling or whole page format on student test performance.

To control for effects that might result from differences in the computers available within each school, the research team brought into each school a set of Macintosh 14-inch iBooks with traditional hand-held mice. All students performing the reading test on computer used one of the research team’s laptop computers. In addition to performing the same four-passage reading test, all students also completed a computer fluidity test, a computer literacy test, and a computer use survey. The computer fluidity test was administered to all students on a laptop. The computer literacy and the computer use surveys were administered to all students on paper. The purpose of administering these three additional instruments was to collect multiple measures of students’ computer skills, knowledge, and use so that we could examine the extent to which any modal affects are related to differences in students’ ability or familiarity with using a computer – constructs that are not intended to be measured by the reading comprehension test.

Data Collection

Data was collected from students in the eight Vermont schools between January 20th and January 30th, 2004. Within each school, researchers first configured the classroom so that desks were spread out and the laptops could be connected to a power source. As students entered the room, they were asked to find their place by looking for their name card on desks, which were set up with either the paper and pencil assessment or with the launched assessment application on a laptop. Researchers then explained the purpose of the research and briefly described the reading comprehension assessment, fluidity exercises, computer literacy test, and survey to the students. Students were given one hour to complete the reading comprehension assessment and an additional hour to complete the computer fluency, literacy, and use instruments.

The number of 4th grade students per school ranged from 13 to 45. In one school, the parents of six students requested through an informed consent letter that their children be excluded from the research. Additionally, two students in another school did not cooperate in taking the assessment. The scores of all remaining 219 students are used in the analysis section of this report.

Table 1 summarizes demographic information about students participating in the study. Specifically, Table 1 indicates that 99 students (45%) attended rural schools, 83 (38%) attended suburban schools, and 37 (17%) attended urban schools. In addition, 54 students (25%) reported that they spoke a language other than English at home.

Table 1: Demographic Summary for Participating Students

Geography	Number of Students
Rural	99
Suburban	83
Urban	37
Multilingual Status	
Multilingual	54
Non multilingual	165

Instruments

Students participating in this study completed four data collection instruments in the following order: 1) Reading Comprehension Test; 2) Computer Fluidity Test; 3) Computer Literacy Test; and 4) Computer Use Survey. Below, each instrument is described in greater detail. In addition, we explain in detail how scale scores were developed for the fluidity, literacy, and use instruments and present information on the reliability and validity of each scale.

Reading Comprehension Test

One test form, containing 4 reading passages and 19 associated multiple choice items was administered to all students participating in this study. Two passages and nine multiple choice items were taken directly from released National Assessment of Educational Progress (NAEP) tests, one passage and five items were taken directly from a released Progress in International Reading Literacy Study (PIRLS) test, and one passage and five items were from a released New Hampshire test. Passages and items were chosen based on their grade level appropriateness and the item characteristics. Specifically, for the vast majority of items, the item difficulty was such that between 40% and 60% of students had answered the item correctly during the item's original administration. Given that these items were of moderate difficulty, we anticipated that they would be most sensitive to modal effects.

Both computer-based assessments began by requiring students to enter an identification number, which was pre-assigned to students based on their mode of delivery. The program then led the students through a 2 minute and 20 second passive tutorial. The tutorial first showed students where to locate the reading passages and items. Then, students were shown how to select an answer to an item and how to navigate through the items. Students were allowed to skip items and change answers to items. Next, a feature called "mark for review" was explained to students. This feature allowed students to indicate whether they want to return to a specific item to work on it at a later time. Students had the ability to answer an item and mark it for review, not answer the item and mark it for review, or not answer the item and not mark it for review. After marking an item for review, the box surrounding the item turned yellow to remind the student that they had marked it. Students had the ability to move independently through the test, both between items and passages. During the tutorial, students were also shown how to highlight and un-highlight sections of text in the passages. This feature allowed students to emphasize portions of the reading text by selecting the text and pressing the highlight button. Lastly, students were shown how to navigate through the reading passages and how to end the test. Depending on whether the student was assigned to the scrolling or whole page format, the tutorial would show students how to navigate using the scroll bar or the "next" and "back" buttons. Figure 1 displays a screen shot of the test interface for the scrolling passage and the whole page formats.

Figure 1: Computer-Based Reading Test Interfaces

Scrolling

The screenshot shows a reading passage titled "Penny's Scrapbook" by Rebecca A. Alter. The passage is divided into sections: an introductory paragraph, "Penny's Scrapbook", "My First Months", and "Back to School". The interface includes a vertical scrollbar on the right side of the passage. To the right of the passage, there are two questions with radio button options. At the bottom of the passage area, there are "highlight text" and "remove highlight" buttons. At the bottom of the question area, there is a "more questions" button and a progress indicator with numbers 1 through 19 and an "End Test" button.

Whole Page

The screenshot shows the same reading passage and questions as the scrolling view, but with a "Back" button on the left and a "Next" button on the right at the bottom of the passage area. The "highlight text" and "remove highlight" buttons are also present. The question area includes the "more questions" button and the progress indicator with numbers 1 through 19 and an "End Test" button.

The paper-based assessment had a cover sheet, where students were presented with brief instructions and were asked to write their name and identification number. The passages and multiple-choice items were presented on double sided pages and the number of words on each line, number of lines per page, and text wrap were identical to the whole page computer based format. The font size and style were nearly identical across all three test forms in order to minimize differences that result from minor changes in the presentation of passages and items. As mentioned earlier, all students who performed the computer versions of the test used the same type of 14-inch screen laptop computer so that we could control for differences in the size of text that may result from different screen sizes and/or resolution.

Computer Fluidity Test

After completing the reading comprehension assessment, all students were asked to complete a computer fluidity test that consisted of four sets of exercises. The purpose of the computer fluidity test was to measure students' ability to use the mouse and keyboard to perform operations similar to those they might perform on a test administered on a computer. In this report, we refer to these basic mouse and keyboard manipulation skills as "computer fluidity."

The first exercise focused on students' keyboarding skills. For this exercise, students were allowed two minutes to keyboard words from a given passage. The passage was presented on the left side of the screen and students were required to type the passage into a blank text box located on the right side of the screen. The total number of characters that the student typed in the two-minute time frame was recorded. A few students typed words other than those from the given passage. These students' were excluded from analyses that focused on the relationship between reading comprehension test performance and computer fluidity.

After completing the keyboarding exercise, students performed a set of three items designed to measure students' ability to use the mouse to click on a specific object. For these items, students were asked to click on hot air balloons that were moving across the computer screen. The first item contained two large hot air balloons. The second item contained five medium-sized hot air balloons that were moving at a faster rate. The third item contained 10 small hot air balloons that were moving at an even faster rate. In each set of hot air balloons, the amount of time and the number of times the mouse button was clicked while clearing the screen were recorded.

The third computer fluidity exercise measured students' ability to use the mouse to move objects on the screen. For this exercise, students were presented with three items each of which asked students to drag objects from the left hand side of the screen to a target on the right hand side of the screen. For the first item, students were asked to drag books into a backpack. The second item asked students to drag birds into a nest. The third item asked students to drag ladybugs onto a leaf. As students advanced through the drag and drop levels, the size of both

the objects and the targets decreased, making the tasks progressively more difficult. Similar to the clicking exercise, for each item the amount of time and the number of times the mouse was clicked were recorded.

Finally, the fourth exercise was designed to measure how well students were able to use the keyboard's arrow keys to navigate on the screen. For this exercise, students were asked to move a ball through a maze by using the arrow keys. Students were shown where on the keyboard to find the arrow keys. The first half of the maze consisted of 90-degree turns and the second half contained turns with curves. The time required to reach the end of the maze were recorded. As described in the analysis section, the data from the keyboarding, clicking, drag and drop, and arrow key exercises was combined into a single scale to produce a computer fluidity score for each student.

Computer Literacy Test

After finishing the computer fluidity exercises, students were asked to complete a short paper-based computer literacy test. The purpose of this test was to measure students' familiarity with computing terms and functionality. Virginia and North Carolina have administered multiple-choice computer literacy tests to students at lower grade levels. Eleven released multiple-choice items from previously administered VA and NC assessments were used in the computer literacy test as part of this research. Items were chosen based on their alignment with the International Society for Technology in Education standards.

Computer Use Survey

Lastly, students were asked to complete a paper-based survey. This survey was adapted from the grade five student survey constructed for the Use, Support, and Evaluation of Instructional Technology (USEIT) study (see Russell, Bebell, and O'Dwyer, 2003). Students were asked questions focusing on their specific uses of technology in school and at home, their comfort level with technology, as well as some demographic information. Students who took the assessment on laptops were asked four additional open ended questions that focused on whether they believed that taking the reading test on computer was easier or more difficult than taking it on paper and pencil, whether they had any problems while taking the test on the computer, and whether they used the highlight or mark for review features.

Scale Development

As described above, three instruments were administered to students in order to measure their computer fluidity, literacy, and use. Each of these instruments was developed specifically for this study. While items that comprised the literacy and use instruments were taken directly from instruments that have been used in previous research and/or state test administrations, the specific set of items that comprise each instrument had not previously been used in tact. In addition, the items that formed the computer fluidity test were developed by the research team and had not previously been administered to a large number of students. Thus, before information from these three instruments could be used for analytic purposes, scale scores had to be developed and the reliability of these scales was examined. To this end, two sets of analyses were conducted to create and then examine the reliability of these scales. First, principal component analyses were performed on each instrument to examine the extent to which the items could be grouped to form a single score. In cases where all items could not be combined to form a single scale, principal component analyses were used to identify a subset of items that formed a unidimensional scale. Scale scores were then created for each student. Second, Cronbach's alpha was calculated for each scale to examine the reliability of the scale. In cases where the scale had unacceptably low reliability (below .60), item to total score correlations were examined to identify items that were contributing to low reliability. These items were then dropped from the scale, new scale scores were created, and the reliability analysis was repeated. Below, we describe the specifics for each scale creation.

Computer Fluidity Scale

The computer fluidity test consisted of four sets of tasks. As described in the instrument section, the four tasks included: keyboarding, clicking, drag and drop, and navigating with the arrow keys. The keyboarding and arrow key tasks consisted of a single item and the only data recorded pertained to the amount of time required to complete each item. The two other tasks each consisted of three items. For each item, two pieces of information were collected: a) the amount of time required to complete the item, and b) the number of mouse clicks required to complete the item. All of this data was initially employed to compute a scale score. However, through an initial principal component analyses, it was clear that some of this information was not useful. To maximize the variance retained by combining information from each of the computer fluidity items into a single score, a two-step procedure was used. First, information provided by each item was weighted and combined into an item score. Second, the item scores were weighted and combined into a single fluidity test score. Specifically, a single variable was created for the clicking exercise by combining the time required in seconds to complete the exercise with the number of clicks used to complete the exercise. The final clicking variable equaled the click-time (in seconds) + (number of clicks)*0.9. This formula penalized students for each extra click that they needed to complete

the exercise. The number of times that students dragged objects to complete the drag exercise was found to not positively contribute to the fluidity factor score and therefore only the drag time was used in the final factor solution. The raw arrow-key time was weighted by 0.6 and the reversed keyboarding speed was weighted by 0.3 in the final factor solution. When a factor analysis of the new click variable, the drag times, the weighted arrow time, and weighted reversed keyboarding times was conducted, a one factor solution which accounts for 45.1% of the total variance with an alpha reliability of 0.76 was achieved. Factor loadings of each variable used range from 0.46 – 0.78.

Computer Literacy Scale

The computer literacy test consisted of 11 multiple-choice tests that asked students about specific aspects of a computer. These aspects included terminology, software, hardware, and tasks typically performed with a computer. When a principal components analysis was run on the 11 items, a three-factor solution emerged. The factor that accounted for the most variance consisted of five items, whose content was based on word processing skills and computer terminology. When a principal component factor analysis was run on these five items, a one-factor solution that accounted for 38.4% of the variance and had an alpha reliability coefficient of 0.60 was achieved. Factor loadings on the five items ranged from 0.57 – 0.74. This one factor solution was used to create scaled scores of students' computer literacy.

Home Computer Use Scale

To measure the extent to which students used a computer at home, a series of questions on the student computer use survey asked how frequently they use computers at home to play games, chat/instant message, email, search the Internet for school, search the Internet for fun, play mp3s/music, write papers for school, and/or create/edit digital photos or movies. Students were asked to choose one of the following responses for each activity: never, about once a month, about once a week, a couple of times a week, and every day.

When a principal components factor analysis was run on the eight home computer use items, a two-factor solution emerged. Only one item was correlated more strongly with the second factor than the first. This item asked students how frequently they use a computer outside of school to write papers for school. Since this item focused on use of home computers for a purpose related specifically to school and was not something that most fourth grade teachers are currently likely to ask students to do, this item was removed from the scale analyses. A principal components factor analysis was then run on the remaining seven home computer use items, yielding a one factor solution that accounted for 49.7% of the variance and had an alpha reliability of 0.83. Factor loadings on the seven items ranged from 0.53 – 0.80.

School Computer Use Scale

To measure the extent to which students use computers in school, a series of questions on the student computer use survey asked how frequently they use computers in school to email, write first drafts, edit papers, find information on the Internet, create a Hyperstudio or Powerpoint presentation, play games, and/or solve problems. Students were asked to choose one of the following responses for each activity: never, about once a month, about once a week, a couple of times a week, and every day. When a principal components factor analysis was run on the seven school computer use items, a two-factor solution emerged. One factor contained four items, which focus on writing, multimedia, and research. A principal components factor analysis was then run on these four school computer use items, yielding a one factor solution that accounted for 50.8% of the variance and had an alpha reliability of 0.68. Factor loadings on the four items ranged from 0.68 – 0.74.

The four factor analyses and creation of scale scores allows for the analysis of reading comprehension assessment scores based on four measures of students' computer skills, knowledge, and use. There is a positive correlation of 0.43 between the fluidity scale score and the computer literacy scale score. There is also a positive correlation of 0.31 between the home-use scale score and the school-use scale score. These statistically significant correlations provide further evidence of the validity of the computer measures.

Beta Test of Computer-Based Reading Test

Prior to using the computer-based versions of the reading test for actual data collection purposes, a beta study was conducted in a local Massachusetts 4th grade classroom to identify any assessment design and operational issues that interfere with student performance on the computer-based test. We describe the beta study here to help the reader better understand why we used a passive rather than an interactive tutorial system and to provide insight into design issues for other test developers and state testing agencies.

For this beta study, ten 4th grade students were administered the reading comprehension assessment and computer fluidity exercises on the laptop computers. Students first were given an interactive tutorial, which explained features of the computer-based test and required students to complete tasks. The interactive tutorial was designed to both explain the functionality and provide students with an opportunity to practice using the test features. Students were then given 30 minutes to read the passages and answer the multiple-choice items. Finally, students were asked to complete the computer fluidity exercises.

This beta study indicated that the basic functionality of the assessment worked well with 4th grade students, but revealed several key design issues with the tutorial and the fluidity exercises. When students sat down at the tables with the laptops,

the sense of excitement in using the laptops interfered with students' concentration and thoroughness in understanding the tutorial. Most students sped through the tutorial, without reading and understanding the assessment features. As evidence of this, several students marked all of the items for review. When they were asked why they marked the items for review, they said that they thought they had to use this feature on every item. Similarly, students did not thoroughly read the instructions to the computer fluidity exercises before starting the tasks. In essence, by providing students with control of the tutorial, many students appeared to explore the interface by applying their intuition rather than following the instructions. While this worked for most aspects of the interface, there were some features that required direct instruction. As a result, some students did not fully understand these features and how to use them while performing the actual test.

Lessons learned from this pilot study were used to make adjustments to the tutorial and fluidity instructions. The tutorial was changed from being interactive to being passive. Students were told that the computer would show them how to take the assessment and that they did not need to press any buttons during the tutorial. As explained in the instrumentation section, the passive tutorial showed students how to select answers, read the passage, highlight text, and mark answers for review without requiring them to practice the tasks. In both the tutorial and the fluidity exercises, descriptions of the tasks were shortened to encourage students to read and understand them. Before each fluidity exercise, the shortened description of the task was presented to the student and the program allowed for a 15 second delay before the "start" button appeared. Again, this delay encouraged students to read the instructions before starting the exercise.

After completing the assessment and fluidity exercises, students were asked if they had any problems and if they liked taking the assessment on the laptops. Students expressed interest and excitement in using the laptop computers and indicated that they had not encountered any problems. Changes made to the computer-based design were further tested by the research team prior to the assessment administration.

Results

To examine the extent to which the mode of testing and the way in which extended text passages presented on a computer affect student performance, a series of analyses were performed. These analyses include a comparison of completion rates and a comparison of mean percent of correct responses across the three modes. To examine the extent to which prior experience and skill using a computer interacted with the presentation mode, analyses were conducted using the computer fluidity, literacy, and use measures. For each of these measures, students were divided into three groups representing high, medium, and low levels of fluidity, literacy, or use. The modal effect was then examined within each group by comparing performance across the three presentation modes. In addition, regression analyses were conducted to examine the modal effect after adjusting for differences in prior fluidity, literacy, and use. Finally, to examine the extent to which the modal effect differed between boys and girls, a gender analysis was conducted. Findings from each of these analyses are presented below.

Completion Rate

Of the 219 participating students, nine students did not reach the last item on the test in the given one hour timeframe. Three of the students who did not reach the last item were administered the assessment on paper, three were administered the assessment on laptop with scrolling text, and three were administered the assessment on laptop with whole page text. Some students also skipped items while taking the test. A total of 194 students (88.6%) answered all 19 items. Of the 25 students who left at least one item unanswered, 14 students were administered the assessment on paper, 5 students were administered the assessment on laptop using scrolling text, and 6 students were administered the assessment on laptop using whole page text. Three of the paper students who left items unanswered did not answer items that were presented on the back of at least one page. It also is noteworthy that students who took the assessment on laptops were prompted at the end of the assessment with a summary of the items that they left unanswered. This automatic prompt may have led to a lower number of students leaving items unanswered.

Reading Comprehension Percent Correct

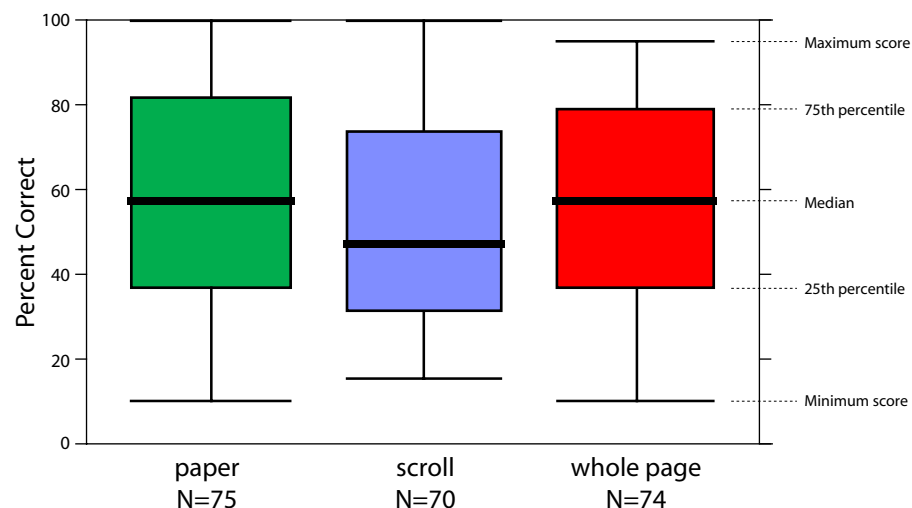
The average student score on the reading comprehension test was 55.8%. The lowest score was 10.5%, the highest score was 100%, and the standard deviation of scores was 23.4. If one of the testing modes unfairly advantaged students, we would expect a significant difference in scores across modes of delivery. As seen in Table 2, there is nearly a six point difference in the mean score between the paper and scrolling groups. A one-way Analysis of Variance (ANOVA), however, indicates that the differences among groups were not statistically significant ($p = .278$).

Table 2: Comparison of Scores Among the Three Presentation Modes

Group	N	Mean	St. Dev.	F	DF	p
Paper	75	58.1%	.243	1.289	218	.278
Whole Page	74	56.9%	.227			
Scrolling	70	52.2%	.230			

The box-plot in Figure 2 displays the differences in test scores across the three presentation modes. Although there is a visible difference between the median for the scrolling condition and the two other presentation modes, this score difference is relatively small compared to the standard deviation for each presentation mode¹.

Figure 2: Box-plot Depicting Score Differences Among the Three Presentation Modes



To examine whether the presentation mode may have had a significant impact on the performance for specific test items, the percent of students within each group responding correctly to each item was compared across the three presentation modes. As seen in Table 3, there is only one item (#5) for which performance differed significantly ($p < .05$) across the three presentation modes. Two additional items (#2 and 18) show significant differences at the $p < .10$ level. In all three cases, scores were similar between the paper and whole page condition, but were noticeably lower for the scrolling condition.

Table 3: Comparison of Item Level Performance Across the Three Presentation Modes

Item	Paper	Scroll	Whole Page	F	p
1	51%	53%	53%	0.044	.957
2	83%	67%	76%	2.370	.096
3	69%	54%	62%	1.746	.177
4	57%	56%	68%	1.259	.286
5	64%	41%	64%	4.990	.008
6	68%	59%	62%	0.703	.496
7	63%	49%	50%	1.791	.169
8	61%	53%	61%	0.659	.518
9	80%	79%	73%	0.575	.564
10	60%	64%	66%	0.322	.725
11	53%	59%	47%	0.916	.402
12	43%	47%	47%	0.204	.816
13	75%	64%	66%	1.034	.357
14	39%	37%	51%	1.829	.163
15	44%	49%	47%	0.162	.851
16	39%	31%	34%	0.435	.648
17	49%	51%	55%	0.280	.756
18	53%	37%	53%	2.429	.091
19	53%	41%	46%	1.052	.351

After reviewing these three items, we could not detect anything unusual about these items that might lead to differences in performance across the three modes. It is also important to note that since multiple comparisons were made across these items, one would expect to find significant difference at the .05 level one out of twenty times and two out of twenty times at the .10 level by chance alone.

As an additional measure of differences in item level scores, an effect size statistic (Glass' delta) was computed for both paper/scroll and paper/whole page for each item. Four items (#2, #3, #5, #18) have effect sizes greater than 0.3, which is

considered by Cohen (1988) to be in the medium effect size range. Again, there does not appear to be anything unusual about these items that may lead to these score differences.

Examining Score Differences by Prior Computer Skills and Use

To examine the relationship between students' performance on the reading test and their computer fluidity, computer literacy, and prior computer use, we divided students into three groups based on their scale scores. This grouping was conducted separately for the fluidity, literacy, home use, and school use scales. For each scale, the three groups represented high, medium, and low level fluidity, literacy, or use. For each scale, students whose scores were .5 standard deviations above the mean scale score were placed into the high group. Students whose scale score was within .5 standard deviations of the mean scale score were placed into the medium group. And students whose scale scores were .5 standard deviations below the mean were placed into the low group. After these three groups were formed, reading comprehension scores were compared across the three presentation modes based on the high-medium-low classification of each computer measure.

Although students were randomly assigned to the three mode of presentation groups, differences in students' prior computer experiences may still exist. If so, these differences may mask differences that result due to the presentation mode. To examine differences in prior computer experience, mean scores for the fluidity, literacy, home use, and school use scales were compared among the three mode of presentation groups. As seen in Table 4, computer fluidity scores were higher for the whole page group and lowest for the paper group. A similar pattern occurred with computer literacy, such that the whole page group had higher literacy scores (on average) and the paper group had the lowest scores. With respect to the home use scale, the whole page group had the lowest scores while the scrolling group had the highest scores. Finally, for school use, the scrolling group again had the highest school use scale scores while the paper group had the lowest scale scores. Despite these visible differences, computer fluidity is the only measure for which score differences are statistically significant at the $p < .05$ level.

Table 4: Comparison of Scale Scores Across Presentation Modes

Scale	Paper Mean	Whole Page Mean	Scroll Mean	Pooled SD	F	Sig.
Fluidity	-0.19	0.32	-0.12	1.00	5.195	.006
Comp. Literacy (%)	50.3	57.5	54.1	22.9	1.780	.171
Comp Literacy Scale Score	-0.22	0.16	0.06	1.00	3.014	.051
Home Use	0.01	-0.06	0.06	1.00	0.199	.820
School Use	-0.13	-0.09	0.24	1.00	2.340	.099

To control for differences among the three groups for each of these measures, regression analyses were performed. For each regression analysis, the scale score was entered as an independent variable and two dummy variables were entered to estimate the effect of group membership on test scores controlling for differences in scale scores. Specifically, the first dummy variable represented the whole page group (i.e., students in the whole page group were assigned a 1 and students in the two other groups were assigned a 0). The second dummy variable represented the scrolling group (i.e., students in the scrolling group were assigned a 1 and students in the two other groups were assigned a 0). Students reading test score was the dependent variable in all the regression analyses.

The full set of analyses for the computer fluidity, computer literacy, home use, and school use scales are presented separately below.

Computer Fluidity Analysis

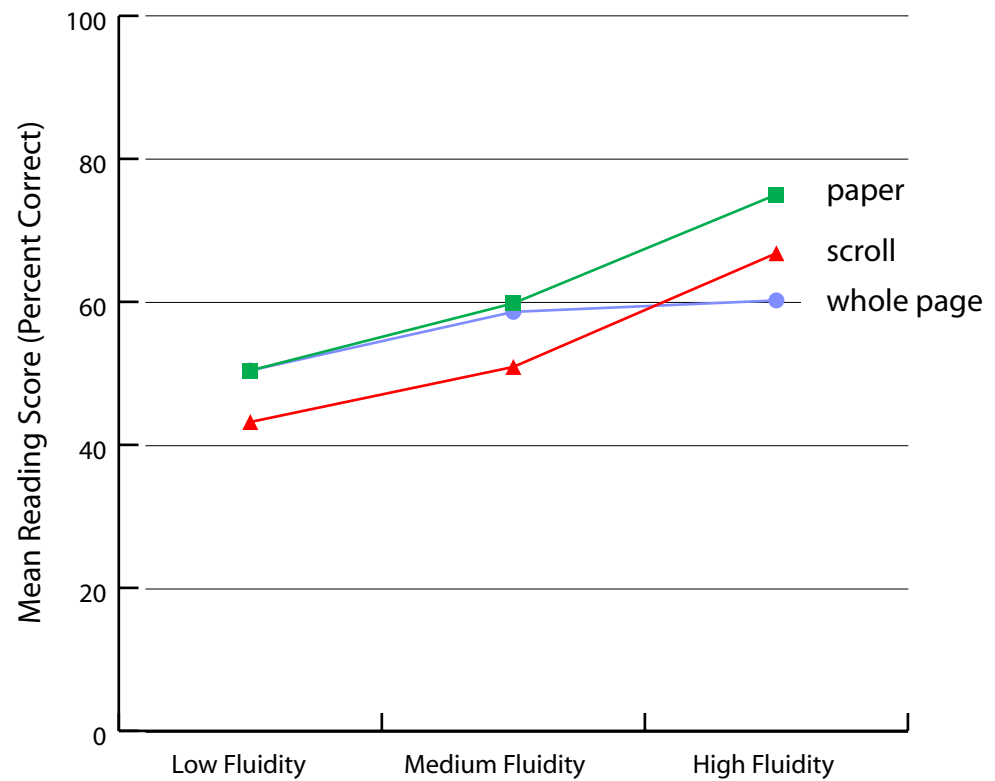
As described above, students were assigned to a high, medium, or low computer fluidity groups based on their fluidity scaled score. Table 5 displays the average reading comprehension test score for students based on their computer fluidity categorization, across each of the three modes.

Table 5: Comparison of Mean Reading Test Scroll across Computer Fluidity Groups and Presentation Mode

	N	All Students	Paper	Scroll	Whole Page	F	Sig
High	66	65.8%	75.0%	66.8%	60.2%	2.429	.096
Medium	66	56.5%	59.8%	50.9%	58.6%	1.045	.358
Low	66	47.8%	50.4%	43.2%	50.4%	0.764	.470

In general, students with higher computer fluidity scores tended to score higher on the reading comprehension assessment across all three modes. Similarly, students in the low computer fluidity group performed lower, on average, on each of the presentation modes. Since this relationship holds across all three testing modes, it does not provide evidence that students taking the assessment using the two computer-based modes were advantaged or disadvantaged by their computer fluidity.

To examine differences within each computer fluidity group across presentation modes, a one-way ANOVA was run across presentation modes for each fluidity categorization. Within each of the three computer fluidity groups, no significant differences were found at the $p < .05$ level. It is interesting to note, however, that for both the medium and the low computer fluidity groups, performance in the scrolling condition was lower than the paper or whole page conditions. In addition, for both the medium and low computer fluidity groups, performance for the paper and whole page conditions differed minimally. Finally, as seen in Figure 3 (*shown on the following page*), there appears to be no meaningful interaction between fluidity and presentation mode.

Figure 3: Mean Reading Score for Fluidity Groups and Presentation Mode

As described above, a regression analysis was performed in order to control for differences in computer fluidity that existed among the presentation modes. In addition to computer fluidity scale scores, the regression model included two dummy variables which represented the scroll and whole page group. For this regression model students' reading comprehension score was the dependent variable.

As seen in Table 6, the regression model accounted for 14.2% of the variance in reading scores. As indicated by the standardized coefficient (beta), computer fluidity was the strongest predictor of students' reading scores. Both scrolling and whole page conditions had a negative effect on test performance, after controlling for differences in computer fluidity. However, the coefficients for both scrolling and whole page were not statistically significant at the .05 level.

Table 6: Regression Model for Computer Fluidity and Group Membership Predicting Reading Scores

R² = .142
 F = 11.88
 p < .001

Variable	Coefficient	SE	Beta	T ratio	Sig.
Intercept	.609	.026		23.054	<.001
Fluidity	.090	.016	.387	5.713	<.001
Scroll	-.069	.038	-.139	-1.832	.068
Whole Page	-.060	.038	-.121	-1.560	.120

Computer Literacy Analysis

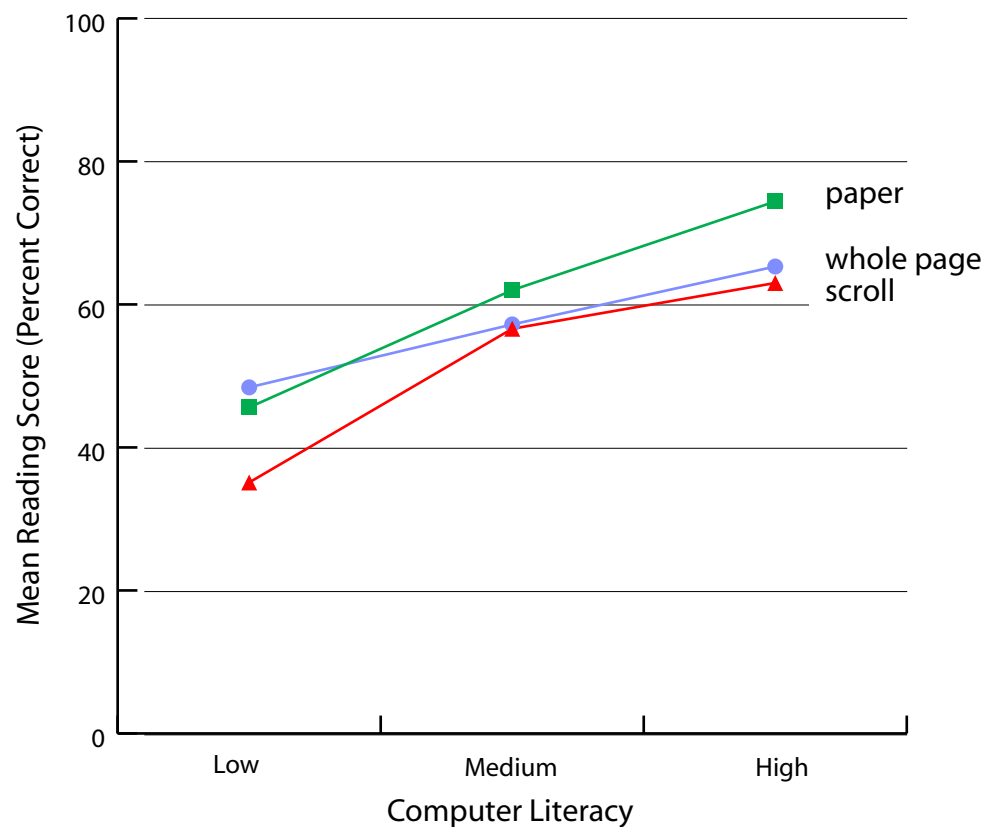
As described above, students were assigned to a high, medium, or low computer literacy group based on their literacy scaled score. Table 7 displays the average reading comprehension test score across the three presentation modes for students based on their computer literacy categorization.

Table 7: Comparison of Mean Reading Test Scroll Across Computer Literacy Groups and Presentation Mode

	N	All Students	Paper	Scroll	Whole Page	F	Sig
High	70	66.4%	74.4%	63.0%	65.3%	1.564	.217
Medium	71	58.9%	62.0%	56.6%	57.2%	0.435	.649
Low	72	43.4%	45.6%	35.1%	48.4%	2.495	.090

On average, students who scored higher on the computer literacy test also scored higher on the reading comprehension test. As depicted in Figure 4, this relationship holds across all three presentation modes and therefore does not indicate that there is an interaction between presentation mode and computer literacy. One-way ANOVAs performed within each level of computer literacy also indicate that there are no significant differences among mean test scores for each presentation mode within each computer literacy group. Similar to the fluidity measure, it is noteworthy that as students' computer literacy decreases, the scrolling group seems to be more disadvantaged than the whole page group when compared to the paper group.

Figure 4: Mean Reading Score for Computer Literacy Groups and Presentation Mode



As described above, a regression analysis was performed to control for differences in computer literacy that existed among the presentation modes. As seen in Table 8, the regression model accounted for 21.9% of the variance in reading scores. As indicated by the standardized coefficient (beta), computer literacy was the strongest predictor of students' reading scores. Both scrolling and whole page conditions had a negative effect on test performance, after controlling for differences in computer literacy. The coefficient for scrolling was statistically significant at the .05 level. This means that after controlling for differences in computer lit-

eracy, students who took the test using the scrolling mode achieved a statistically significant lower reading comprehension score than students who took the assessment on paper and on laptop using whole page text.

Table 8: Regression Model for Computer Literacy and Group Membership Predicting Reading Scores

$$R^2 = .219$$

$$F = 20.82$$

$$p = <.001$$

Variable	Coefficient	SE	Beta	T ratio	Sig.
Intercept	.603	.024		24.755	<.001
Literacy	.111	.014	.474	7.696	<.001
Scroll	-.086	.035	-.171	-2.446	.015
Whole Page	-.042	.035	-.086	-1.225	.222

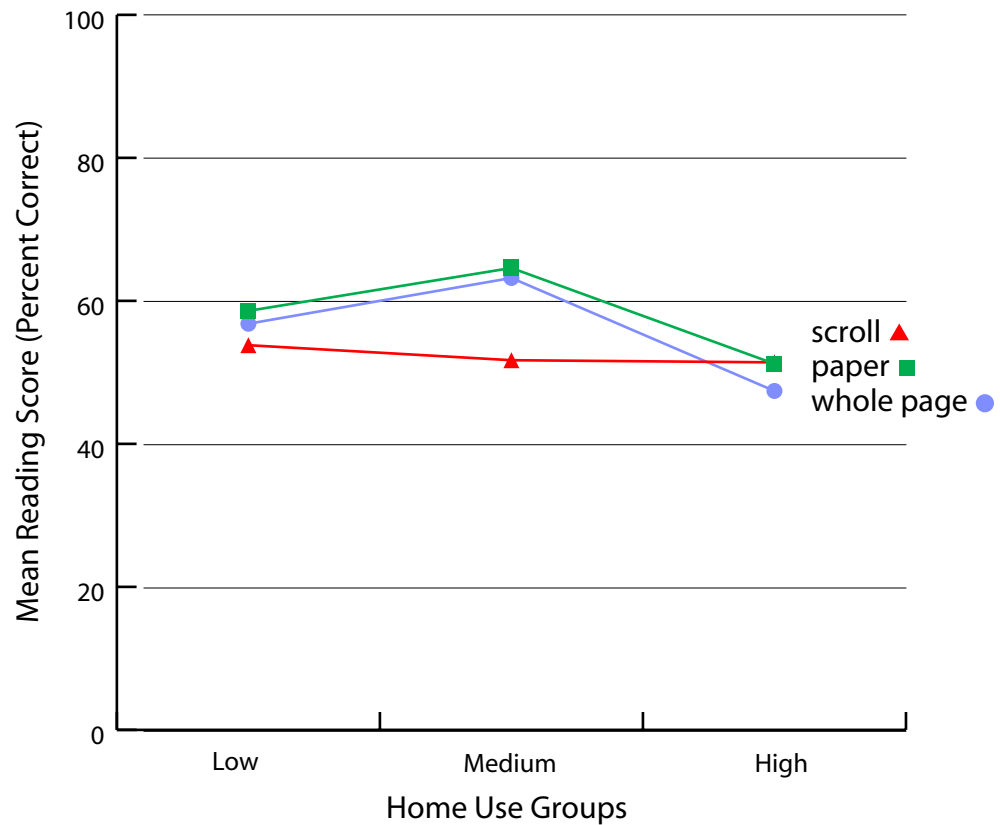
Home Use Analysis

Students' home-use scale scores were used to divide students into three groups which represented high, medium, and low levels of reported use of computers at home. Table 9 displays the average reading comprehension test score across the three presentation modes for students based on their home computer use categorization.

Table 9: Comparison of Mean Reading Test Scroll Across Home Computer Use Groups and Presentation Mode

	N	All Students	Paper	Scroll	Whole Page	F	Sig
High	56	50.1%	51.2%	51.4%	47.4%	0.191	.827
Medium	56	61.3%	64.6%	51.7%	63.2%	1.168	.319
Low	56	56.6%	58.6%	53.8%	56.8%	0.188	.829

On average, students who were categorized as high home computer users scored lower on the reading comprehension test while students who were categorized as medium level home computer users scored highest. However, as depicted in Figure 5 (*shown on the following page*), this relationship holds for the paper and whole page presentation modes, but does not hold for the scroll mode. Specifically, there are only minor visible differences in performance between the three home use groups for the scrolling condition. One-way ANOVAs performed within each level of home computer use indicate that there are no significant differences among mean test scores for each presentation mode within each home computer use group.

Figure 5: Mean Reading Score for Home Use Groups and Presentation Mode

Although there were only small differences in mean home computer use scale scores among the three presentation modes, a regression model was developed to examine the effect of presentation mode controlling for differences in home computer use. As seen in Table 10, the model accounts for less than 3% of the variance in reading comprehension. As indicated by the standardized coefficient (beta), home computer use was the strongest predictor of students' reading scores and the only independent variable to be statistically significant at the .05 level. This model shows a negative relationship between home computer use and reading score.

Table 10: Regression Model for Home Computer Use and Group Membership Predicting Reading Scores

$R^2 = .025$

$F = 2.409$

$p = .069$

Variable	Coefficient	SE	Beta	T ratio	Sig.
Intercept	.579	.031		18.663	<.001
Home Use	-.042	.018	-.178	-2.331	.021
Scroll	-.053	.045	-.104	-1.175	.242
Whole Page	-.008	.043	-.017	-0.190	.850

School Use Analysis

Students' school-use scale scores were used to divide students into three groups which represented high, medium, and low levels of reported use of computers in school. Table 11 displays the average reading comprehension test score across the three presentation modes for students based on their school computer use categorization.

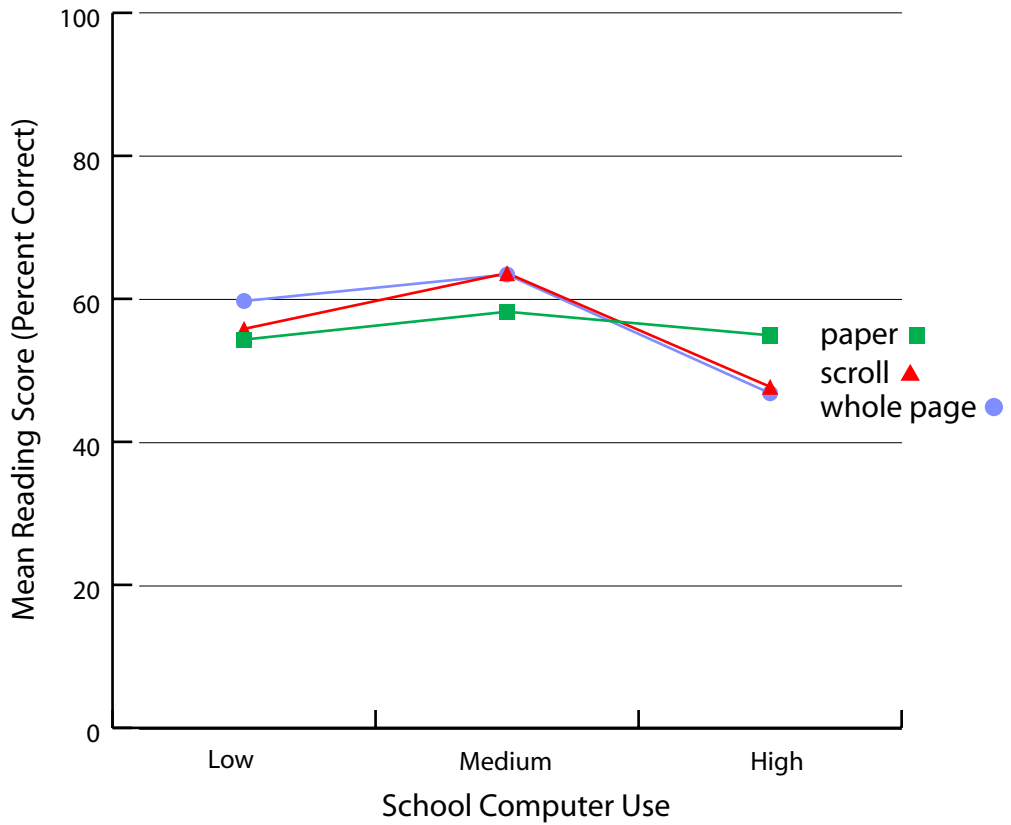
Table 11: Comparison of Mean Reading Test Scroll Across School Computer Use Groups and Presentation Mode

	N	All Students	Paper	Scroll	Whole Page	F	Sig
High	57	49.0%	54.9%	47.7%	46.8%	0.877	.422
Medium	57	61.5%	58.2%	63.6%	63.4%	0.325	.724
Low	57	56.9%	54.3%	55.8%	59.7%	0.297	.744

As seen in Table 11 and in Figure 6, there is not a linear relationship between students reported school computer use and reading comprehension test scores. Within the paper group, reading scores differ only slightly between the high, medium, and low school-use groups. For the scrolling and whole page groups, the medium-level school use group performed noticeably higher than did the high use group. It is difficult to hypothesize why this pattern occurs. Within each level of school computer use, a one-way ANOVA indicates that there are not statistically significant differences in reading scores across the mode of presentation.

(Figure 6 is shown on the following page.)

Figure 6: Mean Reading Score for School Use Groups and Presentation Mode



A regression analysis was performed to control for differences in school computer use that existed among the presentation modes. As seen in Table 12, the model accounts for less than 2% of the variance in reading comprehension scores. As indicated by the standardized coefficient (beta), school computer use was the strongest predictor of students’ reading scores and the only independent variable to be statistically significant at the .05 level. This model shows a negative relationship between school computer use and reading score.

Table 12: Regression Model for School Computer Use and Group Membership Predicting Reading Scores

$R^2 = .019$

$F = 2.080$

$p = .105$

Variable	Coefficient	SE	Beta	T ratio	Sig.
Intercept	.554	.031		18.151	<.001
School Use	-.041	.017	-.180	-2.335	.021
Scroll	-.007	.044	-.014	-.161	.872
Whole Page	.016	.041	.035	.389	.698

Gender Analysis

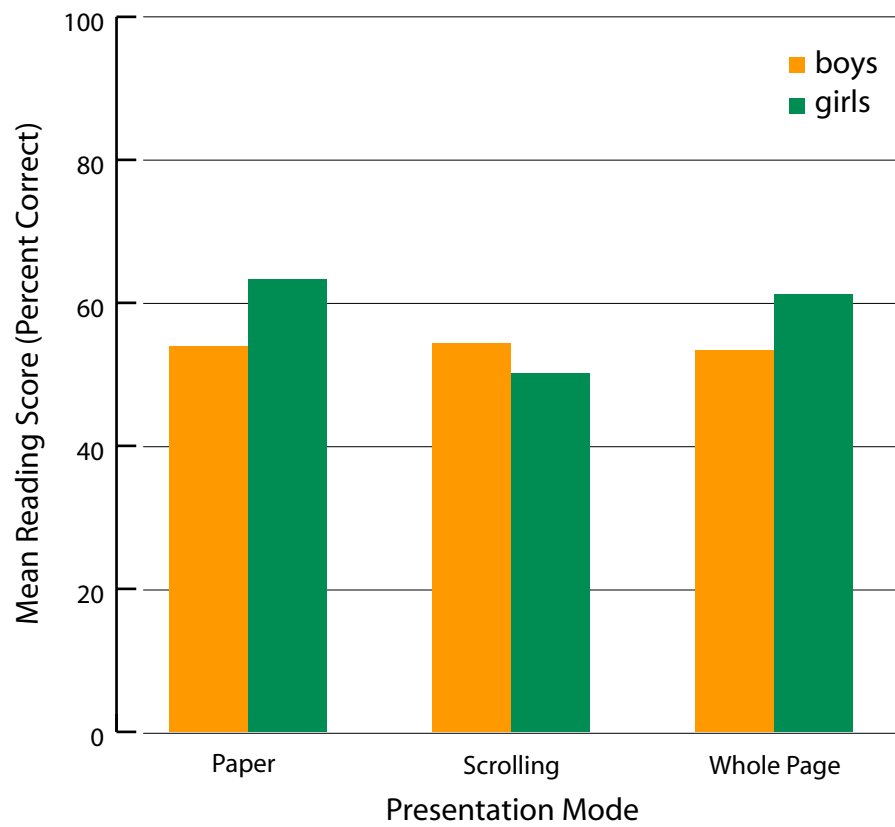
To examine the extent to which the effect of presentation mode differed for boys and girls, performance was compared between gender within and across each presentation mode. As seen in Table 13, overall girls received higher reading, computer literacy, and school use scale scores than did boys. Conversely, boys received higher computer fluidity and home use score. Overall, however, these differences were small and not statistically significant.

Table 13: Gender Comparison Across All Measures and Presentation Modes

	Reading Comprehension	Computer Literacy	Fluidity	Home Use	School Use
All					
Boys (112)	53.9%	52.9%	0.102	0.056	-0.075
Girls (106)	58.2%	55.1%	-0.101	-0.057	0.091
Paper					
Boys (37)	53.9%	46.0%	-0.22	0.034	-0.311
Girls (37)	63.3%	54.5%	-0.16	-0.005	0.097
Scroll					
Boys (34)	54.3%	57.2%	0.17	0.142	0.317
Girls (36)	50.1%	51.1%	-0.35	-0.020	0.162
Whole Page					
Boys (41)	53.4%	55.5%	0.35	0.011	-0.182
Girls (33)	61.2%	59.8%	0.28	-0.141	0.023

As seen in Table 13, the differences in computer literacy, fluidity, home use, and school use are generally consistent within the paper and whole page groups. For the paper group, however, boys had lower fluidity scores than did girls. Similarly, for the scroll group, the pattern reversed for two measures, namely computer literacy and school use. Again, these differences are relatively small and are not statistically significant at the .05 level. As seen in Table 13 and depicted in Figure 7, girls performed noticeably higher on the paper mode. Although the performance difference was slightly smaller for the whole page mode, the same pattern held. For the scroll mode, however, the pattern reversed such that the boys scored slightly higher than the girls when performing the reading test under the scrolling condition. This difference may result in part from differences across all four computer scales. Where as girls performed higher on half of the computer scales within the paper and whole page groups, boys performed higher on all four scales within the scrolling group. Thus, it is difficult to estimate whether the pattern reversal for the scroll group is a function of gender or prior computer experiences.

(Figure 7 is shown on the following page.)

Figure 7: Mean Performance for Boys and Girls by Presentation Mode

Exploratory Analysis of IEP and Multi-Lingual Students

After the data collection phase of this research was completed, Boston College researchers asked the participating teachers and principals to provide information about whether students in their classrooms had language-related IEPs (Individualized Education Plans). One teacher did not respond to the request and one principal responded that none of the students in her school had an IEP. The remaining teachers and principals provided the names of 25 students who had language-related IEPs. Based on this information, all students were coded as either “IEP” or “Non-IEP.” Overall, students with language-related IEPs scored significantly lower as compared to non-IEP students on the reading comprehension assessment, fluidity exercises, and computer literacy assessment. Tables 14–16 show that IEP students scored more than 20% points lower than non-IEP students on the reading comprehension assessment and also scored lower on the fluidity exercises and literacy assessment.

Table 14

	Reading Comprehension Assessment		
	N	Mean Score	Std Dev
IEP	25	36.2%	0.18
Non-IEP	194	58.3%	0.23

Table 15

	Fluidity Exercises		
	N	Scaled Score	Std Dev
IEP	21	-0.66	1.06
Non-IEP	177	0.08	0.97

Table 16

	Literacy Assessment		
	N	Scaled Score	Std Dev
IEP	23	-0.72	0.95
Non-IEP	190	0.09	0.97

These results are not surprising, since the IEP students have been identified as having language-related difficulties. We next examined the relationship between the IEP students' performance on the reading comprehension assessment and the mode of delivery. Table 17 shows that IEP students scored similarly across the three modes of delivery. In other words, the computer based testing modes neither advantaged nor disadvantaged the IEP students.

Table 17

	Reading Comprehension Assessment		
	Paper	Scroll	Whole Page
IEP	36.3%	34.2%	37.4%
Non IEP	61.5%	53.9%	59.6%

One of the items on the student survey asked students "What language other than English do you speak at home?" This item had not been previously pilot tested and many students had a difficult time answering the question. Due to how the question was phrased, students thought that they had to write in a language other than English, even if they only spoke English at home. Proctors explained to students that they could leave this question blank or write in "none" or "English only." A total of 54 students reported that they spoke a language other than English at home. Although the data collected for this item may be problematic, students' responses were used to classify them as either "multilingual" or "non-multilingual." Tables 18–20 show that multilingual students scored slightly higher as compared to non-multilingual students on the reading comprehension assessment, fluidity exercises, and computer literacy assessment.

Table 18

	Reading Comprehension Assessment		
	N	Mean Score	Std Dev
Multilingual	54	58.1%	0.23
Non-Multilingual	165	55.1%	0.23

Table 19

	Fluidity Exercises		
	N	Scaled Score	Std Dev
Multilingual	48	0.05	0.77
Non-Multilingual	150	-0.02	1.06

Table 20

	Literacy Assessment		
	N	Scaled Score	Std Dev
Multilingual	53	0.07	0.93
Non-Multilingual	160	-0.02	1.02

These results are surprising, since multilingual students often struggle with reading comprehension assessments in English. We next examined the relationship between the multilingual students' performance on the reading comprehension assessment and the mode of delivery. Table 21 shows that the multilingual students scored highest when taking the assessment using the whole page format and lowest when taking the assessment using scrolling. Although there is almost an 11-point difference in scores, this difference is not statistically significant due to the small number of multilingual students.

Table 21

	Reading Comprehension Assessment		
	Paper	Scroll	Whole Page
Multilingual	57.9%	53.5%	64.3%
Non-Multilingual	58.2%	51.7%	55.2%

Open-Ended Survey Items

The above analyses provide insight into the extent to which performance differed across the three presentation modes and how prior computer uses and experiences may have affected test performance within each condition. To collect information from the students' perspective on their experiences performing the reading test on computer, students in the scrolling and whole page groups were asked to respond to four open-ended survey items. A content analysis was performed on students' responses to these four items. Patterns that emerged from this content analysis are summarized for each open-ended item.

1. Overall, was taking the test on a computer easier or more difficult than taking the test with paper and pencil? Why?

Of the 135 students who responded to this item, 82.2% reported that they thought it was easier to take the test on computer. The most common reason provided for why the test was easier on computer was that students did not have to write/their hand did not get tired from writing (this was cited even though students were responding to multiple-choice items). The data below details student responses:

Easier

- Didn't have to use your hand as much, hand doesn't get tired (n=36)
- No reason given (n=34)
- More fun (n=8)
- Easier to click instead of fill in (n=5)
- Like working on a computer (n=5)
- Don't have to erase (n=5)
- Can work faster on a computer (n=3)
- Can see the words more clearly on a computer (n=3)
- Use a computer a lot (n=3)
- Answers were more clear (n=2)
- Writing takes longer (n=2)
- Easier to click than flip pages (n=2)
- Questions and story were on the same page (n=1)
- At the beginning there was a tutorial of the test (n=1)
- Computer explained more to you (n=1)

Harder

- No reason given (n=6)
- Can't type fast (n=3)
- Easier to write on paper (n=3)
- Easier to read on paper (n=2)
- Don't have to read the passage when you are taking the test on paper (n=2)
- Could not control the mouse (n=1)
- Not comfortable with computers (n=1)
- Paper would never freeze (n=1)
- Took longer on computer (n=1)
- Don't have to type on paper (n=1)
- Didn't know what to press (n=1)
- Computers can be difficult to use for tests (n=1)
- Computer hurt eyes (n=1)

2. Describe any problems that you had when taking the test on the computer.

Twenty-six students indicated that they had a problem while taking the test on computer. Some of the reported problems were not understanding the directions, having problems using the mouse, not understanding how to end the test, and having the computer freeze. The data below details student responses:

- Computer froze (n=2)
- Could not easily move the mouse (n=2)
- Had problems scrolling (n=2)
- Had problems changing the answer (n=2)
- Didn't know what to do at the end of the test (n=2)
- Sometimes skipped things (n=1)
- Too slow to take the test on computer (n=1)
- Hard to figure out (n=1)
- Had to restart (n=1)
- Had problems understanding what the directions meant and what buttons to push (n=1)
- Had trouble answering some of the questions (n=1)
- Not a fast typer (n=1)
- Didn't know when to end the test or when to press start (n=1)
- Couldn't find the arrow keys (n=1)
- Didn't know what to press (n=1)

- The words were not clear (n=1)
- Had problems with the mouse (n=1)
- Couldn't find exit box (n=1)

3. When reading the passage, did you underline or highlight any text? Why or why not?

Of the 137 students who responded to this item, 16.8% reported using the highlighting feature. The most common reason for using highlighting was to remember important information and the most common reason for not using the highlighting was because they didn't need to use this feature. The data below details student responses:

Yes

- No reason given (n=7)
- To remember information (n=6)
- So that you can go back (n=3)
- Because it was important (n=3)
- To memorize things (n=1)
- It helped me follow the words (n=1)
- Tried, but it didn't work (n=1)
- Because kept losing place (n=1)

No

- Didn't need to (n=57)
- No reason given (n=27)
- Didn't understand how (n=11)
- Forgot to (n=10)
- Nothing interesting/important to highlight (n=6)
- Couldn't go back easily, so didn't need to highlight (n=1)
- Didn't want to lose place (n=1)
- Didn't have enough time (n=1)

4. When taking the test, did you mark any questions that you were unsure about and wanted to work on later? Why or why not?

Of the 132 students who responded to this item, 17.4% said that they did mark items for review. The most common reported reason for marking for review was because they wanted to go back to it later and the most common reason for not marking for review was because they were confident in their answer or did not need to use this feature.

Yes

- No reason given (n=11)
- Didn't understand some things, so went back to it (n=6)
- Marked them all and then looked them over (n=2)
- But didn't go back to them later (n=2)
- Because problems were tough (n=1)
- Was unsure of answer (n=1)

No

- Didn't need to/confident in answer (n=74)
- No reason given (n=22)
- Forgot (n=5)
- Didn't understand it (n=3)
- Wanted test to be done (n=2)
- Guessed if unsure (n=1)
- Would have taken longer (n=1)
- Didn't know you could go back (n=1)

Discussion

The pilot study presented here was limited to 219 fourth grade students from one state who performed a reading comprehension test of moderate length. Findings from a pilot study should not be interpreted as definitive. With this caution in mind, the results highlight some interesting findings and suggest that more research on a larger scale should be conducted in the area of computer based delivery of passage dependent assessments. Below, we summarize the major findings.

1. There were no significant differences in reading comprehension scores across testing modes.

On average, students in the paper group (n=75) answered 58.1% of the items correctly, students in the scrolling group (n=70) answered 52.2% of the items correctly, and students in the whole page group (n=74) answered 56.9% of the items correctly. Although there is almost a 6% point difference in scores between the paper and scrolling groups, this difference was not significant at the $p < .05$ level, meaning that the differences may be due to chance.

Without examining other student measures, finding no significant differences in test scores across paper and computer delivered modes implies that overall students were neither advantaged nor disadvantaged by the mode of test delivery. However, characteristics of the sample and differences that may occur for students with different levels of prior computer experience must also be considered.

Although the sample of schools that participated in this study represent a range of schools within Vermont, the schools and their students were not randomly selected. Rather they were recruited based on their location, composition of their student body, and, perhaps most importantly, their willingness to participate. As part of the recruitment process, principals received recruitment letters that detailed the research questions and described the data collection procedures. Although this recruitment process was necessary in order to obtain school-level support for the data collection, this process may have resulted in the selection of schools in which principals were more interested in the use of technology or in computer-based testing specifically. Although efforts were made to attain a representative sample of Vermont 4th grade students, sample bias is a concern that limits the generalizability of the results.

As a measure of socio-economic status, students were asked the following question on the survey: “About how many books of your own do you have at home, not counting your school books or comic books?” The response options were: a) 5 or fewer b) 6–25 c) 26–50 d) 51–100 e) more than 100. Only nine students reported that they have 5 or fewer books of their own at home and the most common response (n=86, 39.3%) was “more than 100.” Students were also asked to report how many computers they have at home. The response options were a) 0 b) 1 c) 2 d) 3 or more. Again, only nine students reported they do not have a computer at home. The most common response (n=105, 48.6%) was 1, while 54 (24.7%) students reported that they have 2 computers at home, and 48 (21.9%) students reported that they have 3 or more computers at home. Students’ answers to these two items indicate that the sample of 219 students does not include many students from a low socio-economic and low computer access population. Although the results suggest that, across all students, the modal effect is not statistically significant, this finding may be due in part to the atypically high computer access and higher socio-economic status of the sample.

2. There were no statistical differences in reading comprehension scores based on computer fluidity and computer literacy, but a pattern in performance may indicate that students are disadvantaged by the scrolling text mode, particularly students with lower computer skills.

Computer based testing experts hypothesize that scrolling may negatively impact students’ test scores in passage based assessments (Choi & Tinkler, 2002, Pommerich, 2004). Students who are not as familiar with using a computer may have difficulty navigating through a passage for which they are required to use the scroll bar. Although this hypothesis was not supported in a statistically significant way, there is evidence of modal differences in student performance based on their computer skills and knowledge.

As seen in Table 5, students’ reading comprehension scores decrease as their computer fluidity decreases. For all three fluidity groups, scores are noticeably lower for the scrolling mode than on the paper version. For both the medium-

and low-level fluidity groups, the lowest scores occur on the scrolling version (see Figure 3). Similarly, for all three computer literacy groups, scores are lower for the scrolling mode than on the paper version. For the lowest computer literacy group, this difference is substantial (Figure 4). This pattern may indicate that students generally perform worse under the scrolling condition than under the paper or whole page conditions and that the score differences are largest for students with lower computer fluidity and/or literacy. Again, while these patterns are not statistically significant, they warrant further investigation with a larger and more diverse sample of students.

3. The majority of students who took the reading test on a computer indicated that they would prefer to take the test on computer.

After completing the selected response portion of the survey, students who completed the assessment on computer were asked to answer four open-ended questions. One question asked students whether they thought it was easier or harder to take the test on computer or paper. Of the 135 students who responded to this item, 82% reported that it was easier to take the test on computer. In addition, students were asked in a selected response format if they would have preferred to take the test on computer or on paper. Of the 161 students who responded to this item, 87% reported that they would prefer to take the test on computer. In addition, when asked about problems, very few students noted that they experienced difficulty taking the test on computer.

Completion rates also provide evidence of students' ability to perform the reading test on the computer. Only nine students did not reach the last item on the reading comprehension assessment. These nine students were evenly distributed across the paper, scrolling, and whole page groups. Also, the number of students who answered all of the items was higher for the computer groups than for the paper group. Although this sample did not include many students who had limited prior computer experience, the survey responses, completion rates, and student observations are all evidence that computer anxiety generally did not interfere with students' ability to take the assessment.

4. Providing highlighters and review markers is useful for some students.

One concern raised when transitioning a test from paper to computer focuses on the belief that examinees may be disadvantaged when features such as highlighting and being able to skip and return to items are not available. Choi and Tinkler (2002) hypothesized that "providing such accommodations as an electronic marker, which allows students to highlight a selection of text in a passage, will not only simulate the natural pencil-and-paper test-taking practice, but also reduce the administration mode effect to a certain extent" (p. 10). In an effort to create similar testing environments in the paper and computer groups, highlight and mark for review tools were built into the computer-based test interface for this research. Through a survey administered after the completion of the test, less

than 20% of students indicated that they used these features. However, based on open-ended survey responses, it appears that those students who used the features did so in a manner that was consistent with the intended purposes.

Students who reported that they used the highlight feature listed reasons such as remembering information, noting important parts of the passage, and wanting to go back to the highlighted portions as reasons for using this feature. The majority of students who listed reasons for using the mark for review feature reported that they were unsure of the answer to an item and wanted to return to the item later. These appropriate reasons for highlighting text and marking items for review suggest that a small proportion of students may benefit from having access to these test taking tools in the computer-based environment. However, more than 20 students who reported that they did not use the highlighter and 8 students who reported that they did not use the mark for review feature noted that they either did not understand how to use the tools or forgot to use them. This finding indicates that further training and practice may have resulted in higher usage of the highlighter and mark for review.

Conclusion

In this study of the impact of transitioning from paper to two computer based testing formats for passage based reading comprehension assessments, no significant differences in test scores were found between the three groups. Taking multiple measures of students' computer skills allowed rich data analysis to understand the relationship between computer use and performance on computer-based assessments. Although the size and scope of this study were not large enough to show definitively that scrolling negatively affects students with low computer skills and knowledge, the results suggest that further research is warranted to more fully understand the impact of scrolling on passage-based assessments.

Future research should be conducted on a larger and more diverse sample of students and should be expanded to include students in other grade levels. The amount of passages and items should also be increased. The high completion rate and observation of some students finishing the test in approximately 20 minutes is evidence that the test length could be increased. Adding more well constructed items would increase test reliability and provide a better opportunity to examine if there are differences in completion rate across test modes. This future study with a larger and more diverse sample and more items may show more subtle differences in modal performance overall and in student sub-groups. This research could lend further insight into the use of computer-based test features, and reveal relationships between student performance and their computer skills, knowledge, or use.



Endnote

- 1 To ensure that the scores of the three paper students who missed items on the back of pages did not affect the overall research results, a sensitivity analysis was conducted which assumed that the students would have answered the missed items correctly. This analysis showed no changes in the study finding of non-significant differences in performance across the three testing modes.

References

- Bangert-Downs, R.L. (1993). The word processor as an instructional tool: A meta-analysis of word processing in writing instruction. *Review of Educational Research*, 63(1), 69–93.
- Bunderson, C., Inouye, D. & Olsen, J. (1989). The four generations of computerized educational measurement. In Linn, R.L., *Educational Measurement* (3rd ed.), Washington, D.C.: American Council on Education, pp. 367–407.
- Choi, S.W., & Tinkler, T. (April, 2002). *Evaluating comparability of paper-and-pencil and computer-based assessment in a K–12 setting*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.
- 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(5).
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.
- The Editors (2003). Tech's Answer to Testing. *Education Week*. XXII(35), 8–10.
- Goldberg, A., Russell, M., & Cook, A. (2003). The effect of computers on student writing: A meta-analysis of studies from 1992–2002. *Journal of Technology, Learning, and Assessment*, 2(1). Available from <http://www.jtla.org>
- Haas, C. & Hayes, J. (1986). What did I just say? Reading problems in writing with the machine. *Research in the Teaching of English*, 20(1), 22–35.
- Haertel, E. & Wiley, D. (2003, August). *Comparability issues when scores are produced under varying testing conditions*. Paper presented at the Psychometric Conference on Validity and Accommodations in College Park, MD.
- Mazzeo, J. & Harvey, A.L. (1988). *The equivalence of scores from automated and conventional educational and psychological tests: A review of the literature* (College Board Rep. No. 88-8).
- Pommerich, M. (2004). Developing computerized versions of paper-and-pencil tests: Some mode effects for passage-based tests. *Journal of Technology, Learning, and Assessment*, 2(6). Available from <http://www.jtla.org>

- Pommerich, M. & Burden, T. (2000, April). *From Simulation to Application: Examinees react to computerized testing*. Paper presented at the National Council on Measurement in Education conference in New Orleans. Retrieved on August 6, 2003, from <http://home.att.net/~pommie/NCME00.pdf>
- Russell, M. (1999). Testing Writing on Computers: A Follow-up Study Comparing Performance on Computer and on Paper. *Educational Policy Analysis Archives*, 7(20).
- Russell, M., Bebell, D., & O'Dwyer, L. (2003). *Use, support, and effect of instructional technology study: An overview of the USEIT study and the participating districts*. Boston, MA: Technology and Assessment Study Collaborative. Available for download at http://www.intasc.org/PDF/useit_r1.pdf

