



inTASC

TECHNOLOGY AND ASSESSMENT
STUDY COLLABORATIVE

**Examining the Feasibility and Effect
of a Computer-Based Read-Aloud
Accommodation on Mathematics
Test Performance**

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

Helena Miranda, Michael Russell, & 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 a Computer-Based Read-Aloud Accommodation on Mathematics Test Performance

Helena Miranda, Michael Russell & Thomas Hoffmann
Technology and Assessment Study Collaborative
Boston College
Released November 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.necompact.org)

Copyright © 2004 Technology and Assessment Study Collaborative, Boston College



Examining the Feasibility and Effect of a Computer-Based Read-Aloud Accommodation on Mathematics Test Performance

Helena Miranda, Michael Russell, and Thomas Hoffmann
Technology and Assessment Study Collaborative
Boston College

Released November 2004

Background/Purpose

The number of learning disabled students (LD) and English language learners (ELL) in American schools has increased over past thirty years. During the same period, a number of laws (e.g., IDEA, the Civil Rights Act, and ESEA) were passed that required states to include exceptional students in their statewide testing programs and to provide those students with the necessary testing accommodations. More recently, the accountability system implemented by the No Child Left Behind Act (NCLB) requires that all students in grades three to eight be tested each year in English Language Arts and mathematics.

Increased testing required by the NCLB not only places pressure on states from administrative and financial standpoints, but it also attaches consequences to test performance at the state, district, school, and, in some cases, student levels. At the school level, sanctions for underperformance in statewide assessments may include withholding federal funds and restructuring administrative functions. At the student level, poor performance in state assessments may result in grade retention or, at the high school level, failure to receive a diploma. Faced with an increasing number of LD and ELL students, schools must find alternatives to make state assessments accessible to exceptional students to avoid sanctions. Accommodations provide opportunities to include a higher percentage of LD and ELL students in state testing programs.

Accommodations are defined as changes made to test administration procedures that remove irrelevant access barriers without changing the construct measured by the assessment (Thurlow, Ysseldyke, & Silverstein, 1993; Thurlow, Elliot,

& Ysseldyke, 1998; Kosciolk, & Ysseldyke, 2000). Although widely used and required by law, accommodations are controversial and problematic both from a logistic and from a psychometric standpoint. From a logistic perspective, testing LD and ELL students without accommodations simplifies test administration procedures, but the resulting test scores may not reflect what students know or can do. On the other hand, providing accommodations to some students and not to others may give students tested with accommodations an unfair advantage. From a psychometric perspective, it is difficult to remove irrelevant barriers (Schulte, Elliott, & Kratochwill, 2000) to the construct being measured through changes in test administration while ensuring that the construct of interest is not changed as a result of the accommodations (Elliott, McKeivitt, & Kettler, 2002).

The delicate balance between providing accommodations to some students while preserving the underlying construct being measured is at the crux of the controversy surrounding accommodations. Specifically, the accommodations' debate stems from the validity of inferences made from accommodated test scores (Schulte, Elliott, & Kratochwill, 2000). Validity evidence is usually associated with standardized testing conditions that do not preclude testing with accommodations. Given that NCLB requires test results to be reported for all student groups, test data for LD and ELL students must be disaggregated and additional validity evidence must be provided (Bielinski, Thurlow, Ysseldyke, Freidebach, & Freidebach, 2001). By definition, accommodations must not alter the construct being measured by the test. Therefore, validity evidence for accommodated test scores is obtained when accommodations result in score gains for LD or ELL students but have little or no effect on the general student population (Bielinski, Thurlow, Ysseldyke, Freidebach, & Freidebach, 2001; Sireci, Li, & Scarpati, 2003; Elliot, McKeivitt, & Kettler, 2002; Koenig, 2002). That is, there should be a status (i.e., disability, language) by testing condition (i.e., accommodated or standard) interaction (Shepard, Taylor, & Betebenner, 1998; Koenig, 2002).

Of all types of accommodations provided to students, the read-aloud accommodation is perhaps the most controversial yet most widely used accommodation (Bielinski, Thurlow, Ysseldyke, Freidebach, & Freidebach, M., 2001). Essentially, the accommodation consists of reading test directions and items aloud to a group of students. Similar to other accommodations, the premise for providing the read-aloud accommodations is that the accommodation removes irrelevant barriers (e.g., poor reading ability, reading disability or coding difficulties) to the construct being measured (i.e., content area knowledge) and gives students with reading disabilities and ELL students more access to the test material (Bielinski, Thurlow, Ysseldyke, Freidebach, & Freidebach, M., 2001; Abedi, Hofstetter, & Baker, 2001; Tindall, Heath, Hollenbeck, Almond, & Harniss, 1998).

Given the frequency with which it is used, the read-aloud accommodation is one of the most investigated accommodations and it is generally thought to be effective in increasing test scores of ELL and LD students (Tindal, Heath, Hollenbeck, Almond, & Harniss, 1998; Helwig, Tedesco, Heath, Tindal, & Almond, 1999; Kosciolk, & Ysseldyke, 2000). However, the read-aloud accommodation

is problematic both from an administration perspective and from a psychometric perspective. First, the pace of administration is set by the human reader. Therefore, the accommodation may affect students' ability to take the test at their own pace. Although students are able to request changes in pace, they very seldom do so. Second, this type of administration may introduce additional sources of variation into the testing environment due to difficulties controlling the delivery of the accommodation. Specifically, since the accommodation is delivered by a human reader, it is difficult to control the quality of reading (e.g., diction, accent, inaccurate reading of words, interpretation problems). Most importantly, the delivery of the read-aloud accommodation using human readers may introduce a "cuing" effect (Tindall, Heath, Hollenbeck, Almond, & Harniss, 1998). That is, proctors may influence students' responses by emphasizing a particular word or phrase or by using facial expressions to indicate approval or disapproval of students' answers (Landau, Russell, Gourgey, Erin, & Cowan, 2003). Essentially, using a human reader to deliver the read-aloud accommodation may not be the most adequate way of removing irrelevant construct barriers in testing LD and ELL students while ensuring the validity of accommodated test results.

Computer-based text-to-speech tools offer an alternative to delivering the read-aloud accommodation using a human reader. For LD and ELL students who have trouble decoding, computer-based text to speech tools can provide the support they need to demonstrate content knowledge while retaining the independence experienced by general education students in testing situations. Computer-based text-to-speech technology can allow students to read and reread test passages and responses at their own pace and as many times as they need.

Research on the effect of the read-aloud accommodation consists predominantly of research on the effect of delivery using a human reader. In addition, a handful of studies have investigated the effect of delivering the accommodation using audiocassettes or video equipment. However, very few studies have examined the effect of providing a read-aloud accommodation using computers. Given the lack of research on the effect of providing a computer-based read aloud accommodation, the summary of prior research on read-aloud accommodations that follows focuses on all modes of delivering the accommodation.

Evidence collected from most studies on the effect of the read-aloud accommodation support the notion that the accommodation may help increase the scores of ELL and LD students and have a small effect on the performance of general education students (Tindal, Heath, Hollenbeck, Almond, & Harniss, 1998; Helwig, Tedesco, Heath, Tindal, & Almond, 1999; Kosciulek, & Ysseldyke, 2000). For instance, in a study that examines the effect of the read-aloud accommodation delivered by a proctor on mathematics test scores, Tindal, Heath, Hollenbeck, Almond, and Harniss (1998) reported that LD students who took the test with the read-aloud accommodation scored statistically significantly higher than LD students who took the test without accommodations. Conversely, general education students who took the test with the accommodation scored slightly higher

than general education students who took the test without the accommodation but the difference was not statistically significant. Moreover, this study confirmed the presence of a significant status by accommodation interaction effect, providing validity evidence for accommodated test scores.

To confirm whether previous findings would hold in standardized administrations of the read-aloud accommodation – i.e., by removing the effect of “cuing” – Helwig, Tedesco, Heath, Tindal, and Almond (1999) conducted a follow-up study in which the read-aloud accommodation was provided via videotape. Although no differences were found between accommodated and non-accommodated scores, further analyses revealed significant differences for students with low reading ability and for items with complex language, indicating that the accommodation was effective for poor readers and for test items with complex language structures.

Similarly, Kosciolk and Ysseldyke (2000) studied the effect of a standardized administration of the read-aloud accommodation using audiocassettes. Specifically, Kosciolk and Ysseldyke investigated the effect of the read-aloud accommodation on the reading comprehension performance of elementary students. Although no significant interaction was found between accommodation and student status, the read-aloud accommodation had a medium size effect on the scores of LD students, and a minimal effect on the scores of general education students. Moreover, when surveyed about their preferences in test taking, LD students preferred taking the test with accommodations, while general education students preferred to take the test without accommodations.

Studies conducted by Tindal et. al (1998), Helwig et. al (1999) and Kosciolk and Ysseldyke (2000) indicated that the read-aloud accommodation may have a positive effect on the performance of LD students. However, other studies indicated that the read aloud accommodation may not benefit all LD students. For instance, Bielinski, J., Thurlow, M., Ysseldyke, J., Freidebach, J., & Freidebach, M. (2001) examined data from the Missouri Assessment Program to study the effect of the read-aloud accommodation on item characteristics of third and fourth grade math and reading multiple-choice items. Results for the math test indicated that item difficulty estimates for LD students who took the math test using the read-aloud accommodation were not significantly different from the reference group – i.e., non-disabled students taking the test without accommodations. Furthermore, the authors found that test scores for the LD students measured the same trait as they did for the general education students, regardless of whether LD students received the accommodation, thus providing evidence for the validity of accommodated test scores. However, the authors concluded that the results from the math test provided a mixed picture: on the one hand, findings indicated that the use of the read-aloud accommodation did not significantly alter item difficulty estimates; on the other hand, only a small subgroup of LD students appeared to benefit from the read-aloud accommodation indicating that for some LD students, the read-aloud accommodation may have been unnecessary.

Studies investigating the effect of computer-based delivery of the read-aloud accommodation also convey mixed results. A study reported by Brown and Augustine (2001) did not find a significant effect of computer-based delivery of the read-aloud accommodation. Specifically, Brown and Augustine (2001) compared the effect of using screen-reading software to paper-and-pencil administration on the of high school students' science and social studies testing performance. Students in this study took two versions of the assessments: a paper-and-pencil version and a computer-based version using screen-reading software. Results revealed that students' reading ability had a significant effect on both their social studies and science test performance. However, the authors did not find a significant effect of mode of administration (paper-and-pencil vs. computer-based) after controlling for reading ability. Nonetheless, Brown and Augustine (2001) caution that providing the read-aloud accommodation to students with poor reading skills may be an effective way of testing students with LD since the lack of significant results from this study may have been a factor of inappropriate instruction in the content area for students with poor reading skills.

Although Brown and Augustine (2001) did not find significant differences when comparing paper-and-pencil test administration without accommodations to computer-based test delivery with read-aloud accommodations, other studies indicated that computer-based delivery of the read-aloud accommodation tends to increase test scores of LD students and have a minimal effect on the scores of general education students. For instance, in a follow-up study to Tindal et. al's (1998) and Helwig et. al's (1999) prior studies, Hollenbeck, Rozek-Tedesco, Tindal, and Glasgow (2000) compared teacher-paced video accommodation (TPV) to student-paced computer accommodation (SPC). Results indicated that LD students performed significantly better with the SPC accommodation than with the TPV accommodation, and that the effect of SPC was much higher than the effect of TPV.

Similarly, Calhoun, Fuchs, and Hamlett (2000) conducted a study to compare the effects of various modes of delivering the read-aloud accommodation to test administration without accommodations on secondary students' math performance. Students took four parallel forms of a math performance assessment over a period of four weeks under a different testing condition each time they took the test. The conditions included: (a) standard administration without accommodations (SA), (b) teacher read (TR), (c) computer-based (CR), and (d) CR with video (CRV). Results from this study indicated that, in general, providing a reading accommodation increased students' test scores but no significant differences were found between the three modes of delivering the read-aloud accommodation. Nonetheless, Calhoun, Fuchs, and Hamlett (2000) advise that given that computer-based administration of the read-aloud accommodation contributes to test score increases for LD students, and given the medium's ability in repeating readings and video representations, using computers may compel practitioners to provide LD students the necessary accommodations in testing situations.

As the evidence culled from the literature on the effects of the read-aloud accommodation conveys, results are mixed and, at times, contradictory. Although computer-based text-to-speech technology is a promising alternative to using human readers in the delivery of the read-aloud accommodation, we must first understand how well the technology works for students, and the technology's strengths and limitations in ensuring testing fairness for all students. That is, we need to investigate the technology's ability to deliver tests that accurately measure intended constructs while ensuring the validity of test scores.

The purpose of the pilot study presented here was to compare student performance using a human reader to deliver the read aloud accommodation versus using computer-based text-to-speech technology. Specifically, the research questions for this study were:

1. How does computer-based digital speech compare with human readers as a means of providing the read-aloud accommodations?"
2. What is the effect of delivering the read aloud technology through CBT on students' test scores?
3. What is the effect of delivering the read-aloud accommodation using a human reader on students' test scores?
4. How is the accommodation effect related to students' computer skills, computer literacy, and computer use?

Results from this research will provide further evidence about the effect of computer delivery of the read-aloud accommodation on students' math performance compared with the effect of delivering the read aloud-accommodation using a human reader. This research is 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 the effect of computer-based delivery and human delivery of the read-aloud accommodation, 274 students in grades 10 and 11 from 3 New Hampshire high schools were randomly assigned to perform the same 10th grade multiple choice mathematics test in one of three modes: 1) computer-based delivery with digital read-aloud accommodation (CBD), 2) paper-and-pencil with delivery of the read-aloud accommodation by a human reader (HD), and 3) paper-and-pencil delivery with no accommodation (NA). The participating schools were selected with the cooperation of the state Director of Assessment. When selecting schools, we aimed to achieve a sample representing suburban and urban student populations including a mix of general education students (i.e., non-learning disabled English proficient students) (GE), English Language Learners (ELL), and learning disabled students (IEP). However, since the ELL population was of particular interest for this study, sample selection was limited to high schools with large ELL populations. Consequently, the schools used in this study consisted of two urban high schools, and one suburban high school. 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. Moreover, since the content area being evaluated for this study was math, students were randomly assigned to treatment groups within geometry and algebra classes. Through this three-group randomized design, this study compared the effect of computer-based delivery of the read-aloud accommodation to human delivery of the accommodation, and both accommodation delivery modes to testing without the accommodation.

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 iBooks (laptop computers, 12-inch screens) with traditional hand-held mice. All students taking the math test with computer delivery of the read-aloud accommodation took the test on one of the research team's laptop computers. Students in the remaining two groups took the same math test administered on paper. In addition to taking the same math 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 provided by the research team. 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 accommodation modal effects were related to differences in students' ability or familiarity with using a computer – constructs that are not intended to be measured by the math test.

Data Collection

Data was collected from 274 students in 3 New Hampshire high schools between March 22 and April 2, 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 math assessment, fluidity exercises, computer literacy test, and survey to the students. Students were given one hour to complete the math assessment and an additional hour to complete the computer fluidity, literacy, and use instruments.

A total of 217 10th grade students and 31 11th grade ELL students participated in the study and the number of students per school ranged from 31 to 126. Eleventh grade ELL students were included to increase the number of ELL students in the sample. An additional 30 11th grade students were tested because they were in classrooms that were being tested but their scores are not included in the analyses presented here. Additionally, 13 students were deleted from analyses because they did not complete 2 or more instruments. The scores of all remaining 235 students are used in the analysis section of this report.

Table 1: Summary of Demographic Information

Demographic Variable	Categories	Frequency	Percentage
School	Suburban	78	33.2
	Urban	157	66.8
Grade	9 th	2	0.8
	10 th	202	86
	11 th	31	13.2
Gender	Boy	128	54.5
	Girl	107	45.5
Ethnicity	African American or Black	6	2.6
	Asian Pacific Islander	7	3
	Hispanic or Latino	60	25.5
	Native American	4	1.7
	White	135	57.4
	Other	10	4.3
	Multiple	12	5.1
	Unknown	1	0.4
Language spoken at home other than English	Spanish	58	24.7
	Portuguese	14	6
	Cape Verdean Creole	1	0.4
	Vietnamese	1	0.4
	None (English only)	70	29.8
	Other	11	4.7
	Unknown	80	34

Instruments

Students participating in this study completed four data collection instruments in the following order: 1) Multiple-Choice Math Test; 2) Computer Fluidity Test; 3) Computer Literacy Test; and 4) Computer Use Survey. Below, each instrument is described in greater detail. In addition, an explanation of how scale scores were developed for the fluidity, literacy, and use instruments are presented as well as information on the reliability and validity of each scale.

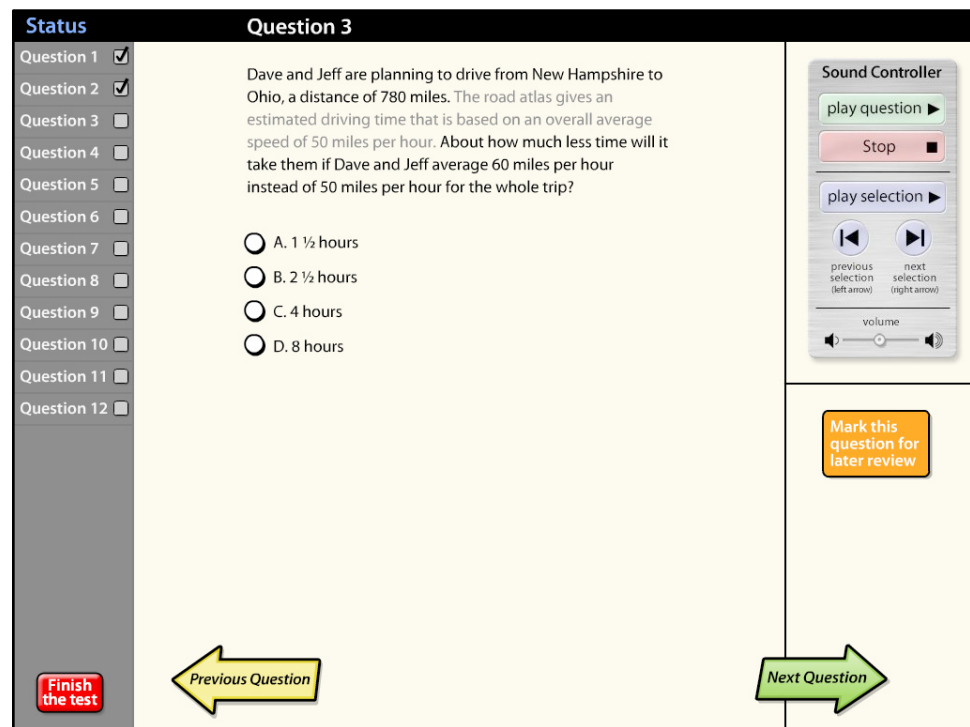
Math Test

One test form containing 12 multiple-choice items was administered to all students participating in this study. All test items were released items from previous New Hampshire 10th grade math assessments. Items were chosen based on their grade level appropriateness and item characteristics. Specifically, items with

medium difficulty (.40 to .60 difficulty level) that had particularly long stems were selected for the assessment.

The 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 students through a 2 minutes and 20 second passive tutorial. The tutorial first showed students how to use the digital read-aloud tool to play the entire question and answer options or just sections of the question. Additionally, the tutorial showed students how to navigate through items and how to select and change answers. 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 wanted to return to a specific question 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. Figure 1 displays a screen shot of the test interface for the math assessment.

Figure 1: Computer-based Math Test Interface



Note that the section of the prompt that is colored grey was being read-aloud when the screen image was captured.

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 multiple-choice math items were presented on double-sided pages and the number of words on each line and number of questions per page was identical to the computer-based format. The font size and style were also identical in both test forms in order to minimize differences that result from minor changes in the formatting of items. All students who took the computer-based 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. Additionally, students did not have access to a calculator. Although not having a calculator may have resulted in lower test scores – the mean obtained for the entire sample was only 33.58% – the research team decided at the onset of the study that it would be best not to allow calculators in order to standardize the test administration conditions. All students were allowed one hour to complete 12 multiple-choice questions.

Computer Fluidity Test

After completing the math 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 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 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 were 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 computer-related terms and functionality. Virginia and North Carolina administer multiple choice computer literacy tests to students at the 8th grade level. Fourteen released multiple-choice items from previously administered Virginia and North Carolina 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 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 test on computer was easier or more difficult than taking it with paper and pencil, whether they had any problems while taking the test on the computer, and whether they used the mark for review features.

Scale Development

As described above, three instruments were administered to students in order to measure their computer fluidity, computer literacy, and computer 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 comprised each instrument had not previously been used. 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 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 alphas were 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 re-calculated. A description of each scale's development is presented below.

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. Computer fluidity data were analyzed using principal components analysis.

Two criteria were used to retain items in the scale: first, items had to have high loadings on the factor (i.e., .5 or better); second, items had to improve the reliability of the scale. Consequently, after initial principal components analysis, a number of items were dropped from the scale, as they did not contribute significantly to maximize the variance explained by the factor. Subsequently, after conducting reliability analyses on remaining items in the scale, a number of items were eliminated as they decreased the scale's reliability. The final factor solution consisted of 3 items with loadings ranging between .828 and .923 accounting for 75% of the variance in the computer fluidity data. After obtaining a factor

solution, items on the scale were weighted by a factor that would maximize the reliability of the scale. For instance, time spent for click exercise 1 was multiplied by 0.9, and times for cases 1 and 2 in the drag-drop exercise were multiplied by 0.65 and 0.8 respectively. An alpha reliability coefficient of .82 was obtained for the computer fluidity scale.

Computer Literacy Scale

The computer literacy test consisted of 14 multiple-choice items that asked students about specific aspects of computer applications and/or skills. These aspects included terminology, software, hardware, and tasks typically performed with a computer. When a principal components analysis was run on the 14 items, a two-factor solution emerged. The factor that accounted for the most variance consisted of 10 items, whose content was based on understanding of electronic communications and application software. When a principal component factor analysis was run on these ten items, a one-factor solution that accounted for 37% of the variance and had an alpha reliability coefficient of .81 was achieved. Factor loadings on the ten items ranged from .52 to .72. 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, mp3/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. Specifically, those items that focused on school-related uses of computers at home formed one factor and those items that focused on recreational uses of computers at home formed a second factor. To capture home use of computers for purposes unrelated to school, a principal components factor analysis was then run on the remaining 5 home computer use items, yielding a one factor solution that accounted for 57% of the variance and had an alpha reliability of 0.80. Factor loadings on the seven items ranged from 0.66 to .82.

School Computer Use Scale

To measure the extent to which students used computers in school, a series of questions on the student computer use survey asked how frequently they used 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. A principal components analysis of the seven school computer-use items resulted in a two-factor solution. One factor contained 3 items, which focused on writing, editing, and multimedia. The remaining 4 items either loaded equally on both factors or loaded only on the second factor. Given the results obtained from the initial principal components analysis, it seemed that the construct underlying the first factor was related to computer use for writing and presenting information, two of the most common uses of computers at the high school level. Therefore, a principal components analysis was run on these 3 school computer use items, yielding a one-factor solution that accounted for 66.5% of the variance and had an alpha reliability of .75. Factor loadings for the 3 items ranged from .72 to .88.

Results

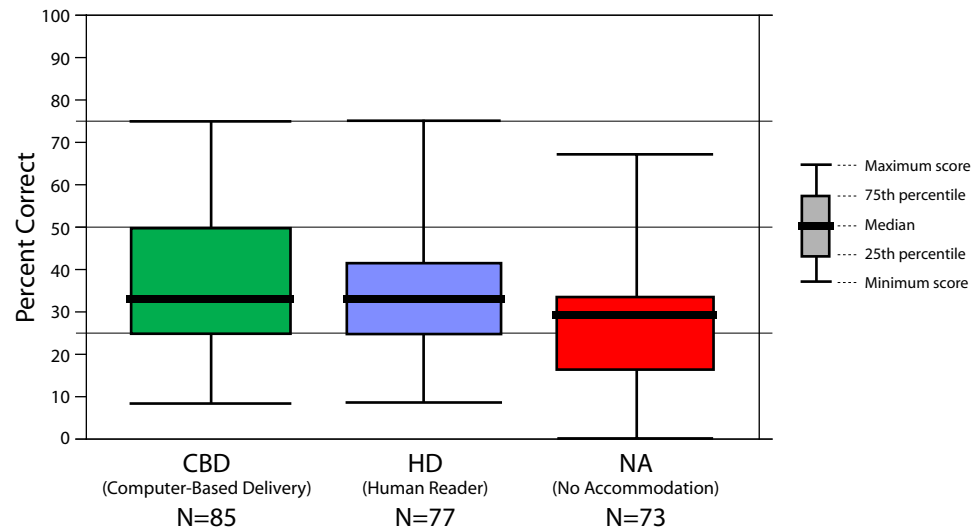
To examine the extent to which the three modes of testing – computer-based delivery of the read-aloud accommodation (CBD), paper and pencil with human delivery of the read-aloud accommodation (HD), and paper and pencil delivery with no accommodation (NA) – affected student math performance, a series of analyses were performed. These analyses included a comparison of descriptive statistics 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 (UNIANOVA) 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 computer 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 Accommodation Mode

To determine the extent to which the mode of accommodation delivery affected students' math performance, the initial stage of data analysis consisted of examining descriptive statistics for all students across the three testing modes – CBD, HD, and NA. To start, a box-plot was formed to visually compare performance across the three groups and descriptive statistics were calculated. The box-plot displayed in Figure 2 indicates that the median for the CBD and the HD groups were nearly identical. Additionally, it appears that the scores for the NA and HD groups have less variance (i.e., narrower band) than the CBD group. Moreover, the lowest and highest score for the NA group also appear to be lower than the lowest and highest scores obtained with CBD and HD. Perhaps most importantly, it is clear

that all three groups performed relatively poorly. Although all the median score for all three groups was above the chance level (25%), a substantial proportion of students in each group was at or below the chance level.

Figure 2: Box-Plot of Students Performance Across the Three Testing Modes



Descriptive statistics were obtained for the three groups and results are presented in Table 2. As seen in Table 2, the group mean for the NA group was lower than that of the two groups tested with accommodations – 29.3% for the NA group compared to 35.4% for the CBD group and 35.6% for the HD group. Additionally, the highest score obtained for the NA group was lower than the highest scores obtained for the CBD and HD groups – 67% for the NA group compared to 75% for the CBD and HD groups. However, as it was depicted in the box-plot, the standard deviation for the NA group was smaller than either the CBD or HD groups. This is likely an artifact of lower performance, on average, by students in the NA group.

Table 2: Descriptive Statistics by Accommodation Mode

Accommodation mode	Mean	N	Standard Deviation	Minimum	Maximum	Range
CBD	.354	85	.150	.08	.75	.67
HD	.356	77	.161	.08	.75	.67
NA	.293	73	.134	.00	.67	.67
Total	.336	235	.151	.00	.75	.75

An ANOVA was used to determine whether the differences detected between means in the box-plot and descriptive statistics were statistically significant. ANOVA results are presented in Table 3. Since ANOVA results indicated that

there were statistically significant differences among groups ($F=4.3$, $p=.015$), the analysis proceeded with a UNIANOVA with pairwise comparisons to determine which group mean differences were statistically significant. Results for the UNIANOVA are presented in Table 4.

Table 3: ANOVA for Performance by Accommodation Mode

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.191	2	.095	4.301	.015
Within Groups	5.148	232	.022		
Total	5.339	234			

Specifically, pairwise comparisons compared group mean differences between CBD and HD, CBD and NA, and HD and NA. The results indicated that the mean difference between CBD and HD (-0.002) was not statistically significant ($p=.927$). However, mean differences between CBD and NA (0.061), and between HD and NA (0.063) were statistically significant ($p=.012$ and $p=.011$ respectively). Essentially, the results indicated that on average, students who were tested with the read aloud accommodation, regardless of whether the accommodation was delivered by a computer or by a human reader, performed better on the math assessment than students who were tested without accommodations and that the mean differences were statistically significant.

Table 4: Pairwise Comparisons (mean_i - mean_j)

Comparison	Mean Difference (I-J)	Std. Error	Sig.
CBD-HD	-.002	.023	.927
CBD-NA	.061	.024	.012
HD-NA	.063	.024	.011

In addition to examining difference in mean scores, effect sizes were computed to examine the magnitude of the effect of each accommodation mode on students' math performance. Effect sizes were computed using Glass's delta¹ and the NA group was used as the control group. Effect sizes are presented in Table 5.

Table 5: Effect Sizes for Accommodation Modes

Accommodation mode	Effect size
CBD	.45
HD	.47

The effect size obtained for CBD was .45 while the effect size obtained for HD was .47. Although human delivery of the accommodation had a slightly higher effect on students' test scores, the effect size for both modes of delivering the read aloud accommodation were moderate in size (Cohen, 1988).

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 examined descriptive statistics for special education students and general education students (not including ELL students) across the accommodation modes. Second, an ANOVA was conducted to determine whether there were statistically significant differences among group means. Finally, a UNIANOVA with pairwise 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 52 special education students – 16 in the CBD group, 24 in the HD group, and 12 in the NA group. Additionally, there were 128 GE students tested – 53 in the CBD group, 37 in the HD group, and 38 in the NA group.

Table 6: Descriptive Statistics by Accommodation Mode and Test Type

IEP status	Accommodation mode	Mean	Standard Deviation	N
IEP	CBD ¹	.3750	.115	18
	HD ²	.3403	.159	24
	NA ³	.3333	.112	12
	Total	.3503	.135	54
GE ⁴	CBD	.3491	.152	53
	HD	.3829	.166	37
	NA	.3092	.113	38
	Total	.3470	.147	128

1. CBD = Computer-based delivery of read aloud accommodation

2. HD = Human delivery of read aloud accommodation

3. NA = Paper-based test with no accommodation

4. GE = General education students

As Table 6 illustrates, on average special education students performed slightly better on the computer-based test (CBD) than general education students – special education students obtained a mean of 37.5% compared to a mean 34.9% for general education (GE) students. However, on average GE students performed better on the HD test than special education students – GE students obtained a mean of 38.3% compared to a mean of 34.0% for special education students. Additionally, the table shows that GE students scored lower on the NA mode than on both CBD and HD.

To determine the extent to which each type of accommodation affected test performance for special education and GE students, effect sizes were computed for CBD and HD using Glass's delta. As seen in Table 7, computer-based delivery had a small effect size (.37) on the math performance of special education students compared to special education students tested without the accommodation. However, the effect size for HD was only .06, indicating that the effect of delivering the read-aloud accommodation via computer is almost five times larger than HD for special education students. Additionally, HD had a much larger effect on the performance of GE students than on the performance of special education students. Specifically, the effect of HD for GE students was .65 whereas the effect for special education students was only .06. Moreover, the effect of HD on the performance of GE students is almost twice the effect of computer-based delivery (.65 compared to .35).

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

Accommodation delivery mode	Student status	Effect size
CBD	IEP	.37
	GE	.35
HD	IEP	.06
	GE	.65

Analysis of test performance descriptive statistics indicated that group means differed by student status (IEP versus GE) and that an interaction between accommodation delivery mode and student status was possible. 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 is shown on the following page.)

Table 8: ANOVA for Performance by IEP Status

IEP Status	Sum of Squares	df	Mean Square	F	Sig.
IEP	.017	2	.008	.454	.638
GE	.102	2	.051	2.4	.095

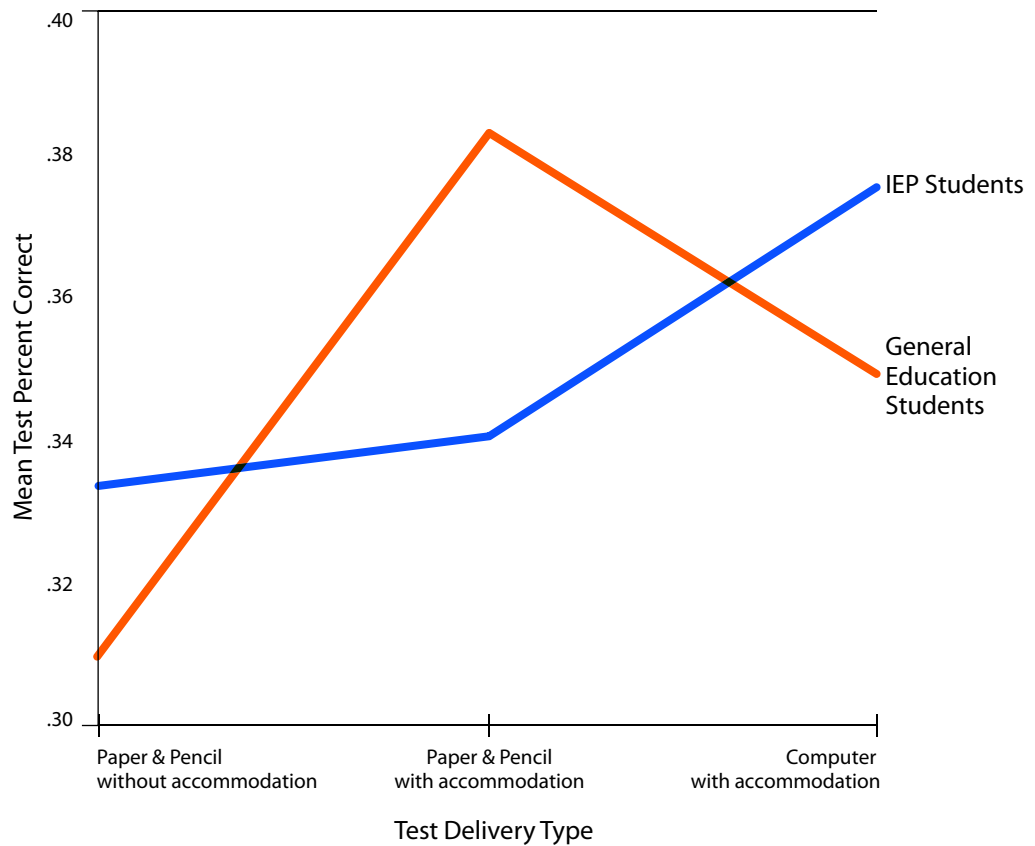
As seen in Table 8, the analysis of variance indicated that mean differences among accommodation delivery groups were not statistically significant for the IEP group ($F=.454$, $p=.638$) or for the GE group ($F=2.4$, $p=.095$).

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

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	.119	5	.024	1.166	.328
Intercept	17.277	1	17.28	843.394	.000
Accommodation mode	.045	2	.023	1.103	.334
IEP status	.000	1	.000	.011	.918
Accommodation mode × IEP status	.041	2	.020	.993	.372
Error	3.605	176	.020		
Total	25.764	182			
Corrected Total	3.725	181			

As seen in Table 9, analysis of between subjects' effects revealed that, although the effects of accommodation delivery were moderate in size, they were not statistically significant. Moreover, the interaction of mode of accommodation delivery by student status (accommodation mode × IEP status) was not statistically significant. Nonetheless, the lack of statistical significance may be an artifact of small sample sizes and, as Figure 3 shows, there is a pronounced interaction between mode of accommodation delivery and student status. That is, special education students performed better on CBD and NA, but GE students performed noticeably better on HD.

(Figure 3 is shown on the following page.)

Figure 3: Graph of the Accommodation by Student IEP Status Interaction

Differences between English Language Learners and General Education Students

In addition to examining performance differences by student IEP status, analyses were also conducted to examine whether there were differences in test performance by student language status. For this purpose, ELL students' performance was compared to English proficient (EP) students (excluding students with IEPs) for each accommodation delivery mode. The first analysis consisted of comparing test performance descriptive statistics for ELL students to EP students in each mode of accommodation. Second, an ANOVA was conducted to determine whether overall group means (i.e., the mean across all three modes of delivery) differed significantly by language status (ELL vs. EP). Finally, ANOVAs were conducted with the file split by language status to determine whether performance differed significantly within each language status group by mode of accommodation. Table 10 contains descriptive statistics for group performance by student language status.

(Table 10 is shown on the following page.)

Table 10: Group Performance by Language Status

Language status	Accommodation mode	Mean	Std. Deviation	N
ELL	CBD	.3452	.187	14
	HD	.3177	.150	16
	NA	.2464	.166	23
	Total	.2940	.170	53
English Proficient	CBD	.3491	.152	53
	HD	.3829	.166	37
	NA	.3092	.113	38
	Total	.3470	.147	128

The final ELL sample contained 53 students – 14 in the CBD group, 16 in the HD group, and 23 in the NA group. The EP sample consisted of 128 students – 53 in the CBD group, 37 in the HD group, and 38 in the NA group. The means for ELL students on the CBD, HD, and NA tests were 34.5%, 31.8%, and 24.6% respectively. Thus, on average ELL students performed better on the CBD test than on the HD and NA tests, and performed slightly better on the HD test than on the NA test. The mean scores for English proficient students on the CBD, HD, and NA tests were 34.9%, 38.3%, and 30.9% respectively. Consequently, English proficient students performed better on the HD test than on the CBD test, and better on the CBD test than on the NA test. Again, the pattern of student performance by language status is similar to the pattern found for performance by IEP status. That is, general education students appear to benefit more from HD than from CBD, whereas ELL and IEP students appear to benefit more from CBD than HD.

Table 11: Effect Sizes of Accommodation Modes by Language Status

Accommodation mode	Student status	Effect size
CBD	ELL	.6
	EP	.35
HD	ELL	.43
	EP	.65

To determine the extent to which the type of accommodation affected test performance for ELL and GE students, effect sizes were computed for CBD and HD using Glass's delta. Computer-based delivery had a large effect (.60) on the math performance of ELL students while HD had a medium size effect (.43). However, for GE students HD had a large effect (.65) and CBD had a small size effect (.35). Thus, it appears that CBD may be more beneficial to ELL students than human-delivery of the read-aloud accommodation. Additionally, HD had a

larger effect on the performance of GE students than on the performance of ELL students. Specifically, the effect of HD for GE students was .65 whereas the effect for ELL students was .43.

An ANOVA was conducted to determine whether group mean differences were statistically significant by language status. Essentially, this ANOVA compared the overall group mean for the ELL group to the overall group mean for the EP group. ANOVA results for performance by language status are presented in Table 12.

Table 12: ANOVA for Performance by Language Status

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.105	1	.105	4.425	.037
Within Groups	4.255	179	.024		
Total	4.360	180			

The analysis of variance of performance by language status indicated that mean differences between ELL and GE student groups were statistically significant ($F=4.425$, $p=.037$). Essentially, the analysis of variance supported earlier evidence found in descriptive statistics of performance by language status that indicated that EP students' math performance was better than the performance of ELL students. Therefore, further analyses were conducted to determine if group performance varied by type of accommodation within each language status group.

To determine whether math performance varied within each language group by type of accommodation, ANOVAs were conducted with the file split by language status. The results for these ANOVAs are presented in Table 13.

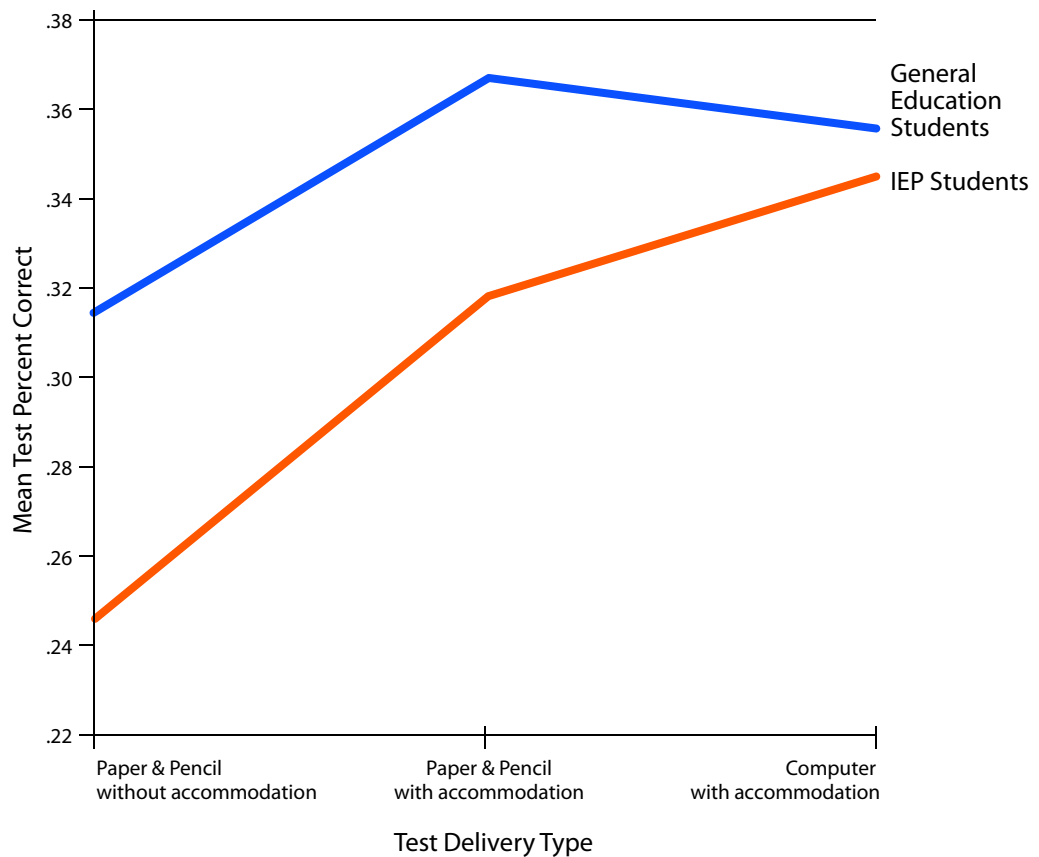
Table 13: ANOVA of Performance by Accommodation Mode within Language Groups

Language Status	Sum of Squares	df	Mean Square	F	Sig.
ELL	.098	2	.049	1.75	.184
EP	.079	2	.039	1.93	.148

Analyses of performance by type of accommodation within each language group indicated that, although performance varied by type of accommodation provided as indicated in Table 10, no statistically significant differences were detected within each language group. The F statistic obtained for the ELL group was 1.75 with a probability of .184, which indicated that similar differences were possible in the population merely by chance. Likewise, the F statistic obtained for the EP group was 1.93 with a probability of .148. Therefore, evidence suggests that no statistically significant differences in performance were detected within

each of the language groups by type of accommodation. However, the lack of significance could have been due to small sample sizes, particularly in the case of the ELL group. Moreover, the effect size for computer delivery of the read-aloud accommodation was large for the ELL population (.60) and, therefore, suggesting that it is an effective alternative to test ELL students. In continuing, the graph depicting the interaction between type of accommodation and language status (Figure 4) indicates that the CBD accommodation maybe the most effective mode of accommodation in closing the performance gap between ELL and EP students in mathematics.

Figure 4: Graph of the Accommodation by Language Status Interaction Examining Score Differences by Prior Computer Skills and Use



Examining Accommodation Effect by Prior Computer Experiences

To examine whether performance on the mathematics test under each of the three conditions may be influenced by students' prior computer experiences, a series of analyses were performed in which students were divided into three groups based upon their level of computer fluidity, computer literacy, and computer use.

Analyses were performed separately for the computer fluency, computer literacy, home computer use, and school computer use scale scores. For all four computer measures, students were divided into three groups based on their scale scores: students in the bottom third (one to 33.3 percentile rank) were classified as low level, students in the middle third of the scale (33.4 to 66.6 percentile rank) were classified as medium level; and students in the top third of the scale (66.7 to 99th percentile rank) were classified as high level. After these three groups were formed, math scores were compared across the three presentation modes based on the high-medium-low classification of each computer measure.

Examining Performance by Computer Fluency Level

Analysis of student performance by computer fluency consisted of two steps. First, an analysis of variance was conducted for all students to determine whether students' performance overall varied by computer fluency level. Second, to determine whether computer fluency affected student performance by accommodation mode, ANOVAs were conducted with the dataset split by computer fluency level. Table 14 presents results for the performance of all students by computer fluency level.

Table 14: Performance By Computer Fluency Level For All Students

Computer Fluency Level	N	Mean	Standard Deviation	Standard Error	df	Mean Square	F	Sig.
Low	77	.3106	.1563	.0178	2	.036	1.6	.203
Medium	83	.3484	.1432	.0157				
High	75	.3478	.1528	.0176				
Total	235	.3358	.1510	.0099				

As Table 14 shows, students in the medium fluency group, on average, scored better on the math test than did students in the high and low fluency level. The analysis of variance, however, indicated that there were no statistically significant differences among computer fluency group means ($F=1.6$, $p=.203$).

To investigate whether computer fluidity affected math performance by accommodation type, analyses of variance were conducted within each of the computer fluidity levels. Table 15 presents results for these analyses.

Table 15: Performance by Computer Fluency Level

Fluency Level	Accommodation Mode	Mean	N	Standard Deviation	df	Mean Square	F	Sig.
Low	CBD	.3091	31	.1277	2	.049	2.069	.134
	HD	.3598	22	.1789				
	NA	.2674	24	.1611				
	Total	.3106	77	.1563				
Medium	CBD	.3763	31	.1457	2	.031	1.53	.223
	HD	.3506	29	.1467				
	NA	.3080	23	.1316				
	Total	.3484	83	.1432				
High	CBD	.3841	23	.1736	2	.041	1.8	.173
	HD	.3590	26	.1663				
	NA	.3045	26	.1079				
	Total	.3478	75	.1528				

Although the ANOVA indicated that there are no statistically significant differences among group means, it is interesting to note that low computer fluidity students performed better on HD than on computer-based delivery of the read aloud accommodation – on average, low fluidity students answered 31% of the math questions correctly on the CBD version of the test, whereas they answered 36% of the questions correctly on the HD version. Conversely, medium and high fluidity students performed better on the CBD version of the test than they did on the HD version.

It appears that the effect of CBD increases as the computer fluidity of students increase. Not surprisingly, CBD had a high effect ($d = .74$) on the performance of high fluidity students, and a medium effect ($d = .52$) on the performance of medium fluidity students. On the other hand, CBD had a small effect on the performance of low fluidity students ($d = .26$). Thus, it seems that high and medium fluidity students may benefit more from CBD than from HD whereas low fluidity students may benefit more from HD than from CBD.

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 accommodation mode, ANOVAs were conducted with the dataset split by computer literacy level.

Table 16 contains results for the performance of all students by computer literacy rank.

Table 16: Performance By Computer Literacy Level For All Students

Computer Literacy Level	N	Mean	Standard Deviation	Std. Error	df	Mean square	F	Sig.
Low	78	.3141	.1427	.0162	2	.029	1.23	.28
Medium	79	.3513	.1594	.0179				
High	78	.3419	.1499	.0170				
Total	235	.3358	.1510	.0099				

Similar to the computer fluidity analysis, on average, students in the medium literacy group performed better on the math test than did students in the high and low literacy levels. Likewise, as was the case in the computer fluidity analysis, lower literacy level students did not perform as well as their medium and high literacy level counterparts – on average, low literacy level students answered 31.4% of the items correctly, whereas medium and high literacy level students answered 35.1% and 34.2% percent of the items correctly. However, an analysis of variance revealed that group differences were not statistically significant ($F=1.23$, $p=.28$).

To investigate whether computer literacy affected math performance by accommodation type, analyses of variance were conducted within each computer literacy level. Table 17 presents results for these analyses.

Table 17: Performance by Computer Literacy Level

Computer Literacy Level	Accommodation Mode	Mean	N	Standard Deviation	df	Mean square	F	Sig.
Low	CBD	.3173	26	.1415	2	.017	.85	.43
	HD	.3363	28	.1530				
	NA	.2847	24	.1321				
	Total	.3141	78	.1427				
Medium	CBD	.3993	24	.1554	2	.076	3.17	.048
	HD	.3661	28	.1732				
	NA	.2932	27	.1335				
	Total	.3513	79	.1594				
High	CBD	.3500	35	.1482	2	.026	1.14	.33
	HD	.3690	21	.1594				
	NA	.3030	22	.1421				
	Total	.3419	78	.1499				

Similar to fluidity, low literacy level students performed better on the math test with human-delivery of the read aloud accommodation than on CBD. However, low literacy students performed better on CBD and HD than on the test without accommodation – on average, low literacy students answered 31.7% of the items correctly with CBD, 33.6% correct with HD, but only answered 28.5% of the items correctly when tested without accommodations. On the other hand, medium literacy students performed better on the CBD test than on the HD test. Medium literacy students answered on average 40% of the items correctly with CBD whereas they only answered 37% of the items correctly with HD. Interestingly, high literacy level students performed better on HD than on CBD. One might expect higher literacy level students to perform better with CBD than with HD.

Although the analysis of means indicated an interesting pattern of performance across fluidity level groups, analysis of variance indicated that no statistically significant differences existed between accommodation modes for the low and high fluidity groups ($F = .85$, $p = .43$, and $F = 1.14$, $p = .33$ respectively). However, there were statistically significant differences across accommodation modes for the medium fluidity group ($F = 3.17$, $p = .048$). Thus, further investigation was required to determine where the significant difference existed.

To determine which pairs of means differed significantly in the medium fluidity group, pairwise comparisons were conducted. The results are presented in the Table 18.

Table 18: Pairwise Comparisons By Accommodation Mode for Medium Literacy Level

Comparison	Mean Difference	Standard Error	Sig.
CBD-HD	.033	.043	.444
CBD-NA	.106	.044	.017
HD-NA	.073	.042	.086

Pairwise comparisons indicated that no statistically significant differences existed between CBD and HD means. However, there were statistically significant differences between CBD and NA means at $\alpha = .05$, and between HD and NA means at $\alpha = .10$. Thus, it appears that medium fluidity students benefited from both CBD and HD, with the effect for CBD being significant at the .05 level.

Examining Performance by School Computer Use

Further analyses were conducted to determine whether math performance varied by school and home computer use. School use levels and home use levels created from school use and home use scales were used for these analyses. Analyses of variance were conducted first to examine whether test performance varied by school and home use. Analyses were then conducted to examine performance by computer use level across accommodation modes. The analysis for school use will be presented first followed by home use analysis.

Table 19: Performance By Computer School-Use Level For All Students

School use Level	N	Mean	Standard Deviation	Standard Error	df	Mean square	F	Sig.
Low	79	.3365	.1383	.0156	2	.004	.193	.825
Medium	77	.3431	.1588	.0181				
High	79	.3281	.1570	.0177				
Total	235	.3358	.1510	.0099				

Analysis of math test performance by level of school computer use for all students indicated that students in the medium school use category, on average, scored slightly higher than students in the high and low level categories. On average, students in the medium use category responded correctly to 34.3% of the items, students in the high use category answered 32.8% of the items correctly, and students in the low use category answered 33.7% of the items correctly. Analysis of variance indicated that that no statistically significant differences existed across school use groups ($F = .193$, $p = .825$). To examine whether performance varied by level of school use across the three accommodation modes, an analysis of variance was conducted within each school computer use level. The results are presented in Table 20.

(Table 20 is shown on the following page.)

Table 20: Performance by School-Use Level

School use rank		N	Mean	Standard Deviation	Standard Error	df	Mean square	F	Sig.
Low	CBD	23	.3406	.1397	.0291	2	.011	.55	.578
	HD	30	.3528	.1527	.0279				
	NA	26	.3141	.1209	.0237				
	Total	79	.3365	.1383	.0156				
Medium	CBD	30	.3472	.1609	.0294	2	.042	1.71	.189
	HD	25	.3800	.1702	.0340				
	NA	22	.2955	.1354	.0289				
	Total	77	.3431	.1588	.0181				
High	CBD	32	.3698	.1496	.0264	2	.07	3.001	.056
	HD	22	.3333	.1647	.0351				
	NA	25	.2700	.1469	.0294				
	Total	79	.3281	.1570	.0177				

The analysis of performance by school use across accommodation modes showed that, although the school-use groups were not statistically significantly different across accommodation modes ($F = .55$, $p = .578$, and $F = 1.71$, $p = .189$ for the low and medium categories respectively) students in the low and medium use categories, on average, performed better on HD than CBD. On average, students in the medium use category answered 38.0% of the math items correctly on the HD version of the test but only answered 34.7% of the items correctly on the CBD version. Likewise, on average, students in the low school use category answered 35.3% of the math items correctly on the HD version and answered 34.1% of the items on the CBD version. Conversely, students in the high use category, performed better on the CBD version of the test than on the HD version. On average, students in the high use category answered 37.0% of the math items correctly, whereas they only answered 33.3% of the items on the HD version and 27.0% of the items on the NA version. Additionally, the mean differences across accommodation modes were statistically significant for students in the high use category. Thus, pairwise comparisons were conducted for the high school use group to determine which accommodation group means differed. The results are presented in Table 21.

Table 21: Pairwise Comparisons By Accommodation Mode for High Level School Users

School use rank	(I) TEST TYPE	Mean Difference (I-J)	Std. Error	Sig.
High	CBD-HD	.036	.042	.393
	CBD-NA	.100	.041	.017
	HD-NA	.063	.045	.161

Pairwise comparisons for the high use category indicated that, although students performed better on the CBD version than on the HD test, the differences were not statistically significant ($p = .393$). However, CBD had a larger effect than HD on the math performance of students in the high school use group. The effect size for CBD was .68, a large effect size, whereas HD had a .43 effect. Additionally, the difference between CBD and NA means were statistically significant ($p = .017$) indicating that CBD improved the math performance of students in the high-school use category as compared to no accommodation.

Examining Performance by Home Computer Use

Finally, analyses were conducted to determine whether math performance varied by home computer use. Analyses of variance were conducted first to determine whether test performance varied by home use. Analyses were then conducted within each level of home use to examine the relationship between home use and the mode of accommodation. Table 22 presents results for all students by home computer use.

Table 22: Performance by Home Computer Use for All Students

Home Use Rank	N	Mean	Standard Deviation	Standard Error	df	Mean square	F	Sig.
Low	78	.3280	.1597	.0181	2	.008	.343	.71
Medium	79	.3470	.1434	.0161				
High	78	.3323	.1511	.0171				
Total	235	.3358	.1510	.0099				

Similar to the analysis for school use, analysis for home use for all students indicated that there were no statistically significant differences in math performance across home use levels ($F = .343$, $p = .71$). Likewise, analysis by home use indicated that students in the medium use category, on average, performed better than students in the high and low categories. On average, students in the medium category answered 34.7% of the math items correctly whereas students in the low category answered 32.8% of the items correctly, and students in the high category answered 33.2% of the items correctly. To determine whether performance varied by home use across accommodation modes, ANOVAs were conducted within each by home use level. Results are presented in Table 23.

(Table 23 is shown on the following page.)

Table 23: Performance by Home Use Level Across Accommodation Modes

Home Use Rank	Test Type	N	Mean	Standard Deviation	Standard Error	df	Mean square	F	Sig.
Low	CBD	31	.3468	.1682	.0302	2	.017	.652	.524
	HD	24	.3333	.1454	.0297				
	NA	23	.2971	.1644	.0343				
	Total	78	.3280	.1597	.018				
Medium	CBD	25	.3800	.1297	.0259	2	.082	4.3	.017
	HD	25	.3833	.1559	.0312				
	NA	29	.2874	.1271	.0236				
	Total	79	.3470	.1434	.0161				
High	CBD	29	.3391	.1476	.0274	2	.018	.797	.454
	HD	28	.3512	.1791	.0338				
	NA	21	.2976	.1105	.0241				
	Total	78	.3323	.1511	.017				

On average, students in the low use category answered 34.7%, 33.3%, and 29.7% of the items correctly on the CBD, HD, and NA versions of the test, respectively. Thus, students in the low use category performed better on CBD than on HD, and performed better on both accommodated versions of the test than on the non-accommodated version. However, analysis of variance indicated that there were no statistically significant differences among group means. On the other hand, students in the medium and high categories performed better on the HD version of the test than on the CBD version. On average, students in the medium category answered 38.3% of the items correctly on the HD version, and answered 38.0% of the items correctly on the CBD version. Likewise, students in the high use category answered 35.1% of the items correctly on the HD version, and answered 33.9% of the items correctly on the CBD version. However, analysis of variance indicated that there were no statistically significant differences between group means in the high use category ($F = .797$, $p = .454$) but there were statistically significant differences between group means in the medium use category ($F = 4.3$, $p = .017$). Thus, pairwise comparisons were conducted to determine which accommodation group means differed in the medium home use category. Results are presented in Table 24.

Table 28: Pairwise Comparisons By Accommodation Mode for Medium Level Home Users

School use rank	(I) TEST TYPE	Mean Difference (I-J)	Standard Error	Sig.
Medium	CBD-HD	-.003	.039	.932
	CBD-NA	.093	.038	.016
	HD-NA	.096	.038	.013

Pairwise comparisons for medium level home users indicated that group mean differences between CBD and NA were statistically significant ($p = .016$). Likewise, group mean differences between HD and NA were statistically significant ($p = .013$). Thus, on average, students in the medium home use category performed better on the CBD and HD versions of the math test than on the non-accommodated version.

Discussion

Reliance on large-scale tests to make decisions about students and schools has increased as a result of federal requirements. At the same time, efforts are being made to include a larger percentage of students in externally mandated test accountability systems. In order to decrease construct irrelevant variance, test accommodations are necessary for some students. The read-aloud accommodation is one of the most common forms of test accommodations and is believed to decrease variance that results from differences in reading skills when reading ability is not the construct being measured by a given test (Bielinski, Thurlow, Ysseldyke, Freidebach, & Freidebach, M., 2001). Despite the frequency with which students are provided with a read-aloud accommodation, this accommodation can be labor intensive to provide since it requires an adult to read the entire test aloud to students. Concerns also have been raised about the consistency with which read-aloud accommodations are provided to students across settings and the possibility that the readers may inadvertently provide clues to students through subtle changes in their voice (Tindall, Heath, Hollenbeck, Almond, & Harniss, 1998). In some cases, the read-aloud accommodation is provided individually to students while in other cases the accommodation is provided to group of students. In both cases, however, concerns have been raised about the extent to which students are able to or feel comfortable asking proctors to re-read portions of the test so that the student fully understands what is being asked (Landau, Russell, Gourgey, Erin, & Cowan, 2003)

The study presented here focused on the feasibility and effect of providing a read-aloud accommodation using a computer. In this study, tenth grade students were randomly assigned to perform the same mathematics test under one of three conditions. In the first condition, students performed the test on paper without

any accommodations. In the second condition, students performed the test on paper in a group setting with a proctor reading aloud all of the directions, the prompts, and the options. In this condition, students were able to ask the proctor to re-read any section of the test, which the proctor would then do to the entire group. Since the proctor read all parts of the test aloud, in this condition the pace of testing was effectively controlled by the proctor. In the third condition, students performed the test individually on a computer and the computer read aloud all parts of the test. In this third condition, students could request that the computer re-read any section of the test multiple times and students could stop the computer from reading a section of the test at any time. In this condition, students had complete control over the pace and order of the testing and were in control of which parts of the test were read and how often they were read.

Recognizing that a given accommodation may be helpful for students who are in need of the accommodation, the accommodation also may inadvertently provide assistance such that the measurement of the construct of interest is altered. In the case of this study, the construct being measured was mathematical achievement. It was believed that for some students who have Individual Education Plans related to verbal and/or reading skills and for students who are English Language Learners, requiring students to read directions and items may interfere with the measure of their mathematical achievement. Yet, allowing students to have items read aloud may also inadvertently benefit students who do not have IEPs related to language skills or who are native English speakers. Thus, one criteria that must be met in order to establish the validity of a given accommodation is that a differential group effect is found. That is, the provision of the accommodation has a positive effect on those students who are believed to need the accommodation and that the accommodation does not have a positive effect on the performance of those students who do not need the accommodation (Bielinski, Thurlow, Ysseldyke, Freidebach, & Freidebach, 2001; Sireci, Li, & Scarpati, 2003; Elliot, McKevitt, & Kettler, 2002; Koenig, 2002). In the study presented here, this requirement would be met if the read-aloud accommodation provided by the human reader or by the computer had a positive effect on students with language IEPs and ELL students and if the accommodation did not have a positive effect on general education students. Of interest, too, is the extent to which the computerized read-aloud accommodation results in a larger or smaller increase in performance for the targeted students (i.e., students with IEPs and ELL students) as compared to the human read-aloud accommodation.

As is presented in greater detail above, the results of this study are mixed. Across all students, both the human- and computer-delivered read aloud accommodation resulted in significantly higher test scores than did the normal test administration conditions. Under both read-aloud accommodations, students' test scores increased by approximately .45 standard deviation units as compared to the traditional paper-based test administration procedures.

Focusing specifically on IEP and general education students, the computer-based read-aloud accommodation resulted in higher performance for IEP students than did the human-delivered accommodation or the traditional administration condition. Conversely, the human-delivered accommodation resulted in higher performance for general education students than did the computer-delivered accommodation or the traditional administration conditions. That is, computer-delivery of the read-aloud accommodation appears to have been more effective in closing the performance gap between special education students and general education students. While these differences were not statistically significant, they may signal an interaction effect between mode of accommodation and IEP status. It should also be noted that although the higher performance by general education students under the human-delivered accommodation mode is not desirable, it is consistent with past research on this form of the read-aloud accommodation. As Sireci (2003) speculates, the controlled pace of testing under the human-delivered read aloud accommodation may help some general education students focus on each item individually and may result in students investing more time working on each individual item. Again, while speculative, the computer-based read aloud accommodation may be beneficial to IEP students due their ability to control the frequency with which portions of the test are read aloud.

Although the ELL students generally scored lower than did the IEP and the general education students, the effect of the accommodation modes for the ELL students is similar to that of the IEP students. Specifically, ELL students received higher scores when the mathematics test was performed under the computer-delivered accommodation condition and received the lowest scores under the traditional paper-based condition. While these differences were not statistically significant, perhaps due to small sample sizes, the effect size (.60) for the computer-based accommodation mode was of moderate to large size while the effect size (.43) for the human-delivered accommodation was of moderate size.

Given the potential construct-irrelevant variance related to prior computer experiences that may be introduced by the computer-based accommodation, a series of analyses was performed to examine the extent to which the accommodation effects varied across prior computer experiences. Specifically, four measures of prior computer experiences were analyzed. These measures included computer fluidity (how well students could use a mouse and keyboard to perform the actions typically required for a computer-delivered test), computer literacy, computer use in school, and computer use at home. As was anticipated, students with low computer fluidity, computer literacy and/or in-school computer use performed worse with the computer-delivered accommodation as compared to the human delivered accommodation. With respect to home use, students with low levels performed slightly higher on the computer version than the human version, but these differences were relatively small. Although the pattern is less consistent for students with medium and high levels of computer fluidity, literacy, school and home use, the effect of computer delivery of the accommodation was generally the same or

larger than the effect of the human delivery of the accommodation. Specifically, students with medium fluidity and literacy performed noticeably higher on the computer version than the human-delivered accommodation while students with medium levels of home computer use performed about the same on the computer and human accommodation modes. Conversely, students with medium school computer use performed better on the human-delivered accommodation mode. For students with high levels of computer fluidity and school use, performance was higher on the computer version, while students with high levels of computer literacy and home computer use performed slightly better under the human delivered accommodation mode. While only a few of these differences were statistically significant, many of the effects were of moderate size. Thus, while further research that involves larger samples is clearly required, there is some evidence that the effect of the computer-delivered read-aloud accommodation may interact with students' prior computer experiences.

In summary, the research presented here was intended to be a pilot study that examined the feasibility and effect of providing a read-aloud accommodation via a computer. Based on the fact that all students who were assigned to perform the test on computer were able to do so without any problems and that none of the students indicated that they experienced any technical issues with the computer-based read-aloud version of the test, it seems feasible to deliver a read-aloud accommodation via computer. While findings are mixed with respect to the effects of the read-aloud accommodation on student performance, we believe there is sufficient evidence to support further larger scale research on the effect of computer-based read-aloud accommodations. Although the read-aloud accommodation appeared to benefit special education students, English Language Learners, and general education students, and thus does not meet the criteria of differential group effect, this finding is consistent with past research on read-aloud accommodations. Interestingly, however, the computer-based read-aloud accommodation appears to come closer to meeting this requirement than did the human delivered accommodation. Specifically, the computer-based accommodation resulted in higher scores than the human-delivered accommodation for both IEP and ELL students while the human-delivered accommodation resulted in higher scores for the general education students. As speculated above, this difference may result from the computer-based accommodation encouraging IEP and ELL students to replay sections of an item multiple times without having to ask the proctor to re-read text while the human delivered accommodation may have forced general education students to slow down their pace of testing and thus increase the amount time they focus on each item.

This study also provides preliminary evidence that the effect of the computer-based accommodation may interact with prior computer experience such that students who are less familiar with using a computer did not benefit as much from the computer-based accommodation as compared to students who were more familiar with working with a computer.

Although a randomized experimental design was employed for this study, there were several shortcomings to this study. First, although the sample included approximately 235 students, disaggregating students into IEP, ELL and general education groups resulted in relatively small sub-group samples. Similarly, dividing students into high, medium and low level groups based on prior computer experiences also resulted in relatively small sample sizes. Clearly, it would be highly desirable to increase the sample size in future studies.

Second, although we attempted to examine the effect separately for students with IEPs and general education students, student IEPs varied widely. Future studies would be strengthened by grouping students by specific categories of need related to language rather than simply clustering all students with IEPs into a single category. Doing so may reveal differential effects for students with different needs. Similarly, ELL students' language proficiency levels also varied considerably. Therefore, future studies should classify ELL students by language proficiency level to examine the differential effects that the accommodation may have on students of different language abilities.

Third, budgetary issues and sample size limited the study to a three group design. Two of the groups performed the test on paper while the third group performed the test on computer. Two of the groups also were provided with a form of the read-aloud accommodation while one group was not. As a result of this crossed design and the use of three groups, we were not able to separate the effect of taking the math test on computer from the effect of receiving the computer-based read-aloud accommodation. A stronger study design would employ four groups, with the fourth group taking the same test on computer but without the read-aloud accommodation.

Fourth, although students within each participating school were randomly assigned to groups, differences in the prior achievement of students may have existed among groups. To improve the random assignment process and to statistically control for any differences in prior achievement, future studies should collect and use information about students' prior achievement. This information might include grades or standardized test scores.

Fifth, by design, the mathematics test employed for this study included items that were of moderate difficulty for the sample that originally performed the items. For all of the items, the original sample included all students in New Hampshire. Thus, the items used for this study were of average or moderate difficulty for all students in New Hampshire. A large portion of the sample used for this study, however, was comprised of IEP and ELL students. Typically, IEP and ELL students perform below average. For many of the students included in this study, the items were relatively difficult. As a result, total test scores were low. While it is highly desirable to employ items of moderate difficulty when examining differential performance across modes, future studies would be strengthened by selecting items that are of moderate difficulty for the sample of students employed in the study rather than the original sample of students who performed the set of items.

Despite these shortcomings, the study presented here provides preliminary evidence that a computer-delivered read-aloud accommodation is both feasible and effective for the targeted population. Based on the findings of this pilot study, we strongly recommend that additional research be conducted that focuses specifically on the effects of computer-based delivery of the test versus computer-based delivery of the read-aloud accommodation. While this research is being conducted, we also suggest that state testing programs begin exploring the feasibility of incorporating a computer-based read-aloud accommodation into their testing programs. Finally, given that the read-aloud accommodation had a positive effect on the performance of general education students, we also recommend that research be conducted on whether validity would be increased by allowing all students to have test items be read aloud, if they desire, when the construct measured is anything other than reading skills or reading comprehension. While such a flexible delivery platform might challenge conventional conceptions of standardized test administration conditions, given the current focus on standards-based criterion-reference testing, this flexibility may increase the validity of the inferences made about student achievement based on test scores.

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.
- Bielinski, J., Thurlow, M., Ysseldyke, J., & Fieidebach, M. (2001). *Read aloud accommodations: Effects on multiple-choice reading and math items*. Minneapolis, MN: National Center on Educational Outcomes.
- Brown, P. J. & Augustine, A. (2001). *Screen reading software as an assessment accommodation: implications for instruction and student performance*. Paper presented at the Annual Meeting Educational Research Association (Seattle, WA, April 10–14, 2001).
- Calhoun, M. B., Fuchs, L. S., & Hamlett, C. L. (2000). Effects of computer-based test accommodations on mathematics performance assessments for secondary students with learning disabilities. *Learning Disability Quarterly*, 23(4), 271–282.
- Elliott, S. N., McKeivitt, B. C., & Kettler, R. J. (2002). Testing accommodations research and decision making: the case of “good” scores being highly valued but difficult to achieve for all students. *Measurement and Evaluation in Counseling and Development*, 35(3), 153–166.
- Helwig, R., Tedesco, M., Health, B., Tindal, G., & Almond, P. (1999) Reading as an access to mathematics problem solving on multiple-choice tests for sixth grade students. *The Journal of Educational Research*, 93(2), 113–125
- Hollenbeck, K., Rozek-Tedesco, M. A., Tindal, G., & Glasgow, A. (2000). An exploratory study of student-paced versus teacher-paced accommodations for large-scale math tests. *Journal of Special Education Technology*, 15(2), 27–36.
- Koenig, J. A. (2002). *Reporting test results for students with disabilities and English-language learners. summary of a workshop* (Washington, DC, November 28, 2001). U.S.; District of Columbia: National Academies Press.
- Kosciolek, S. & Ysseldyke, J. E. (2000). *Effects of a reading accommodation on the validity of a reading test. Technical report 28*. U.S.; Minnesota: National Center on Educational Outcomes.
- Landau, S., Russell, M., Gourgey, K., Erin, J., & Cowan, J. (2003). Use of the Talking Tactile Tablet in mathematics testing. *Journal of Visual Impairment and Blindness*, 97(2) 85–96.
- Shepard, L., Taylor, G., & Betebenner, D. (1998). *Inclusion of limited-English-proficient students in Rhode Island's grade 4 mathematics performance assessment*. Los Angeles: University of California, Center for the study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.

- Sireci, S.G., Li, S., & Scarpati, S. (2003). *The effects of test accommodations on test performance: A review of the literature* (Research Report 485). Amherst, MA: Center for Educational Assessment.
- Tindal, G., Heath, B., Hollenbeck, K., Almond, P., & Harniss, M. (1998). *Accommodating students with disabilities on large-scale tests: an empirical study of student response and test administration demands*. U.S.; Oregon.
- Tindal, G., Heath, B., Hollenbeck, K., Almond, P., & Harniss, M. (1998). Accommodating students with disabilities on large-scale tests: an empirical study of student response and test administration demands. *Exceptional children*, 64(4), 439–450.
- Thurlow, M. L., Ysseldyke, J. & Silverstein, B. (1995). Testing accommodations for students with disabilities. *Remedial and Special Education*, 16(5), 260–270.
- Thurlow, M. L., Ysseldyke, J. & Silverstein, B. (1993). *Testing accommodations for students with disabilities: A review of the literature*. (Synthesis Report No. 4). Minneapolis, MN: National Center on Educational Outcomes.
- Thurlow, M. L., Hurley, C., Spicuzza, R., & El Sawaf Hamdy. (1996). *A review of the literature on testing accommodations for students with disabilities*. (State Assessment Series Report 9). Minneapolis, MN: National Center on Educational Outcomes.
- Thurlow, M. L., Ysseldyke, J. & Silverstein, B. (1998). *The research basis for the need for research on accommodations*. Paper presented at the American Education Research Association, San Diego, CA.
- Shulte, A.A., Elliot, S.N, & Kratochwill, T.R. (2000). *Effects of testing accommodations on standardized mathematics test scores: An experimental analysis of the performances of students with and without disabilities*. Madison, WI: University of Wisconsin.

