# Education Policy Analysis Archives

**Volume 5 Number 3** **January 15, 1997** **ISSN 1068-2341**

A peer-reviewed scholarly electronic journal.

Editor: Gene V Glass Glass@ASU.EDU.

College of Education

Arizona State University,Tempe AZ 85287-2411




## Testing Writing on Computers: An Experiment Comparing Student Performance on Tests Conducted via Computer and via Paper-and-Pencil


**Michael Russell**
**Boston College**

**Walt Haney**
**Boston College**



### Abstract

Computer use has grown rapidly during the past decade. Within the educational community, interest in authentic assessment has also increased. To enhance the authenticity of tests of writing, as well as of other knowledge and skills, some assessments require students to respond in written form via paper-and-pencil. However, as increasing numbers of students grow accustomed to writing on computers, these assessments may yield underestimates of students' writing abilities. This article presents the findings of a small study examining the effect that mode of administration -- computer versus paper-and-pencil -- has on middle school students' performance on multiple-choice and written test questions. Findings show that, though multiple-choice test results do not differ much by mode of administration, for students accustomed to writing on computer, responses written on computer are substantially higher than those written by hand (effect size of 0.9 and relative success rates of 67% versus 30%). Implications are discussed in terms of both future research and test validity.


## Introduction

Two of the most prominent movements in education over the last decade or so are the introduction of computers into schools and the increasing use of "authentic assessments." A key assumption of the authentic assessment movement is that instead of simply relying on multiple choice tests, assessments should be based on the responses students generate for open-ended "real world" tasks. "Efforts at both the national and state levels are now directed at greater use of performance assessment, constructed response questions and portfolios based on actual student work" (Barton & Coley, 1994, p. 3). At the state level, the most commonly employed kind of non-multiple-choice test has been the writing test (Barton & Coley, 1994, p. 31) in which students write their answers long-hand. At the same time, many test developers have explored the use of computer administered tests, but this form of testing has been limited almost exclusively to multiple-choice tests. Relatively little attention has been paid to the use of computers to administer tests which require students to generate responses to open-ended items.

The consequences of the incongruities in these developments may be substantial. As the use of computers in schools and homes increases and students do more of their writing with word processors, at least two problems arise. First, performance tests which require students to produce responses long-hand via paper-and-pencil (which happens not just with large scale tests of writing, but also for assessments of other skills as evidenced through writing) may violate one of the key assumptions of the authentic assessment movement. For people who do most of their writing via computer, writing long-hand via paper-and-pencil is an artificial rather then real world task. Second, and more importantly, paper-and-pencil tests which require answers to be written long-hand to assess students' abilities (in writing or in other subjects) may yield underestimates of the actual abilities of students who are accustomed to writing via computer.

In this article, we present the results of a small study on the effect of computer administration on student performance on writing or essay tests. Specifically, we discuss the background, design and results of the study reported here. However, before focusing on the study itself, we present a brief summary of recent developments in computerized testing and authentic assessment.

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). Such visions of computerized testing, however, 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/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. Although exceptions have been found, Bunderson, Inouye & Olsen (1989) summarize the general pattern of findings from several studies which examined the equivalence of scores acquired through computer or paper- and-pencil test forms as follows: "In general it was found more frequently that the mean scores were not equivalent than that they were equivalent; that is the scores on tests administered on paper were more often higher than on computer-administered tests." However, the authors also state that "[t]he score differences were generally quite small and of little practical significance" (p. 378). More recently, Mead & Drasgow (1993) reported on a meta-analysis of 29 previous studies of the equivalence of computerized and paper-and-pencil cognitive ability tests (involving 159 correlations between computerized and paper-and-pencil test results). Though they found that computerized tests were slightly harder than paper-and-pencil tests (with an overall cross-mode effect size of -.04), they 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" (Mead & Drasgow, 1993, p. 457).

Yet, as previously noted, standardized multiple-choice tests, which have been the object of comparison in previous research on computerized versus paper-and-pencil testing, have been criticized by proponents of authentic assessment. Among the characteristics which lend authenticity to an assessment instrument, Darling-Hammond, Ancess & Falk (1995) argue that the tasks be "connected to students' lives and to their learning experiences..." and that they provide insight into "students' abilities to perform 'real world' tasks" (p.4-5). Unlike standardized tests, which may be viewed as external instruments that measure a fraction of what students have learned, authentic assessments are intended to be closely linked with daily classroom activity so that they seamlessly "support and transform the process of teaching and learning" (Darling-Hammond, Ancess & Falk, 1995, p. 4; Cohen, 1990).

In response to this move towards authentic assessment, many developers of nationally administered standardized tests have attempted to embellish their instruments by including open-ended items for which students have to write their answers. These changes, however, have occurred during a period when both the real-world and the school-world have experienced a rapid increase in the use of computers.

The National Center for Education Statistics report that the percentage of students in grades 1 to 8 using computers in school has increased from 31.5 in 1984, to 52.3 in 1989 and to 68.9 in 1993 (Snyder & Hoffman, 1990; 1994). In the workplace, the percentage of employees using computers has risen from 36.0 in 1989 to 45.8 in 1993. During this period, writing has been the predominant task adult workers perform on a computer (Snyder & Hoffman, 1993; 1995). Given these trends, tests which require students to answer open-ended items via paper-and-pencil may decrease the test's "authenticity" in two ways: 1. Assessments are not aligned with students' learning experiences; and 2. Assessments are not representative of 'real-world' tasks. As the remainder of this paper suggests, these shortcomings may be leading to underestimates of students' writing abilities.

**Background to this Study**

In 1993, the Advanced Learning Laboratory School (ALL School, (http://nis.accel.worc.k12.ma.us) of Worcester, Massachusetts decided to adopt the Co-NECT school design (or Cooperative Networked Educational Community for Tomorrow, http://co-nect.bbn.com). Developed by BBN, Inc., a Boston- based communications technology firm, Co-NECT is one of nine models for innovative schooling funded by the New American Schools Development Corporation. Working with BBN, the ALL School restructured many aspects of its educational environment. Among other reforms, the traditional middle school grade structure (that is, separately organized grade 6, 7 and 8 classes) was replaced with blocks which combined into a single cluster students who otherwise would be divided into grades 6, 7 and 8. In place of traditional subject-based classes (such as English Class, Math Class, Social Studies, etc.), all subjects were integrated and taught through project-based activities. To support this cooperative learning structure, several networked computers were placed in each classroom, allowing students to perform research via the Internet and CD-ROM titles, to write reports, papers and journals, and to create computer based presentations using several software applications.

To help evaluate the effects the restructuring at the ALL School has on its students as a whole, the Center for the Study of Testing, Evaluation and Educational Policy (CSTEEP) at Boston College helped teachers gather baseline data in the fall of 1993 with plans to perform follow-up assessments in the spring of 1994 and each spring thereafter. To acquire a broad picture of students' strengths and weaknesses, the forms of tests included in the baseline

assessment ranged from multiple choice tests to short and long answer open-ended assessments to hands-on performance assessments covering a wide range of reading, writing, science and math skills. To acquire insight into how cooperative projects affected the development of group skills, some of the performance assessments required students to work together to solve a problem and/or answer specific questions. Finally, to evaluate how the Co-NECT Model, as implemented in the ALL School, affected students' feelings about their school, a student survey was administered. Assessments and surveys were administered to representative samples of the whole school's student population.

In the spring of 1994, the same set of assessments was re-administered to different representative samples of students. While a full discussion of the results is beyond the scope of this paper, many of the resulting patterns of change were as expected. For example, performance items which required students to work cooperatively generally showed more improvement than items which required students to work independently. On items that required students to work independently, improvement was generally stronger on open-ended items than on multiple-choice items. But there was one notable exception: open-ended assessments of writing skills suggested that writing skills had declined.

Although teachers believed that the Co-NECT Model enhanced opportunities for students to practice writing, performance on both short answer and long answer writing items showed substantial decreases. For example, on a short answer item which asked students to write a recipe for peace, the percentage of students who responded satisfactorily decreased from 69% to 51%. On a long answer item which asked students to imagine a superhero, describe his/her powers, and write a passage in which the superhero uses his/her powers, the percentage of satisfactory responses dropped from 71% to 41%. On another long answer item that asked students to write a story about a special activity done with their friends or family, student performance dropped from 56% to 43%. And on a performance writing item which first asked students to discuss what they saw in a mural with their peers and then asked them to write a passage independently that described an element in the mural and explain why they selected it, the percentage of satisfactory responses decreased from 62% to 47%. These declines were all statistically significant, and more importantly were substantively troubling.

Since writing was a skill the school had selected as a focus area for the 1993-94 school year, teachers were surprised and troubled by the apparent decrease in writing performance. During a feedback session on results in June 1994, teachers and administrators discussed at length the various writing activities they had undertaken over the past year. Based on these conversations, it was evident that students were regularly presented with opportunities to practice their writing skills. But a consistent comment was that teachers in the ALL School were increasingly encouraging students to use computers and word processing tools in their writing. As several computers were present in all classrooms, as well as in the library, teachers believed that students had become accustomed to writing on the computer. When one teacher suggested that the decrease in writing scores might be due to the fact that all writing items in spring 1994 were administered on paper and required students to write their responses by hand, the theory was quickly supported by many teachers. With a follow-up assessment scheduled to occur a year later, several teachers asked if it would was possible for students to perform the writing items on a computer.

After careful consideration, it was decided that a sub- sample of students in spring 1995 would perform a computer- administered version of the performance writing item and items from the National Assessment of Educational Progress (NAEP) (items were mostly multiple-choice with a few short answer items included). But, to preserve comparisons with results from 1993-94, the majority of the student population would perform these assessments as they had in that year -- via the traditional pencil-and-paper medium. Hence, we undertook an experiment to compare the effect that the medium of administration (computer versus paper-and-pencil) has on

student performance on multiple-choice, short-answer and extended writing test items.

**Study Design and Test Instruments**

To study the effect the medium of administration has on student performance, that is taking assessments on computer versus by hand on paper, two groups of students were randomly selected from the ALL School Advanced Cluster (grades 6, 7 and 8). For the experimental group, which performed two of three kinds of assessments on computer, 50 students were selected. The control group, which performed all tests via pencil-and-paper, was composed of the 70 students required for the time-trend study described above. The three kinds of assessments performed by both groups were:

1. An open-ended (OE) assessment comprising 14 items, which included two writing items, five science items, five math items and two reading items.
2. A test comprised of NAEP items which was divided into three sections and included 15 language arts items, 23 science items and 18 math items. The majority of NAEP items were multiple-choice. However, 2 language arts items, 3 science items and 1 math item were open-ended and required students to write a brief response to each item's prompt.
3. A performance writing assessment which required an extended written response.

Both groups performed the open-ended (OE) assessment in exactly the same manner, by hand via paper-and pencil. The experimental group performed the NAEP and writing assessment on computer, whereas the control group performed both in the traditional manner, by hand on paper.

The performance writing assessment consisted of a picture of a mural and two questions. Students formed small groups of 2 or 3 to discuss the mural. After 5 to 10 minutes, students returned to their seats and responded to one of two prompts:

1. Now, it is your turn to pick one thing you found in the mural. Pick one thing that is familiar to you, that you can recognize from your daily life or that is part of your culture. Describe it in detail and explain why you chose it.
2. Artists usually try to tell us something through their paintings and drawings. They may want to tell us about their lives, their culture or their feelings about what is happening in the neighborhood, community or world. What do you think the artists who made this mural want to tell us? What is this mural's message?

Due to absences, the actual number of students who participated in this study was as follows:

**Experimental (Computer) Group:** 46

**Control (Paper-and-Pencil) Group:** 68

It should be noted that the study described in this paper was performed as part of a larger longitudinal study which relied heavily on matrix sampling. For this reason, not all of the students in the control group performed all three tests. However, all students included in the analyses reported here performed at least two tests, one of which was the open-ended assessment. Table 1 shows the actual number of students in each group that performed each test.

## Table 1
## Number of Students Performing Each Test

| Test | Experimental | Control | Total |
|---|---|---|---|

| | | | |
|---|---|---|---|
| Open-ended | 46 | 68 | 114 |
| NAEP | 44 | 42 | 86 |
| Perf. Writing | 40 | 46 | 86 |

To be clear, we emphasize that the treatment, in terms of which the experimental and control groups differed, had nothing to do with educational experience of the two groups. The groups were receiving similar -- albeit quite unusual in comparison to most middle schools -- educational experiences in the ALL school. The treatment, in terms of which the two groups differed, was simply that the experimental group took the NAEP and performance writing tests on computer, whereas the control group took these tests in the traditional manner, by hand with paper-and-pencil.

**Converting Paper Tests to Computer**

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, 1986) 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, each page of the paper version of the NAEP items and the performance writing item was replicated on the computer screen as precisely as possible. To that end, the layout of text and graphics on the computer version matched the paper version, including the number of items on a page, the arrangement of response options, and the positioning of footers, headers and directions. Despite these efforts, not every screen matched every page. Since the computer screen contained less vertical space, it was not always possible to fit the same number of questions on the screen as appeared on the page. In addition, to allow the test taker to move between screens (e.g., to go on to the next screen, back to a previous screen, or to flip to a passage or image to which an item referred), each screen of the computer versions contained navigation buttons along its bottom edge. Finally, to decrease the impact of screen fatigue, a larger font was used on the computer version than on the paper version.

To create a computerized version of the NAEP and performance writing tests, the following steps were taken:

1. An appropriate authoring tool was selected. To fully integrate the several graphics used in the multiple-choice items and the full-color photograph of a mural used in the performance writing item, as well as to track students' responses, Macromedia Director was used.
2. All graphics and the photograph of the mural were scanned. Adobe Photoshop was used to retouch the images.
3. A data file was created to store student input, including name, ID number, school name, birth date, gender, date of administration and responses to each item.
4. A prototype of each test was created, integrating the graphics, 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 biographical information.

5. The prototype was tested on several adults and students to assure that all navigational buttons functioned properly, that data was stored accurately, and that items and graphics were easy to read.
6. Finally, the prototype was revised as needed and the final versions of the computer tests were installed on twenty-four computers in the ALL School.

As described above, the addition of navigational buttons along the lower edge of the computer screen was the most noticeable difference between the computer and paper versions of the tests. To allow students to review their work and make changes as desired, a "Next Page" and "Previous Page" button appeared on all pages (or screens) of the computer tests (except the first and last page). To allow students to review their work, student responses were not recorded until the student reached the last page of the assessment and clicked a button labeled "I'm Finished." When the "I'm Finished" button was clicked, the student's biographical information and responses to each item were recorded in a data file before the program terminated. For all multiple-choice items, students clicked the option they felt best answered the question posed. For both short- and long-answer questions, examinees used a keyboard to type their answers into text boxes which appeared on their screen. Though they could edit using the keyboard and mouse, examinees did not have access to word processing tools such as spell-checking.

## Scoring

A combination of multiple choice and open-ended items were performed by both groups of students. Multiple-choice NAEP items were scored as either correct or incorrect based upon the answer key accompanying the NAEP items. To prevent rater bias based on the mode of response, all short-answer NAEP responses were entered verbatim into the computer. Responses of students who had taken the NAEP questions on computer and via paper-and-pencil were then randomly intermixed. Applying the rating rubrics designed by NAEP, two raters independently scored each set of six short answer items for each student. As part of an overall strategy to summarize results on all items in terms of percent correct, the initial ratings (which ranged from 1 - 5) were converted to a dichotomous value: 1 or 0; to denote whether student responses were adequate or inadequate. The two raters' converted scores were then compared. Where discrepancies occurred, the raters re-evaluated responses and reached consensus on a score.

To score the performance writing item, all hand written responses were entered verbatim into the computer -- again so as to prevent raters from knowing which responses were originally written by hand. The hand-written and computer- written responses were randomly intermixed. Three independent raters then scored each written response, using the following four-point scoring rubric:

1. Too brief to evaluate: Student did not make an attempt; indicates that student either did not know how to begin, or could not approach the problem in an appropriate manner.
2. Inadequate Response: Student made an attempt but the response was incorrect, reflected a misconception and/or was poorly communicated.
3. Adequate Response: Response is correct and communicated satisfactorily, but lacks clarity, elaboration and supporting evidence.
4. Excellent Response: Response is correct, communicated clearly and contains evidence which supports his/her response.

Initial analyses of the three raters' ratings showed that there was only a modest level of inter-rater reliability among the three (inter-rater correlations ranged from 0.44 to 0.62, across the total of 89 performance writing responses). Although these correlations were lower than expected, research on the assessment of writing has shown that rating of writing samples, even

among trained raters, tends to be only modestly reliable (Dunbar, Koretz, & Hoover, 1991). Indeed, that is why we planned to have more than one rater evaluate each student response to the performance writing task. Hence for the purpose of the study reported here we created composite performance rating scores by averaging the three ratings of each student's response (which we call PWAvg).

Since the open-ended assessment was performed by paper- and-pencil by all students, student responses were not entered into the computer. A single rater, who did not know which students had performed other assessments on the computer, scored all responses using a 4 point scale. Although each of the 14 items had its own specific scoring criteria, the general meaning of each score was the same across all 14 open-ended items, as well as the performance writing item. The raw scores were then collapsed into a 0, 1 scale, with original scores of 1 or 2 representing a 0, or inadequate response, and original scores of 3 or 4 representing a 1, or adequate response. For the purpose of the study reported here, total open-ended response scores were calculated by summing across all 14 OE items.

## Results

In presenting results from this study, we discuss: 1) assessment results overall; 2) comparative results from the two groups that took assessments via computer or via paper-and-pencil; 3) results of regression analyses; and 4) separate analyses of performance on the short-answer and multiple-choice NAEP items.

We present descriptive data summaries before results of statistical tests. Regarding the latter, we note that this experiment involved multiple comparisons of results based on just two random samples of students. While the literature on how to adjust alpha levels to account for multiple comparisons (e.g. Hancock & Klockars, 1996) is too extensive to review here, let us simply summarize how we dealt with this issue. We planned to compare results for the experimental and control groups on five different measures: OE, performance writing, and three NAEP subtests, in science, math, and language arts. The Dunn approach to multiple comparisons tells us that the $\alpha$ for *c* multiple comparisons, $\alpha_{pc}$, is related to simple $\alpha$ for a single comparison, as follows:

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

Hence for five comparisons, the adjusted value of a simple 0.05 alpha level becomes 0.0102. Analogously, a simple alpha level of 0.01 for a single comparison becomes 0.0020 for five planned comparisons. We use these alpha levels in discussing the statistical significance of comparisons between experimental and control group results. In discussion, we address not just the statistical significance, but also the substantive significance of our findings.

## Overall Results

The actual raw data on which all analyses are based is being made available to the reader. From this point, the data files can be accessed in ASCII or EXCEL Spreadsheet (binary) form.

Table 2 presents a summary of overall results, that is, combined results for all students who took any of the three assessments in Spring 1995.

### Table 2
### Summary Statistics for All Assessments

| | Scale Range | n | Mean | SD |
|---|---|---|---|---|
| OE | 0-14 | 114 | 7.87 | 2.96 |

| | | | | |
|---|---|---|---|---|
| NAEP Lang Arts | 0-15 | 86 | 9.84 | 3.79 |
| NAEP Science | 0-23 | 86 | 9.70 | 4.37 |
| NAEP Math | 0-18 | 86 | 6.21 | 3.39 |
| Perf Writing Avg | 1-4 | 86 | 2.53 | 0.62 |

These data indicate that the assessments were relatively challenging for the students who performed them. Mean scores were in the range of 56-66% correct for the OE and NAEP Language Arts tests, but considerably below 50% correct for the NAEP science and NAEP math subtests. In this regard, it should be noted that all of these assessments were originally designed to be administered to eighth graders, but in the study reported here they were administered to 6th, 7th and 8th grade level students who in the ALL school are intermixed in the same clusters.

Table 3 presents Spearman rank order intercorrelations of all assessments, again across both groups. The OE results correlated only slightly higher with the PWAvg results, possibly reflecting the fact that both of these assessments were open-ended requiring students to produce rather than select an answer. The three NAEP item subtests showed moderate intercorrelations (0.56-0.62) which might be expected for multiple-choice tests in the different subject areas (despite the fact that none of the NAEP subtests contained as many as two dozen items). The PWAvg results showed modest correlations with the NAEP subtests. Of the three NAEP sub-tests, the PWAvg was most strongly correlated with the Science sub-test. Although the NAEP science results were based largely on multiple choice items, of the three NAEP subtests, the Science section contained the largest number of short answer items (3 out of 23 items). The NAEP subtest that correlated least with the PWAvg scores (0.37) was the NAEP Math subtest, which contained only one open-ended item.

**Table 3**
**Intercorrelations of Assessment Results**

| | OE | NAEP Lang Arts | NAEP Science | NAEP Math | Perf. Writing |
|---|---|---|---|---|---|
| OE | 1.00 | | | | |
| NAEP Lang Arts | 0.46 | 1.00 | | | |
| NAEP Science | 0.44 | 0.62 | 1.00 | | |
| NAEP Math | 0.40 | 0.56 | 0.57 | 1.00 | |
| Perf Writing | 0.48 | 0.49 | 0.54 | 0.37 | 1.00 |

$p < .01$ for all intercorrelations

## Computer versus Paper-and-Pencil Results

Table 4 presents results separately for the experimental and control groups, namely the group which took NAEP and performance writing assessments on paper and the one that took them on computer. The table also shows results of t-tests (for independent samples, assuming equal variances for the two samples and hence using a pooled variance estimate). As an aid to interpretation, the table also shows the effect of computer administration in terms of Glass's delta effect size, that is the mean of the experimental group minus the mean of the control group divided by the standard deviation of the control group. While other methods for calculating effect size have been proposed (Rosenthal, 1994, p. 237), note that results would not differ dramatically if a pooled standard deviation were used instead of the control group standard deviation.

Results indicate that, after adjusting for the planned multiple comparisons, the effect of computer administration was significant only for the PWAvg. The effect size of computer administration on the performance writing task was 0.94.

The four tests which did not show a statistically significant difference between the two groups were the OE test and the NAEP Language Arts, Science, and Math tests. The absence of a statistically significant difference on the OE test was, of course, expected since the OE test was the one test that was administered in the same form (paper-and- pencil) to the two groups. Similarly, since the NAEP tests were primarily composed of multiple-choice items, which previous research suggests are affected minimally by the mode of administration, differences between the two groups on the NAEP tests were not expected. Note however that the size of the difference in OE scores between the two groups was surprisingly large, given that the two groups had been randomly selected. The absence of four students randomly selected for the experimental group who did not take any tests may partially explain this difference. Nevertheless to explore the possibility that group differences may partially account for apparent mode of administration effects (and also, of course, to estimate effects more precisely), regression analyses were conducted.

**Table 4**
**Summary Results by Group**

| | *Control* | | | *Experimental* | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | n | Mean | SD | n | Mean | SD | Effect Size (df) | t | Sig |
| OE | 68 | 7.62 | 3.14 | 46 | 8.24 | 2.66 | 0.20 (112) | 1.10 | 0.27 |
| Lang Arts | 42 | 9.24 | 3.96 | 44 | 3.58 | 0.30 | 0.30 (84) | 1.44 | 0.15 |
| Science | 42 | 8.67 | 4.17 | 44 | 10.68 | 4.39 | 0.48 (84) | 2.18 | 0.03 |
| Math | 42 | 6.00 | 3.30 | 44 | 6.41 | 3.51 | 0.12 (84) | 0.56 | 0.58 |
| Perf Writ. | 46 | 2.30 | 0.55 | 40 | 2.81 | 0.59 | 0.94 (84) | 4.16 | <.0001** |

** statistically significant at the 0.01 level after taking multiple comparisons into account

## Regression Analyses

As a further step in examining the effects of mode of administration, regression analyses were conducted using the OE scores as a covariate and then introducing a dummy variable (0= paper/pencil group; 1= computer administration group) to estimate the effects of mode of administration on the NAEP Language Arts, Science and Math subtests and on the PWAvg scores. Results of these regression analyses are shown in Table 5.

**Table 5**
**Results of Regression Analyses**

| Dependent Variable | Coeff | SE | t-ratio | Sig |
|---|---|---|---|---|
| NAEP Lang Arts | | | | |
| Constant | 5.03 | 1.09 | 4.60 | <.0001** |
| OE | 0.57 | 0.13 | 4.40 | <.0001** |
| Group* | 0.66 | 0.75 | 0.89 | 0.38 |
| NAEP Science | | | | |

| | | | | |
|---|---|---|---|---|
| Constant | 3.72 | 1.23 | 3.02 | .0033 |
| OE | 0.67 | 0.15 | 4.59 | <.0001** |
| Group* | 1.42 | 0.84 | 1.69 | 0.09 |
| NAEP Math | | | | |
| Constant | 1.99 | 0.97 | 2.04 | <.0445 |
| OE | 0.54 | 0.12 | 4.70 | <.0001** |
| Group* | -0.07 | 0.67 | 0.11 | 0.91 |
| Perf Writing | | | | |
| Constant | 1.59 | 0.16 | 9.73 | <.0001** |
| OE | 0.09 | 0.02 | 4.88 | <.0001** |
| Group* | 0.44 | 0.11 | 3.98 | .0001** |

* (1=computer)

** statistically significant at the 0.01 level after taking multiple comparisons into account

These results confirm the findings shown in Table 4, namely that even after controlling for OE scores, the effect of mode of administration was highly significant on the PWAvg. However, for the largely multiple-choice NAEP subtests, results indicate no difference for mode of administration.

## Performance on Multiple Choice and Short-Answer NAEP Items

Although the regression analysis suggested that mode of administration did not significantly influence performance on the NAEP subtests, further analysis was performed on the NAEP subtest items to examine the effect of administration mode on the two forms of items contained in the NAEP subtest -- multiple-choice and short answer. Table 6 shows the mean score for the two groups on both the multiple-choice items and the short-answer items for the three subtests. Although slight differences between the means were found for the multiple-choice items, none were significant. However, for the science and language arts short answer items, those students who responded on computer performed significantly better than the paper-and-pencil group. While it was expected that performance on multiple-choice items would not differ, the differences detected on the short answer items suggest that even for items that require a brief written response, the mode of administration may affect a student's performance.

The question arises as to why the mode of administration affected performance on the short answer Language Arts and Science questions, but not on the one short-answer Math item. It is likely that the nature of the open-ended Math item accounts for similar performance between the two groups. The open-ended Math question required a short answer which could not be provided without correctly answering the multiple-choice question that preceeded it. In contrast, the three short answer Science items asked students to interpret data in a table, explain their process and respond to a factual item. In particular, the second short answer Science item provided a fair amount of space for a response and many students wrote at least one complete sentence. Although the three Science items were related to the same set of data displayed in a table, response to these items were not dependent on answers to previous items.

### Table 6: Results of Analysis of NAEP Subtest Item formats: Multiple-choice versus Short Answer

| Items | n | Control | SD | n | Experimental | SD | Effect | t | Sig |
|---|---|---|---|---|---|---|---|---|---|

| | | Mean | | | Mean | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Lang. Arts** | | | | | | | | | |
| Mult. Choice | 42 | 8.6 | 3.47 | 44 | 9.0 | 3.03 | 0.12 | 0.64 | .522 |
| Short Answer | 42 | 0.6 | 0.73 | 44 | 1.4 | 0.75 | 0.99 | 4.52 | <.0001** |
| **Science** | | | | | | | | | |
| Mult. Choice | 42 | 8.0 | 3.97 | 44 | 9.0 | 3.99 | 0.26 | 1.22 | .226 |
| Short Answer | 42 | 0.7 | 0.77 | 44 | 1.7 | 0.98 | 1.25 | 5.06 | <.0001** |
| **Math** | | | | | | | | | |
| Mult. Choice | 42 | 5.8 | 3.07 | 44 | 6.1 | 3.33 | 0.10 | 0.44 | .660 |
| Short Answer | 42 | 0.2 | 0.41 | 44 | 0.3 | 0.47 | 0.25 | 1.08 | .282 |

** statistically significant at the 0.01 level after taking multiple comparisons into account

To inquire further into the apparent effect of mode of administration on short answer Language Arts and Science items, we conducted regression analyses, using OE scores as a covariate. Results, shown in Table 7 indicate that the mode of administration had a significant effect on the students' performances on the NAEP Language Arts and Science short-answer items.

## Table 7: Results of Regression Analyses on NAEP Language Arts and Science Short-Answer Items

| Dependent Var | Coef. | s.e. | beta | s.e. | t-ratio | Sig |
|---|---|---|---|---|---|---|
| **NAEP Lang Arts** | | | | | | |
| Constant | -0.08 | 0.22 | | | -.38 | .71 |
| OE | 0.10 | 0.03 | 0.35 | 0.09 | 3.77 | .0003** |
| Group* | 0.63 | 0.15 | 0.39 | 0.09 | 4.23 | .0001** |
| **NAEP Science** | | | | | | |
| Constant | 0.20 | 0.28 | | | 0.74 | .4645 |
| OE | 0.07 | 0.03 | 0.20 | 0.09 | 2.09 | .0397 |
| Group* | 0.91 | 0.19 | 0.45 | 0.09 | 4.77 | <.0001** |

* (1=computer)
** statistically significant at the 0.01 level after taking multiple comparisons into account

## Discussion

The experiment described here was a small inquiry aimed at investigating a particular question. Motivated by a question as to whether or not performance on an extended writing task might be better if students were allowed to write on computer rather than on paper, the study aimed at estimating the effects of mode of administration on test results for two kinds of assessments, namely the largely multiple-choice NAEP subtests and the extended writing task previously described. Unlike most previous research on the effects of computer administered tests, which has focused on multiple-choice tests and has generally found no or small differences due to mode of administration, our results indicate substantial effects due to mode of administration. The size of the effects was found to be 0.94 on the extended writing task and .99

and 1.25 for the NAEP language arts and science short answer items. Effect sizes of this magnitude are unusually large and of sufficient size to be of not just statistical, but also practical significance (Cohen, 1977; Wolf, 1986). An effect size of 0.94, for example, implies that the score for the average student in the experimental group exceeds that of 83 percent of the students in the control group.

A number of authors have noted the difficulty of interpreting the practical significance of effect sizes and have suggested that one useful way of doing so is with a "binomial effect size display" showing proportions of success and failure under experimental and control conditions (Hedges & Olkin, 1985; Rosenthal & Rubin, 1982). While there are a number of ways in which effect sizes, expressed as either Glass's delta or a correlation coefficient, can be converted to a binomial effect size display, in the case of our PWAvg scores, we have a direct way of showing such a display. Recall that student responses to the performance writing item were scored on a 4-point scale in which scores of 1 and 2 represented a less than adequate response and scores of 3 and 4 represented an adequate or better response. Using the cut-point of 2.5 as distinguishing between inadequate (failure) and adequate (success) responses in terms of PWAvg scores, we may display results as shown in Table 8.

## Table 8: Binomial Effect Size Display of Experimental Results: In Terms of Inadequate vs. Adequate PWAvg Scores

| **Control (Paper)** | Inadequate | Adequate |
|---|---|---|
| N | 32 | 14 |
| Percent | 69.6% | 30.4% |
| **Experimental (Computer)** | Inadequate | Adequate |
| N | 13 | 27 |
| Percent | 32.5% | 67.5% |

This display indicates that the computer mode of administration had the effect of increasing the success rate on the performance writing item (as judged by the average of three independent raters) from around 30% to close to 70%.

As a means of inquiring further into the source of this large effect, we conducted a variety of analyses to explore why and for whom the mode of administration effect occurred. To explore why the mode of administration effect may have occurred, we first undertook a textual analysis of student responses to the extended writing task. Specifically we calculated the average number of words and paragraphs contained in the responses of both groups. As Table 9 below indicates, those students who performed the assessment on the computer tended to write almost twice as much and were more apt to organize their responses into more paragraphs.

## Table 9: Characters, Words and Paragraphs on Performance Writing Task by Mode of Administration

| | Characters | Words | Paragraphs |
|---|---|---|---|
| **Control (Paper)** | | | |
| Mean | 586.9 | 111.6 | 1.457 |
| Std | 275.58 | 52.47 | 1.069 |
| n | 46 | 46 | 46 |
| **Experimental (Computer)** | | | |

| | | | |
|---|---|---|---|
| Mean | 1022.2 | 204.7 | 2.625 |
| Std | 549.55 | 111.32 | 2.306 |
| n | 40 | 40 | 40 |
| observed t with pooled variance | 4.73 | 5.07 | 3.08 |
| sig | <.0001** | <.0001** | <.0001** |

** statistically significant at the 0.01 level after taking multiple comparisons into account

In some ways, this pattern is consistent with the findings of Daiute (1985) and Morocco and Neuman (1986), who have shown that teaching writing with word processors tends to lead students to write more and to revise more than when they write with paper-and-pencil. Not surprisingly, the length of students' written responses (in terms of numbers of characters and words correlated significantly with PWAvg scores, 0.63 in both cases). Although this suggests that longer responses tended to receive higher scores, the fact that length of response explains less than half of the variance in PWAvg scores suggests that rated quality is not attributable simply to length of response.

Second, we considered the possibility that motivation might help explain the mode of administration effect. This possibility was suggested to us by spontaneous comments made by students after the testing. For example, after taking the writing assessment on computer, one student commented, "I thought we were going to be taking a test." In contrast, a student in the control group, who had not taken any tests via computer, inquired of us, "How come we didn't get to take the test on computer?" Such comments raised the possibility that motivation and the simple novelty of taking tests on computer might explain the mode of administration effect we found.

Two lines of thought suggest that simple motivation cannot explain our results. If differential motivation arising from the novelty of taking tests on computer was the main cause of our results, it is hard to explain why mode of administration effects were absent on the multiple-choice NAEP subtests, but were prevalent on the performance writing test and the NAEP open-ended items. Furthermore, recent research on the effects of motivation on test performance, suggests that the effects of motivation are not nearly as large as the mode of administration effect we found on the performance writing test. Recently, Kiplinger & Linn (1996) reported on the effects of an experiment in which "low- stakes" NAEP items were embedded in a "high stakes" state testing program in Georgia. Though results from this experiment were mixed, the largest effects of "high stakes" motivation occurred for nine NAEP items designed for eighth grade students. For these nine items, however, the effect size was only 0.18. (Kiplinger & Linn, 1996, p.124). In a separate study, O'Neill, Sugrue & Baker (1996) investigated the effects of monetary and other incentives on the performance of eighth grade students on NAEP items. Again, though effects of these motivational conditions were mixed, the largest influence of motivation ranged from an effect size of 0.16 to 0.24 (O'Neill, Sugrue & Baker, 1996, p. 147). With the largest effects of motivation on eighth grade students found to be in the range of 0.16 to 0.24, these results suggest that motivation alone cannot explain the magnitude of mode of administration effects we found for written responses.

To examine for whom the mode of administration effects occurred, we also inquired into whether the mode of administration effect appeared to be different for different students. First we inquired into whether the mode of administration effect seemed to be different for students performing at different levels on the OE test. One simple way of testing this possibility was to calculate PWAvg scores predicted on the basis of OE scores and see if there was a statistically

significant correlation between residuals (actual minus predicted PWAvg scores) and OE scores among the experimental group students. No significant correlation was found, suggesting that the mode of administration effect was not different for students of different ability levels as indicated by their OE scores. A graphical presentation of this pattern is shown in Figure 1, which depicts the line of PWAvg scores regressed on OE scores, with the experimental cases represented with X's and the control group with dots. As can be seen in Figure 1, the actual PWAvg scores for the experimental group tended to exceed the predicted scores across ability levels as represented by the OE scores.

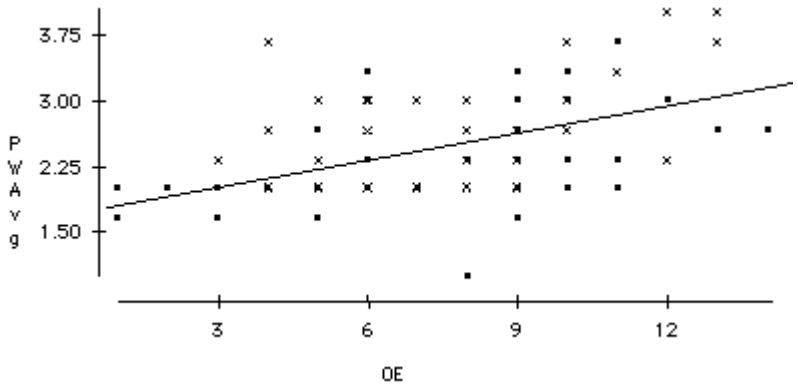


**Figure 1: Regression of PWAvg scores on OE Scores**

Finally, we explored whether mode of administration effect seemed to differ for males versus females. Table 10 shows PWAvg scores by gender for both control and experimental groups.

## Table 10: PWAvg Scores by Gender and Group

| | Female | Male | Total |
|---|---|---|---|
| **Control (Paper)** | | | |
| Mean | 2.33 | 2.27 | 2.30 |
| SD | 0.58 | 0.47 | 0.55 |
| n | 21 | 25 | 46 |
| **Experimental (Computer)** | | | |
| Mean | 2.92 | 2.60 | 2.81 |
| SD | 0.53 | 0.63 | 0.59 |
| n | 26 | 14 | 40 |
| **Total** | | | |
| Mean | 2.66 | 2.38 | 2.53 |
| SD | 0.65 | 0.54 | 0.62 |
| n | 46 | 39 | 86 |

Within the control groups, females performed only slightly better on PWAvg scores than did males (with means 2.33 and 2.27 respectively). However within the experimental group females scored considerably better than males (with means of 2.92 and 2.60). Thus it appears that the effect of computer administration may have been somewhat larger for females than for males.

Nonetheless the males who took the extended writing task on computer still performed considerably better than the females who took the writing task on paper (with respective means of 2.60 and 2.33). A two way analysis of variance (PWAvg by gender and group) showed group but not gender to be significant (this was the case whether or not an interaction term was included). This general pattern was confirmed by regression analyses of PWAvg scores on OE scores, sex and group. Though OE scores and the group variable were significant, the sex variable was not.

We should note that post hoc, we were surprised that the proportion of males in the control group (54%) differed by nearly 19 percentage points from the proportion of males in the experimental group (35%). Although the two groups were selected randomly, the probability that this difference would occur is less than .08. However, as can be calculated based on the data in Table 10, even after controlling for gender, the average effect size is 0.86.

Although the experiment reported here had several weaknesses--only one extended writing task was used, no other variables on academic achievement beyond the OE test results were used as covariates in regression analyses, and information on students' extent of experience working on computers was not collected--further research into this topic clearly is warranted.

Increasingly, schools are encouraging students to use computers in their writing. As a result, it is likely that increasing numbers of students are growing accustomed to writing on computers. Nevertheless, large scale assessments of writing, at state, national and even international levels, are attempting to estimate students' writing skills by having them use paper-and-pencil. Our results, if generalizable, suggest that for students accustomed to writing on computer for only a year or two, such estimates of student writing abilities based on responses written by hand may be substantial underestimates of their abilities to write when using a computer.

This suggests that we should exercise considerable caution in making inferences about student abilities based on paper-and-pencil/handwritten tests as students gain more familiarity with writing via computer. And more generally it suggests an important lesson about test validity. Validity of assessment needs to be considered not simply with respect to the content of instruction, but also with respect to the medium of instruction. As more and more students in schools and colleges do their work with spreadsheets and word processors, the traditional paper-and- pencil modes of assessment may fail to measure what they have learned.

We suspect that it will be some years before schools generally, much less large scale state, national or international assessment programs, develop the capacity to administer wide-ranging assessments via computer. In the meantime, we should be extremely cautious about drawing inferences about student abilities when the media of assessment do not parallel those of instruction and learning.

## Acknowledgment

We wish to acknowledge our great appreciation of Carol Shilinsky, Principal, and the teachers of the ALL School of Worcester, MA, who first suggested the idea of the study reported here. In particular we thank Rich Tamalavich and Deena Kelly who helped arrange the equipment needed and oversaw administration of the computerized testing. We also wish to thank five anonymous EPAA reviewers who, through editor Gene Glass, provided us with some very helpful suggestions on a previous version of this manuscript with a rapidity that was absolutely astonishing. Five reviews were received within one week of submission of the manuscript! We also thank Gene Glass for suggesting that we take better advantage of the electronic medium via which this journal is published, for example, by appending to this article the full data set for the study so that others might conduct secondary analyses, and by providing electronic links to institutions mentioned in the study. So now, the reviewer who wondered about

the unusual school in which this study was conducted can visit the ALL Schools WWW site. Any such visitors will find that students at this remarkable school are now not just writing via computer, but also publishing via the WWW.

## References


Barton, P. E. & Coley R. J. (1994) Testing in America's schools. Princeton, NJ Educational Testing Service Policy Information Center.

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.

Cizek, G. J. (1991). The Effect of Altering the Position of Options in a multiple-choice Examination. Paper presented at NCME, April 1991. (ERIC)

Cohen, D. (1990). Reshaping the Standards Agenda: From an Australian's Perspective of Curriculum and Assessment. In P. Broadfoot, R. Murphy & H. Torrance (Eds.), *Changing Educational Assessment: International Perspectives and Trends*. London: Routledge.

Cohen, J (1977). *Statistical power analysis for the behavioral sciences* (rev. ed.) NY: Academic Press.

Daiute, C. (1985). *Writing and computers*. Reading, MA: Addison-Wesley Publishing Co.

Darling-Hammond, L., Acness, J. & Falk, B. (1995). *Authentic Assessment in Action*. New York, NY: Teachers College Press.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality Control in the Development and Use of Performance Assessments. *Applied Measurement in Education,* 4(4), 289-303.

Green, B. F., Jr. (1970). Comments on tailored testing. In W. H. Holtzman (Ed.), *Computer-assisted instruction, testing and guidance*. New York: Harper and Row.

Haas, C. & Hayes, J. R. (1986). What Did I Just Say? Reading Problems in Writing with the Machine. *Research in the Teaching of English*, 20(1), 22-35.

Hancock, G. R. & Klockars, A. J. (1996). The quest for ∝: Developments in multiple comparison procedures in the quarter century since Games (1971). *Review of Educational Research*, 66(3), 269-306.

Hedges, L. V. & Olkin, I. (1985) *Statistical methods for meta-analysis*. San Diego: Academic Press.

Kiplinger, V. L. & Linn, R. L. (1996). Raising the stakes of test administration: The impact on student performance on the National Assessment of Educational Progress. *Educational Assessment*, 3(2), 111-133.

Mead, A. D & Drasgow, F. (1993) Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta- analysis. *Psychological Bulletin*, 114(3), 449-458.

Morocco, C. C. & Neuman, S. B. (1986). Word processors and the acquisition of writing strategies. *Journal of Learning Disabilities* 19(4), 243-248.

Mourant, R. R, Lakshmanan, R. & Chantadisai, R. (1981). Visual Fatigue and Cathode Ray Rube Display Terminals. *Human Factors*, 23(5), 529-540.

O'Neil, H. F. Jr., Sugrue, B. & Baker, E. L. (1996). Effects of motivational interventions on the National Assessment of Educational Progress mathematics performance. *Educational Assessment*, 3(2), 135-157.

Rosenthal, R. (1994) Parametric measures of effect size. In Cooper, H. & Hedges, L. *The handbook of research synthesis*. NY: Russell SAGE, pp. 231-244

Rosenthal, R. & Rubin, D. B. (1982) A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology*, 74, 166-169.

Snyder, T. D. & Hoffman, C. M. (1990). Digest of Education Statistics. Washington, DC: U. S. Department of Education.

Snyder, T. D. & Hoffman, C. M. (1993). Digest of Education Statistics. Washington, DC: U. S. Department of Education.

Snyder, T. D. & Hoffman, C. M. (1994). Digest of Education Statistics. Washington, DC: U. S. Department of Education.

Snyder, T. D. & Hoffman, C. M. (1995). Digest of Education Statistics. Washington, DC: U. S. Department of Education.

Wolf, F. (1986) *Meta-analysis: Quantitative methods for research synthesis*. SAGE University series on quantitative applications in the social sciences, series no. 07-059. Newbury Park, CA: SAGE.

## About the Authors

**Michael Russell**

russelmh@bc.edu

Boston College
School of Education
Center for the Study of Testing, Evaluation and Educational Policy
323 Campion Hall
Chestnut Hill, MA 02167
Ph. 617/552-4521
Fax 617-552-8419

**Walt Haney**

haney@bc.edu
Boston College
School of Education
Center for the Study of Testing, Evaluation and Educational Policy
323 Campion Hall
Chestnut Hill, MA 02167

---



The World Wide Web address for the *Education Policy Analysis Archives* is
http://olam.ed.asu.edu/epaa

General questions about appropriateness of topics or particular articles may be addressed to the Editor, Gene V Glass, glass@asu.edu or reach him at College of Education, Arizona State University, Tempe, AZ 85287-2411. (602-965-2692). The Book Review Editor is Walter E. Shepherd: shepherd@asu.edu . The Commentary Editor is Casey D. Cobb: casey@olam.ed.asu.edu .

## EPAA Editorial Board