



TECHNOLOGY AND ASSESSMENT STUDY COLLABORATIVE

Testing on Computers: A Follow-Up Study Comparing Performance on Computer and on Paper

Michael Russell
Technology and Assessment Study Collaborative
Boston College
332 Champion Hall
Chestnut Hill, MA 02467

www.intasc.org



Testing on Computers: A Follow-Up Study Comparing Performance on Computer and on Paper

MICHAEL RUSSELL

Technology and Assessment Study Collaborative

Released July 2002

Originally published in *Educational Policy Analysis Archives*, 7(20), 1999.

Introduction

Recently, Walt Haney and I (Russell & Haney, 1997) reported that open-ended tests administered on paper to students accustomed to working on computer may seriously underestimate students' achievement. Although previous research on multiple-choice items suggests that the mode of administration, that is paper versus computer administration, does not significantly affect the test taker's performance (Bunderson, Inouye & Olsen, 1989), our study suggests that the mode of administration may have an extraordinarily large effect on students' performance on open-ended items.

Focusing on students participating in a project that placed heavy emphasis on computers, the study indicates that approximately 60% of the students in the Advanced Learning Laboratory (ALL School) were performing adequately on writing tests before the project began. Nearly two years after the program was implemented the same writing tests taken on paper indicated that only 30% of the students were writing adequately, an apparent decline of approximately 30% points. Yet, when the same tests were administered on computer (without student access to word processing tools such as spell checking or grammar checking), nearly 70% of students performed adequately. This significant and startling difference also occurred on National Assessment of Educational Progress (NAEP) reading and science items, which required students to respond to open-ended items (similar to items used in the Third International Math and Science Study, the Massachusetts Comprehensive Assessment System and other state level testing programs). The study concludes that for the students in the ALL School, most of whom are accustomed to working on computers, open-ended test questions administered on paper severely underestimated students' achievement.

Although our findings raise questions about the validity of open-ended test results for students accustomed to working on computer but who completed tests on paper, our study had several shortcomings. As we noted, only one extended writing item was used. Furthermore, no information regarding the extent to which students used computers or the proficiency with which students used computers was available. All of the examinees included in the study were accustomed to working on computers. Thus it was not possible to study the mode of administration effect across varied levels of previous computer use. Finally, beyond scores for a set of open-ended items performed by both groups on paper, no other information about prior academic achievement, such as standardized test scores or grades, was considered.

Despite these shortcomings, the results raise important questions about the extent to which scores for open-ended items administered on paper can be used to make inferences about individual students (or their schools) who are accustomed to working on computers. Moreover, if test scores are used to evaluate the effect increased expenditures for computers have on student achievement, the use of open-ended items administered on paper may also undermine the growing emphasis on educational technology.

In this study, I build on our prior work and overcome the shortcomings of our previous study. Specifically, I improved the study design in five ways. First, the sample of students was broadened to cover a range of prior computer experience. Second, information about students' prior use of computers, preference for writing on computer or on paper, and an indicator of students' keyboarding skill was collected. Third, rather than focusing on one extended writing task, several items were administered. Fourth, rather than focusing specifically on writing, open-ended math, language arts, and science items were examined. And fifth, all items included in this study had been used in state or national testing programs and had been validated previously. Thus, for this study I analyze the effect of computer administration across levels of computer use/proficiency using several open-ended items in the areas of language arts, math, and science. Specifically, the following research questions are addressed:

1. Does the effect reported by Russell and Haney (1997) hold across levels of computer use/proficiency?
2. Does the mode of administration effect occur within and across subject areas and if so, is the magnitude of the effect consistent across subject areas?
3. Do students who prefer to write on paper perform better than predicted on open-ended items administered on paper and do students who prefer to write on computer perform better than predicted on open-ended items administered on computer?

Background

For three decades, educational theorists have proposed many ways in which computers might influence education. Although it was not until the 1970's that computers began having a presence in schools, since then the use of computers in education has increased dramatically (Zandvliet & Farragher, 1997). The National Center for Education Statistics reports that the percentage of students in grades 1 to 8 using computers in school more than doubled from 31.5 in 1984 to 68.9 in 1993 (Snyder & Hoffman, 1990; 1994). Similarly, the availability of computers to students in school increased from one computer for every 125 students in 1983 to one computer for every 9 students in 1995 (Glennan & Melmed, 1996). As the number of computers has increased, theories about how computers might benefit students' writing have proliferated. To a lesser extent, some researchers have carried out formal studies to examine whether writing on computer actually leads to better writing. Many of these studies have reported that writing on computers leads to measurable increases in students' motivation to write, the quantity of their work and the number of revisions made. Some of these studies also indicate that writing on computers improved the quality of writing. In a meta-analysis of 32 computer writing studies, Bangert-Drowns (1993) reports that about two-thirds of the studies indicated improved quality for text produced on computer. However, the extent to which writing on computers leads to higher quality writing seems to be related to the type of students examined. Generally, improvements in the quality of writing produced on a computer are found for learning disabled students, early elementary students, low-achieving students and college-aged students. Differences generally are not found for middle school and high school students.

Learning Disabled, Early Elementary Students and College-Aged Students

Although neither Kerchner and Kistingner (1984) nor Sitko and Crealock (1986) included a comparison group in their studies, both noted significant increases in motivation, quantity and quality of work produced by learning disabled students when they began writing on the computer. After teaching learning disabled students strategies for revising opinion essays, MacArthur and Graham (1987) reported gains in the number of revisions made on computer and the proportion of those revisions that affected the meaning of the passage. They also noted that essays produced on computer were longer and of higher quality. In a separate study, MacArthur again reported that when writing on a computer, learning disabled students tended to write and revise more (1988). At the first grade level, Phoenix and Hannan (1984) report similar differences in the quality of writing produced on computer.

Williamson and Pence (1989) found that the quality of writing produced by college freshman increased when produced on computer. Also focusing on college age students, Robinson-Stavely and Cooper (1990) report that sentence length and complexity increased when a group of remedial students produced text on the computer. Hass and Hayes (1986a) also found that experienced writers produced papers of greater length and quality on computer as compared to those who created them on paper.

Middle and High School Students

In a study of non-learning disabled middle school students, Dauite (1986) reported that although writing performed on the computer was longer and contained fewer mechanical errors, the overall quality of the writing was not better than that generated on paper. In a similar study, Vacc (1987) found that students who worked on the computer spent more time writing, wrote more and revised more, but that holistic ratings of the quality of their writing did not differ from text produced with paper-and-pencil.

At the middle school level, Grejda (1992) did not find any difference in the quality of text produced on the two mediums. Although Etchison (1989) found that text produced on computer tended to be longer, there was no noticeable difference in quality. Nichols (1996) also found that text produced on computer by sixth graders tended to be longer, but was not any better in quality than text produced on paper. Yet, for a group of eighth grade students, Owston (1991) found that compositions created on computer were rated significantly higher than those produced on paper.

Focusing on high school freshman, Kurth (1987) reports that there was no significant difference in the length of text produced on computer or on paper. Hawisher (1986) and Hawisher and Fortune (1989) also found no significant differences in the quality of writing produced by teenagers on paper and on computer. Hannafin and Dalton (1987) also found that for high achieving students, writing on computer did not lead to better quality writing. But for low-achieving students, texts produced on the computer were of a higher quality than those produced on paper.

Summary of Studies

The research summarized above suggests many ways in which writing on computer may help students produce better work. Most formal studies report that when students write on computer they tend to produce more text and make more revisions. Studies that compare student work produced on computer with work produced on paper find that for some groups of students, writing on computer also has a positive effect on the quality of student writing. This positive effect is strongest for students with learning disabilities, early elementary-aged students and college-aged students. All of the studies described above focus on student work produced in class under untimed conditions. These studies also focus on work typically produced for English or Language Arts class, such as short stories or essays. However, the study presented here focuses on writing produced under formal timed testing conditions in three subject areas, namely language arts, math and science. Specifically, this study addresses the extent to which producing open-ended responses on computer or on paper effects students' performance, particularly for students with different levels of computer use.

Study Design

To better understand whether open-ended test items administered on paper underestimate the achievement of students accustomed to working on computers, six open-ended math, six science, and six language arts items were converted to a computer format and then administered in two modes, paper and computer. In addition, all students completed a computer use survey and performed a short keyboarding test. Finally, an indicator of prior achievement, namely Grade 7 Stanford Achievement Test version 9 (SAT 9) scores, was collected for each student. As is explained in detail below, the indicator of achievement was used to stratify and randomly assign representative sample groups and is used as a covariate for some analyses.

The study focuses on three subject areas: math, language arts, and science. For each subject area, a total of six open-ended items were administered. To decrease the amount of testing time required for each student, students were divided into four groups. Two of these four groups performed the six science items and three of the language arts items, only. For ease of reference I call these groups of students SL and LS. The remaining two groups of students performed the six math items and the other three language arts items, only. These groups of students are referred to as ML and LM. All students completed the computer use survey and performed the keyboarding test. In addition, an indicator of prior achievement was collected for each student.

The study occurred in three stages. During stage 1, SAT 9 scores were collected for each student. In total, four SAT 9 scores were collected for each student: Comprehensive Normal Curve Equivalent (NCE), Math NCE, Language Arts NCE and Science NCE. Once collected, the Comprehensive NCE was used to stratify and randomly assign four groups of students. Two of these groups formed the SL and LS students while the remaining two groups formed the ML and LM students.

During stage 2, all students completed the computer use survey and performed the keyboarding test. During stage 3, a crossed design was used to administer the open-ended items to each group. In this crossed design, the SL students first performed the science items on computer and then performed three language arts items on paper. The LS students first performed the three language arts items on computer and then performed the science items on paper. Similarly, the ML students first performed the math items on computer and then performed the three language arts items on paper. Finally, the LM students first performed the language arts items on computer and then performed the math items on paper. Below, the instruments, sampling method and scoring method are discussed in greater detail.

Instruments

The instruments used in this study fall into three categories: indicators of prior achievement; computer experience; and open-ended tests.

Indicator of Prior Achievement

As described in greater detail below, an indicator of prior achievement was used to assign students to experimental groups and as a covariate during analyses. Since the sample of students was limited to students in grade eight, the students' grade 7 SAT 9 NCE scores were used as the indicator of prior achievement.

Computer Experience

Two instruments were used to estimate students' level of computer experience. First, a survey that focused on prior computer use was administered to all students. Second, all students completed a brief keyboarding test administered on computer.

Student Questionnaire

The survey was designed to collect information about how much experience students had working with computers and, in particular, how they used computers during their writing process. The survey included questions that asked:

- a. how long the student has had a computer in his/her home;
- b. how many years they have used a computer;
- c. how often they currently use a computer in school and at home;
- d. how often they use a computer during different stages of their writing process (e.g., brainstorming, outlining, composing a first draft, editing, writing the final draft); and
- e. whether they prefer to write papers on paper or on computer.

In addition, the survey asked students to draw a picture of their writing process and to then describe what they had drawn. The purpose of the drawing prompt was to collect information about if and when computers enter the student's writing process. Finally, the student questionnaire asked students to indicate their gender and their race/ethnicity.

To code student drawings, the following guide was used:

- 0 – No computer visible
- 1 – Computer used for final draft only
- 2 – Computer used prior to creating the final draft

When coding drawings, both the drawing and the student's description of their drawing were reviewed prior to assigning a score. All drawings were coded by one rater. However, to examine inter-rater reliability, a sample of 20 drawings was coded by a second rater. For these 20 drawings, there was no discrepancy between the two raters' scores.

Keyboarding Test

To measure keyboarding skills, all students performed a computer based keyboarding test. The keyboarding test contained two passages which students had two minutes apiece to type verbatim into the computer. Words per minute unadjusted for errors was averaged across the two passages and was used to estimate students'

keyboarding speed. Both keyboarding passages were taken directly from encyclopedia articles to assure that the reading level was not too difficult.

Although there is considerable debate about how to quantify keyboarding ability (see West, 1968, 1983; Russon & Wanous, 1973; Arnold, et al, 1997; and Robinson, et al, 1979), for the purposes of this study, students average words per minute (WPM) uncorrected for errors was recorded. In each of the scoring guidelines used to rate student responses to the open-ended test items, spelling was not explicitly listed as a criterion raters should consider when scoring student responses. For this reason, students keyboarding errors did not seem to be directly relevant to this study.

Open-Ended Tests

This study examines the mode of administration effect on student performance in three subject areas: science, math, and language arts. To restrict testing time to 60 minutes per test, 6 science items, 6 math items and two sets of 3 language arts items were administered. All items included in this study were taken directly from open-ended test instruments used previously. Sources for items include the National Assessment of Educational Progress (NAEP) and the Massachusetts Comprehensive Assessment System (MCAS).

Language Arts Items

In total, six language arts items were used in this study. Three of the language arts items were taken from the 1999 Spring administration of MCAS. Two of the items were taken directly from the 1988 Grade 8 NAEP Writing Assessment. And the final language arts item was taken from the 1992 Grade 8 NAEP Writing Assessment.

The three MCAS language arts items focus on reading comprehension. For each of these items, students read a brief passage and then answers an open-ended question about the passage. The passages include a poem titled “The Caged Bird,” a speech titled “Sojourner Truth’s Speech From the 1850s,” and a short story titled “The Lion’s Share.”

The three NAEP language arts items focus on writing. The first writing item asks students to create a narrative piece that describes an embarrassing experience they have had. The second writing item prompt focuses on creative writing and asks students to write a good, scary ghost story. The final writing item tests students’ expository writing skills and asks students to write about their favorite story, telling why they like it and what it means to them.

When selecting the items, two criteria were used. First, the time required to respond to the item could not exceed 30 minutes. Second, the amount of reading (if any) students had to complete before responding to the item could not exceed 1 page. The reason for this second criterion was to maximize the amount of time students spent actually responding to the item. It should be noted that all three MCAS items required students to read a short body of text before responding to a question while none of the NAEP items required students to read any text. For this reason, the MCAS items can be classified as primarily measuring reading comprehension and the NAEP items can be classified as measuring writing ability.

After the six items were selected, they were placed into one of two booklets. Two MCAS items and the 1992 NAEP item formed the test booklet titled Language Arts 1. The remaining MCAS item and the two 1988 NAEP items formed the second test booklet titled Language Arts 2.

Mathematics

The mathematics test booklet contained six items. Three of the items were taken from the 1998 grade 8 spring MCAS test and three items were taken from the 1996 grade 8 NAEP Assessment. Two of the math items tested fractions and proportions. Two items focused on students' ability to read and interpret a graph. One item tested students' ability to calculate and interpret means and medians. And the final item focused on students' problem solving skills.

When selecting mathematics items, two criteria were applied. First, the item had to require students to generate an extended (a minimum of one sentence) written response. Second, the item could not require students to draw a picture, diagram or graph. The first criterion was used to assure that students had to compose text in order to perform well on the item. The second criterion was used to prevent students working on computer from having to access drawing or graphing programs.

Science

Like the mathematics items, three of the science items came from the 1998 grade 8 spring MCAS test and three items came from the 1996 grade 8 NAEP assessment. Similarly, all of the items required students to generate a substantial amount of text (more than a sentence) in order to succeed and none of the items required students to draw pictures or graphs. Two of the items tested students understanding of the physical sciences. Two items focused on human biology. One item tested students understanding of electricity. And the final item tested students' ability to design an experiment.

Scoring Criteria

For all of the items, the scoring criteria developed by MCAS or NAEP were used.

All of the MCAS scoring guidelines used a scale that ranged from 0 to 4. For the MCAS items, a score of 0 indicated that the item was left blank or that the student's response was completely incorrect. Scores of 1 to 4 represented increasingly higher levels of performance.

The scales for the NAEP scoring guidelines varied from 1–3, 1–4, 1–5, and 1–6. A code of 9 was awarded to items that were left blank. For all items, a 1 indicated that the student's response was completely incorrect. Scores of 2 to 6 represented increasingly higher levels of performance. To make the scores for the MCAS and the NAEP items more comparable, all blank responses were re-coded as a zero. The resulting NAEP scales ranged from 0–3, 0–4, 0–5, or 0–6.

Converting Paper Versions to Computer Versions

Before the tests could be administered on computer, the paper versions were converted to a computerized format. Several studies suggest that slight changes in the appearance of an item can affect performance on that item. Something as simple

as changing the font in which a question is written, the order items are presented, or the order of response options can affect performance on that item (Beaton & Zwick, 1990; Cizek, 1991). Other studies have shown that people become more fatigued when reading text on a computer screen than when they read the same text on paper (Mourant, Lakshmanan & Chantadisai, 1981). One study (Haas & Hayes, 1986b) found that when dealing with passages that covered more than one page, computer administration yielded lower scores than paper-and-pencil administration, apparently due to the difficulty of reading extended text on screen. Clearly, by converting items from paper to computer, the appearance of items is altered.

To minimize such effects, students taking a test on computer were given a hard copy of the test booklet. The only difference between the hard copy of the test booklets received by students taking the test on computer and the original paper version was that the blank lines on which students recorded their responses were replaced by instructions to write answers on the computer.

Prior to beginning a test, students in the computer group launched a computer program that performed four tasks. First, the program prompted students to record their name and identification number. Second, the program presented the same directions that appeared in their hard copy. Third, the program allowed students to navigate between text boxes in which they recorded their responses to the open-ended questions presented in their test booklet. Finally, after a student completed the test, the program presented two questions about the taking the test on computer (described more fully below).

To assist students in recording their responses in the proper text box, the program placed the question number and accompanying prompt above each text box. To help avoid confusion, only one text box appeared on the screen at a time. To move between text boxes, two buttons appeared on the bottom of the screen. The button labeled "Next" allowed students to navigate to the text box for the next question and the button labeled "Back" allowed students to move to the previous question. Below the last text box, a button labeled "I'm Finished" appeared. Once students felt they were finished with the test, they clicked on the "I'm Finished" button. To assure that they were in fact finished, students were asked again if they were done. If so, they clicked the "I'm Finished" button again. Otherwise, they clicked the "Back" button to continue working on their responses.

When students were finished taking the test and had selected the "I'm Finished" button twice, they were prompted with two questions about the test. The first asked students: "Do you think you would have done better on this test if you took it on paper. Why?" The second question asked: "Besides not knowing the answer to a question, what problems did you have while taking this test on computer?" Students were required to answer these questions before they could quit the program.

To create a computerized version of the test booklets, the following steps were taken:

1. An appropriate authoring tool, namely Macromedia Director, was selected to create software that would allow students to navigate between questions and to write data to a text file.
2. A data file was created to store student input, including name, ID number, and responses to each item.
3. A prototype of each test was created, integrating the text and database into a seamless application. As described earlier, navigational buttons were placed along the lower edge of the screen. In addition, a “cover” page was created in which students entered their name and id numbers.
4. The prototype was tested on a class of ninth grade students to assure that all navigational buttons functioned properly, that data was stored accurately, and that items were easy to read.
5. Finally, the prototype was revised as needed and the final versions of the computer tests were installed on twenty computers in the ALL School and twenty-four computers in the Sullivan Middle School.

For all questions, examinees used a keyboard to type their answers into text boxes that appeared on the screen. To enable students to write as much as they desired, scrolling text boxes were used for all items. Although they could edit using the keyboard and mouse, examinees did not have access to word processing tools such as spell-checker or grammar-checker.

Sampling Method

The sample of students was drawn from two Worcester Public Middle Schools, namely The Advanced Learning Laboratory (ALL School) and the Sullivan Middle School. Since the analyses focus on how the mode of administration effect varies across achievement levels and, more importantly, computer use/proficiency levels, the population of students was pooled across the two schools. Before sampling began, a list of all grade eight students in the ALL School and all grade eight students from one team in the Sullivan Middle School was generated. In total, this yielded 327 students. For each student on the list, an indicator of prior achievement, namely grade 7 SAT 9 scores, was collected. Since some of the students were new to the district or had not taken the SAT 9 the previous year, SAT 9 scores were only available for 287 students. Using a stratified random assignment procedure, students were then assigned to one of two groups. Group 1 was then assigned to the Language Arts 1 and Math tests and group 2 was assigned to the Science and Language Arts 2 tests. For each group, this process was repeated again, this time assigning half of the students in group 1 to take the Language Arts 1 test on computer first and the remaining half to take the Math test on computer first. Similarly, half of the second group of students was assigned to take the Language Arts 2 test on computer first and the remaining half took the Science test on computer first.

Those students for whom SAT 9 scores were not available were randomly assigned to one of the four groups. Although their scores are not included in the analyses below,

their responses were used to train raters prior to scoring the test booklets for students included in the analyses.

Due to absences and refusals to perform one or more instruments, complete data records were available for 229 students. To be clear, a complete data record was defined as one containing a student's SAT 9 scores, their responses to the student questionnaire, the results of the keyboarding test, and results from at least one of the open-ended tests.

Scoring

To reduce the influence hand writing has on raters' scores (Powers, Fowles, Farnum & Ramsey, 1994), all responses to the open-ended items administered on paper were transcribed verbatim into computer text. The transcribed responses were randomly intermixed with the computer responses. All student responses were formatted with the same font, font size, line spacing and line width. In this way, the influence mode of response might have on the scoring process was eliminated.

Scoring guidelines designed for each item were used to score student responses. To increase the accuracy of the resulting scores, all responses were double-scored. When discrepancies between raters' scores arose, an adjudicator awarded the final score. At the conclusion of the scoring process, one score was recorded for each student response.

To estimate inter-rater reliability, the original scores from both raters were used. The resulting scores were compared both via correlation and percent agreement methods. Table 1 shows that for most items the correlation between the two raters' scores was above .8 and for many items the correlation was above .9. For two of the items on the first language arts test, however, correlations were closer to .7. Nonetheless, this represents an adequate level of inter-rater reliability.

Table 1: Inter-rater Reliability for Open-Ended Items

	Correlation	% Exact Agreement	% Within 1 Point
Language Arts 1			
Item 1	.80	.68	1.00
Item 2	.74	.50	1.00
Item 3	.72	.59	.95
Language Arts 2			
Item 1	.95	.84	1.00
Item 2	.94	.88	1.00
Item 3	.91	.76	1.00
Math			
Item 1	.91	.60	1.00
Item 2	.83	.65	.90
Item 3	.88	.80	.95
Item 4	.94	.80	1.00
Item 5	.84	.90	1.00
Item 6	.70	.75	.95
Science			
Item 1	.80	.64	.95
Item 2	.86	.73	1.00
Item 3	.88	.82	1.00
Item 4	.88	.86	1.00
Item 5	.92	.82	1.00
Item 6	.85	.73	1.00

To estimate intra-rater reliability, one rater double-scored 20% of the responses. The resulting scores for this rater were compared via correlation and percent agreement methods. Table 2 shows high correlations between the two sets of scores. Moreover, where discrepancies occurred, the difference between the two scores was never more than one point.

Table 2: Intra-rater Reliability for Open-Ended Items

	Correlation	% Exact Agreement	% Within 1 Point
Language Arts 1			
Item 1	.92	.86	1.00
Item 2	.95	.95	1.00
Item 3	.88	.77	1.00
Language Arts 2			
Item 1	.91	.82	1.00
Item 2	.93	.91	1.00
Item 3	.94	.86	1.00
Math			
Item 1	.92	.77	1.00
Item 2	.97	.91	1.00
Item 3	.98	.95	1.00
Item 4	.96	.82	1.00
Item 5	.88	.91	1.00
Item 6	.84	.82	1.00
Science			
Item 1	.94	.86	1.00
Item 2	.96	.91	1.00
Item 3	.93	.91	1.00
Item 4	.92	.91	1.00
Item 5	.98	.95	1.00
Item 6	.94	.86	1.00

Note that the adjudicated scores were produced for all students and that the adjudicated scores were used for all analyses described below.

Results

This study explores the relationships between prior computer use and performance on four open-ended test booklets. To examine this relationship, three types of analyses were performed. First, independent samples t-tests were employed to compare group performance. Second, total group regression analyses were performed to estimate the mode of administration effect controlling for differences in prior achievement. And third, sub-group regression analyses were performed to examine the group effect at different levels of keyboarding speed. However, before the results of these analyses are described, summary statistics are presented.

Summary Statistics

Summary statistics are presented for each of the instruments included in this study. For the student questionnaire, keyboarding test, and the SAT 9 scores, summary statistics are based on all 229 students included in the study. For the language arts, math and science open-ended tests, summary statistics are based on the sub-set of students that performed each test. When between group analyses are presented, summary statistics for select variables are presented for each sub-set of students that performed a given test.

Keyboarding Test

The keyboarding test contained two passages. Table 3 shows that the mean number of words typed for passage 1 and passage 2 was 31.2 and 35.0, respectively. As described above, the number of words typed for each passage was summed and divided by 4 to yield the number of words typed per minute for each student. Across all 229 students included in this study, the mean WPM was 16.5. Considering that the minimum WPM required by most employers when hiring a secretary is at least 40, an average of 16.5 WPM suggests that most students included in this study were novice keyboarders.

Table 3: Summary Statistics for the Keyboarding Test

N=229	Mean	Std Dev	Min	Max
Passage 1	31.2	11.2	5	71
Passage 2	35.0	10.6	9	80
WPM	16.5	5.3	4.8	37.8

Student Questionnaire

The student questionnaire contained 11 questions. The maximum score for the Survey was 46 and the minimum score was 10. The scale for each item varied from 0 to 2, 1 to 2, 1 to 3 and 1 to 6. To aid in interpreting the summary statistics presented in table 4, the scale for each item is also listed. In addition to the Survey total score, summary statistics are presented for the Comp-Writing sub-score.

Although comparative data is not available, table 4 suggests that on average students included in this study do not have a great deal of experience working with

computers. The average student reports using a computer for between two and three years, having had a computer in the home for less than a year, and using a computer in school and in their home less than 1–2 hours a week. Furthermore, most students report that they do not use a computer when brainstorming, creating an outline or writing a first draft. Slightly more students report using a computer to edit the first draft. Most students, however, report using a computer at least sometimes to write the final draft. Similarly, most students indicate that if given the choice, they would prefer to write a paper on computer than on paper. Yet, when asked to draw a picture of their writing process, less than half the students included a computer in their drawing.

Again, the divergence between students' preference and their reported use of a computer in the writing process may indicate that when recording their preference some students provided a socially desirable response. If students did provide socially desirable responses, estimating the effect preference had on students' performance will be less precise.

Table 4: Summary Statistics for the Student Questionnaire

N=229	Scale	Mean	Std Dev	Min	Max
Years having computer at home	1–6	2.81	1.86	1	6
Years using computer	1–6	4.75	1.51	1	6
Use computer in school	1–6	2.74	1.10	1	6
Use computer at home	1–6	2.75	1.77	1	6
Brainstorm with computer	1–3	1.35	.55	1	3
Outline with computer	1–3	1.50	.57	1	3
First draft with computer	1–3	1.72	.75	1	3
Edit with computer	1–3	1.85	.73	1	3
Final draft with computer	1–3	2.50	.65	1	3
Preference	1–2	1.80	.40	1	2
Computer in drawing	0–2	.59	.72	0	2
Survey	10–43	24.37	5.68	12	41
Comp-Writing	5–17	9.51	2.77	5	17

Indicator of Prior Achievement

Four indicators of prior achievement were collected prior to the study. Specifically, SAT 9 Composite, Reading, Math and Science NCE scores were collected for each student. Note that the NCE scores provided for this study had been multiplied by 10 before they were supplied by the district office. Thus, the range for the NCE scores was 10 to 990 with a mean of 500 and a standard deviation of approximately 210 (see Crocker and Algina, 1986 for a fuller description of NCE scores). Table 5 displays the mean and standard deviation for each SAT 9 score. The mean score for each subject area and for the composite score for students included in this study is approximately .5 standard deviations below the national average. However, within this sample of students there is a substantial variation.

Table 5: Summary Statistics for Indicators of Prior Achievement

N=229	Mean	Std Dev	Min	Max
Composite NCE	402	150.2	119	888
Reading NCE	391	178	10	990
Math NCE	389	174	10	990
Science NCE	434	172	67	896

Open-Ended Tests

Four open-ended tests were administered in three subject areas: math, science and language arts. As is described more fully above, two versions of the language arts test were administered. Each of the tests was administered to a sample of students. The number of students who performed each test ranged from a high of 117 for Language Arts 1 to a low of 100 for Language Arts 2. Within each sample, approximately half of the students performed the test on computer and half of the students performed the test on paper.

The summary statistics for the total sample of students who performed each test are presented in tables 6 through 9. Since each test contained some items from NAEP and some items from MCAS, it is not possible to directly compare the total test scores to the performance of students in other settings. However, to aid in interpreting the test scores, summary statistics are presented for each item along with the national or state average performance for each item. Note that comparison data for the MCAS items represents the mean score on a 0–4 point scale for all students in the state. Comparison data for the NAEP items represents the percentage of students nationally performing adequately or better on the item.

Table 6 presents the summary statistics for Language Arts 1 and table 7 presents the summary statistics for Language Arts 2. For all items but one item included in the language arts tests, a score below 3 indicates inadequate performance. For item 2 on language arts test 1, a score below 4 indicates inadequate performance. For all items, many students failed to perform adequately. For all items, the mean performance was below 3 and for four items the mean performance was below 2. This low level of performance suggests these items were difficult for these samples of students.

Table 6: Summary Statistics for Language Arts 1

N=117	Scale	Mean	Std Dev	% Adequate*	Mean on MCAS	% Adequate on NAEP
Item 1	0-4	1.42	1.04	19	1.99	NA
Item 2	0-4	1.50	1.05	16	1.73	NA
Item 3	0-6	2.91	1.41	31	NA	29
Total	0-14	5.84	2.85			

* For items 1 and 2, a score of 3 or higher was considered adequate performance.
For item 3, a score of 4 or higher was considered adequate performance.

Table 7: Summary Statistics for Language Arts 2

N=100	Scale	Mean	Std Dev	% Adequate*	Mean on MCAS	% Adequate on NAEP
Item 1	0-4	1.23	0.98	10	1.74	NA
Item 2	0-4	1.67	1.07	31	NA	51
Item 3	0-4	2.12	0.81	22	NA	25
Total	0-12	5.02	2.29			

* For all three items, a score of 3 or higher was considered adequate performance.

Table 8 displays the summary statistics for the Math test. Again, for all items except number 4, a score below 3 indicates inadequate performance. For item 4, a score below 4 indicates inadequate performance. For all items, the mean performance for this sample of students indicates that on average students performed below the adequate level.

Table 8: Summary Statistics for Math

N=110	Scale	Mean	Std Dev	% Adequate*	Mean on MCAS	% Adequate on NAEP
Item 1	0-4	1.18	1.17	15	2.05	NA
Item 2	0-4	1.26	1.23	18	1.83	NA
Item 3	0-4	1.64	1.03	22	1.84	NA
Item 4	0-5	2.45	1.28	33	NA	28
Item 5	0-3	1.44	.53	2	NA	11
Item 6	0-4	1.99	.89	30	NA	26
Total	0-24	9.96	4.18			

* For items 1, 2, 3 and 6, a score of 3 or higher was considered adequate.
For item 4, a score of 4 or higher was considered adequate.
For item 5, a score of 3 was considered adequate.

Table 9 displays the summary statistics for the Science test. For all items, a score below 3 indicates inadequate performance. For all items, the mean performance was below the adequate level.

Table 9: Summary Statistics for Science

N=102	Scale	Mean	Std Dev	% Adequate*	Mean on MCAS	% Adequate on NAEP
Item 1	0–4	1.89	1.02	32	1.49	NA
Item 2	0–4	1.71	1.21	26	1.70	NA
Item 3	0–3	1.50	.75	8	NA	19
Item 4	0–3	1.21	.67	3	NA	9
Item 5	0–3	1.78	.90	22	NA	52
Item 6	0–4	1.57	1.09	20	1.81	NA
Total	0–24	9.66	4.14			

* For all items, a score of 3 or higher was considered adequate performance.

Clearly, students had difficulty with all four of these tests. For all MCAS items, this sample of students performed at a level below that of other students in the state of Massachusetts. For the NAEP items, students performed about as well or worse than other students in the nation.

Comparing Performance on Computer and on Paper

For each test, approximately one half of the sample of students was randomly assigned to perform the test on computer while the other half performed the test on paper. Tables 10 through 13 present the results of between group comparisons for each test. For each test, an independent samples t-test (assuming equal variances for the two samples and hence using a pooled variance estimate) was performed for the total test score. The null hypothesis for each of these tests was that the mean performance of the computer and the paper groups did not differ from each other. Thus, these analyses test whether performance on computer had a statistically significant effect on students' test scores.

To examine whether prior achievement, computer use or keyboarding skills differed between the two groups of students who performed each test, independent samples t-tests were also performed for students' SAT 9 Composite score, the corresponding SAT 9 sub-test score, Survey, Comp-Writing and WPM. The results of these tests are also presented in tables 10 through 13.

Table 10 shows that on average students who performed the first language arts test on paper performed the same as students who performed the test on computer. Similarly, differences between the two groups' SAT 9 Comprehensive scores, SAT 9 Reading scores, Survey scores, and WPM were not statistically significant. However, Table 10 shows that the mean Comp-Writing score for students who performed the language arts 1 test on paper was larger than the mean for students who performed the test on computer.

Table 10: Between Group Comparisons for Language Arts 1

Paper N = 57 Computer N = 60	Mean	Std Dev	SE of Mean	t-value	Sig.
LA 1					
Paper	5.84	2.65	.35		
Computer	5.83	3.04	.39	.02	.99
SAT 9 Comp.					
Paper	379	145	19		
Computer	421	159	21	-1.49	.14
SAT 9 Reading					
Paper	360	168	22		
Computer	426	189	24	-1.98	.05
Survey					
Paper	24.6	5.6	.75		
Computer	23.9	5.9	.76	.67	.51
Comp-Writing					
Paper	10.1	2.8	.37		
Computer	9.0	2.6	.33	2.09	.04*
WPM					
Paper	17.5	4.3	.56		
Computer	16.4	5.3	.68	1.17	.24

*Significant at the .05 level.

On the second language arts test, table 11 shows that Comp-Writing was the only measure on which the two groups differed. However, for the second language arts test, students who performed the test on computer had higher Comp-Writing scores on average than did those students who performed the test on paper. For all other instruments, the two groups did not differ significantly.

Table 11: Between Group Comparisons for Language Arts 2

Paper N = 45 Computer N = 55	Mean	Std Dev	SE of Mean	t-value	Sig.
LA 2					
Paper	5.07	1.70	.25		
Computer	4.98	2.70	.36	.18	.86
SAT 9 Comp.					
Paper	413	138	20		
Computer	393	146	19	.59	.55
SAT 9 Reading					
Paper	402	173	26		
Computer	376	172	23	.74	.46
Survey					
Paper	24.4	5.8	.87		
Computer	25.1	5.7	.76	-.63	.53
Comp-Writing					
Paper	8.9	3.1	.46		
Computer	10.1	2.6	.36	-2.09	.04*
WPM					
Paper	15.7	4.9	.73		
Computer	17.2	6.5	.88	-1.23	.22

*Significant at the .05 level.

With a few exceptions, the students who performed the first language arts test also performed the math test. However, those students who performed the first language arts test on computer performed the math test on paper and vice versa. For this reason, table 12 indicates that the mean Comp-Writing score for the computer group was higher than that of the paper group. Again, this difference is statistically significant. For all other instruments, Table 12 indicates that differences between the two groups' scores were not statistically significant.

Table 12: Between Group Comparisons for Math

Paper N = 54 Computer N = 56	Mean	Std Dev	SE of Mean	t-value	Sig.
Math					
Paper	10.70	4.34	.59		
Computer	9.25	3.90	.52	1.84	.07
SAT 9 Comp.					
Paper	414	155	21		
Computer	407	154	21	.23	.82
SAT 9 Math					
Paper	401	179	24		
Computer	406	190	25	-.16	.87
Survey					
Paper	23.6	5.98	.81		
Computer	25.1	5.72	.76	-1.29	.20
Comp-Writing					
Paper	9.0	2.66	.36		
Computer	10.1	2.65	.35	-2.22	.03*
WPM					
Paper	16.0	4.7	.64		
Computer	17.9	5.2	.69	-2.01	.05

*Significant at the .05 level.

The open-ended science test was the only test for which there was a statistically significant difference in the two groups' test performance. Table 13 shows that on average the computer group performed better than the paper group. There were no other statistically significant differences between the two groups.

Table 13: Between Group Comparisons for Science

Paper N = 51 Computer N = 51	Mean	Std Dev	SE of Mean	t-value	Sig.
Science					
Paper	8.55	3.88	.54		
Computer	10.76	4.14	.58	-2.79	.006*
SAT 9 Comp.					
Paper	388	134	19		
Computer	426	152	21	-1.33	.19
SAT 9 Science					
Paper	414	154	21		
Computer	466	181	25	-1.57	.12
Survey					
Paper	25.4	5.5	.77		
Computer	24.0	5.7	.80	1.22	.23
Comp-Writing					
Paper	9.9	2.6	.37		
Computer	9.1	3.0	.43	1.39	.17
WPM					
Paper	17.1	6.3	.88		
Computer	15.9	5.1	.72	1.01	.32

*Significant at the .05 level.

Note that statistical significance for the t-tests reported above was not adjusted to account for multiple comparisons. Given that six comparisons were made for each group, there is an increased probability that reported differences occurred by chance. Employing the Dunn approach to multiple comparisons (see Glass & Hopkins, 1984), α for c multiple comparisons, α_{pc} , is related to simple α for a single comparison as follows:

$$\alpha_{pc} = 1 - (1 - \alpha)^{1/c}$$

Hence, for six comparisons the adjusted value of a simple 0.05 alpha level becomes 0.009. Analogously, a simple alpha level of 0.01 for a simple comparison becomes 0.001.

Once the level of significance is adjusted for multiple comparisons, the open-ended science test is the only instrument for which there is a statistically significant group difference. This difference represents an effect size of .57 (Glass's delta effect size

was employed). Although this effect size is about half of that reported by Russell and Haney (1997), it suggests that while half of the students in the computer group scored above 10.76, approximately 30% of students performing the test on paper scored above 10.76. The difference between the two groups' open-ended science scores, however, may be due in part to differences in their prior achievement as measured by SAT 9 Science scores.

To control for differences in prior achievement, a multiple regression was performed for each open-ended test. Tables 14 through 17 present the results of each test score regressed on the corresponding SAT 9 score and group membership. For all four regression analyses, the regression coefficient (B) for group membership indicates the effect group membership has on students' performance when the effect of SAT 9 scores is controlled. Group membership was coded 0 for the paper group and 1 for the computer group. A positive regression coefficient indicates that performing the test on computer has a positive effect on students' test performance. A negative regression coefficient suggests that on average students who performed the test on computer scored lower than students who performed the test on paper.

Table 14 indicates that SAT 9 Reading scores are a significant predictor of students' scores on the first open-ended language arts test. For each one standard score unit increase in SAT 9 Reading scores, on average students experience a .42 standard score increase in their test score. Table 14 also indicates that after controlling for differences in SAT 9 Reading scores, performing the first language arts test on computer has a negative impact on students scores. This effect, however, is not statistically significant.¹

Table 14: Language Arts 1 Regressed on SAT 9 Reading and Group Membership

	B	SE B	Beta	T	Signif.
SAT 9 Reading	.007	.001	.42	4.87	<.0001
Group	-.443	.492	-.08	-.90	.37
F	11.85				<.0001
N	117				
R²	.17				
Adjusted R²	.16				

The results for the second language arts test are similar to those for the first language arts test. Table 15 shows that a one point standard score increase in SAT 9 Reading score is associated with a .4 point standard score increase in language arts 2 score and that this effect is statistically significant. Controlling for SAT 9 Reading scores, group membership does not have a significant effect on students' test score.

Table 15: Language Arts 2 Regressed on SAT 9 Reading and Group Membership

	B	SE B	Beta	T	Signif.
SAT 9 Reading	.005	.001	.40	4.3	<.0001
Group	.051	.428	.01	.1	.91
F	9.12				.0002
N	100				
R²	.16				
Adjusted R²	.14				

For both the math and science tests, SAT 9 scores and group membership have statistically significant effects on students' scores. The direction of the effect, however, is different for each test. Table 16 indicates that performing the open-ended math test on computer has a negative effect on students' test scores when SAT 9 Math scores are controlled. For science, this effect is reversed. Table 17 shows that after controlling for differences in SAT 9 Science scores, performing the open-ended science test on computer leads to higher scores than performing the same test on paper. For both tests, the effects are equivalent to just less than .2 standard score units.

Table 16: Math Regressed on SAT 9 Math and Group Membership

	B	SE B	Beta	T	Signif.
SAT 9 Math	.016	.001	.72	11.03	<.0001
Group	-1.546	.541	-.19	-2.86	.005
F	64.41				<.0001
N	110				
R²	.54				
Adjusted R²	.54				

Table 17: Science Regressed on SAT 9 Science and Group Membership

	B	SE B	Beta	T	Signif.
SAT 9 Science	0.014	.002	.59	7.50	<.0001
Group	1.466	.645	.18	2.28	.025
F	34.19				<.0001
N	102				
R²	.41				
Adjusted R²	.40				

Sub-Group Analyses

The regression analyses presented above indicate that mode of administration did not have a significant effect on students' performance on either language arts test. For the science test, performing the test on computer had a positive effect on students' scores. And for the math test, performing the test on computer led to lower performance. For all four of these analyses, the effect was examined across levels of computer use. To test whether the effect of mode of administration varied for students with different levels of computer skill, students' WPM was used to form three groups. The first group contained students whose WPM was .5 standard deviations below the mean, or less than 13.8. The second group contained students whose WPM was between .5 standard deviations below the mean and .5 standard deviations above the mean, or between 13.8 and 19.2. The third group contained students whose WPM was .5 standard deviations above the mean or greater than 19.2. For each group, the open-ended test scores were regressed on SAT 9 scores and group membership.

Table 18 displays the results of the three separate regressions for the first language arts test. For students whose WPM is .5 standard deviations below the mean and for students whose WPM is within .5 standard deviations of the mean, performing the test on computer has a negative effect on their scores. However, for these two groups of students, neither SAT 9 Reading nor group membership is a statistically significant predictor of language arts 1 score. However, for students whose keyboarding speed is one-half of standard deviation above the mean, or greater than 19.2 words per minute, performing the test on computer has a statistically significant positive effect on their performance. This effect is also three times stronger than the relationship between their SAT 9 reading score and their performance on the first language arts test. For the first language arts test, performing the test on computer seems to hurt students whose WPM is near or well below the mean and helps students whose WPM is well above the mean.

Table 18: Language Arts 1 Regressed on SAT 9 Reading and Group for Three Sub-Groups

	B	SE B	Beta	T	Signif.
WPM <13.8 N=30					
SAT 9 Reading	0.006	0.003	.36	1.98	.06
Group	-1.115	1.001	-.20	-1.12	.27
Adjusted R ²	.08				
13.8 < WPM < 19.2 N=54					
SAT 9 Reading	0.004	.002	.23	1.66	.10
Group	-1.330	.694	-.27	-1.91	.06
Adjusted R ²	.05				
WPM >19.2 N=33					
SAT 9 Reading	.003	.002	.19	1.32	.20
Group	2.946	.764	.56	3.86	.0006
Adjusted R ²	.38				

The same relationship was found for the second language arts test (Table 19). However, for this test, the negative effect of taking the test on computer was statistically significant for students whose WPM was .5 standard deviations below the mean. For students whose WPM was within .5 standard deviations of the mean, performing the test on computer also had a negative effect on test performance, but this effect was not statistically significant. For students whose WPM was .5 standard deviations above the mean, performing the test on computer had a positive effect of nearly a half standard score on their language arts 2 test scores. This effect was statistically significant.

Table 19: Language Arts 2 Regressed on SAT 9 Reading and Group for Three Sub-Groups

	B	SE B	Beta	T	Signif.
WPM <13.8 N=35					
SAT 9 Reading	-.0001	.002	-.02	-.10	.92
Group	-1.48	.601	-.41	-2.47	.02
Adjusted R ²	.11				
13.8<WPM >19.2 N=37					
SAT 9 Reading	.002	.002	.17	1.02	.31
Group	-.974	.631	-.26	-1.54	.13
Adjusted R ²	.08				
WPM >19.2 N=28					
SAT 9 Reading	.006	.002	.37	2.32	.03
Group	2.068	.71	.46	2.90	.008
Adjusted R ²	.43				

For the open-ended math test, performing the test on computer had a negative effect on students' scores at all levels of keyboarding (Table 20). However, as keyboarding speed increased, this effect became less pronounced. For students whose WPM was .5 standard deviations below the mean, taking the test on computer had an effect of -.39 standard score units. For students whose WPM was within .5 standard deviations of the mean, this effect was -.24 standard score units. Both of these effects were statistically significant. However, for students whose WPM was .5 standard deviations above the mean, the effect was -.17 standard units and was not statistically significant.

Table 20: Math Regressed on SAT 9 Math and Group for Three Sub-Groups

	B	SE B	Beta	T	Signif.
WPM <13.8 N=30					
SAT 9 Math	.015	.003	.68	5.58	<.0001
Group	-3.068	.970	-.39	-3.16	.004
Adjusted R ²	.57				
13.8<WPM >19.2 N=49					
SAT 9 Math	.010	.003	.50	4.09	.0002
Group	-1.55	.796	-.24	-1.96	.05
Adjusted R ²	.30				
WPM >19.2 N=31					
SAT 9 Math	.019	.003	.73	5.96	<.0001
Group	-1.499	1.070	-.17	-1.40	.17
Adjusted R ²	.56				

Conversely, taking the science test on computer had a positive effect on students' scores at all levels of keyboarding speed (Table 21). However, this effect was only statistically significant for students whose WPM was within .5 standard deviations of the mean. For students whose WPM was .5 standard deviation units above the mean, this effect is less pronounced and is not statistically significant.

Table 21: Science Regressed on SAT 9 Science and Group for Three Sub-Groups

	B	SE B	Beta	T	Signif.
WPM <13.8 N=35					
SAT 9 Science	.011	.003	.50	3.33	.002
Group	.909	1.020	.13	.89	.38
Adjusted R ²	.24				
13.8<WPM >19.2 N=40					
SAT 9 Science	.010	.002	.47	4.00	.0003
Group	3.368	.893	.45	3.77	.0006
Adjusted R ²	.48				
WPM >19.2 N=27					
SAT 9 Science	.021	.004	.70	4.71	.0001
Group	.170	1.204	.02	.14	.89
Adjusted R ²	.45				

Discussion

The experiment described here extends the work of Russell and Haney (1997) and improved upon their study in five ways. First, this study included students whose prior computer experience varied more broadly. Second, many more open-ended items in the area of language arts, math and science were administered. Third, all of the open-ended test items included in this study had been used in state or national testing programs and had been validated previously. Fourth, an indicator of academic achievement was collected prior to the study and was used both to randomly assign students to groups and as a covariate during regression analyses. And fifth, information on students' prior computer use and keyboarding speed was collected and used during analyses.

In their study, Russell and Haney (1997) reported large, positive group differences which were consistent for all writing, math and science open-ended items administered on computer. In this study, a significant positive group difference was found only for the open-ended science test. This effect was about half the size reported by Russell and Haney (1997). However, in this study students' level of prior computer use varied more than it did in the previous study. Although Russell and Haney did not collect a formal measure of computer use, the students included in their study were so accustomed to working on computer that when standardized tests were given, the school had difficulty finding enough pencils for all students. Although three years have passed since the previous study, it may be possible to estimate the difference in the level of prior computer use of the students included in both studies.

This study includes students from two schools, one of which was the focus of the previous study. Table 22 compares the WPM and survey scores for students in the ALL School and Sullivan Middle School. For both measures of computer use and for keyboarding speed, students in the ALL School have significantly higher scores. For the ALL School the mean WPM was nearly .5 standard deviations above the mean for the total sample while the mean for students from the Sullivan Middle School was below the total sample mean. By including Sullivan Middle School students in this study, a broader range and lower levels of computer use were represented. Including students with low levels of computer use and poor keyboarding skills seems to have counteracted the effect described in the previous study since these students performed less well on the language arts computer tests than on the paper tests.

Table 22: Comparison of Computer Use across Participating Schools

ALL N = 35 Sullivan N = 194	Mean	Std Dev	SE of Mean	t-value	Sig.
WPM					
ALL	18.9	5.1	.36		
Sullivan	16.1	6.2	1.05	2.94	.004
Survey					
ALL	27.5	5.8	.99		
Sullivan	23.8	5.5	.39	3.69	<.0001
Comp-Writing					
ALL	11.1	2.6	.43		
Sullivan	9.2	2.7	.20	3.84	<.0001

To examine the effect the mode of administration had on student performance at different levels of computer use, sub-group analyses were performed. Figure 1 summarizes the effects found for three sub-groups: a. students whose WPM was .5 standard deviations below the mean; b. students whose WPM was within .5 standard deviations of the mean; and c. students whose WPM was .5 standard deviations above the mean.

Figure 1: Effect of Performing Test on Computer When Prior Achievement is Controlled

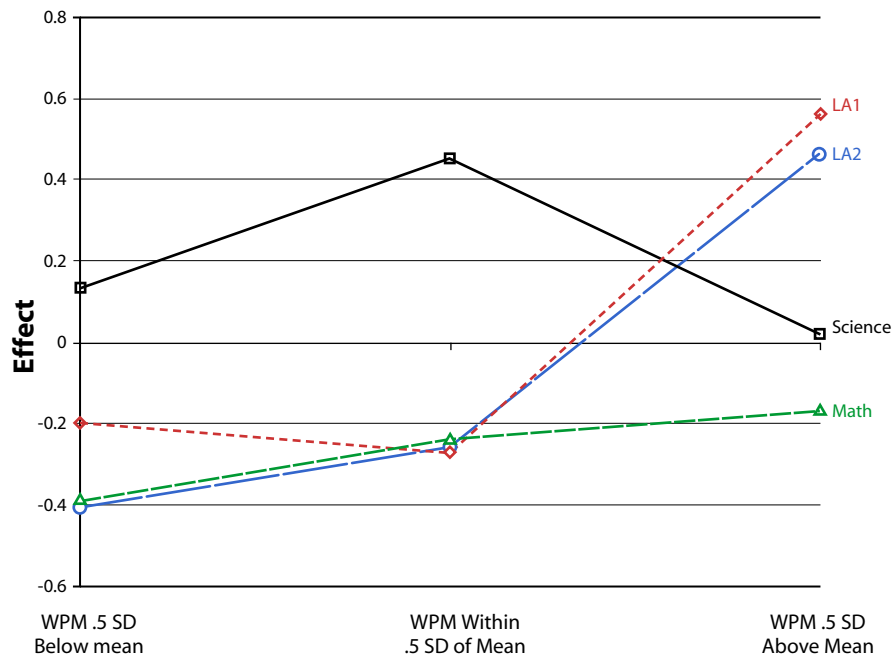


Figure 1 shows that across three of the four tests, performing the test on computer had an adverse effect on the performance of students whose WPM was .5 standard deviations below the mean. Conversely, for students whose WPM was at least .5 standard deviations above the mean, performing the language arts tests on computer had a moderate positive effect. While performing the math test on computer had a negative effect for all students, this negative effect became less pronounced as students' keyboarding speed increased. For the Science test, performing the test on computer had a positive effect across levels of computer use. However, the effect was much larger for students whose WPM was within .5 standard deviations of the mean than it was for students whose keyboarding speed was either .5 standard deviations above the mean or .5 standard deviations below the mean.

Explaining the Effects

To explore the reasons why some students had difficulty working on computers, students were asked to answer the following two questions after they completed the computer version of the test: 1. Do you think you would have done better on this test if you took it on paper? Why?; and 2. Besides not knowing the answer to a question, what problems did you have while taking this test on computer?

Students' responses to these questions were coded in two ways. First, the following numerical code was used:

- 0 – No, I would not perform better on paper
- 1 – I would perform the same or it didn't matter
- 2 – Yes, I would perform better on paper

In addition to these codes, an emergent coding scheme was used to tabulate the reasons students provided for their answers. While coding responses to the post-test questions, it became apparent that when read together, the two questions provided more information about students' experience than reading them separately. Some students would simply write yes or no for the first question, but their reasoning became apparent in their response to the second question. Other students explained the problems they encountered for the first question and wrote little for the second question. For this reason, responses to both questions were read during the emergent coding.

Table 23 presents the numerical codings for the first question. Across all tests, only 10% of students indicated that they would have performed better if they had taken the test on paper. However, over half of those who indicated they would have performed better on paper took the math test on computer. To explore why more students who took the math test on computer felt they would perform better on paper, the full responses to the two follow-up questions were examined.

Table 23: Frequency of Students Responses to Post-Test Question 1—*Do you think you would have done better on this test if you took it on paper?*

	Frequency	Percent
Language Arts 1		
Not better on paper	38	63.3
Same on paper	19	31.7
Better on paper	3	5.0
Language Arts 2		
Not better on paper	36	65.5
Same on paper	15	27.3
Better on paper	4	7.3
Math		
Not better on paper	20	35.7
Same on paper	18	32.1
Better on paper	18	32.1
Science		
Not better on paper	32	62.7
Same on paper	16	31.4
Better on paper	3	5.9

Table 24 presents the frequency of student responses by test. Clearly, the most frequently cited problem related to students' keyboarding skills.² Across all tests, about 25% of the students who performed the test on computer indicated they had difficulty "finding the keys," "pressing the wrong key" or simply said they "couldn't type." Twenty percent of the students who performed the math test on computer also complained that it was difficult to show their work on the computer or that they had to solve problems on paper and then transfer it to the computer. Several students who performed the language arts tests on computer mentioned that they preferred the computer because it was neater and that they didn't have to erase mistakes but could simply delete them. Across all tests, a few students also stated that they preferred the computer because their hands did not get as tired or that it was faster to write on the computer.

Table 24: Frequency of Responses to Post-Test Questions 1 and 2

	LA 1	LA 2	Math	Science	Total
Difficulty typing	12	12	17	18	59
Neater on computer/can delete	8	9	2	2	21
Can't show work/drawings			10		10
Ran out of time on computer	1	2	3	1	7
Hand doesn't get tired	3	1	2	1	7
Faster on computer	3	1	2	1	7
Can take notes/solve problems on paper	1		4		5
Think better on computer/concentrate better	2	1	1	1	5
Write easier on paper			1	2	3
Hard looking back and forth between paper and computer			2		2
Hard to read screen			2		2
Problems with mouse		1	1		2
Easier to concentrate on paper	1	1			2
Write Poorly on paper				1	1
More comfortable on paper		1			1
More space on paper		1			1
Became confused where to put answers			1		1

Examining these responses, it appears that many more students who took the language arts tests recognized the computer's ability to display text that is easy to read and edit as an advantage. Conversely, students who took the math test felt that the inability to present and manipulate numbers in text was a disadvantage. In part, these different reactions to performing the tests on computer may explain the negative group effect for the math test and the positive group effects for the language arts tests. However, students' responses provide little insight into the overall group effect for science.

The Effect of WPM on Student Performance

To further examine the relationship between level of computer use and students' performance on the language arts tests, separate regression analyses were performed for students who performed the tests on paper and those who performed the tests on computer. For each of these regression analyses, the effect of prior computer use on students' performance was estimated controlling for SAT 9 scores. To provide separate estimates for keyboarding speed and for students' survey scores, two sets of regressions were performed for each sub-group. First, the test score was regressed on SAT 9 score, WPM and Survey. Second, the test score was regressed on SAT 9 score, WPM and Comp-Writing. Since Survey is partially composed of Comp-Writing, effects for each variable are estimated through separate regressions to avoid redundancy in the data and hence decrease the effects of colinearity. Figures 2 through 5 display the effects

each variable had on test performance for students who took the test on computer and for those who took the test on paper.

Figure 2 and 3 show that across all tests, WPM is a weak predictor of students' scores when the test is performed on paper. However, for both language arts tests and the science test, WPM is a good predictor of students' scores when the test is performed on computer. This suggests that when these tests are performed on computer, the speed with which a student can type had a significant effect on their performance. However, for the math test, the effect of WPM on students' performance on computer is much less pronounced.

Figure 2: Effect of WPM on Student Performance Controlling for Survey

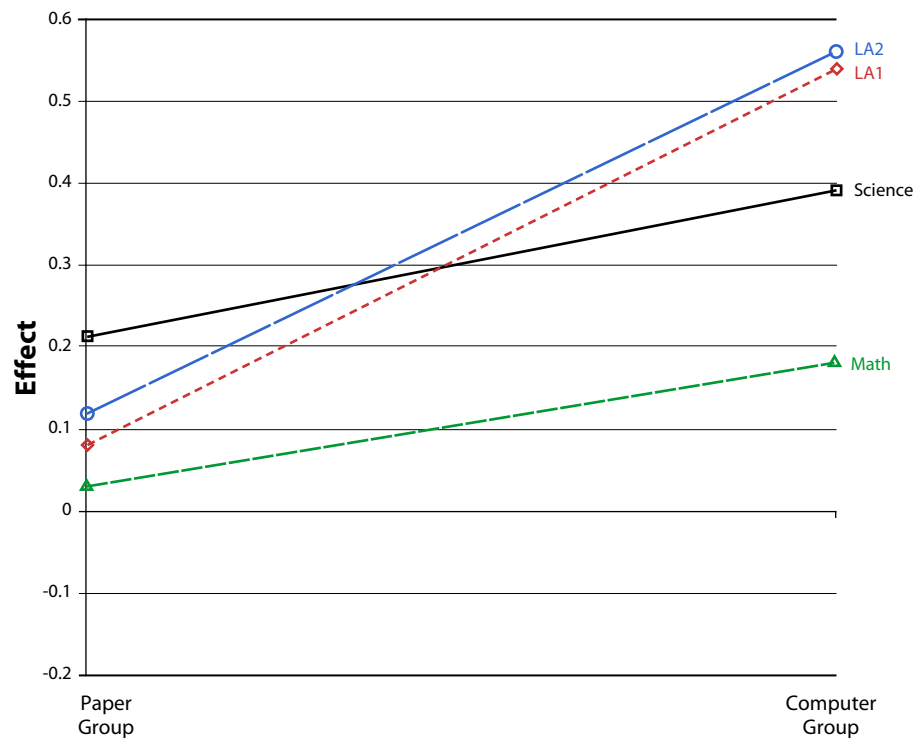
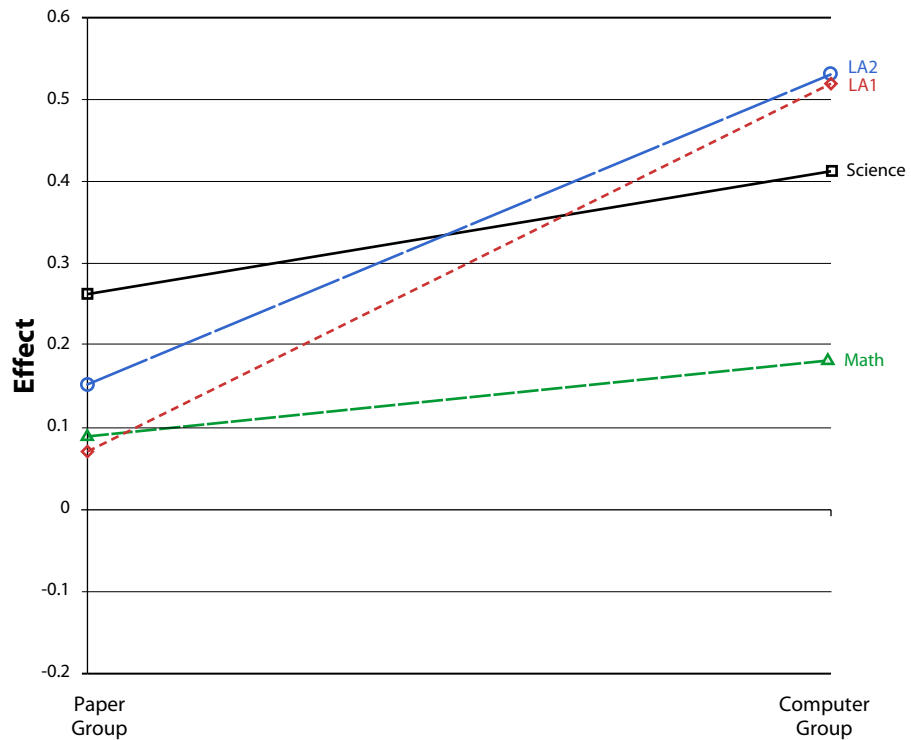


Figure 3: Effect of WPM on Student Performance Controlling for Comp-Writing



Figures 4 and 5 indicate that neither the total Survey score nor the Comp-Writing score had a meaningful effect on the performance of students in either group. In fact, when the effect of WPM is considered, both the amount of prior computer use and use of computers during the writing process have slightly lower effects when the test is taken on computer for the language arts and science tests. Yet, for the math test, the effect is larger and positive. This pattern is difficult to explain. Nonetheless, the weak relationship between either Survey or Comp-Writing and students' performance on computer suggests that students' level of computer use is not as important as their keyboarding proficiency in predicting their performance on open-ended tests. In future studies it is highly recommended that measures of keyboarding speed rather than self-reported levels of computer use are collected and used to examine effects of computer and paper administration.

Figure 4: Effect of Survey on Student Performance Controlling for WPM

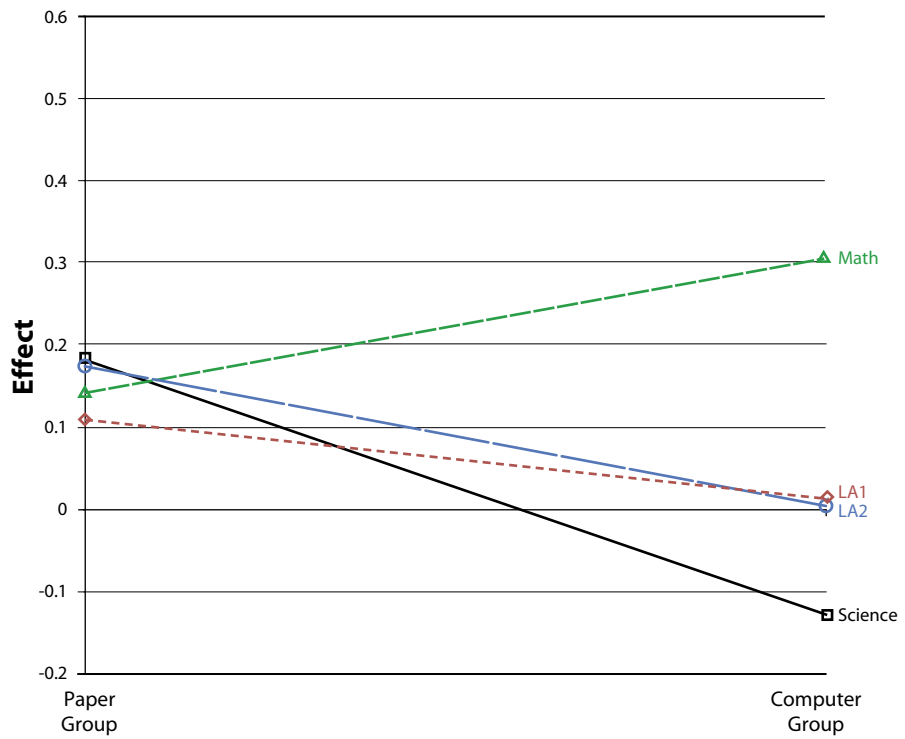
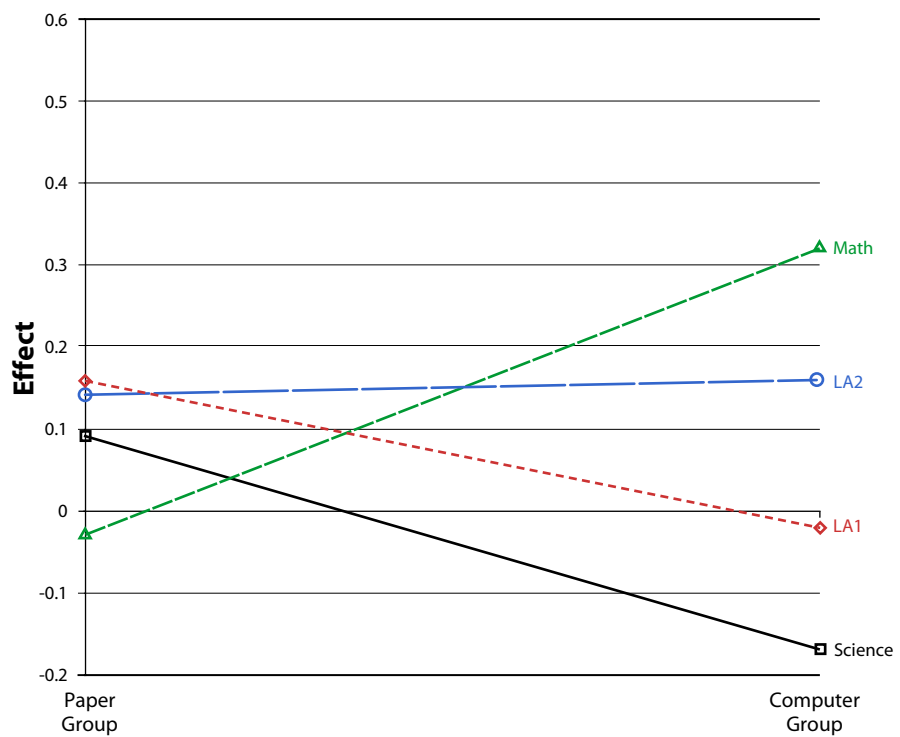


Figure 5: Effect of Comp-Writing on Student Performance Controlling for WPM



Preference and Performance

One of the questions this experiment was designed to address was whether students who performed the test via their preferred medium performed better than predicted and whether those who did not perform the test on their preferred medium performed worse than predicted. Prior to performing either test, students responded to the following survey question: If forced to choose, would you rather write a paper on computer or on paper? To examine the relationship between preference and performance, a dummy variable was coded 1 if the students' preference was the same as the medium on which they performed the test and 0 if their preference and performance medium did not match. For each test, students' test scores were regressed on their SAT 9 scores and Match. Table 25 shows that for the science test, students who took the test on their preferred medium did perform significantly better after controlling for prior achievement. Matching preference with medium of performance did not have a significant effect for the other three tests.

Table 25: Test Score Regressed on SAT 9 and Match

	B	SE B	Beta	T	Signif.
Language Arts 1 Match=43 NoMatch=74					
SAT 9 Reading	.007	.001	.42	4.83	<.0001
Match (1=yes)	-.367	.510	-.06	.72	.47
Adjusted R ²	.15				
Language Arts 2 Match=53 NoMatch=47					
SAT 9 Reading	.005	.001	.40	4.27	<.0001
Match (1=yes)	-.522	.422	-.11	1.24	.22
Adjusted R ²	.15				
Math Match=63 NoMatch=47					
SAT 9 Math	.016	.002	.72	10.80	<.0001
Match (1=yes)	-.948	.561	-.11	-1.69	.09
Adjusted R ²	.52				
Science Match=51 NoMatch=51					
SAT 9 Science	.014	.002	.58	7.46	<.0001
Match (1=yes)	1.487	.646	.18	2.30	.02
Adjusted R ²	.40				

As discussed above, preference for some students seems to have been influenced by social desirability. As a result, the relationship between preference and performance on the preferred medium may be poorly estimated. Simply giving students the alternative to perform open-ended test questions via their "preferred" medium may not reduce the effect of medium found in this study. Rather, before students are given the choice, it might be useful to explain the apparent relationship between keyboarding speed and performance.

Gender, Keyboarding and Performance on Computers

Recent research suggests that females do not use computers in school as frequently as males (ETS, 1998). If this research is accurate, it is possible that the keyboarding skill of females is less developed than males. Given the relationship between WPM and performance on computer, performing tests on computer may have an adverse impact on the scores for females.

To examine the relationship between gender and WPM, an independent samples t-test was performed using all 229 students included in the study. To examine whether there were gender differences on computer use and prior achievement, t-tests were also performed for Survey, Comp-Writing and the SAT 9 comprehensive NCE. Table 26 indicates that WPM was the only variable for which there was a gender difference. However, on average, it was males' keyboarding speed that was 3 words per minute slower than females. This represents an effect size of approximately .68. This difference, however, does not seem to be caused by less computer experience or less use of computers in the writing process since there were negligible differences for either Survey or Comp-Writing.

Table 26: Gender Differences for WPM, Survey, Comp-Writing and SAT 9 Comprehensive

Males=97 Females=132	Mean	Std. Dev.	SE	T-value	Signif.
WPM					
Males	14.8	4.4	.45		
Females	17.8	5.6	.49	4.41	<.001
Survey					
Males	24.2	5.5	.56		
Females	24.5	5.8	.51	.37	.72
Comp-Writing					
Males	9.4	3.0	.31		
Females	9.6	2.6	.22	.56	.58
SAT 9 Comp.					
Males	408	160	16.3		
Females	398	143	12.4	.48	.63

As described above, WPM was a significant predictor for students' performance on computer in all subject areas. But given that males were on average slower keyboarders, one might expect their scores in all tests to be lower when performed on computer. Table 27 shows that this was the case for all four tests but that the difference was only significant for the first language arts test.

Table 27: Gender Differences for Test Performance on Computers

	Mean	Std. Dev.	SE	T-value	Signif.
LA 1					
Males (26)	4.96	2.60	.51		
Females (34)	6.50	3.22	.55	1.99	.05
LA 2					
Males (24)	4.33	2.57	.52		
Females (31)	5.48	.273	.49	1.59	.12
Math					
Males (21)	8.24	4.38	.96		
Females (35)	9.86	3.51	.59	1.52	.13
Science					
Males (21)	10.48	4.62	1.01		
Females (30)	10.97	3.83	.70	.41	.68

Table 28 shows that gender differences were not found for any tests when prior achievement and WPM were controlled. In part, this finding suggests that although males included in this study tended to be slower keyboarders, they performed as well as females with similar keyboarding and SAT 9 scores. This finding provides further evidence that keyboarding skills play an important role in how well students, regardless of their sex, perform on computers.

Table 28: Test Score Regressed on SAT 9, WPM and Gender for Computer Groups Only

	B	SE B	Beta	T	Signif.
Language Arts 1 N=60					
SAT 9 Reading	.004	.002	.28	2.28	.03
WPM	.258	.072	.45	3.573	.0007
Sex (1=Male)	-.888	.648	-.15	1.37	.18
Adjusted R ²	.43				
Language Arts 2 N=55					
SAT 9 Reading	.003	.002	.20	1.51	.14
WPM	.246	.059	.59	4.15	.0001
Sex (1=Male)	.311	.605	.06	.51	.61
Adjusted R ²	.48				
Math N=56					
SAT 9 Math	.011	.002	.53	4.67	.0001
WPM	.177	.089	.23	2.01	.05
Sex (1=Male)	-.605	.833	-.08	.73	.47
Adjusted R ²	.45				
Science N=51					
SAT 9 Science	.011	.003	.49	4.25	.0001
WPM	.266	.095	.33	2.79	.008
Sex (1=Male)	-.379	.935	-.05	.41	.69
Adjusted R ²	.42				

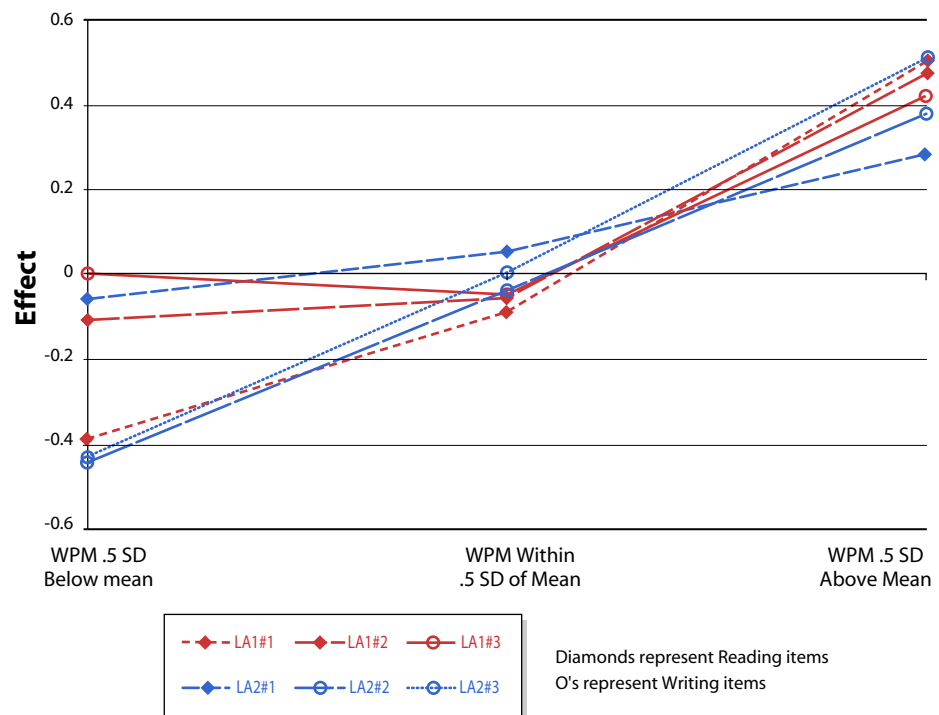
Reading Comprehension vs. Writing Items

For students whose WPM was at least .5 standard deviations below the mean, performing either language arts test on computer had a negative effect on students' test scores. The negative effect was much larger for the second language arts test than it was for the first. To explore why the effect was larger for the second language arts test, the content of the two tests was examined.

Recall that the language arts tests contained two types of items, namely reading comprehension and writing. The first language arts test contained two reading comprehension items and only one writing item while the second language arts test contained two writing items and only one reading comprehension item. To examine the relationship between item type and the effect at each level of keyboarding speed, separate regressions were performed for each item. Figure 6 indicates that the effect for all items are about the same for students whose WPM is within .5 standard deviations of the mean or at least .5 standard deviations above the mean. However, for students whose WPM is at least .5 standard deviations below the mean, there seem to be two different effects. Language arts 1 item 1 and language arts 2 items 2 and 3 all seem to have an effect of about -.4. Language arts 1 item 2 and 3 and language arts 2 item 1

have effects between 0 and $-.1$. This pattern, however, does not seem to be related to item format. Although language arts 1 item 1 addresses reading comprehension, the two other items showing a similar effect test writing skills. Similarly, although language arts 1 item 2 and language arts 2 item 2 are both reading comprehension, the third item in this triad is a writing item. Thus, item format does not seem to explain the differences in the effect sizes for the two language arts tests at the low level of keyboarding speed.

Figure 6: Effects by Language Arts Item



Explaining Smaller Effect Sizes

As noted above, the magnitude of the effects in this study are about half the size reported by Russell and Haney (1997). While these positive effects are still quite large and represent approximately one half of a standard deviation difference in test scores, there are three observations that may shed some light on why the effects in this study were less pronounced than in the 1997 study.

First, the test scores used in Russell and Haney's study were part of a formal testing program. In the study reported here, the tests were described to students as practice for the spring MCAS administration and thus may not have been taken as seriously by students, especially those unaccustomed to working on computer. This was particularly evident during the computer administration. Whereas the author noted only one student being disciplined during four paper administration sessions that he observed, nearly 40 behavioral problems (e.g., students talking, students touching each other,

or students moving around the room without permission) were addressed during the seven computer administration sessions that he observed. This increased level of disruptions may have occurred in part because students were frustrated by their inability to type. These disruptions also may have distracted students who did not experience difficulty keyboarding. Between not being as motivated during a practice test and being distracted more often, students' performance on computer may have suffered. In turn, this may have led to under-estimates of the positive effects and over-estimates of the negative effects.

Second, when the previous study occurred, the ALL school was in its third year of reform and was receiving full external support for its technology reforms. For this reason, there was great enthusiasm for the use of technology by teachers and students. As noted in the previous study, at the time students performed almost all of their work on computer. Since then, three years have passed, there has been a turn-over in teachers, and the external support for the ALL school has largely disappeared. For these reasons, it is possible that students in the ALL School are not using technology as extensively as they did three years ago. This is supported to some extent by the relatively low mean keyboarding speed for students in the ALL school. Although the ALL School's mean WPM was significantly higher than that of the Sullivan Middle School, it was still less than 20 WPM. This low keyboarding speed suggests that although students in the ALL School use computers more often than students in the Sullivan Middle School, their keyboarding skills are not as developed as one might expect if students are using computers on a daily basis. Although there is no direct data to confirm possible decrease use of computers in the ALL School, a decreased use might partially account for the smaller effect.

Finally, given the findings of the previous study and the heavy emphasis the Massachusetts Department of Education has placed on schools' performance on MCAS, it is possible that teachers require students to write more on paper in the ALL School now than three years ago in order to improve their performance on open-ended items. Sadly, after sharing this hypothesis with the ALL School's principal, Carol Shilinsky confirmed that in preparation for MCAS, teachers now require students to perform most of their writing on paper. If this was a successful strategy, then it would have improved students' scores on paper. In turn, the size of the effect of performing the tests on computer would be decreased.

Limitations

Despite efforts to create equivalent groups and to control for the confounding effects of scoring handwritten and computer printed responses, reading extensive passages of text on screen, and only using items that had been formally validated, this study still had several limitations. First, only a small group of students from one urban district were included. Recent research suggests that computers are not used the same way in all schools and that there are meaningful differences in the way students in urban and suburban schools use computers, particularly for math (ETS, 1998). These differences may lead to different effects for students in different settings.

Second, the tests were not administered under formal, controlled testing conditions. This may have decreased motivation, increased distractions and led to under-performance for many students. As noted above, this may be particularly true for

students with better keyboarding skills who performed the tests on computer.

Third, although this study included many more open-ended items than did the previous study, testing time for each test was limited to sixty minutes. In order to increase the number of items included in the study, the time required to respond to items was limited, on average, to 10 minutes for math and science and 20 minutes for language arts. This time limit precluded extended writing and extended math items (requiring more than 10 minutes) from the study. However, MCAS and other testing programs include more extended open-ended items. And the effect of performing these types of items on computer may be larger given that in order to perform well, students generally need to produce more text.

Fourth, the sample of students included in this study had relatively slow keyboarding skills. For this reason, it was not possible to estimate the effect of taking open-ended tests on computer for students who are proficient or advanced keyboarders. Given the sharp increase in the size of the effect as keyboarding speed increases from near the mean to .5 standard deviations above the mean (see figure 5.1), it is possible that the effect of performing tests on computer is even larger for students with more advanced keyboarding skills.

Implications

This study suggests that for students who keyboard about 20 words per minute or more, performing open-ended language arts tests on paper substantially underestimates their level of achievement. However, for slower keyboarders, performing open-ended tests on computer adversely affects their performance. To provide more accurate estimates of students' achievement, these findings suggest that students who can keyboard at a moderate level should be allowed to compose their responses to open-ended items on computers. Conversely, students with weak keyboarding speed should compose their responses on paper.

This study also demonstrates that for math tests, performance on computer underestimates students' achievement regardless of their level of keyboarding speed. This occurred despite efforts to include items that did not require students to draw pictures or graphs to receive credit. Nonetheless, about 20% of the students who performed the math test on computer indicated that they had difficulty showing their work and/or needed scrap paper to work out their solutions. For these reasons, it is likely that the negative effect found in this study underestimates the effect that would occur if a full range of open-ended math items were included.

This study also re-emphasizes the danger of making inferences about students or schools based solely on paper-and-pencil tests. Similarly, as the public investigates the impact computers have on student learning (Oppenheimer, 1997), caution should be taken when student learning is measured by tests containing open-ended items. As found in the previous study, scores on paper and pencil tests for students accustomed to working on computer may substantially under-estimate students' achievement. As computer use in schools and at home continues to increase rapidly, it is likely that more students will develop solid keyboarding skills and, thus, will be adversely affected by taking open-ended tests on paper.

Finally, this study provides further evidence that the validity of open-ended tests should be considered in terms of both content and medium of learning. Until all students have access to and use computers regularly, open-ended tests administered via a single medium, either paper or computer, will likely under-estimate performance of students accustomed to working in the alternate medium. Based on this study, further research on a larger scale into computers and open-ended tests is clearly warranted. Until then, we should exercise caution when drawing inferences about students based on open-ended test scores when the medium of assessment does not match their medium of learning.



Notes

- ¹ Note that for each group, only one regression analyses was performed. Hence, statistical significance was not adjusted for multiple comparisons. For all regression coefficients, simple alpha of .05 was used to determine statistical significance.
- ² Note that several students indicated that they had difficulty typing but did not indicate that they would perform better on paper.

References

- Bangert-Drowns, R. L. (1993). The word processor as an instructional tool: A meta-analysis of word processing in writing instruction. *Review of Educational Research*, 63(1), 69–93.
- Beaton, A. E. & Zwick, R. (1990). *The Effect of Changes in the National Assessment: Disentangling the NAEP 1985–86 Reading Anomaly*. Princeton, NJ: Educational Testing Service, ETS.
- Bunderson, C. V., Inouye, D. K. & Olsen, J. B. (1989). The four generations of computerized educational measurement. In Linn, R. L., *Educational Measurement* (3rd Ed.), Washington, D.C.: American Council on Education, pp. 367–407.
- Burstein, J., Kaplan, R., Wolff, S., & Lu, C. (1997). *Using lexical semantic techniques to classify free-responses*. A report issued by Educational Testing Service. Princeton, NJ.
- Cizek, G. J. (1991). The effect of altering the position of options in a multiple-choice Examination. Paper presented at NCME, April 1991. (ERIC)
- Crocker, L. & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. Orlando, FL: Harcourt Brace Jovanovich College Publishers.
- Daiute, C. (1986). Physical and cognitive factors in revising: insights from studies with computers. *Research in the Teaching of English*, 20 (May), p. 141–59.
- Educational Testing Service. (1998). *Does it compute? The relationship between educational technology and student achievement in mathematics*. Princeton, NJ: Policy Information Center Research Division Educational Testing Service.
- Etchison, C. (1989). Word processing: A helpful tool for basic writers. *Computers and Composition*, 6(2), 33–43.
- Glass, G. & Hopkins, K. (1984). *Statistical Methods in Education and Psychology*. Boston, MA: Allyn and Bacon.
- Glennan, T. K., & Melmed, A. (1996). *Fostering the use of educational technology: Elements of a national strategy*. Santa Monica, CA: RAND.
- Grejda, G. F. (1992). Effects of word processing on sixth graders' holistic writing and revision. *Journal of Educational Research*, 85(3), 144–149.
- Haas, C. & Hayes, J. R. (1986b). What did I just say? Reading problems in writing with the machine. *Research in the Teaching of English*, 20(1), 22–35.
- Haas, C. & Hayes, J. R. (1986a). Pen and paper versus the machine: Writers composing in hard-copy and computer conditions (CDC Technical Report No. 16). Pittsburgh, PA: Carnegie-Mellon University, Communication Design Center. (ERIC ED 265 563).

- Hannafin, M. J. & Dalton, D. W. (1987). The effects of word processing on written composition. *The Journal of Educational Research*, 80 (July/Aug.) p. 338–42.
- Hawisher, G. E. & Fortune, R. (1989). Word processing and the basic writer. *Collegiate Microcomputer*, 7(3), 275–287.
- Hawisher, G. E. (1986, April). The effect of word processing on the revision strategies of college students. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA. (ERIC ED 268 546)
- Kerchner, L. B. & Kistingner, B. J. (1984). Language processing/word processing: Written expression, computers, and learning disabled students. *Learning Disability Quarterly*, 7(4), 329–335.
- Kurth, R. J. (1987). Using word processing to enhance revision strategies during student writing activities. *Educational Technology*, 27(1), 13–19.
- MacArthur C. & Graham, S. (1987). Learning disabled students' composing under three methods of text production: Handwriting, word processing and dictation. *Journal of Special Education*, 21(3), 22–42.
- MacArthur, C. A. (1988). The impact of computers on the writing process. *Exceptional Children*, 54(6), 536–542.
- Mourant, R. R., Lakshmanan, R. & Chantadisai, R. (1981). Visual Fatigue and Cathode Ray Tube Display Terminals. *Human Factors*, 23(5), 529–540.
- Nichols, L. M. (1996). Pencil and paper versus word processing: a comparative study of creative writing in the elementary school. *Journal of Research on Computing in Education*, 29(2), 159–166.
- Oppenheimer, T. (1997). The computer delusion. *The Atlantic Monthly*. 280(1), 45–62.
- Owston, R. D. (1991). Effects of word processing on student writing in a high-computer-access environment (Technical Report 91-3). North York, Ontario: York University, Centre for the Study of Computers in Education.
- Phoenix, J. & Hannan, E. (1984). Word processing in the grade 1 classroom. *Language Arts*, 61(8), 804–812.
- Powers, D., Fowles, M, Farnum, M, & Ramsey, P. (1994). Will they think less of my handwritten essay if others word process theirs? Effects on essay scores of intermingling handwritten and word-processed essays. *Journal of Educational Measurement*, 31(3), 220–233.
- Robinson-Stavely, K. & Cooper, J. (1990). The use of computers for writing: Effects on an English composition class. *Journal of Educational Computing Research*, 6(1), 41–48.
- Russell, M. & Haney, W. (1997). Testing writing on computers: an experiment comparing student performance on tests conducted via computer and via paper-and-pencil. *Education Policy Analysis Archives*, 5(3), <http://olam.ed.asu.edu/epaa/v5n3.html>.

- Sitko, M.C. & Crealock, C. M. (1986, June). A longitudinal study of the efficacy of computer technology for improving the writing skills of mildly handicapped adolescents. Paper presented at the Invitational Research Symposium on Special Education Technology, Washington, DC.
- Snyder, T. D. & Hoffman, C. (1990). *Digest of Education Statistics*. Washington, DC: U. W. Department of Education.
- Snyder, T. D. & Hoffman, C. (1994). *Digest of Education Statistics*. Washington, DC: U. W. Department of Education.
- Vacc, N. N. (1987). Word processor versus handwriting: A comparative study of writing samples produced by mildly mentally handicapped students. *Exceptional Children*, 54(2), 156–165.
- Williamson, M. L. & Pence, P. (1989). Word processing and student writers. In B. K. Briten & S. M. Glynn (Eds.), *Computer Writing Environments: Theory, Research, and Design* (pp. 96–127). Hillsdale, NJ: Lawrence Erlbaum & Associates.
- Zandvliet, D. & Farragher, P. (1997). A comparison of computer-administered and written tests. *Journal of Research on Computing in Education*, 29(4), 423–438.

