



www.intasc.org

Effects of Computer Versus Paper Administrations of a State-Mandated Writing Assessment

**Michael Russell and Tom Plati
Technology and Assessment Study Collaborative
CSTEPP, Boston College
332 Campion Hall
Chestnut Hill, MA 02467**



Effects of Computer Versus Paper Administrations of a State-Mandated Writing Assessment

Michael Russell and Tom Plati
Technology and Assessment Study Collaborative
CSTEED, Boston College
Released June 2000

Introduction

The prominence of educational testing and educational technology have increased rapidly over the past decade. When well implemented, state level testing programs encourage schools to think more critically about their curriculum and may provide guidance on topics and skills students need to further develop. When applied to meet curricular goals, education technologies provide alternative approaches to sustaining students' interest, developing students' knowledge and skills, and provide supplementary materials that teachers can use to extend students' learning. As one example, several studies have shown that writing with a computer can increase the amount of writing students perform, the extent to which students edit their writing, and, in turn, lead to higher quality writing (Dauite, 1985; Etchison, 1989; Hanafin & Dalton, 1987; Hass & Hayes, 1986; Kerchner & Kistingner, 1984; MacArthur, 1988; MacArthur & Graham, 1987; Nichols, 1996; Owston, 1991; Owston & Wideman, 1997; Phoenix & Hannan, 1984).

Recent research, however, suggests that these two strategies for improving education, namely state level testing programs and writing on computers, may work against each other. As Russell and Haney (2000) describe, two previous studies provide evidence that students accustomed to writing on computers perform better on written tests when these students are allowed to compose their responses using a word processor (without access to spell checker or grammar checker). Despite this improved performance on computer, scores from paper tests are used to make judgments about students, schools, and districts (Sacks, 1999), as well as the impact of technology on student learning.

This article builds on two previous studies (Russell, 1999; Russell & Haney, 1997) that explored the effect mode of administration, that is computer versus paper-and-pencil, has on student performance on open-ended items requiring written responses.

Whereas the two previous studies have focused on middle school students, the study presented here focuses on students in eighth and tenth grade. In addition, whereas the previous studies examined the effect on relatively short open-ended items that ranged from two to thirty minutes in length, this article focuses on extended composition items designed to be completed during two 45 to 60 minute blocks of time. Finally, whereas the previous studies focused on students attending an urban school district that had just begun placing large numbers of computers in its schools and which generally performs below the state average on state tests, this article focuses on students attending school in a suburban district in which computers are used regularly in most classrooms and which generally performs well above the state average on state tests. Given the increasing emphasis many states are placing on student performance on state-level tests, the study presented here provides further evidence that state-level tests containing open-ended items are under-estimating the achievement of students accustomed to using computers for writing.

Background

In the past, it was unrealistic to administer tests containing open-ended items on computers in elementary and secondary schools. Until recently, most schools did not have a sufficient number of computers to administer tests in an efficient and timely manner. Moreover, until just a few years ago, most students did not work regularly on computers. The situation, however, is changing rapidly and increasing numbers of students are able to access and write regularly on computers in schools.

While schools had one computer for every 125 students in 1983, they had one for every 9 students in 1995, and 1 for every 6 students in 1998 (Market Data Retrieval, 1999). In 1998, a national survey of teachers also showed that 50 percent of K–12 teachers had students use word processors, 36 percent had them use CD-ROMS, and 29 percent had them use the WWW (Becker, 1999). Although it is unclear how computers are affecting student achievement in schools (see, for example, Fabos & Young, 1999, questioning the efficacy of Internet based telecommunications exchange programs in schools), there is little doubt that more and more students are writing and performing school assignments on computers.

Computers and Writing

Writing is a recursive process that involves four inter-related steps: prewriting, writing, revision, and editing (Calkins, 1983, 1986; Graves, 1983). Over the past two decades, educators have identified several ways computers can be helpful during each of these stages. Researchers have also conducted numerous studies that examine the impact computer use has on students, their writing process and the quality of their products. Below, findings from some of these formal studies are summarized.

When reviewing these studies, it is important to heed MacArthur's warning that computers are not a "magical writing tool" that independently transform students' writing (1988, p 536). As Cochran-Smith, Paris and Khan (1991) and Owston and Wideman (1997) point out, the extent to which computers impact students' writing is dependent upon the context in which students use computers. In addition to

the instructional practices employed by teachers, these contexts include the amount, regularity and longevity of use (Owston & Wideman, 1997), type of word processors employed (Haas, 1989) and students' ability to interact with a computer via a keyboard and commands (Dauite, 1984; MacArthur, 1988, MacArthur & Graham, 1987; Nichols, 1996). As just two examples, MacArthur and Graham (1987) found that keyboarding speed was highly correlated with the length of passages created on a computer. Similarly, Hass (1989) reported that students using a word-processor equipped with a modern graphical user interface (GUI) wrote longer and higher quality essays than did students who wrote using older command-based software.

Despite Bridwell's (as cited in Collier & Werier, 1995, p. 49) assertion that "the real effects of writing with computers probably should be measured long after a writer had made initial adjustments to the medium," many studies have examined the impact of computers on student writing after relatively short periods of use. As summarized below, when students are provided the opportunity to develop basic keyboarding and word processing skills and use computers regularly over an extended period of time, the quality of their writing improves at a faster rate. When these conditions are not met, the impact of computers on students' writing is less pronounced, but is still often positive.

Quality of Writing

In a meta-analysis of 32 computer writing studies, Bangert-Drowns (1993) reported that about two-thirds of the studies indicated improved quality for text produced on computer. Generally, the largest improvements in the quality of writing produced on a computer are found for learning disabled students (Kerchner & Kisting, 1984; MacArthur & Graham, 1987; Sitko & Crealock, 1986), elementary students (Phoenix & Hannan, 1984), low-achieving students (Hannafin & Dalton, 1987; Robinson-Stavelly & Cooper, 1990) and college-aged students (Hass & Hayes, 1986; Hawisher, 1986; Williamson & Pence, 1989). For middle school and high school students, the impact of computers on the quality of students' writing is less consistent. Several studies, for example, report that compositions produced on a computer were rated higher in quality than those produced on paper (Breese, Hackson, & Prince, 1996; Hannafin & Dalton, 1987; Larter, 1987; Moore, 1987, Owston, 1991; Snyder 1993), while other studies have found that computers did not significantly improve the quality of students' writing (Dauite, 1986; Etchinson, 1989; Grejda, 1992; Nichols, 1996; Vacc, 1987). Finally, two studies that allowed late elementary and middle school students to use GUI word processors and were conducted over an extended period of time report large, significant improvements in the quality of students' writing (Owston, Murphy, & Wideman, 1992; Owston & Wideman, 1997).

Over the past fifteen years, educators and researchers have identified several reasons why computers may help improve students' writing. These impacts occur as students generate text, revise text, share their work with others, and, more recently, receive feedback directly from the computer. Below, each of these areas of impact is discussed.

Generating Text

Clearly, the ability to efficiently create neat, formatted text on a computer is advantageous. For young writers, especially those with poor handwriting skills, typing can greatly increase the speed and clarity with which students are able to record their thoughts on paper (MacArthur, 1988). For early adolescents, who tend to want results more quickly, the efficiency with which computers allow students to record their thoughts reduces students' perception that writing is a slow, painful process (Dauite, 1985).

Knowing that errors, transitions, and other problem areas can be addressed later without having to recopy entire bodies of text, students are freed to focus more closely on their ideas during the initial stages of their writing (Dauite, 1985; Wresch, 1984). By focusing on the flow of thought rather than on grammar, spelling and punctuation, computers can help students develop a better sense of fluency (Elbow, 1981).

Many studies report that the ease of generating neat, formatted text increases students' motivation to write (Kerchner & Kistingner, 1984; Sitko & Crealock, 1986; Morrocco and Neuman, 1986). Other studies have documented that students produce more and longer passages when computers are used (Hass & Hayes, 1986; Etchison, 1989; Dauite, 1985). Dauite (1985) also reports that students tend to work more quickly on computer than on paper, producing longer passages in shorter amounts of time.

Revising Text

After learning how to use a computer, Dauite (1985) reports that most children prefer to write on a computer because the mundane task of recopying their work is eliminated. Many writers also find that the neat appearance of text both on the screen and particularly in print makes it easier to critique and revise early drafts. And for experienced writers, neatly formatted text enables them to see problems that may otherwise be obscured by revisions written in margins or scribbled directly over text (Dauite, 1985). In turn, the ability to alter this neat text without permanently damaging the appearance of the medium on which it appears may help encourage young students to make edits to their text (MacArthur, 1988). This theory is echoed by Wresch (1984), who suggests that the impermanence of text on the screen helps students see writing as a fluid process through which ideas can be presented in several different ways and that the text they generate is not 'carved in stone'. "Instead, their words have the quality of light. Their sentences slide back and forth, ripple down the screen, become highlighted, disappear, and reappear" (Marcus, 1984, p. 122). Once cognizant of this impermanence, students may be more willing to make changes to their text after initial drafts are produced.

Several studies have found that students who write on computer return more often to their text and make more revisions to their text than do students who generate text on paper (Nichols, 1996; MacArthur, 1988; MacArthur & Graham, 1987; Vacc, 1987). More focused research, however, suggests that the quality of the revisions made varies by the writer's experience. Bridwell and Ross (1984) report that experienced adult writers revise more extensively when using a word processor as compared to using paper-and-pencil. Dauite (1986), MacArthur and Graham (1987) and

Scardamalia and Bereiter (1986), however, report that word processing has limited impact on the quality of revision by inexperienced writers. Although inexperienced writers may revise more on computer, the revisions they make focus primarily on surface features of their writing whereas more experienced writers revise both to clarify meaning and to make surface corrections. However, when computer use is coupled with on-going instruction that focuses on revising, revisions made by young writers can lead to qualitative improvements in their writing (Owston & Wideman, 1997).

Collaboration and Audience

The computer's ability to display text on a monitor which can be observed directly by others during the writing process and immediately after completing a block of text allows writers to work collaboratively on pieces of writing and to easily share their work in progress (Levin, Riel, Rowe and Boruta, 1985). Similarly, with several computers placed in a single classroom or a computer lab, writing on a computer becomes a public process in which writers can display their work to others and receive immediate feedback (MacArthur, 1988). Dauite (1984) also notes that by interacting with the word processor through program commands, the computer may seem like an audience to some writers. For some writers, this sense of interaction makes writing feel as natural as speaking. For others, the blinking cursor may act as a prod to keep writing (Dauite, 1985).

Feedback

Advances in the computer's ability to provide feedback on specific aspects of students' writing may also increase students' motivation to write and help them target areas for improvement. After writing text on paper, students often must wait days or even weeks before receiving feedback from their teacher. The computer, however, can respond in a matter of seconds. Spelling and grammar checkers can provide important feedback about surface features of the students' work, while more sophisticated algorithms are able to provide feedback on the writer's style and organization. In turn, this quick feedback is believed to motivate students to make further refinements to their work (Wresch, 1984). More recent programs based on latent semantic analysis are also able to compare students' work to models or exemplars and then provide feedback on the strengths and weaknesses of their writing (Burnstein, Kaplan, Wolff Lu, 1997). In studies with college-age students, this immediate feedback is reported to be preferred over and to be more useful than the much-delayed feedback from an instructor (McCollum, 1998).

In-Class versus On-Demand Writing

All of the studies summarized above focus on student work produced in class under un-timed conditions. These studies also focus on work typically produced for English or Language Arts class, such as short stories or essays. The study presented in this article, however, focuses on writing produced under formal timed testing conditions. Specifically, this study examines the extent to which producing open-ended responses on computer or on paper affects the performance of eighth and tenth grade students, particularly for students with different levels of experience writing on a computer.

Computers and Testing

In 1968, Bert Green, Jr., predicted “the inevitable computer conquest of testing” (Green, 1970, p. 194). Since then, other observers have envisioned a future in which “calibrated measures embedded in a curriculum . . . continuously and unobtrusively estimate dynamic changes in student proficiency” (Bunderson, Inouye & Olsen, 1989, p. 387). Although progress has been made in this area, such visions of computerized testing are far from present reality. Instead, most recent research on computerized testing has focused on computerized adaptive testing, typically employing multiple-choice tests. Perhaps the most widely publicized application of this form of testing occurred in 1993 when the Graduate Record Examination (GRE) was administered nationally in both paper-and-pencil and computerized adaptive forms.

Naturally, the introduction of computer administered tests has raised concern about the equivalence of scores yielded via computer versus paper-and-pencil administered test versions. After reviewing several studies that examined the equivalence of scores acquired through computer or paper-and-pencil test forms, Bunderson, Inouye & Olsen concluded that although the mean scores for paper-and-pencil test forms were often slightly higher than for computerized test forms, “[t]he score differences were generally quite small and of little practical significance” (1989, p. 378). Similarly, following their meta-analysis of 29 studies focusing on the equivalence of computerized and paper-and-pencil cognitive ability tests, Mead and Drasgow concluded that their results “provide strong support for the conclusion that there is no medium effect for carefully constructed power tests. Moreover, no effect was found for adaptivity. On the other hand, a substantial medium effect was found for speeded tests” (1993, p. 457).

More recent research, however, shows that young people who have gone to school with computers perform significantly better on open-ended (that is, not multiple choice) questions administered via computer as compared to the same questions administered via paper-and-pencil (Russell, 1999; Russell & Haney, 1997). In both studies, the effect sizes for students accustomed to writing on computer ranged from .57 to 1.25 (effect size was estimated using Glass’s delta, that is the mean of the experimental group minus the mean of the control group divided by the standard deviation of the control group (Glass & Hopkins, 1984)). Effect sizes of this magnitude are unusually large and of sufficient size to be of not just statistical, but also practical significance (Cohen, 1988; Wolf, 1986). Effect sizes of this magnitude, for example, imply that the score for the average student in the experimental group tested on computer exceeds that of 72 to 89 percent of the students in the control group tested via paper and pencil.

Research on the effect of mode of administration on student test performance began with a puzzle. While evaluating the progress of student learning in the Accelerated Learning Laboratory (ALL), a high-tech school in Worcester, MA, teachers were surprised by the results from the second year of assessments. Although students wrote more often after computers were widely used in the school, student scores on writing tests declined in the second year of the new program.

To help solve the puzzle, a randomized experiment was conducted, with one group of sixty-eight students taking math, science and language arts tests, including

both multiple-choice and open-ended items, on paper, and another group of forty-six students taking the same tests on computer (but without access to word processing tools, such as spell-checking or grammar-checking). Before scoring, answers written by hand were transcribed so that raters could not distinguish them from those done on computer. There were two major findings. First, the multiple-choice test results did not differ much by mode of administration. Second, the results for the open-ended tests differed significantly by mode of administration. For the ALL School students who were accustomed to writing on the computer, responses written on computer were much better than those written by hand. This finding occurred across all three subjects tested and on both short answer and extended answer items. The effects were so large that when students wrote on paper, only 30 percent performed at a “passing” level; when they wrote on computer, 67 percent “passed” (Russell & Haney, 1997).

Two years later, a more sophisticated study was conducted, this time using open-ended items from the new Massachusetts state test (the Massachusetts Comprehensive Assessment System or MCAS) and the National Assessment of Educational Progress (NAEP) in the areas of language arts, science and math. Again, eighth grade students from two middle schools in Worcester, MA, were randomly assigned to groups. Within each subject area, each group was given the same test items, with one group answering on paper and the other on computer. In addition, data were collected on students’ keyboarding speed and prior computer use. As in the first study, all answers written by hand were transcribed to computer text before scoring.

In the second study, which included about two hundred students, large differences between computer and paper-and-pencil administration were again evident on the language arts tests. For students who could keyboard moderately well (20 words per minute or more), performance on computer was much better than on paper. For these students, the difference between performance on computer and on paper was roughly a half standard deviation. According to test norms, this difference is larger than the amount students’ scores typically change between grade 7 and grade 8 on standardized tests (Haney, Madaus, & Lyons, 1993, p. 234). For the MCAS, this difference in performance could easily raise students’ scores from the “failing” to the “passing” level (Russell, 1999).

In the second study, however, findings were not consistent across all levels of keyboarding proficiency. As keyboarding speed decreased, the benefit of computer administration became smaller. And at very low levels of keyboarding speed, taking the test on computer diminished students’ performance (effect size of about 0.40 standard deviations). Similarly, taking the math test on computer had a negative effect on students’ scores. This effect, however, became less pronounced as keyboarding speed increased.

State Level Testing Programs

During the past decade, an increasing number of states have sought to hold students, teachers and schools accountable for student learning as measured by state-sponsored tests containing open-ended items administered on paper. According to annual surveys by the Council for Chief State School Officers (CCSSO, 1998), 48 states use statewide tests to assess student performance in different subject areas.

Many of these tests are tied to challenging standards for what students should know and be able to do. Scores on these tests are being used to determine whether to: (1) promote students to higher grades; (2) grant high school diplomas; and (3) identify and sanction or reward low- and high-performing schools (Sacks, 1999). Due to the limitations of multiple-choice tests, many statewide tests include sections in which students must write extended answers or written explanations of their work. As the recent CCSSO report commented, “Possibly the greatest changes in the nature of state student assessment programs have taken place in the 1990s as more states have incorporated open-ended and performance exercises into their tests, and moved away from reliance on only multiple-choice tests” (CCSSO, 1998, p. 17).

Although prior research on computer use and performance on open-ended test items administered on paper does not call into question the value of state level accountability systems, it does suggest that these systems should begin thinking about alternative ways of administering open-ended items. As state level accountability tests begin to explore transitioning from paper administration to computer administration, several issues arise. First, until all students are accustomed to writing on computers, a better understanding of the extent to which the mode of administration affects student performance at different grade levels must be developed. Second, given the many computing devices available, the effect of performing open-ended items on desktop computers need to be contrasted with performance on cheaper and more portable writing devices such as eMates and AlphaSmarts. Third, before testing programs offer students the option of performing tests on paper or on computer, the extent to which handwritten versus computer printed responses influence raters’ scores needs to be explored. Fourth, administrative procedures for administering tests on computers in schools must be developed.

The study presented here focuses on the first issue identified above, namely the extent to which the mode of administration affects student performance at different grade levels. More specifically, this study employs the same methodology as Russell (1999) to examine the mode of administration effect for extended essay items in grades eight and ten.

Study Design

To examine the extent to which student performance on an open-ended writing task differs on paper versus on computer, a controlled experiment was conducted with both grade eight and grade ten students. For both grade levels, students responded to a grade appropriate extended composition item from the 1999 Massachusetts Comprehensive Assessment System (MCAS). In both grades, students were randomly assigned to compose their responses on paper or on computer. In grade eight, students did not have access to spell-checker or grammar-checker. In grade ten, students also did not have access to grammar-checker, but could use spell-checker. All students who worked on paper and on computer had access to a dictionary.

For students included in the study, three types of background information were collected, including: prior grades in English, prior computer use, and keyboarding speed.

The study occurred in three stages. During stage 1, prior English grades were collected for each student. For all students, year-end grades from the previous year and mid-term grades from the current year were collected. The course-end grades were used during the stratified group assignment process and the mid-term grades were used as covariates during some analyses.

During stage 2, all students completed the computer use survey and performed the keyboarding test. During stage 3, students performed the composition item.

To the extent possible, the same administration procedures were employed in this study as occurred during the 1999 MCAS administration. In the actual MCAS composition administration, students completed a composition item during two sessions. During the first session, students composed a first draft. After a fifteen minute break, students then revised and finalized their composition during a second writing session. Both sessions were designed to last for forty-five minutes, but students were given additional time as needed. In some cases, students were reported to take up to an additional hour to complete the composition item. In this study, time constraints and scheduling conflicts challenged efforts to replicate the MCAS composition administration. In grade eight and ten, only two hours of total testing time was available. In grade eight, the two sessions were completed during two consecutive fifty minute blocks. In grades ten, however, the two writing sessions were completed over two days.

Sampling and Group Assignment

All students included in this study attended Wellesley Public Schools, a suburban school district located outside of Boston. Within the district, half of the eighth grade students attending the Middle School participated in the study (the second half of the class participated in a related study that compared performance on a portable word processor with performance on paper). In the tenth grade, the study included students taking English with any one of the four teachers who volunteered their class for this study.

Within each grade level, the process of assigning students to groups was identical. Students' prior grade in English was used to stratify participating students within each grade level. Participating students within each stratum were then randomly assigned to groups. In both grade eight and grade ten, the composition item was administered in two formats, namely paper and desktop computer. Thus, two groups were formed in each grade level. Note that in grade eight, there were not enough computers available for exactly half of the students to work on during a single sitting. For this reason, 60 students were randomly assigned to work on computer and the remaining students were assigned to work on paper.

Table 1 summarizes the studies conducted within each grade level and indicates the actual number of students within each group.

Table 1: Number of Subjects Per Group by Grade Level

	Paper	Computer
Grade 8	85	59
Grade 10	74	71

Prior Computer Use

In addition to performing the composition item, all students completed a computer use questionnaire and a keyboarding test. The computer use questionnaire focused on students' use of computers at home, in school, and during their normal writing process. In addition, the questionnaire collected information about students' use of eMates (a portable word processor capable of displaying up to 20 lines of text and performing basic cut, copy and pasting functions) in school and during the writing process. Finally, the questionnaire queried students about their preference for taking a writing test on: a. paper or on computer, and b. paper or an eMate/AlphaSmart.

Keyboarding Test

To measure keyboarding skills, all students performed a computer based keyboarding test. The keyboarding test contained two passages for which students had two minutes apiece to type verbatim into the computer. Words per minute unadjusted for errors were averaged across the two passages and were 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 Arnold, et al, 1997; Robinson, et al, 1979; Russon & Wanous, 1973; and West, 1968, 1983), 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, 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.

Scoring

All responses were scored independently by two raters. Of the ten raters employed for this study, five were full-time classroom teachers, four were advanced doctoral students in an educational research program and one was an educational researcher. All of the raters were blind to the study design, student identities and the mode on which student responses were created. All raters participated in a one-and-a-half to two hour training session prior to scoring student responses.

For all of the items, the scoring criteria developed by MCAS were used (Massachusetts Department of Education, 2000a). The MCAS scoring guidelines for the composition items focused on two areas of writing, namely Topic/Idea Development and Standard English Conventions. The scale for Topic Development ranged from 1 to 6 and the scale for English Conventions ranged from 1 to 4, with one representing

the lowest level of performance for both scales. Table 2 presents the category descriptions for each point on the two scales.

Table 2: Category Descriptions for MCAS Composition Rubrics

Score	Topic Development	English Standards
1	Little topic/idea development, organization, and/or details Little or no awareness of audience and/or task	Errors seriously interfere with communication AND Little control of sentence structure, grammar and usage and mechanics
2	Limited or weak topic/idea development, organization, and/or details Limited awareness of audience and/or task	Errors interfere somewhat with communication and/or Too many errors relative to the length of the essay or complexity of sentence structure, grammar and usage, and mechanics
3	Rudimentary topic/idea development and/or organization Basic supporting details Simplistic language	Errors do not interfere with communication and/or Few errors relative to length of essay or complexity of sentence structure, grammar and usage, and mechanics
4	Moderate topic/idea development and organization Adequate, relevant details Some variety in language	Control of sentence structure, grammar and usage, and mechanics (length and complexity of essay provide opportunity for students to show control of standard English conventions)
5	Full topic/idea development Logical organization Strong details Appropriate use of language	
6	Rich topic/idea development Careful and/or subtle organization Effective/rich use of language	

In addition to the general descriptions, MCAS also provides anchor papers and benchmark papers for each category. These exemplars are grade level specific and provide a variety of responses the prompt administered at each grade level.

To reduce the influence handwriting 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 reduce rater bias all responses were double scored. To protect against differences in the severity with which raters scored responses and the effects of rater fatigue, all

responses originally produced on paper and on computer were intermixed and then distributed to raters. In this way, all raters scored an equal number of responses originally produced on paper and on computer. At the conclusion of the scoring process, scores awarded by two raters were added together to produce a Topic Development scale that ranged from two to twelve and an English Standards scale that ranged from two to eight.

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 3 shows that for most items the correlation between the two raters' scores ranged from .55 to .64. Agreement within one point ranged from 89% to 100%. Although the inter-rater correlations were lower than desired, they suggest that when discrepancies arose, one set of raters was not consistently more or less lenient than the second set of raters. Although no information has been published regarding inter-rater reliability of composition scores for the actual administration of the MCAS composition items, the extent to which raters were within one point of agreement is similar to the frequency of agreement obtained for the actual MCAS open-ended scoring (Massachusetts Department of Education, 1999).

Table 3: Inter-rater Reliability for Open-Ended Items

	Correlation Coefficient	% Within 1 Point
Grade 8		
Topic Development	.63	92%
English Standards	.57	100%
Grade 10		
Topic Development	.64	89%
English Standards	.55	97%

Results

The study presented here explores the relationships between prior computer use and performance on an extended composition item administered in grades eight and ten. To examine this relationship, three types of analyses were performed within each grade level. 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 separately for grades eight and ten. For the student questionnaire, keyboarding test, and English grades, summary statistics are based on all students included in these two studies within each grade level.

Keyboarding Test

The keyboarding test contained two passages. 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. Note that due to the passage length, the maximum keyboarding speed students could obtain was 59 words per minute. Table 3 indicates that the mean WPM increased from 27 in grade eight to nearly 36 in grade ten.

Table 3: Summary Statistics for the Keyboarding Test

	N	Mean WPM	Std. Dev.	Min.	Max.
Grade 8	144	27.43	9.04	10	55
Grade 10	145	35.77	10.37	10	59

Student Questionnaire

The student questionnaire contained 12 questions. Although comparative data are not available, Tables 4a and 4b suggest that on average students in both grade levels included in this study have substantial experience working with computers. The vast majority of students report using a computer for three or more years, using computers in schools, and using a computer in their home nearly every day. Furthermore, most students report that they use a computer at least once a month when writing a first draft. Slightly more students report using a computer two to three times a month to edit the first draft. And most students report using a computer regularly 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.

Tables 4a and 4b also show that the eighth grade students use portable writing devices about the same as do students in tenth grade. Recall, however, that those eighth grade students whose teachers regularly use portable writing devices participated in a separate study that compared performance on paper with performance on a portable writing device.

Table 4a: Summary Statistics for the Student Questionnaire – Grade 8

	Never	Less than One	One	Two	Three	Four or More
Years using computer	1%	1%	1%	7%	18%	72%
	Never	Less than 1 hour/ week	1-2 hours/ week	2-5 hours/ week	1-2 hours/ day	More than 2 hours/ day
Use computer in school	10%	69%	19%	2%	1%	0%
Use computer at home	4%	2%	11%	35%	38%	11%
Use eMate/AlphaSmart in School	88%	12%				
	Never	1-2 times/ year	3-4 times/ year	Once a Month	2-3 times/ month	About Once a Week
Compose First Draft w/ Computer	15%	6%	5%	20%	18%	36%
Edit w/ Computer	8%	4%	4%	20%	27%	36%
Type Final Draft w/ Computer	1%	1%	1%	17%	25%	56%
Compose First Draft w/ eMate/AlphaSmart	91%	6%	1%	1%		
Edit w/ eMate/ AlphaSmart	91%	5%	2%	1%	1%	
Type Final Draft w/ eMate/AlphaSmart	90%	6%	1%	1%	1%	
	Paper	Computer /eMate				
Paper or Computer Preference	15%	85%				
Paper or eMate/ AlphaSmart Preference	37%	63%				

Table 4b: Summary Statistics for the Student Questionnaire—Grade 10

	Never	Less than One	One	Two	Three	Four or More
Years using computer	1%	1%	1%	5%	7%	86%
	Never	Less than 1 hour/ week	1-2 hours/ week	2-5 hours/ week	1-2 hours/ day	More than 2 hours/ day
Use computer in school	6%	41%	30%	14%	8%	1%
Use computer at home	1%	6%	10%	32%	36%	15%
Use eMate/AlphaSmart in School	71%	26%	3%			
	Never	1-2 times/ year	3-4 times/ year	Once a Month	2-3 times/ month	About Once a Week
Compose First Draft w/ Computer	16%	9%	6%	15%	22%	32%
Edit w/ Computer	11%	6%	6%	20%	21%	36%
Type Final Draft w/ Computer	1%	4%	3%	17%	31%	44%
Compose First Draft w/ eMate/AlphaSmart	87%	12%	1%			
Edit w/ eMate/ AlphaSmart	88%	10%	1%			
Type Final Draft w/ eMate/AlphaSmart	88%	10%	1%			
	Paper	Computer /eMate				
Paper or Computer Preference	32%	69%				
Paper or eMate/ AlphaSmart Preference	45%	55%				

Indicator of Prior Achievement

Mid-year English grades were collected for all students included in this study. In both eighth and tenth grade, alphabetic grades (e.g., A, B-, C+) were awarded. These alphabetic grades were converted to a numeric scale as indicated in Table 5.

Table 5: Letter Grade to Numeric Grade Conversion Chart

Letter Grade	Number Grade
A+	97
A	95
A-	92
B+	87
B	85
B-	82
C+	77
C	75
C-	72
D+	67
D	65
D-	62
F	50

Table 6 displays the mean and standard deviation for mid-year grades for each grade level.

Table 6: Summary Statistics for Mid-Year Grades

	N	Mean	Std Dev	Min	Max
Grade 8	144	86.7	5.83	65	97
Grade 10	145	82.2	6.86	62	95

Composition Scores

One extended composition item was administered to students in each grade level. As is explained more fully above, two scores were awarded to each composition. The first score represents the quality of the composition's Topic Development and the second score indicates the quality of the student's Standard English Conventions. Summary statistics are presented in Table 7. In addition to the mean score for students included in this study, the mean score for students across the state who performed the composition in the spring of 1999 are also presented. On average, students included in this study scored higher than students across the state.

Table 7: Summary Statistics for Composition Scores

	N	Scale	Mean	Std Dev	1999 Statewide Mean Score
Grade 8					
Topic Dev.	144	2-12	8.41	1.84	7.18
Stand. English	144	2-8	6.26	1.19	5.67
Grade 10					
Topic Dev.	145	2-12	7.60	2.12	7.18
Stand. English	145	2-8	6.45	1.24	5.97

Comparing Performance by Mode of Administration

In both grade eight and ten, independent sample t-tests (assuming equal variances for the two samples and hence using a pooled variance estimate) were employed to examine differences between the pre-assigned modes of administration. As described in greater detail below, in three cases the assumption of equal variances was violated. In such cases, Welch's t-test was performed to examine differences between the pre-assigned modes of administration. The null hypothesis for each of these tests was that the mean performance of the computer and the paper groups did not differ. Thus, these analyses test whether performance on computer had a statistically significant effect on students' test scores.

To examine whether prior achievement or keyboarding skills differed between the two groups of students who performed each test, independent samples t-tests were also performed for students' mid-term grades and WPM. In addition, independent samples t-tests were performed to examine differences in the length of students' responses. In grade eight, the amount of time students spent working on their compositions was also collected. Given concerns that the MCAS tests consume a large amount of time, testing time was recorded in grade eight to examine whether drafting and revising on computer might reduce testing time without jeopardizing student's performance. For this reason, testing time was also contrasted in grade eight.

Grade 8

Table 8 displays results for the paper versus computer experiment. Although there was a small difference between the two groups' mid-term grades, this difference was not statistically significant. On average, however, students writing on computer produced passages that were one and a half times longer than those composed on paper. More importantly, students who composed on computer also received higher scores for both Topic Development and English Standards. On average, students composing on computer scored 1.2 points higher on Topic Development and .7 points higher on English Standards. When the two sub-scores are combined, the computer group performed nearly two points higher than the paper group (see Figure 1).

Table 8: Paper versus Computer Group Comparisons—Grade 8

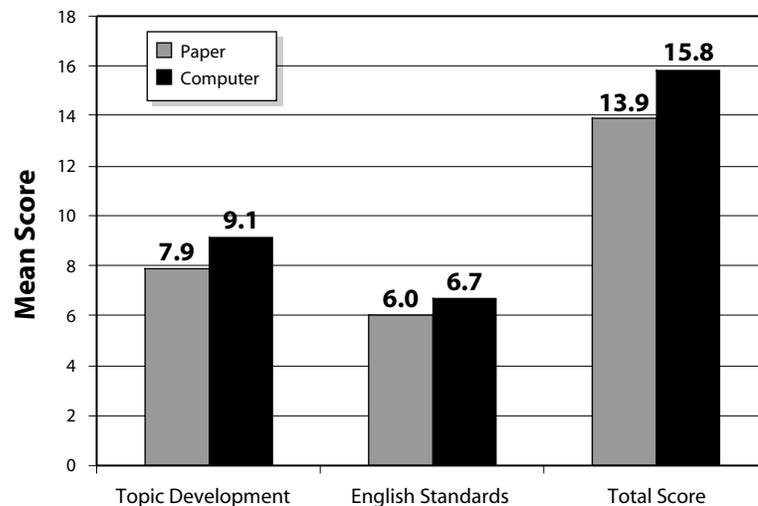
	Mean	Std. Dev.	Std. Error	T-statistic	Sig.	Effect Size	Levene's Test for Equality of Variances	
							F	Sig.
Mid-Term Grade								
Paper	86.3	6.11	.67					
Computer	87.4	5.4	.70	1.09	.28	.18	.07	.79
Topic Development								
Paper	7.9	1.8	.20					
Computer	9.1	1.6	.21	4.05	<.001	.64	1.68	.20
English Standards								
Paper	6.0	1.2	.13					
Computer	6.7	1.1	.15	3.68	<.001	.61	.60	.44
Total Score								
Paper	13.9	2.8	.30					
Computer	15.8	2.5	.32	4.25	<.001	.69	.70	.40
Passage Length								
Paper	457	144	15.7					
Computer	709	243	31.6	7.13*	<.001	1.74	14.1	<.001
Finish Time								
Paper	107.8	8.2	.89					
Computer	101.9	12.4	1.62	3.22*	.001	-.72	11.0	.001
WPM								
Paper	26.8	9.8	1.07					
Computer	28.3	7.7	1.01	.97	.36	.15	2.62	.11

N for Paper Group=85

N for Computer Group=59

* Welch's t-test used due to inequality of variance.

Figure 1: Mode of Administration Effect on MCAS Language Arts Scores—Grade 8



It is also interesting to note that this higher level of performance was produced in less time. Although the difference in Finish Time was statistically significant and represents a sizeable effect size, in reality this six minute difference would have little impact on the total testing time.

Note that statistical significance for the t-tests reported above was not adjusted to account for multiple comparisons. Given that seven comparisons were made, 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 seven comparisons the adjusted value of a simple 0.05 alpha level becomes 0.007. 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 difference in passage length, finish time and all categories of composition scores remain statistically significant. Moreover, as shown in Table 8, these differences in composition scores represent effect sizes of .61 to .69 (Glass's delta effect size was employed (Glass & Hopkins, 1984)). These effect sizes fall in between those reported by Russell and Haney (1997) and by Russell (1999). These effect sizes suggest that while half of the students in the computer group received total scores above 15.8, approximately 25% of students performing the test on paper scored above 15.8.

To control for differences in prior achievement, a multiple regression was performed for each experiment. Table 9 presents the results of each test score regressed on mid-term grades and group membership. For both regression analyses, the regression coefficient (B) for group membership indicates the effect group membership has on students' performance when the effect of mid-term grade is controlled. Similarly, Beta indicates the effect group membership has on students' performance when all

variables are standardized and the effect of mid-term grade 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 had 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 9 indicates that mid-term grades are a significant predictor of students' Topic Development, English Standards and Total scores. For each one standard score unit increase in mid-term grades, on average students experience a .52 standard score increase in their Topic Development score, a .39 standard score increase in their English Standard score, and a .51 increase in their total score. Table 9 also indicates that after controlling for differences in mid-term grades, performing the composition item on computer has a positive impact on student scores. This impact includes a .25 standard score increase in students' English Standards score, a .27 standard score increase in Topic Development score, and a .29 standard score increase in total scores. All of these effects are statistically significant.

Table 9: Regression of Composite Total Score and Sub-Scores on Mid-Term Grades and Computer Group Membership – Grade 8

Topic Development					
R=.61	B	Std. Error	Beta	T	Sig.
Constant	-5.50	1.78			
Mid-Year Grade	0.16	0.02	.52	7.61	<.001
Computer Group	0.96	0.24	.27	3.97	<.001
English Standards					
R=.48	B	Std. Error	Beta	T	Sig.
Constant	-0.50	1.27			
Mid-Year Grade	0.08	0.02	.39	5.17	<.001
Computer Group	0.58	0.17	.25	3.39	.001
Total Score					
R=.60	B	Std. Error	Beta	T	Sig.
Constant	-5.99	2.69			
Mid-Year Grade	0.23	0.03	.51	7.45	<.001
Computer Group	1.55	0.37	.29	4.22	<.001

Grade 10

Table 10 indicates that there was a small difference between the two groups' mid-term grades. This difference, however, was not statistically significant. On average, passages composed on computer were one hundred words longer than those produced on paper. Responses composed on computer also received higher scores than those produced on paper. The effects were larger for English Standards than for Topic Development. After adjusting significance for multiple comparisons, the difference in both the English Standards and Total score were statistically significant. Moreover, the effect sizes for these differences ranged from .32 for Topic Development, .68 for English Standards, and .51 for the Total Score. The effect size for the total score suggests that while about half of the students writing on computer scored above 14.8, approximately 30% of students performing the test on paper scored above 14.8. On average, students who composed on computer scored 1.5 points higher than students who composed on paper (see Figure 2).

Table 10: Paper versus Computer Group Comparisons—Grade 10

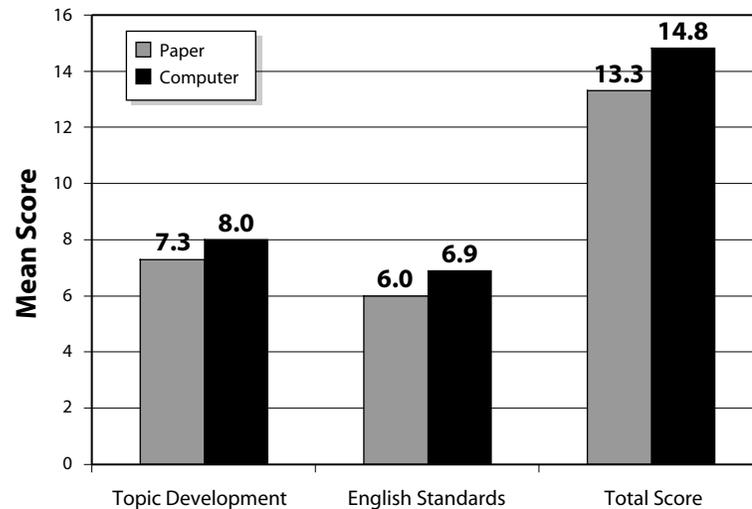
	Mean	Std. Dev.	Std. Error	T-statistic	Sig.	Effect Size	Levene's Test for Equality of Variances	
							F	Sig.
Mid-Term Grade								
Paper	81.9	6.2	.72					
Computer	82.6	7.5	.89	.59	.56	.11	.64	.42
Topic Development								
Paper	7.3	2.2	.25					
Computer	8.0	2.0	.24	2.02	.05	.32	.47	.50
English Standards								
Paper	6.0	1.3	.15					
Computer	6.9	1.1	.13	4.43	<.001	.68	.82	.37
Total Score								
Paper	13.3	3.1	.36					
Computer	14.8	2.8	.33	3.21	.002	.51	.21	.65
Passage Length								
Paper	462	138	16.1					
Computer	567	292	34.6	2.77*	.007	.76	4.55	.04
WPM								
Paper	35.1	10.4	1.21					
Computer	36.4	10.4	1.23	.74	.46	.12	.04	.84

N for Paper Group=74

N for Computer Group=71

* Welch's t-test used due to inequality of variance.

Figure 2: Mode of Administration Effect on MCAS Language Arts Scores—Grade 10



To control for differences in prior achievement, a series of multiple regressions was performed in which each composition score was regressed on mid-term grades and group membership. Table 11 indicates that mid-term grades are a significant predictor of students' scores. For each one standard score unit increase in mid-term grades, on average students experience a .39 standard score increase in their Topic Development score, a .27 standard score increase in their English Standard score, and a .38 increase in their total score. Table 11 also indicates that after controlling for differences in mid-term grades, performing the composition item on computer has a positive impact on student scores. This impact includes a .15 standard score increase in students' Topic Development score, a .33 standard score increase in English Standards score, and a .24 standard score increase in total scores. After adjusting for multiple comparisons, the effect was statistically significant for both English Standards and for the Total Score.

Table 11: Regression of Composite Total Score and Sub-Scores on Mid-Term Grades and Computer Group Membership—Grade 10

Topic Development					
R=.42	B	Std. Error	Beta	T	Sig.
Constant	-2.50	1.94			
Mid-Year Grade	0.12	0.02	.39	5.06	<.001
Computer Group	0.62	0.32	.15	1.93	.05
English Standards					
R=.44	B	Std. Error	Beta	T	Sig.
Constant	1.98	1.13			
Mid-Year Grade	.05	0.01	.27	3.63	<.001
Computer Group	.83	0.19	.33	4.44	<.001
Total Score					
R=.46	B	Std. Error	Beta	T	Sig.
Constant	-0.54	2.71			
Mid-Year Grade	0.17	0.03	.38	5.13	<.001
Computer Group	1.45	0.45	.24	3.22	.002

Examining Keyboarding Speed and Mode of Administration Effect

The regression analyses presented above (Tables 9 and 11) indicate that mode of administration had a significant effect on students' performance. 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. The second group contained students whose WPM was between .5 standard deviations below the mean and .5 standard deviations above the mean. The third group contained students whose WPM was .5 standard deviations above the mean. For each group, the composition total score was regressed on mid-term grades and group membership. Below, these three sub-group regressions are presented separately for grade eight and grade ten.

Grade 8

In grade eight, "slow" keyboarders were defined as students whose keyboard speed was less than 23.8 words per minute. "Average" keyboarders typed between 23.8 and 32.4 wpm. And "fast" keyboarders typed more than 32.4 wpm. Table 12 displays the results of the three separate regressions. For all sub-groups, performing the composition item on computer had a positive effect on their scores. Moreover, the size of this effect was similar at all three levels of keyboarding speed.

Table 12: Total Score Regressed on Mid-Year Grade and Computer Group for Each WPM Sub-group – Grade 8

WPM<23.8 N=48					
R=.54	B	Std. Error	Beta	T	Sig.
Constant	-1.72	4.68			
Mid-Year Grade	0.17	0.06	.38	3.04	.004
Group	1.82	0.71	.32	2.56	.014
23.8>WPM<32.4 N=61					
R=.55	B	Std. Error	Beta	T	Sig.
Constant	-2.14	4.33			
Mid-Year Grade	0.19	0.05	.42	3.82	<.001
Group	1.46	0.05	.32	2.96	.004
WPM>32.4 N=31					
R=.45	B	Std. Error	Beta	T	Sig.
Constant	1.53	6.56			
Mid-Year Grade	0.16	.07	.37	2.22	.035
Group	1.17	.67	.29	1.76	.09

Grade 10

In grade ten, “slow” keyboarders were defined as students whose keyboard speed was less than 30.6 words per minute. “Average” keyboarders typed between 30.6 and 41.0 wpm. And “fast” keyboarders typed more than 41.0 wpm. Table 13 displays the results of the three separate regressions. The largest effect was found for students with slower keyboarding skills. In this case, however, slow keyboarding skills ranged from 20 to 30 words per minute. In essence, no administration effect was found for average keyboarders, but a sizeable effect was found for fast keyboarders.

Table 13: Total Score Regressed on Mid-Year Grade and Computer Group for Each WPM Sub-group—Grade 10

WPM<30.6 N=43					
R=.52	B	Std. Error	Beta	T	Sig.
Constant	1.28	4.62			
Mid-Year Grade	0.13	0.06	.30	2.21	.033
Group	3.03	0.85	.48	3.58	.001
30.6>WPM<41.0 N=57					
R=.55	B	Std. Error	Beta	T	Sig.
Constant	-6.15	4.33			
Mid-Year Grade	0.25	0.05	.54	4.77	<.001
Group	0.38	0.66	.07	0.57	.568
WPM>41.0 N=42					
R=.31	B	Std. Error	Beta	T	Sig.
Constant	13.84	6.35			
Mid-Year Grade	0.002	0.08	.001	0.003	.99
Group	1.55	0.78	.31	1.98	.05

Discussion

The experiment described here extends the work of Russell (1999) and Russell and Haney (1997) in two important ways. In addition to examining the mode of administration effect in grade eight, this study examines the effect in grade ten. In many state testing programs as well as the National Assessment of Educational Progress, these two grade levels are commonly tested. Thus, it is important to understand the extent to which mode of administration affects the performance of students in these commonly tested grade levels. Second, whereas the previous studies primarily focused on relatively short open-ended items, this study focuses specifically on students' performance on extended writing composition items. Again, in many state level testing programs, extended writing items are employed.

As in the two previous studies, this experiment found that students who wrote their compositions on computer produced longer responses that received higher scores. This effect was larger in grade eight than in grade ten, but was statistically and practically significant in both grade levels.

Across both grade levels, the relationship between keyboarding speed and the mode of administration effect was inconsistent. Russell (1999) found that students need to have sufficient keyboarding speed before the mode of administration effect becomes meaningful. In the grade eight experiment presented here, however, this pattern did not emerge. Instead, the effect was positive, nearly the same size, and of practical significance at all levels of keyboarding speed. In grade 10, the pattern was

less consistent. Nonetheless, moderately large effects were found for both slow and fast keyboarders. One reason why these findings differ from that reported by Russell (1999) may stem from the relatively high level of keyboarding speed for all students included in these experiments. Whereas Russell (1999) reported an average keyboarding speed of about 15 words per minute, average speed for the grade eight students included in this study was nearly twice as fast. Moreover, the average speed of students included in the low keyboarding level of this study exceeded the cut point for the high keyboarding level in Russell's (1999) study. It appears that once students achieve keyboarding speed of roughly 20 wpm, the mode of administration effect becomes meaningful. The size of the effect, however, does not seem to increase further as students' keyboarding speed increases.

Limitations

This study focused on students in grades eight and ten attending school within a single district. This district tends to perform well above the state average on standardized and state level tests. In the study presented here, very few students performed at low levels on the composition item. As a result, it was not possible to examine the mode of administration effect across the full range of performance levels.

Similarly, students included in this study were generally accustomed to working with computers. The relatively high level of keyboarding speed complicated efforts to examine the mode of administration effect at low levels of keyboarding speed. Additionally, students' familiarity with computers prevented an examination of the mode of administration effect for students who are not accustomed to working with computers. Despite this high level of access to computers within Wellesley's schools, it should be noted that seventy-five out of approximately three hundred districts across the Commonwealth of Massachusetts have a lower (better) student to computer ratio than does Wellesley (Massachusetts Department of Education, 2000b).

Since students' scores on the composition item did not "count" toward students' grades and did not become part of students' records, many students in grade ten may not have put forth their best efforts. This was particularly true for the computer group which was proctored by a person who was not a teacher in the high school. Students performing the test on paper, however, were proctored by their English teacher. Although there were no severe behavioral problems reported, most students in the computer group were reported to have finished their compositions within the first fifteen minutes of the second writing period. Students in the paper groups, however, worked for nearly the entire period. Moreover, four responses generated on computer failed to address the question posed, whereas all of the responses composed on paper were on topic. For these reasons, data collected in grade ten may not accurately reflect students' actual levels of performance.

Implications

This study provides further evidence that students accustomed to writing on computers perform significantly better when open-ended tests are administered on a computer. For the MCAS Language Arts test, this improved performance translates into approximately two points on the composition item. But in addition to the composition item, the MCAS Language Arts test contains four short answer open-ended items each worth four points. Assuming that the effect found for the composition item and that the effect reported on the shorter open-ended items by Russell (1999) holds across all Language Arts open-ended items, students accustomed to writing with computers may perform better by an additional two points on these open-ended items. Across all short and extended open-ended MCAS items, the mode of administration effect may result in an increase of four raw score points if students were allowed to compose responses on a computer. An increase of four raw score points translates into between four and eight point scale score points, depending upon where on the scale a student's score resides.

School Level Effect

Within Wellesley, eliminating the mode of administration effect for both the composition item and the four shorter open-ended items would have a dramatic impact on district level results. As Figure 3 indicates, based on the 1999 MCAS results, 4% of eighth graders would move from the "Needs Improvement" category to the "Proficient" category and 13% more students would be deemed "Advanced." In grade 10, allowing students to perform written items on computer would move 14% more students into the "Advanced" category and approximately 7% more students into the "Proficient" category (Figure 4).

Figure 3: Mode of Administration Effect on Grade 8 1999 MCAS Results

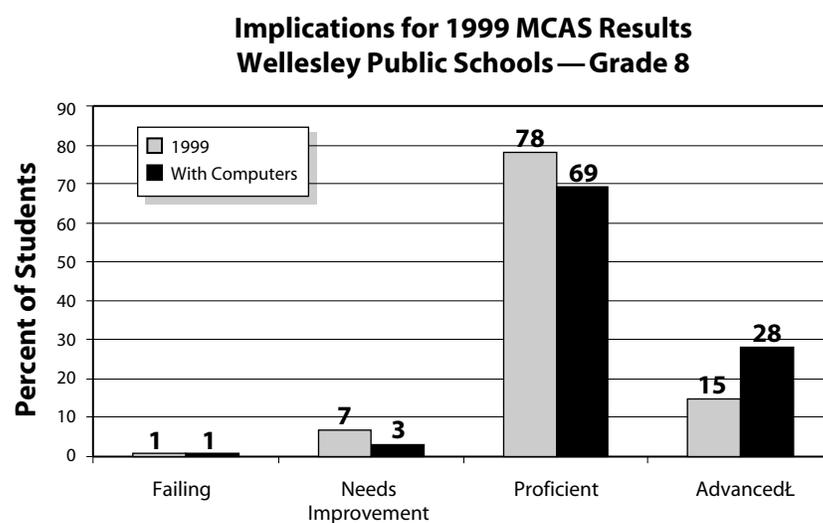
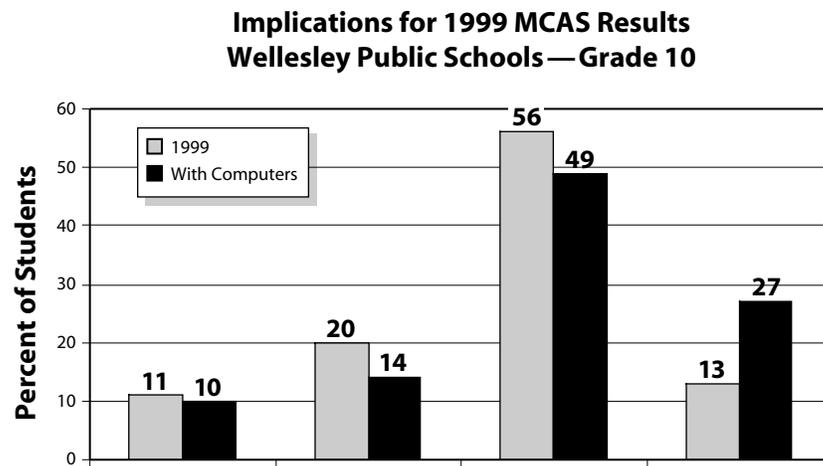


Figure 4: Mode of Administration Effect on Grade 10 1999 MCAS Results

Despite the limitations of the studies performed to date, the mode of administration effects reported here and in previous studies (Russell, 1999; Russell and Haney, 1997) highlight a complicated challenge testing programs must overcome as they continue to use tests containing open-ended items to make inferences about student and school achievement. One solution state testing programs might adopt to reduce the mode of administration effect is to allow students to select the mode in which open-ended responses are composed. For the past decade, the Province of Alberta has employed this strategy for its graduation testing program. Over the past five years, the province has seen the percentage of students opting to perform the English, Social Studies, Biology and French tests on computer increase from 6.7% in 1996 to 24.5% in 2000. Within high schools, the percentage of students opting to perform the test on a computer ranges from 0 to 80% (Sakyi, 2000).

Although this approach adds to the complexity of test administration procedures (see Russell & Haney, 2000 for a fuller review of added complexities), providing students the option of working on paper or computer would create writing conditions that more closely reflect normal practice. In turn, students would be better able to demonstrate their best possible work under the circumstances.

We thank Greg Nadeau and Brenda Thomas of the Massachusetts Department of Education and Karen Toomey of Wellesley Public Schools for their support and assistance. We also acknowledge the National Board on Education Testing and Public Policy for its contributions to this study.



References

- Arnold, V., Joyner, R. L., Schmidt, B. J., & White, C. D. (1997). Establishing electronic keyboarding speed and accuracy timed-writing standards for postsecondary students. *Business Education Forum*, 51(3), p. 33–38.
- 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.
- Becker, H. J. (1999). *Internet Use by Teachers: Conditions of Professional Use and Teacher-Directed Student Use*. Irvine, CA: Center for Research on Information Technology and Organizations.
- Breese, C., Jackson, A. & Prince, T. (1996). Promise in impermanence: Children writing with unlimited access to word processors. *Early Child Development and Care*, 118, 67–91.
- Bridewell, L. & Ross, D. (1984) Integrating computers into a writing curriculum; or buying, begging, and building. In W. Wresch (ed.) *The Computer in composition instruction: A writer's tool*. Urbana, IL: National Council of Teachers of English.
- Bunderson, C., Inouye, D. & Olsen, J. (1989). The four generations of computerized educational measurement. In Linn, R. L., *Educational Measurement* (3rd ed.), Washington, D.C.: American Council on Education, pp. 367–407.
- 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.
- Calkins, L. M. (1983). *Lessons from a Child: On the Teaching and Learning of Writing*. Exeter, N.H. : Heinemann Educational Books.
- Calkins, L. M. (1986). *The Art of Teaching Writing*. Portsmouth, NH: Heinemann.
- Cocran-Smith, M, Paris, C., & Kahn, J. (1991). *Learning to write differently: Beginning writers and word processing*. Norwood, NJ: Ablex.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. (2nd ed.) Hillsdale, NJ: Erlbaum.
- Collier, R. & Werier, C. (1995). When computer writers compose by hand. *Computers and Composition*, 12, 47–59.
- Council of Chief State School Officers (1998). *Key State Education Policies on K–12 Education: Standards, Graduation, Assessment, Teacher Licensure, Time and Attendance*. Washington, DC.
- Daiute, C. (1984). Can the computer stimulate writers' inner dialogues? In W. Wresch (ed.) *The Computer in composition instruction: A writer's tool*. Urbana, IL: National Council of Teachers of English.
- Daiute, C. (1985). *Writing and Computers*. Don Mills, Ontario: Addison-Wesley.
- 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.

- Elbow, P. (1981). *Writing with Power : Techniques for Mastering the Writing Process*. New York : Oxford University Press.
- Etchison, C. (1989). Word processing: A helpful tool for basic writers. *Computers and Composition*, 6(2), 33–43.
- Fabos, B. & Young, M. (1999). Telecommunications in the classroom: Rhetoric versus reality. *Review of Educational Research*. 69:3 217–259.
- Glass, G. & Hopkins, K. (1984). *Statistical Methods in Education and Psychology*. Boston, MA: Allyn and Bacon.
- Graves, D. (1983). *Writing: Teachers and Children at Work*. Portsmouth, NH: Heinemann Educational Books.
- Green, B., Jr. (1970). Comments on tailored testing. In W. H. Holtzman (Ed.), *Computer-assisted instruction, testing and guidance*. New York: Harper and Row.
- Grejda, G. (1992). Effects of word processing on sixth graders' holistic writing and revision. *Journal of Educational Research*, 85(3), 144–149.
- Haas, C. & Hayes, J. (1986). 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).
- Haas, C. (1989). How the writing medium shapes the writing process: Effects of word processing on planning. *Research in the Teaching of English*, 23, 181–49.
- Haney, W., Madaus, G., & Lyons, R. (1993). *The Fractured Marketplace for Standardized Testing*. Boston, MA: Kluwer Academic Publishers.
- Hannafin, M. & Dalton, D. (1987). The effects of word processing on written composition. *The Journal of Educational Research*, 80 (July/Aug.) p. 338–42.
- Hawisher, G. (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)
- Holmes, R. (1999). A gender bias in the MCAS? *MetroWest Town Online*. <http://www.townonline.com/metrowest/archive/022499/>.
- Kerchner, L. & Kistinger, B. (1984). Language processing/word processing: Written expression, computers, and learning disabled students. *Learning Disability Quarterly*, 7(4), 329–335.
- Larter, S. (1987). *Writing with microcomputers in the elementary grades: Process, roles, attitudes, and products*. Toronto: Queens Printer.
- Levin, J. Riel, M., Rowe, R. & Boruta, M. (1985). Muktuk meets jacuzzi: computer networks and elementary school writers. In S. W. Freedman (Ed.), *The Acquisition of Written Language: Response and Revision* (pp. 160–171). Norwood, NJ: Ablex.
- 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.
- Marcus, S. (1984). Real-Time gadgets with feedback: Special effects in computer-assisted writing. In W. Wresh (ed.) *The Computer in Composition Instruction: A Writer's Tool*. Urbana, IL: National Council of Teachers of English.
- Market Data Retrieval. (1999). *Technology in Education 1999*. (A report issued by Market Data Retrieval). Shelton, CN: Market Data Retrieval.
- Massachusetts Department of Education. (1999). *Massachusetts Comprehensive Assessment System 1998 Technical Report*. Malden, MA.
- Massachusetts Department of Education. (2000a). *1999 MCAS Sample Student Work and Scoring Guides*. <http://www.doe.mass.edu/mcas/student/1999/>.
- Massachusetts Department of Education. (2000b). *Educational Technology in Massachusetts, 1999–2000*. Malden, MA.
- McCollum, K. (1998). How a computer program learns to grade essays. *The Chronicle of Higher Education*, September 4, 1998, <http://chronical.com/weekly/v45/i02/02a03701.htm>.
- Mead, A. & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin*, 114:3, 449–58.
- Moore, M. (1987). *The effect of word processing technology in a developmental writing program on writing quality, attitude toward composing, and revision strategies for fourth- and fifth-grade students*. Dissertation Abstracts International, 48, 635A.
- Morocco, C. & Neuman, S. (1986). Word processors and the acquisition of writing strategies. *Journal of Learning Disabilities*, 19, 243–247.
- Nichols, L. (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.
- Owston, R. & Wideman, H. (1997). Word processors and children's writing in a high-computer-access setting. *Journal of Research on Computing in Education*, 30(2), 202–220.
- Owston, R. (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.
- Owston, R., Murphy, S. & Wideman, H. (1992). The effects of word processing on students writing ability. *Research in the Teaching of English*, 26, 249–276.
- 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.

- Robison, J. W., Erickson, L. W., Crawford, T.J. & Ownby, A. C. (1979). *Typewriting: Learning and Instruction*. Cincinnati: South-Western Publishing Company.
- 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>.
- Russell, M. & Haney, W. (2000). Bridging the Gap Between Testing and Technology in Schools. *Education Policy Analysis Archives*, 8(19), <http://epaa.asu.edu/epaa/v8n19.html>.
- Russell, M. (1999). Testing Writing on Computers: A Follow-up Study Comparing Performance on Computer and on Paper. *Educational Policy Analysis Archives*, 7(20).
- Russon, A. & Wanous, S. (1973). *Philosophy and Psychology of Teaching Typewriting*. (2nd ed.) Cincinnati: South-Western Publishing Company.
- Sacks, P. (1999). *Mind Games: How our mental testing culture sustain the privileged, punishes the poor and standardizes the American mind*. Reading MA: Perseus Books,
- Sakyi, A. (2000). Word Processing in Humanities Diploma Examinations. Personal correspondence from the Alberta Learning, Student Evaluation Branch.
- Scardamalia, M. & Bereiter, C. (1986). Research on written composition. In M. C. Wittrock (Ed.), *Handbook of Research on Teaching* (3rd ed., pp. 778–803). New York: Macmillan.
- 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 at the Invitational Research Symposium on Special Education Technology, Washington, DC.
- Snyder, I. (1993). The impact of computers on student writing: A comparative study of the effects of pens and word processors on writing, context and product. *Australian Journal of Education*, 37(1), 5–25.
- Vacc, N. (1987). Word processor versus handwriting: A comparative study of writing samples produced by mildly mentally handicapped students. *Exceptional Children*, 54(2), 156–165.
- West, L. (1968) The vocabulary of instructional materials for typing and stenographic training—research findings and implications. *Delta Pi Epsilon Journal*, 10(3), 13–125.
- West, L. (1983). *Acquisition of Typewriting Skills*, (2nd ed.), Indianapolis: The Bobbs-Merrill Company, Inc.
- Williamson, M. & 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.
- Wolf, F. (1986). *Meta-analysis: Quantitative methods for research synthesis*. Sage University series on quatitative applications in the social sciences, series no. 07-059. Newbury Park, CA: Sage.

Wresch, W. (1984). *The Computer in Composition Instruction: A Writer's Tool*. Urbana, IL: National Council of Teachers of English.

Abstract

This study builds on two previous studies to examine the effect administering extended composition test items on paper and on computer has on student performance. This study employs writing items from the 1999 Massachusetts Comprehensive Assessment System (MCAS) to examine the mode of administration effect in grades eight and ten. Like the previous studies, this article reports that open-ended Language Arts items that require students to generate responses using paper and pencil severely underestimate the achievement of students accustomed to writing using a computer. Combining the effects found in this study with those found in a prior study, this article estimates that students accustomed to writing using a computer under-perform on the MCAS Language Arts test by four to eight points on an eighty point scale. This article concludes by recommending that state testing programs that employ open-ended items in Language Arts provide students with the option of composing responses on paper or on computer.

