



TECHNOLOGY AND ASSESSMENT STUDY COLLABORATIVE

Examining the Feasibility and Effect of Computer-Based Verbal Response to Open-Ended Reading Comprehension Test Items

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

**Helena Miranda, Michael Russell, Kevon Seeley, & Thomas Hoffmann
Technology and Assessment Study Collaborative
Boston College
332 Champion Hall
Chestnut Hill, MA 02467**

www.intasc.org

Examining the Feasibility and Effect of Computer-Based Verbal Response to Open-Ended Reading Comprehension Test Items

Helena Miranda, Michael Russell, Kevon Seeley, & Thomas Hoffmann
Technology and Assessment Study Collaborative
Boston College
Released January 2005

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.necompact.org)

Copyright © 2005 Technology and Assessment Study Collaborative, Boston College



Examining the Feasibility and Effect of Computer-Based Verbal Response to Open-Ended Reading Comprehension Test Items

Helena Miranda, Michael Russell, Kevon Seeley, and Thomas Hoffmann
Technology and Assessment Study Collaborative
Boston College
Released January 2005

Since its inception in the early nineteenth century, the United States public education system has provided services to an increasingly diverse population of students. The end of the nineteenth century saw extensive growth in America's public schools, with enrollment increasing from roughly 7 million in 1870 to nearly 13 million in 1890 (Mondale & Patton, 2001). By the beginning of the twentieth century, "...public schools had opened up a world of promise for children who came off steamships and out of factories and farms" (Mondale & Patton, pp. 118–119, 2001). Although the legacy of this era of massive growth was the development of a system of mass public-education that exists today, the expanding diversity of American students has been a challenging corollary. Today, the roughly 40-million (NCES, 2004) faces of America's K–12 students in public schools across the nation reflect a broad spectrum of races, ethnicities, cultures, nationalities, and abilities. With the present Federal push for greater accountability, tougher standards, and more testing to ensure the success of *all* students, schools across the nation are facing a seemingly major challenge to validly measure the performance of all students.

The current high-stakes testing movement set in motion largely due to the 2001 reauthorization of the Elementary and Secondary Education Act (ESEA) also known as the No Child Left Behind (NCLB) Act, has established unprecedented implications for America's public educational system. As a result of this legislation, schools have become increasingly accountable for the academic progress of all students, regardless of physical ability, academic competency, or English language proficiency. Of growing concern is the validity with which students with disabilities or students whose native language is not English are assessed via traditional test formats. Because some aspects of standardized testing can be unfair to

students with limited English proficiency (LEP) and in particular, students with disabilities, a number of laws have been passed to ensure that states make provisions for reasonable and appropriate accommodations for students with special needs (e.g. Titles I and VII of ESEA, the Individuals with Disabilities Education Act, Equal Educational Opportunities Act, Title II of the Americans with Disabilities Act, and Title VI of the Civil Rights Act).

Special-needs students nationwide have historically been exempted from participating in state-mandated assessments (Allington & McGill-Franzen, 1992; Langenfeld, Thurlow, & Scott, 1997) primarily out of fear that their test scores would be lower and thus adversely impact a school's overall test performance (Salvia & Ysseldyke, 1998). Research has shown that students with special-needs do score lower on standardized tests than do students who require no special accommodations (Chin-Chance, Gronna, & Jerkins, 1996; Ysseldyke, Thurlow, Langenfeld, Nelson, Teelucksingh, & Seyfarth, 1998). Furthermore, the vast majority of research suggests that when students are denied accommodations prescribed by their Individual Education Plans (IEPs), "[their test] performance appears to be impeded" (Tindal, Heath, Hollenbeck, Almond, & Harniss, 1998) and rates of failure can climb "...thirty-five to forty percentage points higher than those for non-disabled [and native English-speaking] students" (Heubert as cited in Orfield & Kornhaber, 2001, p. 181). With the reauthorization of the Individuals with Disabilities Education Act (IDEA) in 1997, Federal lawmakers now demand that "children with disabilities are included in general and state wide assessment programs, with appropriate accommodations, where necessary" [IDEA, section 612(a)(17)(A)]. Unfortunately, very little guidance was provided to help states operationally define what "appropriate accommodations, where necessary" means.

Butler and Stevens (as cited in Abedi, Hofstetter, Baker, & Lord, 2001) define test accommodations as "support provided for a given testing event either through modifications of the test itself or through modification of the testing procedure to help students access the content...and better demonstrate what they know" (p. 11). Further, as acknowledged by the National Research Council (1997) in *Educating One and All*, "accommodations are intended to correct for distortions in a student's true competence caused by a disability unrelated to the construct being measured" (p. 197). Ultimately, the goal of any accommodation is not to make the test easier or to give an advantage to special-needs students, as some opponents allege. Rather, test accommodations are designed to promote fairness in testing so that the test's format or administration procedures or conditions "do not unduly prevent [special needs] students from demonstrating their 'true' knowledge, skills, and abilities" (Sireci, 2004, p.5).

The issue of accommodating students with disabilities and LEP students during testing has generated a great deal of concern, especially at the state level. NCLB stipulates that states must report disaggregated test data for students with disabilities and LEP students and provide these students with appropriate accommodations. Although all states currently have accommodation policies in place,

as mentioned above there seems to be little consensus on, what constitutes an “appropriate” accommodation, when accommodations should and should not be allowed, and how best to score and report the use of accommodations to comply with NCLB stipulations. To make matters worse, insufficient empirical evidence exists to inform states’ decisions; thus, questions remain as to whether accommodations provide an unfair advantage over students not receiving an accommodated assessment; whether accommodations improve the accuracy of the measurement of special-needs students’ knowledge, skills, and abilities; and whether the scores of tests from accommodated administrations are comparable to standard test administrations. Sireci, Li, & Scarpata (2003) argue that “from a psychometric perspective, the issues involved in test accommodations for [ELLs]...are the same as those for students with disabilities. The comparability of scores from accommodated and non-accommodated test administrations is an important validity issue regardless of the reason for the accommodation” (p. 48).

Validity of assessment results is a major concern when testing accommodations are made for special populations because, according to the National Research Council (1999), accommodations “...may provide the wrong correction. They may provide too weak a correction, leaving the scores of individuals with disabilities lower than they should be, or they may provide an irrelevant correction or an excessive one, distorting or changing scores further and undermining rather than enhancing validity [of tests]” (p. 197). Most experts in the field of testing and assessment agree that in order to ensure measurement validity, evidence must exist “...that a test is valid for the particular purpose for which it is being used, that the constructs it measures are relevant in making the decision in question, that the test actually measures those constructs, and that the test is reliable and accurate” (Heubert, as cited in Orfield & Kornhaber, 2001, p. 182). Further, the test must accurately measure the tested content knowledge of the student being assessed.

Tindal, Heath, Hollenbeck, Almond, & Harniss (1998) contend that “[in order to] validate an accommodation, it must work with the targeted subgroup (e.g., students in special education) and must not work for students in general education” (pp. 2-3). In an effort to further examine the effects of test accommodations on the test performance of special needs students (e.g., students with disabilities and ELL students), Sireci, Li, & Scarpata (2003) recently conducted a comprehensive review of literature to investigate whether evidence was available that test accommodations improved students’ performance. Their investigation identified over 150 papers, although subsequent review narrowed this extensive list down to only “...46 studies [that] actually focused on the effects of test accommodations and only 38 involved empirical analysis” (p. 7). Because such a vast array of assessment accommodations exists, Sireci, et al. (2003) chose to “borrow” Thompson, Blount, & Thurlow’s (2002) four general categories of accommodations: presentation, timing/scheduling, response, or setting.

Further, for the purposes of their study, Sireci, et al. (2003)

...defined test presentation accommodations to include oral administration, (i.e., “read aloud” protocols...), changes in test content (e.g., simplified language), and changes in test format (e.g., Braille or large print). Timing/scheduling accommodations included allowing extended or unlimited time and breaking up test administration into separate sessions....Response accommodations included allowing students to write directly into test booklet or dictate their answers. Setting accommodations typically involved administering the tests individually or in a separate room (p. 8).

Results revealed that nearly all of the studies focused on either presentation-type accommodations or timing accommodations (64 out of 75). Only one study focused on setting accommodations, four were not specified, and the remaining 5 focused on the following response-type accommodations: “in-booklet vs. answer sheet response” (1 study); “mark task book to maintain place” (1 study); “transcription” (1 study); and “scribes” (2 studies) (Sireci, et al., 2003). Because of the paucity of research on response accommodations, specifically verbal-response accommodations, the present study adds to the limited experimental evidence by examining the effects of verbal-response accommodations.

Verbal-response accommodations most often allow a student to orally respond to test items via a recording device or a scribe who records the student’s answers instead of having him or her provide a written response. This type of accommodation is usually offered to “students with a variety of disabilities including learning disabilities, behavioral disorders, mild mental retardation, physical impairments, and communication disorders” (Thurlow & Bolt, 2001, p. 17).

Research on the verbal-response accommodations for students with disabilities has been limited and is somewhat inconclusive. Two studies found that even with the response accommodations, all students with disabilities scored *lower* than students without disabilities (Koretz & Hamilton, 2000 and Johnson, Kimball, Brown, & Anderson, 2001 as cited in Sireci, et al., 2003). However, in their comprehensive review of testing accommodations, Tindal and Fuchs (as cited in Schulte, Elliot, and Kratochwill, 1999) revealed that “dictational/oral response appears to be an effective accommodation for a wide range of students and test content” (p.529). Koretz (as cited in Schulte, Elliot, and Kratochwill, 1999) reported that “dictation had the strongest effect on scores across grade levels and subject areas” (p. 529).

In their report for the National Center on Educational Outcomes, Thurlow & Bolt (2001) recommended that verbal-response accommodations should be provided to students only on tests not specifically designed to measure specific skills such as spelling or grammar. They also recommended that if students are incapable of writing by hand but have capable keyboarding skills, computer-response accommodation should be considered the first choice over verbal-response accom-

modations. Lastly, they concluded that “more research should examine the effects of this accommodation on the test scores of students with and without disabilities” (p. 20).

The study presented here sought to address the need for further investigation into the effect of the verbal-response accommodation in testing situations for ELL students and students with disabilities students as well as to investigate the validity of accommodated scores by comparing accommodated and non-accommodated scores for ELL, students with disabilities, and general education (GE) students.

The purpose of this study was to examine the feasibility and effect of allowing students to provide verbal responses that are recorded using digital technologies. Specifically, the study examined the effects of allowing students to respond to items about a reading passage using: a) paper delivery and written response; b) computer delivery and keyboard response; c) computer delivery and verbal response recorded by the computer. The research questions investigated by this study were:

1. How does the computer delivery of a fourth grade reading comprehension test with computer-recorded verbal response (CDVR) compare to paper delivery of the test with written response (PDWR)?
2. How does the computer delivery of a fourth grade reading comprehension test with computer-recorded verbal response (CDVR) compare to computer delivery of the test with computer response (CDCR)?
3. How does computer delivery of a fourth grade reading comprehension test with computer response (CDCR) compare to paper delivery of the test with written response (PDWR)?
4. What is the effect of computer delivery of a fourth grade reading comprehension test with computer-recorded verbal response (CDVR) on the test performance of fourth grade general education students?
5. What is the effect of computer delivery of a reading comprehension test with computer-recorded verbal response (CDVR) on the test performance of fourth grade special education students?
6. What is the effect of computer delivery of a reading comprehension test with computer-recorded verbal response (CDVR) on the test performance of fourth grade English language learners?
7. What is the effect of computer literacy on student performance when a fourth grade reading comprehension test with verbal response is computer-delivered?
8. What is the effect of computer fluidity on student performance when a fourth grade reading comprehension test with verbal response is computer-delivered?
9. What is the effect of computer use on student performance when a fourth grade reading comprehension test with verbal response is computer-delivered?

Results from this pilot study were intended to provide evidence about: a) whether it was feasible to provide a computer-recorded verbal response accommodation, and b) whether there are differences in student reading comprehension performance on a standardized test when verbal response is used as an accommodation. In addition, the pilot study aimed to examine whether there are differences in performance between general education (GE) students, special education students (SpEd), and English language learners (ELL) when fourth grade students respond verbally to open-response questions on a reading comprehension tests. This research was federally funded through the Enhancing State Assessment grant program and conducted collaboratively with Vermont, Rhode Island, New Hampshire, Maine, the Education Development Center (EDC), and CAST.

Design

To examine whether and how the computer delivery of reading passages with computer-recorded verbal-response accommodation affected the test performance of fourth grade students, 300 students from five Rhode Island schools were randomly assigned to perform the same reading test in one of three modes: 1) paper delivery with written response, 2) computer delivery with written response; and 3) computer delivery with computer-recorded verbal response. The participating schools were selected with the cooperation of the state Director of Assessment. When selecting schools, we aimed to maximize the diversity of the sample in terms of English Language Learners (ELL), special education students, and general education students. Since ELL students tended to be clustered within urban schools, and since the location of the school could not be manipulated, all of the schools selected were urban schools and random assignment occurred within rather than across schools. Consequently, students within schools were randomly assigned to one of the 3 modes of delivery and response – CDVR, CDCR, PDWR.

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 12-inch iBook laptops 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 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 five Rhode Island schools between April 22, 2004 and May 6, 2004. Within each school, researchers first configured the classroom where the CDCR was conducted so that desks were spread out and the laptops could be connected to a power source. Additionally, CDVR testing was conducted either in the cafeteria, computer lab, or library of the school to allow researchers to spread students out to minimize noise disturbances and to prevent students from hearing responses by neighboring students. 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. Moreover, in the CDVR group, researchers briefly described the recording technology to students and reiterated to students the importance of understanding the tutorial presented by the testing program prior to the test session. Students were given one hour to complete the reading comprehension assessment and an additional hour to complete the computer fluidity, literacy, and use instruments.

A total of 300 fourth grade students participated in this pilot study. The number of students per school ranged from 14 to 97. Due to a variety of technical problems, data for 117 students were not used because the data were incomplete, students were not able to use the recording tool properly or students did not respond to the open-ended items due to a lack of familiarity with this response mode. Specifically, 38 students were eliminated because they did not complete more than one of the data collection instruments and 79 students were eliminated because they were not able to or opted not to respond to any of the open-ended items. As we discuss in greater detail in the final section of this report, the large number of technical problems and the discomfort many students experienced when attempting to respond aloud using the computer-based audio recorder suggests that the approach used in this study is not feasible for larger-scale testing. Nonetheless, to examine whether the response mode affected performance for those students who did not experience technical problems and did use the technology to respond to the open-ended items, the statistical analyses focused only on those students in the CDVR group who had open-ended item scores greater than 0.

It must be emphasized that focusing only on students who had scores greater than zero likely overestimates the performance of students in the CDVR group. As seen in the analyses that follow, despite this overestimated mean performance, the CDVR group consistently performed worse than the two other response modes. As a result, the analyses that follow likely underestimate the difference in performance between the CDVR and the two other groups.

The final dataset used for analyses presented here consisted of 183 students. Table 1 summarizes demographic information for the study's sample.

Table 1: Demographic Summary for Participating Students

Demographic Variable	Categories	Frequency	Percentage
Gender	Boy	71	39
	Girl	110	60
	Missing	2	1
Ethnicity	African American or Black	49	27
	Asian Pacific Islander	4	2
	Hispanic or Latino	47	26
	Native American	2	1
	White	25	14
	Other	21	11
	Multiple	12	7
	Unknown	23	12
Language spoken at home	Spanish	67	37
	Portuguese	4	2
	Cape Verdean Creole	19	10
	Hmong	3	2
	Haitian Creole	4	2
	English	51	28
	Unknown	35	19

In addition to the demographics provided in Table 1, 17% of students were classified as ELL, 8% had IEPs, and 75% were general education students. Moreover, 20% of students took the CDVR reading comprehension test; 24% took the CDWR test; and 56% took the PDWR test. Although students were randomly assigned to one of the three testing groups, the majority of students whose data was not used for analyses were those who participated in computer-delivered testing because of either missing data points or technical difficulties which resulted in insufficient data. Therefore, although each of the testing groups started out with roughly the same number of students, the CDVR and CDWR groups experienced the highest percentage of attrition, resulting in considerably smaller samples than the less technically demanding PDWR sample.

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 four reading passages, nineteen associated multiple-choice items, and eight open-ended items was administered to all students participating in this study. All of the multiple-choice and open-response items used were released items from previously administered large scale national or state assessments. Specifically, two passages with a total of nine multiple-choice questions, and five open-ended questions were released National Assessment of Educational Progress (NAEP) items; one passage with five multiple-choice questions, and two open-ended questions was a released Progress in International Reading Literacy Study (PIRLS) item; and one passage with five multiple choice questions and one open-ended question was a released item from New Hampshire's state 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 a multiple-choice question and how to navigate through the questions. Additionally, the tutorial also showed students how to use the audio recorder on the students' computers, how to re-record an answer, how to listen to their answers, and how to control the volume of the recorder and the player. Students were able to listen to the tutorial as many times as needed.


Students were allowed to skip items and change answers to items. Students were also able to "mark for review" items that they wanted to return to later. 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. Figure 1 displays a screen shot of the test interface for the scrolling passage, and the two different answer formats.

Figure 1: Computer-based Reading Test Interfaces

Typed Response:

Reading Passage: Blue Crabs

Blue Crabs
By George W. Frame



Nearly every day last summer my nephew Keith and I went crabbing in a creek on the New Jersey coast. We used a wire trap baited with scraps of fish and meat. Each time a crab entered the trap to eat, we pulled the doors closed. We cooked and ate the crabs we caught.

Blue crabs are very strong. Their big claws can make a painful pinch. When cornered, the crabs boldly defend themselves. They wave their outstretched claws and are fast and ready to fight. Keith and I had to be very careful to avoid having our fingers pinched.

Crabs are **arthropods**, a very large group of animals that have an external skeleton and jointed legs. Other kinds of arthropods are insects, spiders, and centipedes. Blue crabs belong to a particular arthropod group called **crustaceans**. Crustaceans are abundant in the ocean, just as insects are on land.

The blue crab's hard shell is a strong armor. But the

There are 7 questions for this passage

previous questions

20. Describe the appearance of a female blue crab that is carrying eggs.

enter your response below

End questions for this passage

Go to the next reading passage


highlight text remove highlight

Navigation bar 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 End Test

Spoken Response (no answer given):

Reading Passage: Blue Crabs

Blue Crabs
By George W. Frame



Nearly every day last summer my nephew Keith and I went crabbing in a creek on the New Jersey coast. We used a wire trap baited with scraps of fish and meat. Each time a crab entered the trap to eat, we pulled the doors closed. We cooked and ate the crabs we caught.

Blue crabs are very strong. Their big claws can make a painful pinch. When cornered, the crabs boldly defend themselves. They wave their outstretched claws and are fast and ready to fight. Keith and I had to be very careful to avoid having our fingers pinched.

Crabs are **arthropods**, a very large group of animals that have an external skeleton and jointed legs. Other kinds of arthropods are insects, spiders, and centipedes. Blue crabs belong to a particular arthropod group called **crustaceans**. Crustaceans are abundant in the ocean, just as insects are on land.

The blue crab's hard shell is a strong armor. But the

There are 7 questions for this passage

previous questions

20. Describe the appearance of a female blue crab that is carrying eggs.

Record your answer

Record

Listen to your answer

Play

volume

End questions for this passage

Go to the next reading passage

highlight text remove highlight

Navigation bar 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 End Test

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, multiple-choice, and open-ended questions were presented on double-sided pages. The number of words on each line and number of lines per passage were identical to the computer-based format. The font size and style and text wrap were also 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 12-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. All students were allowed one hour to complete the four reading passages, nineteen multiple-choice questions, and eight open-ended questions.

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 as many words as possible 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 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 ten 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 two intermediary points as well as the total 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 completing 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 administer multiple choice-computer literacy tests to students at the fourth grade level. 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 fifth grade student survey constructed for the Use, Support, and Effect 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 in order 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, initial principal component analysis revealed that some of the items did not provide useful information to the scale. Items that did not provide useful information to the scale were dropped. One factor emerged consisting of five items: time and number of clicks for case 2 in the clicking exercise, time for case 3 in the clicking exercise, and time for case 1 and 3 in the drag and drop exercise. The fluidity factor accounted for 52% of the variance and had an alpha reliability coefficient of .72.

Computer Literacy Scale

The computer literacy test consisted of 11 multiple-choice items 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 principal component containing 10 items with loadings ranging from .36 to .72 emerged, which accounted for 27% of the variance. Additionally, weights were applied to 3 of the items: item eight's score was multiplied by .85, item 10 was multiplied by .7, and item 5 was multiplied by .5. This was done to better reflect these items' contribution to the variance in computer literacy scores. The reliability coefficient for the computer literacy scale was .69. The factor score was saved to create a computer literacy scaled score.

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, listen to 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 principal components analysis was run on the items, a one factor solution containing 8 items with loadings ranging from .85 to .95 emerged, which accounted for 83% of the variance. The reliability obtained for the home use scale was .97.

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 eleven school computer use items, an eleven item scale emerged. The factor solution accounted for 78% of the variance and had an alpha reliability of .97. Factor loadings on the four items ranged from .82 to .93.

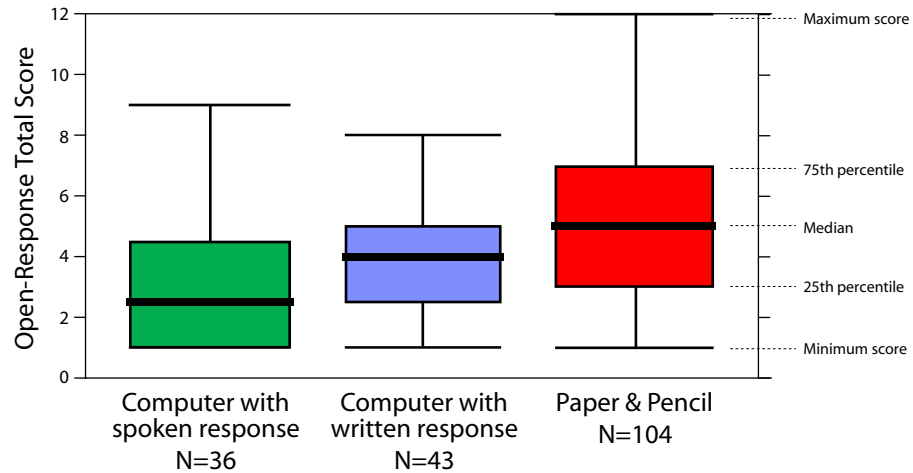
Results

To examine the extent to which the three modes of testing – CDVR, CDWR, and PDWR – affected student performance on a reading comprehension test, a series of analyses were performed. These analyses included descriptive statistics to compare performance across the three treatment groups for all students, analysis of variance (ANOVA) to determine whether group differences were statistically significant, and general linear model (GLM) univariate analyses of variance (UNI-ANOVA) to determine whether group means varied by student status – IEP vs. general education students (GE), and ELL vs. English proficient (GE) students. Additionally, to examine the extent to which prior experience and skill using a computer interacted with the presentation mode, analyses of variance were conducted using the computer fluidity, computer literacy, and computer 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 testing modes.

Examining Test Performance by Delivery Mode

To determine the extent to which the mode of accommodation delivery affected students' reading comprehension performance, the initial stage of data analysis consisted of examining descriptive statistics for all students across the three testing modes – CDVR, CDWR, PDWR. Since the study was affected by both technical difficulties and unfamiliarity with providing responses in a verbal manner, a number of students were not able to respond to the open-response items (OR) and, consequently, obtained a total score of zero for the open-response section. Consequently, given that students who did not respond to any of the items do not provide any valuable information for the analyses reported here, students whose overall score for the open-response section of the test was zero were excluded from analyses for the open-response section of the test. The entire dataset contains data from 300 students. When students with total scores of zero were removed, the remaining number of students in the sample was only 183. Therefore, the analyses here correspond to a sample of 183 students. Additionally, item 20 was removed from analyses as only 13 students responded to the item and it was found to lower the reliability of the open-response section of the test. Finally, it should be noted that since this is a pilot study designed to examine both the feasibility of a computer-based verbal-response mode and the potential effects of such an accommodation, statistical significance at both the .05 and .10 levels will be discussed. We opt to discuss both levels of significance in order to identify patterns or issues that should be examined through future research that would employ larger samples of students.

A box-plot was created to visually compare performance across the three groups (see Figure 2). In addition, descriptive statistics are presented in Table 2.

Figure 2: Box-Plot of Students Performance Across the Three Delivery Modes**Table 2: Descriptive Statistics by Delivery Mode**

Test Mode	N	Mean	Standard Deviation	Standard Error
CDVR	36	3.14	2.43	.40
CDWR	43	4.05	2.24	.34
PDWR	104	5.14	2.97	.29
Total	183	4.49	2.81	.21

The descriptive statistics indicate that the paper-and-pencil group performed better on the open-response section of the test than did the CDWR and the CDVR groups. Specifically, the mean for the PDWR was 5.14 (out of a maximum of 12 points) while the means for CDWR and CDVR were 4.05 and 3.14 respectively. Essentially, both groups responding in writing to the test performed better than the group responding orally.

An ANOVA was run to determine whether the differences observed between group means were statistically significant.

Table 3: ANOVA for Performance by Delivery Mode

	Sum of Squares	df	Mean Square	F	Sig
Between Groups	118.69	2.00	59.34	8.07	.000
Within Groups	1323.05	180.00	7.35		
Total	1441.74	182.00			

As seen in Table 3, the ANOVA indicated that there were statistically significant differences between groups ($F= 8.074, p<.0005$). To determine which group mean differences were statistically significant, an ANOVA with group contrasts was performed. Specifically, three contrasts compared: (1) PDWR and the CDVR group means; (2) CDWR and the CDVR group means; and (3) PDWR and CDWR group means. Results for the ANOVA with contrasts are presented in Table 4.

Table 4: Contrast Tests for Performance by Delivery Mode

Contrast	Value of Contrast	Standard Error	t	df	Sig. (2-tailed)
PDWR - CDVR	2.01	.52	3.83	180.00	.000
CDWR - CDVR	.91	.61	1.48	180.00	.140
PDWR - CDWR	1.10	.49	2.23	180.00	.027

As the contrasts presented in Table 4 indicate, on average, students taking the PDWR test performed 2 points better on the reading comprehension test than did students taking the CDVR test and the difference was statistically significant ($t=3.83, p<.0005$). Similarly, on average, students taking the PDWR test performed 1.1 points better on the reading comprehension test than did students taking the CDWR test and the difference was statistically significant ($t=2.23, p=.027$). Although students performed 0.9 points better on the CDWR as compared to the CDVR, this difference was not statistically significant ($t=1.48, p=.140$).

To examine the magnitude of the differences between group means, effect sizes were computed using Glass's delta¹ and the PDWR group was used as the control group. As seen in Table 5, students in the CDVR performed approximately two-thirds of a standard deviation below the PDWR group while students in the CDWR performed .39 standard deviations below the PDWR group. Thus, taking a the open-ended items using the CDVR version seems to have a larger negative effect while taking the test using the CDWR interface seemed to have a moderate negative effect on the performance of students participating in this study.

Table 5: Effect Sizes for Delivery Modes

Accommodation mode	Effect size
CDVR	-.67
CDWR	-.39

Examining Performance Differences by Student Status

Differences Between Special Education and General Education Students

Analyses of variance were conducted to examine whether there were differences in test performance by student IEP status in each of the testing groups. The first analysis consisted of comparing test performance descriptives for special education students and general education students (not including ELL students) for each of the accommodation modes. Second, an ANOVA was conducted to determine whether there were statistically significant differences between group means. Finally, a UNIANOVA with pair-wise comparisons was conducted to determine which group means were statistically significantly different.

Table 6 contains descriptive statistics for group performance by student IEP status. Although students were randomly assigned to groups of equal sizes, the majority of the students scheduled to participate in the study were from an urban high school with a high absenteeism rate. Therefore, some of the students scheduled to participate were absent on the day of testing, and some refused to take the test. Additionally, a number of students left one or more testing instruments blank, did not write their identification number or used the wrong number. Thus, the final sample contained 14 special education students – 4 in the CDVR group, 5 in the CDWR group, and 5 in the PDWR group. Additionally, there were 137 GE students tested – 26 in the CDVR group, 28 in the CDWR group, and 83 in the PDWR group.

Table 6: Descriptive Statistics by Accommodation Mode and Test Type
Descriptive Statistics(a)

Student Status	Test Type	Mean	Standard Deviation	N
General Education	CDVR	3.62	2.61	26
	CDWR	4.57	2.23	28
	PDWR	5.47	3.01	83
	Total	4.93	2.87	137
Special Education	CDVR	2.25	1.50	4
	CDWR	2.20	.84	5
	PDWR	5.60	3.58	5
	Total	3.43	2.74	14

As Table 6 illustrates, general education students performed better overall in the reading comprehension test than special education students. The overall mean for the GE students was 4.93 whereas the overall mean for SpEd students was 3.43. However, SpEd students performed better on the PDWR test (5.60) than did GE students (5.47). On the other hand, GE students performed better on the computer delivered tests (CDVR and CDWR) than did SpEd students. The mean for GE students on the CDVR test was 3.62 whereas the mean for SpEd

students was 2.25. Likewise, the mean for GE students on the CDWR rest was 4.57 whereas the mean for SpEd students was 2.20. In addition, the GE students and SpEd students performed better on the paper-and-pencil version of the test than they did on the computer delivered tests.

To determine the extent to which each type of response mode affected test performance for special education and GE students, effect sizes were computed for CDVR and CDWR using Glass’s delta. Effect sizes for IEP and GE students are presented in Table 7.

Table 7: Effect Sizes for Accommodation Delivery Modes by IEP Status

Accommodation delivery mode	Student status	Effect size
CDVR	IEP	-.94
	GE	-.62
CDWR	IEP	-.95
	GE	-.3

The computer delivery of the reading comprehension test appears to have had a negative effect on the performance of both GE and IEP students. However, the negative effect may have been more pronounced for IEP students. Specifically, IEP students who took either the CDVR or the CDWR performed on average approximately a standard deviation below the average performance of IEP students who took the PDWR test. Additionally, it appears that responding orally to reading comprehension questions had more of a negative effect on GE students than did responding in writing. Specifically, the effect of CDVR for GE students was twice that of CDWR. The effect for CDVR was -.62 while the effect for CDWR was only -.30.

Having found that performance differences did exist between IEP and GE students, an ANOVA was conducted to determine whether group mean differences were statistically significant. This was followed by univariate analysis of variance (UNIANOVA) to determine which mean differences were statistically significant and whether the accommodation delivery mode by student status interaction was statistically significant. Results for the initial ANOVA are presented in Table 8.

Table 8: ANOVA for Performance by IEP Status

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	107.799	2	53.90	7.00	.001
Within Groups	1138.836	148	7.69		
Total	1246.636	150			

The analysis of variance revealed that group mean differences were statistically significant ($F=7.00, p<.05$). Therefore, one can conclude that student performance on the reading comprehension test varied by students' status. To determine which group means differed significantly, an UNIANOVA was conducted.

Table 9: Tests of Between-Subjects Effects for Test Type by IEP Status

Student Status	Source	Type III Sum of Squares	df	Mean Square	F	Sig.
General Education	Corrected Model	36.68(a)	2	18.34	3.32	.074
	Intercept	155.39	1	155.39	28.14	.000
	Response mode effect	36.68	2	18.34	3.32	.074
	Error	60.75	11	5.52		
	Total	262.00	14			
Special Education	Corrected Model	72.72(a)	2	36.36	4.65	.011
	Intercept	2163.03	1	2163.03	276.65	.000
	Response mode effect	72.72	2	36.36	4.65	.011
	Error	1047.69	134	7.82		
	Total	4456.00	137			

As seen in Table 9, analysis of between subjects' effects revealed that, although GE students performed differently on the 3 modes of test response – CDVR, CDWR, and PDWR – the effect of response mode was not statistically significant for GE students ($F=3.32, p=.074$). The effect of response mode, however, was statistically significant for IEP students ($F=4.65, p<.05$). This suggested that further investigation was needed to determine which response mode means differed significantly for the IEP group. The results for response mode mean comparisons for the IEP group are presented in Table 10.

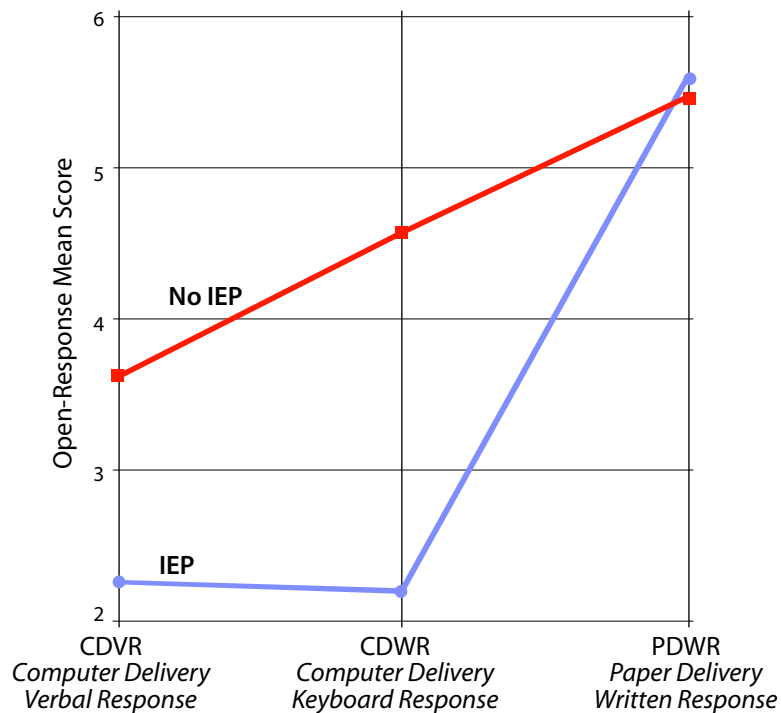
Table 10: Tests of Group Means Differences for Response Mode by IEP Status

Student Status	(I) Test Type	(J) Test Type	Mean Difference (I-J)	Std. Error	Sig.
IEP	CDVR	CDWR	.05	1.58	.98
		PDWR	-3.35	1.58	.06
	CDWR	CDVR	-.05	1.58	.98
		PDWR	-3.40	1.49	.04

The analysis of group mean differences for the IEP group revealed that the only mean difference that was statistically significant at .05 alpha was between CDWR and PDWR. Additionally, IEP students performed better of the PDWR than on the CDVR test; however, the difference between group means was only statistically significant at .10 alpha. Therefore, evidence suggests that IEP students

appear to perform better on reading comprehension tests in the traditional mode of test delivery and response (i.e., paper delivery with written response) than on computer-delivered tests with either written or verbal responses.

Figure 3: Graph of the Test Mode by Student IEP Status Interaction



Differences Between English Language Learners and General Education Students

To examine differences in test performance by student language status, ELL students' performance was compared to English proficient (EP) students (excluding students with IEPs) for each accommodation delivery mode. The final ELL sample contained 32 students – 6 in the CDVR group, 10 in the CDWR group, and 16 in the PDWR group. The EP sample consisted of 137 students – 26 in the CDVR group, 28 in the CDWR group, and 83 in the PDWR group. Table 11 presents descriptive statistics for group performance by student language status.

Table 11: Group Performance by Language Status

Student Language Status	Test Type	Mean	Standard Deviation	N
English Proficient Or GE	CDVR	3.62	2.61	26
	CDWR	4.57	2.23	28
	PDWR	5.47	3.01	83
	Total	4.93	2.87	137
English Language Learner	CDVR	1.67	1.21	6
	CDWR	3.50	2.22	10
	PDWR	3.31	1.82	16
	Total	3.06	1.93	32

Similar to what was found in the analysis for performance by IEP status, descriptive statistics for performance by language status indicate that, overall, EP students performed better on the reading comprehension test than did ELL students. Specifically, the overall mean for EP students was 4.93 whereas the mean for ELL students was 3.06. EP students also outperformed ELL students in all 3 modes of test delivery. As seen in Table 11, ELL students performed worse on the CDVR version of the test (1.67) than they did either on the CDWR (3.50) or the PDWR (3.31) versions of the test. However, ELL students performed better on the CDWR test (3.50) than they did on the PDWR test (3.31). That is, descriptive statistics indicate that, although computer delivery with verbal response hampered ELL students’ performance on the reading comprehension test, computer delivery appears to have slightly increased the performance of ELL students when compared to paper delivery.

Table 12: Effect Sizes of Accommodation Modes by Language Status

Accommodation mode	Student status	Effect size
CDVR	ELL	-.62
	EP	-.30
CDWR	ELL	-.90
	EP	.10

To determine the extent to which the type of response mode affected test performance for ELL and GE students, effect sizes were computed for CDVR and CDWR using Glass’s delta. Effect sizes for CDVR and CDWR are presented in table 11. Computer delivery with verbal response had a negative effect for both ELL and EP students as compared to PDWR. Specifically, on average, ELL students taking the CDVR test performed .62 standard deviations lower than ELL students taking the PDWR test. Likewise, EP students taking the CDVR test performed .30 standard deviations below the average performance for EP students

taking the PDWR test. Similarly, CDWR had a negative effect on the reading comprehension performance of ELL students; on average, English language learners who took the CDWR test performed .90 standard deviations below the average performance of ELL students who took the PDWR test. Conversely, CDWR had a positive effect, albeit small (.10) on the reading comprehension performance of EP students as compared the performance of EP students taking the PDWR test.

Descriptive statistics for groups by language status indicated that differences appear to exist between EP and ELL groups. Therefore, an ANOVA was conducted to determine whether group mean differences were statistically significant and an UNIANOVA was employed to determine which group mean differences were statistically significant. Results for the initial ANOVA are presented in Table 13.

Table 13: ANOVA for Performance by Language Status

Dependent variable: OR

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	88.81	2	44.41	5.95	.003
Within Groups	1238.36	166	7.46		
Total	1327.17	168			

The analysis of variance revealed that mean differences between ELL and GE students were statistically significant ($F=5.95, p<.05$). Therefore, further analyses were conducted to determine if the effect of mode of response to the reading comprehension test, language status, and the interaction between mode of response and language status were statistically significant. The results for performance by response mode and language status are presented in Table 14.

Table 14: ANOVA of Performance by Response Mode and Language Status

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	178.22	5.00	35.64	5.06	.000
Intercept	1179.60	1.00	1179.60	167.35	.000
Response mode	44.07	2.00	22.03	3.13	.047
Language status	64.53	1.00	64.53	9.16	.003
Response mode by Language Status	5.73	2.00	2.87	.41	.667
Error	1148.96	163.00	7.05		
Total	4872.00	169.00			
Corrected Total	1327.17	168.00			

The analysis of variance of performance by response mode and language status for EP and ELL students revealed that the effect of response mode was statistically significant ($F=3.13, p<.05$). Additionally, the analysis of variance revealed that the effect of language status was statistically significant ($F=9.16, p<.05$). On the other hand, the interaction between response mode and language status was not statistically significant ($F=.41, p=.667$). Therefore, one can conclude that the type of response mode had a statistically significant effect on the reading comprehension performance of EP and ELL students and that students' reading comprehension performance varied by language status. To determine which response-mode group means differed significantly, pair-wise mean comparisons were conducted.

Table 15: Pair-wise Comparisons for Group Means by Test Mode

Dependent Variable: Open response total score

Pairwise comparison	Mean Difference	Std. Error	Sig.
CDVR-CDWR	-1.39	.78	.074
CDVR-PDWR	-1.75	.70	.014
CDWR-PDWR	-.36	.61	.560

As Table 15 shows, the only mean difference that was statistically significant was between CDVR and PDWR ($p<.05$). That is, on average students performed 1.75 points better on the PDWR test and that the difference between means is statistically significant.

To determine whether the effect of response mode was statistically significant within each language status group, an ANOVA was conducted with the file split by language status.

Table 16 indicates that the effect of response mode is only statistically significant for the ELL group ($F=4.65, p<.05$). Therefore, one can conclude that the reading comprehension performance of ELL students varied by response mode. Pair-wise comparisons were then conducted to determine which group means differed significantly within the ELL language group.

Table 16: ANOVA Effect of Response Mode by Language Status

Student Status	Source	Type III Sum of Squares	df	Mean Square	F	Sig.
English Language Learner	Corrected Model	72.72	2.00	36.36	4.65	.011
	Intercept	2163.03	1.00	2163.03	276.65	.000
	Response mode	72.72	2.00	36.36	4.65	.011
	Error	1047.69	134.00	7.82		
	Total	4456.00	137.00			
	Corrected Total	1120.41	136.00			
English Proficient	Corrected Model	14.60	2.00	7.30	2.09	.14
	Intercept	218.42	1.00	218.42	62.55	.00
	Response mode	14.60	2.00	7.30	2.09	.14
	Error	101.27	29.00	3.49		
	Total	416.00	32.00			
	Corrected Total	115.87	31.00			

As seen in Table 17, pair-wise comparisons indicated that the only statistically significant mean difference occurred between CDVR and PDWR. This suggests that ELL students performed better in the reading comprehension test administered by paper-and-pencil and responded to in writing when compared to ELL students' performance on the test with computer delivery and verbal response.

Table 17: Performance by Language Status and Response Mode

Student Status	Pairwise Comparison	Mean Difference (I-J)	Standard Error	Sig.
English Proficient	CDVR-CDWR	-.96	.76	.212
	CDVR-PDWR	-1.85	.63	.004
	CDWR-PDWR	-.90	.61	.144

Examining Score Differences by Prior Computer Skills and Use

The next step in exploring how the response mode affected students' reading comprehension scores was to examine the relationship between students' performance on the reading comprehension test with their performance on the computer fluidity test, computer literacy test, and computer use survey. To this end, students were divided into three groups based on their scaled scores. This grouping was conducted separately for the computer fluidity, computer literacy, home use, and school use scales. For each scale, the three groups represented high, medium, and low levels for computer fluidity, computer literacy, or computer use. For each scale, students were divided into three groups according to their scale scores: students in the bottom third (zero to 33.3 percentile) were classified as low level, students in the middle third of the scale (33.4 to 66.6 percentile) were classified as medium level; and students in the top third of the scale (66.7 to 100th percentile). After these three groups were formed, reading comprehension scores were compared across the three response modes based on the high-medium-low classification of each computer measure.

Examining Performance by Computer Literacy Level

Analysis of student performance by computer literacy consisted of two steps. First, analysis of variance was conducted for all students to determine whether students' performance overall varied by computer literacy level. Second, to determine whether computer literacy affected student performance by response mode, ANOVAs were conducted with the dataset split by computer literacy level. Table 18 presents results for the performance of all students by computer literacy level.

Table 18: Performance by Computer Literacy Level for All Students

Literacy Level	N	Mean	Standard Deviation	Standard Error	df	Mean Square	F	Sig.
Low	62	4.19	2.54	.322	2	5.32	.67	.514
Medium	60	4.78	2.77	.357				
High	61	4.50	3.12	.399				
Total	183	4.49	2.81	.210				

Comparisons of descriptive statistics for reading comprehension performance by computer literacy revealed that there was no statistically significant difference between low, high, and medium literacy groups ($F=.67$, $p=.514$). To examine whether the differences in performance across modes within the three levels of computer literacy were statistically significant, analyses of variance were conducted with the file split by computer literacy level.

Table 19: Performance by Computer Literacy Level

Literacy Level	Response Mode	N	Mean	Std. Deviation	Std. Error	df	Mean Square	F	Sig
Low	CDVR	13	3.23	2.13	.59	2	3.68	2.204	.119
	CDWR	10	3.85	2.70	.60				
	PDWR	37	4.86	4.86	2.49	.46			
	Total	60	4.19	4.19	2.54	.32			
Medium	CDVR	9	3.78	2.95	.98	2	11.03	1.462	.24
	CDWR	13	4.15	1.99	.55				
	PDWR	38	5.23	2.90	.47				
	Total	60	4.78	2.77	.36				
High	CDVR	14	2.64	2.41	.64	2	35.32	3.980	.024
	CDWR	10	4.30	1.57	.50				
	PDWR	37	5.27	3.40	.56				
	Total	61	4.51	3.12	.40				

As seen in Table 19, differences in performance across modes was statistically significant for the high literacy group, only, although the low literacy group appeared to have been negatively affected by CDVR, the differences in performance between computer literacy groups are not statistically significant ($F=2.2$, $p=.12$). On the other hand, reading comprehension performance appears to have been affected by computer literacy for students classified as high literacy ($F=4.0$, $p<.05$). Consequently, further analyses are needed to investigate which mode of delivery affected the reading comprehension performance of high literacy students. Pair-wise comparison results for the high literacy group are presented in Table 20.

Table 20: Pair-wise Comparisons for High Literacy Level

Literacy Level	Comparisons	Mean Difference	Std. Error	t	df	Sig.
High	PDWR-CDVR	2.63	.85	13.08	33.28	.004
	PDWR-CDWR	.97	.75	1.29	33.13	.203
	CDWR-CDVR	1.66	.81	2.04	21.88	.053

Analyses of mean differences for the high computer literacy group revealed that the difference between PDWR and CDVR was statistically significant ($t=13.08$, $p=.004$). Similarly, the difference between the means of CDWR and CDVR was statistically significant at an alpha level of .10 ($t=2.04$, $p=.053$). The difference between PDWR and CDWR, however, was not statistically significant ($t=1.29$, $p=.203$). Thus, it appears that students with high computer literacy were not significantly affected by performing the open-ended items on computer using a keyboard, but were significantly negatively affected by responding orally using the computer-based audio recorder.

Examining Performance by Computer Fluidity Level

Analysis of student performance by computer fluidity consisted of two steps. First, an analysis of variance was conducted to determine whether students' performance difference across the three levels of computer fluidity. Second, to determine whether computer fluidity affected student performance by response mode, ANOVAs were conducted with the dataset split by computer fluidity level.

Table 21: Performance By Computer Fluidity Level For All Students

Fluidity Level	N	Mean	Standard Deviation	Standard Error	df	Mean Square	F	Sig.
Low	66	4.09	2.79	.343	2	11.5	1.4	.254
Medium	56	4.82	3.09	.413				
High	61	4.84	2.79	.358				
Total	183	4.56	2.89	.214				

As seen in Table 21, students in the high and medium fluidity groups, on average, performed better than did students in the low fluidity group. However, analysis of variance indicated that no statistically significant differences existed among the groups ($F=1.4$, $p=.25$). To examine whether performance varied by test type within each fluidity level, additional analyses of variance were conducted with the file split by computer fluidity level.

Table 22: Performance by Computer Fluidity Level

Fluidity Level	Response Mode	N	Mean	Standard Deviation	Standard Error	df	Mean Square	F	Sig.
Low	CDVR	12	2.67	1.72	.497	2	15.65	2.08	.133
	CDWR	15	4.13	1.85	.477				
	PDWR	39	4.51	3.22	.515				
	Total	66	4.09	2.79	.343				
Medium	CDVR	12	2.67	2.77	.801	2	35.53	4.14	.021
	CDWR	10	5.30	3.529	1.116				
	PDWR	34	5.44	2.798	.480				
	Total	56	4.82	3.093	.413				
High	CDVR	12	4.17	2.588	.747	2	35.35	5.16	.009
	CDWR	18	3.50	1.689	.398				
	PDWR	31	5.87	3.030	.544				
	Total	61	4.84	2.794	.358				

As seen in Table 22, reading comprehension performance on the open-ended items did not vary significantly between response modes for students with low computer fluidity ($F=2.08, p=.133$). However, analyses of variance for the medium and high fluidity groups indicate that student reading comprehension performance varied among response modes. To examine which testing mode differed significantly within each of the two fluidity levels, pair-wise comparisons were conducted within the medium and the high fluidity levels.

Table 23: Pairwise Comparisons for Response Mode by Fluidity Level

Fluidity Level	Contrast	Value of Contrast	Standard Error	t	df	Sig
Medium	PDWR-CDVR	2.77	.984	2.820	53	.007
	PDWR-CDWR	.14	1.054	.134	53	.894
	CDWR-CDVR	2.63	1.255	2.1	53	.041
High	PDWR-CDVR	1.70	.890	1.914	58	.060
	PDWR-CDWR	2.37	.776	3.056	58	.003
	CDWR-CDVR	-.67	.976	-.683	58	.497

As seen in Table 23, pair-wise comparisons for the medium fluidity group indicate that the means of PDWR and CDVR differed significantly ($t=2.8, p<.05$). In addition, the means of CDWR and CDVR differed significantly ($t=2.1, p<.05$). Thus, it appears that medium fluidity students performed significantly worse when using the computer-based audio recorder than when they wrote their responses by hand or entered them using a keyboard. For high fluidity students, however performance was significantly better on paper than on either computer-based

response mode. For the high fluidity group, the means of PDWR and CDVR differed significantly at the alpha 0.1 level ($t=1.9, p<.10$). The means of CDWR and CDVR differed significantly ($t=3.1, p<.05$) at the alpha 0.05 level.

Examining Performance by School Computer Use

To examine the relationship between school computer use and performance on the open-ended reading comprehension items under the three response modes, an analysis of variance was conducted to determine whether test performance varied by school computer use.

Table 24: Performance by School-Use Level for All Students

School Use Level	N	Mean	Standard Deviation	Standard Error	df	Mean Square	F	Sig.
Low	60	4.55	2.94	.379	2	11.3	1.44	.241
Medium	63	4.05	2.60	.326				
High	60	4.90	2.90	.374				
Total	183	4.49	2.82	.208				

As seen in Table 24, students reporting high levels of school computer use performed better on the reading comprehension open-ended items than did students in the low use level while students in the medium use level performed the poorest. However, differences among the three school computer use groups were not statistically significant ($F=1.44, p=.241$). To determine whether reading comprehension performance differed across the three modes within each level of school use, analyses of variance were performed within each level of school use.

Table 25: Performance within each School-Use Level by Test Type

School Use Level	Response Mode	N	Mean	Standard Deviation	Standard Error	df	Mean Square	F	Sig.
Low	CDVR	13	2.54	1.71	.475	2	42.47	5.71	.005
	CDWR	16	4.25	2.72	.680				
	PDWR	31	5.55	3.04	.546				
	Total	60	4.55	2.94	.379				
Medium	CDVR	11	2.73	2.10	.634	2	16.95	2.67	.078
	CDWR	12	3.50	1.83	.529				
	PDWR	40	4.58	2.77	.438				
	Total	63	4.05	2.59	.326				
High	CDVR	12	4.17	3.13	.903	2	11.31	1.36	.264
	CDWR	15	4.27	2.02	.521				
	PDWR	33	5.45	3.10	.540				
	Total	60	4.90	2.89	.374				

As seen in Table 25, performance differed significantly across modes within the low school computer use group ($F=5.71, p=.005$). Specifically, within the low school computer use group, students performed worse in the CDVR mode than they did in the CDWR and the PDWR modes. Similarly, performance by test mode also varied for the medium school use level albeit at a .10 alpha level ($F= 2.67, p=.08$). To determine which test type group means differed significantly within these levels of school use, pair-wise comparisons were conducted.

Table 26: Pairwise Comparisons by Response Mode for Low School Use Level

School Use Level	Contrast	Value of Contrast	Standard Error	t	df	Sig.
Low	PDWR-CDVR	3.01	.90	3.340	57	.001
	PDWR-CDWR	1.30	.84	1.547	57	.127
	CDWR-CDVR	1.71	1.02	1.681	57	.098
Medium	PDWR-CDVR	1.85	.86	2.154	60	.035
	PDWR-CDWR	1.08	.83	1.296	60	.200
	CDWR-CDVR	.77	1.05	.735	60	.465

As seen in Table 26, differences between PDWR and CDVR were statistically significant within the low school computer use group ($t=3.34, p=.001$) and the medium school computer use group ($t=2.15, p=.035$). For the low school computer use group, the difference between CDWR and CDVR was statistically significant at a .10 level of significance ($t=1.68, p=.098$). For both the low and medium computer use groups, it appears that students performed significantly worse when using the CDVR mode as compared to the PDWR mode.

Examining Performance by Home Computer Use

To examine whether performance on the reading comprehension open-ended items varied by home computer use, an analysis of variance was conducted. As seen in Table 27, students who reported higher levels of home computer use performed better, on average, than students with medium or low levels of home computer use.

Table 27: Performance by Home Computer Use for All Students

DV: Open Response Total Score

Home Use Level	N	Mean	Standard Deviation	Standard Error	df	Mean Square	F	Sig.
Low	61	3.98	2.86	.3663	2	18.049	2.311	.102
Medium	61	4.43	2.51	.3209				
High	61	5.07	2.99	.3833				
Total	183	4.49	2.82	.2081				

To examine whether differences across the mode of administration occurred within each level of home computer use, an analysis of variance was performed within each level of home computer use.

Table 28: Performance by Test Type for Home Computer Use Levels

School Use Level	Response Mode	N	Mean	Standard Deviation	Standard Error	df	Mean Square	F	Sig.
Low	CDVR	13	2.00	1.15	.32	2	33.43	4.57	.014
	CDWR	14	4.21	2.64	.70				
	PDWR	34	4.65	3.10	.53				
	Total	61	3.98	2.86	.37				
Medium	CDVR	8	3.00	2.07	.73	2	14.81	2.47	.093
	CDWR	13	3.85	2.44	.68				
	PDWR	40	4.90	2.51	.40				
	Total	61	4.43	2.51	.32				
High	CDVR	15	4.20	3.00	.78	2	27.72	3.33	.043
	CDWR	16	4.06	1.77	.44				
	PDWR	30	6.03	3.26	.60				
	Total	61	5.07	2.99	.38				

As seen in Table 28, significant differences among testing modes were found within the low ($F=4.57, p=.014$) and high ($F=3.33, p=.043$) home computer use groups. Differences at the .10 level of significance were also found within the medium home computer use group ($F=2.47, p=.093$). Within all three groups,

students performed better on the PDWR version. Within the low and medium home computer use groups, students performed worse on the CDVR mode. To examine which differences between modes within each group were statistically significant, pair-wise comparisons were performed within each level of home computer use.

Table 29: Pair-wise Comparisons by Test Type for Home Use Levels

School Use Level	Contrast	Value of Contrast	Standard Error	t	df	Sig.
Low	PDWR-CDVR	2.65	.62	4.26	45.00	.000
	PDWR-CDWR	.43	.88	.49	28.41	.628
	CDWR-CDVR	2.21	.77	2.86	18.10	.010
Medium	PDWR-CDVR	1.90	.95	2.01	58	.050
	PDWR-CDWR	1.05	.78	1.35	58	.183
	CDWR-CDVR	.85	1.10	.77	58	.445
High	PDWR-CDVR	1.83	.98	1.87	30.30	.071
	PDWR-CDWR	1.97	.74	2.66	43.96	.011
	CDWR-CDVR	-.14	.89	-.15	22.37	.879

As seen in Table 29, significant differences were found between the PDWR and CDVR for the low ($t=4.26$, $p<.05$) and medium ($t=2.01$, $p=.05$) home computer use groups. Similar differences were found at the .10 level for the high home computer use group ($t=1.87$, $p=.071$). Within the low home computer use group, a significant difference was also found between the CDWR and CDVR modes ($t=2.86$, $p=.01$). Within the high home computer use group, the only other significant difference occurred between the PDWR and CDWR modes ($t=2.66$, $p=.011$). Within each of the levels of home computer use, it appears that students performed significantly worse when using the computer-based voice recorder as compared to paper-and-pencil. Within the low home computer use group, performance was also worse when using the computer-base voice recorder than when using the keyboard.

Discussion

The pilot study presented here was conducted to examine the feasibility and effect of allowing students to respond orally to open-ended reading questions. To record students' verbal responses, a computer-based voice recorder was employed. As described above, the use of a computer-based voice recorder presented major technological challenges. In an effort to reduce problems that might arise by using older, less stable computers that existed in each of the participating schools, the research team brought a set of newly purchased laptops that were optimized for the computer-based voice recorder. Despite this, however, several technical problems occurred. In some cases, students' attempts to record responses to the same item multiple times caused the computer to freeze. At other times, students' failure to stop the recorder after they spoke their response resulted in very large audio files. This occurred even though the test delivery interface was programmed to stop recording a given response after two minutes. In some cases, these large files caused a slow down in the presentation of items. In still other cases, memory problems occurred when students attempted to play back responses repeatedly or when they attempted to play back responses to multiple items in rapid succession. In short, despite several pilot tests with the interface and the use of modern, specially configured computers, the test delivery interface proved to be unstable under the actual testing conditions. It should be noted that this lack of stability would likely have been even more problematic if the test was delivered via the Internet rather than from the hard drive of each laptop.

In addition to the technical challenges we encountered, students' lack of familiarity with a verbal-response option proved problematic for many students. Based on observations and informal interviews with students, a substantial portion of students felt extremely uncomfortable providing responses orally. In some cases, students invested considerable time constructing a response in their head before attempting to record a response. In one case, a student was observed staring at one open-ended item for approximately ten minutes. When asked if he needed help, the student indicated that he was trying to figure out exactly how to say his response before he began recording. In another case, a student who skipped all of the open-ended responses indicated that he did so because he felt that he could not say what he was thinking. When asked if he could write what he was thinking, the student indicated that he could. In other cases, students invested considerable time recording their response, listening to it, then recording it again multiple times. In short, for most of the students in the computer-based voice recording group, the process of responding aloud to a test item seemed very unfamiliar. Moreover, using a computer to record their voice appeared to be a foreign experience.

Despite the technical problems encountered and the challenges many students appeared to have responding aloud to open-ended test items, analyses of students responses to the open-ended items was performed to examine the extent to which the response mode seemed to affect students' performance. To reduce the effect of

technical problems and unfamiliarity of students with the respond aloud option, students who received zero points for their open-ended responses were removed from the analyses. Despite the removal of these students, however, students who used the computer-based voice recorder performed worse than students using the paper-based or keyboard response options. These differences consistently occurred across IEP and ELL status and within each level of computer fluidity, computer literacy and computer use.

Given that the technology we used to deliver the computer-based voice recorder version of the test is more modern and stable than most computer equipment in elementary schools, yet significant technical problems still occurred, we do not believe that this response mode will be feasible for several years. Moreover, until students become more accustomed to responding aloud to open-ended test items, it seems that the added construct of responding verbally will continue to add construct irrelevant variance that negatively affects students' performance. For these reasons, we do not believe that further research on a computer-based voice recorded response for reading comprehension tests should be a high priority for state testing programs. Nonetheless, should future research be conducted, attention should be paid to the interactions between prior computer use, prior experiences with verbal responses, and test performance. As we found in our analyses with relatively small sample sizes, prior computer experiences do seem to affect student performance when computer-based tools are employed. It seems logical that prior experiences with verbal response to test items would have similar affects.

Endnote

- 1 Glass's delta= Mean of treatment group-Mean of control group/standard deviation of control



References

- Abedi, J., Hofstetter, C., Baker, E., & Lord, C. (2001). *NAEP math performance and test accommodations: Interactions with student language background*. (CSE technical report No. CSETR536). U.S.; California: University of California.
- Allington, R. & McGill-Franzen, A. (1992). Improving school effectiveness or inflating high-stakes test performance? *ERS Spectrum*, 10(2), 3–12.
- Chin-Chance, S. A., Gronna, S. S., & Jenkins A. A. (1996, March). Assessing special education students in a norm-referenced statewide testing program: Hawaii State Department of Education. Paper presented at the meeting of the State Collaborative on Assessment and Student Standards (SCASS) Assessing Special Education Students consortium, Washington, DC.
- Langenfeld, K. L., Thurlow, M. L., & Scott, D. L. (1997). *High-stakes testing for students: Unanswered questions and implications for students with disabilities* (Synthesis Report 26). Minneapolis: University of Minnesota, National Center on Educational Outcomes.
- Mondale, S. & Patton, S. B. (Eds.). (2001). *School: The story of American public education*. Boston: Beacon Press.
- National Center for Education Statistics (2004).
Available from: <http://nces.ed.gov>
- National Research Council (1999). *High Stakes: Testing for Tracking, Promotion, and Graduation*, J.P. Huebert and R.M. Hauser, eds. Committee on Appropriate Test Use, Board on Testing and Assessment, Commission on Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.
- National Research Council (1997). *Educating One and All: Students with Disabilities and Standards-Based Reform*, L.M. McDonnell, M.L. McLaughlin, and P. Morison, eds. Committee on Goals 2000 and the Inclusion of Students with Disabilities, Board on Testing and Assessment. Washington, DC: National Academy Press.
- Orfield, G. & Kornhaber, M. (Eds.). (2001). *Raising standards or raising barriers?: Inequality and high-stakes testing in public education*. New York: Century Foundation Press.
- Salvia, J. & Ysseldyke, J. E. (1998). *Assessment* (Seventh Edition). Boston: Houghton Mifflin Company.
- Schulte, A.G., Elliot, S., & Kratochwill, T. (2001). Effects of testing accommodations on standardized mathematics test scores: An experimental analysis of the performance of students with and without disabilities. *School Psychology Review*, 30(4), 527–547.

- Sireci, S. G., Li, S., & Scarpati, S. (2003). *The effects of test accommodations on test performance: A review of the literature*. Center for Educational Assessment (Research Report No. 485), Amherst, MA: School of Education, University of Massachusetts Amherst.
- Sireci, S. G. (2004). *Validity issues in accommodating NAEP reading tests*. Center for Educational Assessment (Research Report No. 515), Amherst, MA: School of Education, University of Massachusetts Amherst.
- Thompson, S., Blount, A., & Thurlow, M. (2002). *A summary of research on the effects of test accommodations: 1999 through 2001* (Technical Report 34). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved September 2004, from the World Wide Web: <http://education.umn.edu/NCEO/OnlinePubs/Technical34.htm>
- Thurlow, M., & Bolt, S. (2001). *Empirical support for accommodations most often allowed in state policy* (Synthesis Report 41). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Tindal, G., Heath, B., Hollenbeck, K., Almond, P., & Harniss, M. (1998). Accommodating students with disabilities on large-scale tests: An experimental study. *Exceptional Children*, 64(4), 439–450.
- Ysseldyke, J. E., Thurlow, M. L., Langenfeld, K., Nelson, J. R., Teelucksingh, E., & Seyfarth, A. (1998). *Education results for students with disabilities: What do the data tell us?* (Technical Report 23). Minneapolis University, National Center on Education Outcomes.

