J·T·L·A

The Journal of Technology, Learning, and Assessment

Volume 5, Number 2 · November 2006

Does it Matter if I Take My Writing Test on Computer? An Empirical Study of Mode Effects in NAEP

Nancy Horkay, Randy Elliot Bennett,
Nancy Allen, Bruce Kaplan, & Fred Yan

www.jtla.org

A publication of the Technology and Assessment Study Collaborative
Caroline A. & Peter S. Lynch School of Education, Boston College

J·T·L·A

Volume 5, Number 2

Does it Matter if I take My Writing Test on Computer? An Empirical Study of Mode Effects in NAEP

Nancy Horkay, Randy Elliot Bennett, Nancy Allen, Bruce Kaplan, & Fred Yan

Editor: Michael Russell
russelmh@bc.edu
Technology and Assessment Study Collaborative
Lynch School of Education, Boston College
Chestnut Hill, MA 02467

Copy Editor: Kevon R. Tucker-Seeley
Design: Thomas Hoffmann
Layout: Aimee Levy

JTLA is a free on-line journal, published by the Technology and Assessment Study Collaborative, Caroline A. & Peter S. Lynch School of Education, Boston College.

Preferred citation:

Horkay, N., Bennett, R. E., Allen, N., Kaplan, B., & Yan, F. (2006). Does it Matter if I Take My Writing Test on Computer? An Empirical Study of Mode Effects in NAEP. *Journal of Technology, Learning, and Assessment, 5*(2). Retrieved [date] from http://www.jtla.org.

J·T·L·A

Abstract:

This study investigated the comparability of scores for paper and computer versions of a writing test administered to eighth grade students. Two essay prompts were given on paper to a nationally representative sample as part of the 2002 main NAEP writing assessment. The same two essay prompts were subsequently administered on computer to a second sample also selected to be nationally representative. Analyses looked at overall differences in performance between the delivery modes, interactions of delivery mode with group membership, differences in performance between those taking the computer test on different types of equipment (i.e., school machines vs. NAEP-supplied laptops), and whether computer familiarity was associated with online writing test performance. Results generally showed no significant mean score differences between paper and computer delivery. However, computer familiarity significantly predicted online writing test performance after controlling for paper writing skill. These results suggest that, for any given individual, a computer-based writing assessment may produce different results than a paper one, depending upon that individual's level of computer familiarity. Further, for purposes of estimating population performance, as long as substantial numbers of students write better on computer than on paper (or better on paper than on computer), conducting a writing assessment in either mode alone may underestimate the performance that would have resulted if students had been tested using the mode in which they wrote best.

J·T·L·A

Does it Matter if I Take My Writing Test on Computer? An Empirical Study of Mode Effects in NAEP

Nancy Horkay
Randy Elliot Bennett
Nancy Allen
Bruce Kaplan
Fred Yan
Educational Testing Service

Introduction

Over the past several years, numerous states have begun offering components of their K–12 assessment programs on computer. In the 2005–2006 school year, for example, 22 states were reported to offer some type of online assessment ("Computer-based testing," 2006). In some state testing programs, online assessment is already well-established: Oregon is reported to have administered over one million tests online in the 2004–2005 school year ("State: Online testing helped raise scores," 2005), and Virginia to have given over 650,000 examinations during its Spring 2005 Standards-of-Learning testing window (Virginia Department of Education, undated). In both these instances, and in most other state online testing programs, multiple-choice items are exclusively used because the test delivery software for their presentation is more evolved than that for constructed-response delivery and because of concern that some groups of students would be unfairly disadvantaged by having to answer constructed-response questions on computer.

Anticipating the movement of state assessments to computer, the National Center for Education Statistics funded three field studies to explore the implications of electronic delivery for the National Assessment of Educational Progress (NAEP). NAEP differs from most state testing programs in that it is a sample survey that does not report individual student scores and in that it relies heavily on the use of constructed-response items. Studies were carried out in mathematics, writing, and problem solving in technology-rich environments. This paper reports selected results from the Writing Online (WOL) study, in particular, those results pertaining to the comparability and fairness of scores.

Four main questions are addressed:

- Do students perform differently on computer-based versus paper-based writing assessments?
- Does test mode differentially affect the performance of NAEP reporting groups (e.g., those categorized by gender or by race/ethnicity)?
- Does performance vary as a function of the type of computer used to take the test (i.e., school computers vs. NAEP-supplied laptops)?
- Do students who are relatively unfamiliar with computers perform differently from students who are more familiar with them?

With respect to the first question, very few studies of the effect of delivery mode on writing test performance have been conducted at the K–12 level. Moreover, the studies that are available generally use small, non-representative samples. Even so, the results suggest that mode does have an impact on test score. For example, in two studies, Russell (Russell & Haney, 1997; Russell & Plati, 2001) found that middle school students who took an essay test on computer not only wrote longer essays but also performed better than a randomly assigned group taking the same test on paper. This performance advantage persisted even after controlling for score on a broad test of academic skills in one case and for English mid-year course grades in the other. A similar effect for increased essay length was detected by Wolfe, Bolton, Feltovich, and Niday (1996) for secondary school students, each of whom wrote one essay on computer and one with paper and pencil. Finally, MacCann, Eastment, and Pickering (2002) found that students randomly assigned to test on computer received higher scores than those taking the same test on paper for either one or two of three essays, depending upon whether the essays were graded in their original forms or transcribed to the other form before being graded.

Two studies with older students taking postsecondary admissions tests also show evidence of overall mode effects. For TOEFL® (Test of English as a Foreign Language) examinees given a choice of administration mode, Wolfe and Manalo (2004) found scores to be marginally higher on paper versus computer forms of that test's essay section, after controlling for English-language proficiency. Similarly, in a large group of business-school applicants who wrote GMAT (Graduate Management Admission Test) essays in each mode, students performed better on the paper than on the computer versions of the test (Bridgeman & Cooper, 1998).

As for overall mode differences, only a few studies have investigated the performance of population groups on computer compared to paper writing tests. In a small-sample study, Russell and Haney (1997) found that the differences in performance on computer versus paper writing tests were similar for middle-school boys and girls. Among a large sample of prospective business school students, Bridgeman and Cooper (1998) found no interactions between delivery mode and population groups by gender, race/ethnicity, or whether English was the first language.

The current study's third question dealt with performance on NAEP-supplied laptops vs. on the school computers with which students routinely worked. At the time this study was conducted, school instructional technology inventories were composed almost entirely of desktop machines. The research literature on the comparability of scores between laptop and desktop computers is almost non-existent. One study, conducted by Powers and Potenza (1996), assessed the performance of 199 first-year graduate students and upper-division undergraduates. Each participant took two parallel verbal and quantitative test forms, one on desktop and one on laptop, with order of administration of the computing platforms and the test forms counterbalanced across participants. Each form contained one essay. Results showed a mode-by-order interaction, with study participants who wrote first on desktop and then on laptop performing less well by a small amount on their second essay (taken on laptop) than on their first essay (taken on desktop). Those who took the test on laptop first showed no difference in performance between essays.

Finally, does familiarity with computers affect online writing test performance in unwanted ways? Several studies have looked at the relationship of computer familiarity to writing test performance. Although the results are not entirely consistent, they suggest that computer and paper writing tests may not measure the same skill for all students. For example, Wolfe, Bolton, Feltovich, and Bangert (1996) and Wolfe, Bolton, Feltovich, and Niday (1996) found that secondary school students with less experience writing on computer were disadvantaged by having to test that way. In the first study, tenth-grade students with little or no experience using computers outside of school scored higher on pen-and-paper essays than on computer-written ones, whereas students with a lot of computer experience showed no difference in performance across modes. In the second study, less experienced students achieved lower scores, wrote fewer words, and wrote more simple sentences when tested on computer than when they tested on paper. Students with more experience writing on computer achieved similar scores in both modes, but wrote fewer words and more simple sentences on paper than on computer. Another research team led by Russell (1999) found that, after controlling for reading performance, middle-school students with low keyboarding speed were disadvantaged by a computer-writing test relative to students with similar low levels of keyboarding skill taking a paper test. The opposite effect was detected for students with high keyboarding speed, who fared better on the computer than on paper examinations. In a subsequent investigation, however, Russell and Plati (2001) found that eighth-and tenth-grade students performed better on the computer-writing test regardless of whether their keyboarding speed was high or low.

Method

Participants

The WOL study samples were selected to comprise nationally representative groups of eighth-grade students. Samples were drawn from the main NAEP 2002 assessments, which were administered between the end of January and the beginning of March 2002. The group taking the WOL computer test consisted of two subsamples tested from the beginning of April through the end of May 2002, following the conclusion of the main NAEP assessments. One subsample of 715 students was drawn from the main NAEP 2002 *writing* assessment. This subsample was selected from among students who had been administered any one of 10 predetermined main NAEP writing test books. These books included informative or persuasive prompts other than the ones used in WOL. The second subsample taking the WOL computer test consisted of 593 students from the main NAEP 2002 *reading* assessment who had taken any one of nine predetermined reading books. Since these students did not participate in the main NAEP writing assessment, their performance was used to help determine if taking the main NAEP writing assessment prior to WOL affected the WOL score in any way. The performance of the main NAEP writing and reading students taking WOL was compared to a third group of 2,983 students who, as part of the 2002 main NAEP writing assessment, were administered the same two essay tasks on paper in the same order as presented in WOL.

Students were sampled for taking the WOL computer test in the following way.[1] From the 5,368 schools sampled for the main NAEP 2002 writing and reading assessments, 236 schools were randomly selected. One hundred and fifty-eight of these schools participated. The weighted school response rate, which reflects the accumulated effect of main NAEP *and* WOL study attrition, is 67 percent. Within the 158 participating schools, 1,859 students were identified as eligible for WOL by reason of their having been assigned one of the 19 targeted writing or reading assessment booklets during the main NAEP 2002 assessment.

Of those 1,859 students, 1,313 participated in WOL. Reasons for non-participation include absence from the main NAEP or WOL administrations (344), withdrawal from school or otherwise ineligible (85), exclusion because of disability or limited English proficiency (65), failing to complete the WOL test (29), or participating in an associated WOL substudy not reported here (23). In addition to these nonparticipating students, five individuals who did participate were not included in the analysis because they were incorrectly classified as not taking part in main NAEP. After accounting for non-participants and these five misclassified individuals, the weighted student response rate reflecting both main NAEP and WOL attrition is 77 percent.

For most of the analyses conducted, data were used only from those students who responded to both essay tasks, regardless of whether those tasks were taken on paper or on computer. This restriction was imposed because it allowed for a more powerful statistical test, repeated-measures analysis of variance (ANOVA), to be used in the investigation of mode effects. In addition, this technique permitted testing relevant interactions with essay, including the interaction of essay and delivery mode, and of essay, delivery mode, and population group. If shown to be significant statistically, such interactions imply that delivery mode may not be consistent in its effects across essays. After eliminating those who only responded to one essay, a very high percentage of participating students – more than 95 percent – was retained in each of the samples.

Table 1 (next page) shows the characteristics of the final study samples, including both the total WOL student sample and the two subsamples that comprise it.

Table 1: Characteristics of Study Samples

		WOL Students		
Characteristic	Main NAEP writing students responding to both paper-and-pencil essays in the same order as WOL	Total sample of students responding to both essays on computer	Subsample of students drawn from main NAEP writing and responding to both essays on computer	Subsample of students drawn from main NAEP reading and responding to both essays on computer
Number of students	2,878	1,255	687	568
NAEP writing mean	156	—	157	—
	Percent of Students			
Exclusion rate	3	5	5	4
Gender				
Male	45	52	52	51
Female	54	47	47	48
Race/ethnicity				
White	65	69	69	69
Black	16	14	15	14
Hispanic	15	12	11	13
Asian/Pacific	4	3	4	2
Islander/Other	1	2	2	2
Type of school				
Public	90	92	92	91
Nonpublic	10	8	8	9
Parents' highest education level				
Less than HS	6	5	6	4
Graduated HS	17	15	17	13
Some education after HS	19	21	20	24
Graduated college	46	49	48	50
Unavailable	13	10	10	10
Student eligibility for free/reduced price school lunch				
Eligible	30	28	28	29
Not eligible	54	58	58	57
Unavailable	15	14	14	14
School location				
Central city	28	28	28	27
Urban fringe/ large town	43	38	38	39
Rural/ small town	29	34	34	35

The study samples shown in Table 1 diverge in relatively minor ways from the main NAEP samples from which they were drawn. (See Horkay, Bennett, Allen, & Kaplan, 2005, for complete details.) For the study sample taking the essays on paper, the largest divergence from main NAEP was in having a significantly higher percentage of female students (by 4 points) and a lower percentage of male students (by 5 points). For the WOL sample drawn from the 2002 main NAEP writing assessment, the largest divergence from main NAEP was in having a greater percentage of White students (by 4 points) and of rural students (by 5 points). Finally, the largest divergences for the WOL sample drawn from main NAEP reading were in having more White students (by 4 percentage points), more students with one or more parents having some education after high school (by 5 percentage points), more rural students (by 6 percentage points), and fewer students whose parents' highest education level was graduation from high school (by 4 percentage points).

Not surprisingly, the study samples taking the test on paper and those taking it on computer also differed in relatively minor ways from one another (Table 1, page 9). To deal with these differences, the first question of whether delivery mode affects test performance was analyzed with each demographic characteristic included in turn as an independent variable to control for its effects. Similarly, many of the analyses conducted to address the other study questions were run with gender as one of the independent variables, as the largest demographic difference between the paper and computer samples appeared to be in the distribution of this characteristic.

Instruments

As noted, all sampled students participated in one of two main NAEP paper-and-pencil assessments, each of which was completed in a single session. During these proctored sessions, students responded to a booklet of questions from either a main NAEP reading test or writing test, and to a background questionnaire.

At least three weeks after the 2002 main NAEP tests were administered, NAEP field staff returned to a subset of schools to test students sampled for the Writing Online (WOL) study. Thirty-five percent of the participating students completed the study instruments via the Internet on school computers. The remainder, 65 percent, were administered the instruments on NAEP laptops brought into schools. Regardless of computer type, all sessions were proctored by NAEP field staff.

Each student took the following WOL components in a single session:

Online tutorial

The online tutorial showed how to use the computer to respond to the essay tasks. The tutorial provided instruction and practice in the use of the mouse and scrolling, presented information about the test interface and how to navigate from one question to the next, and described the functions of the WOL word processor. Students were given two minutes to practice typing and to try out the word processing tools. A portion of the WOL tutorial can be viewed at http://nces.ed.gov/nationsreportcard/studies/tbatutorial.asp#wol.

Online computer-skills measure

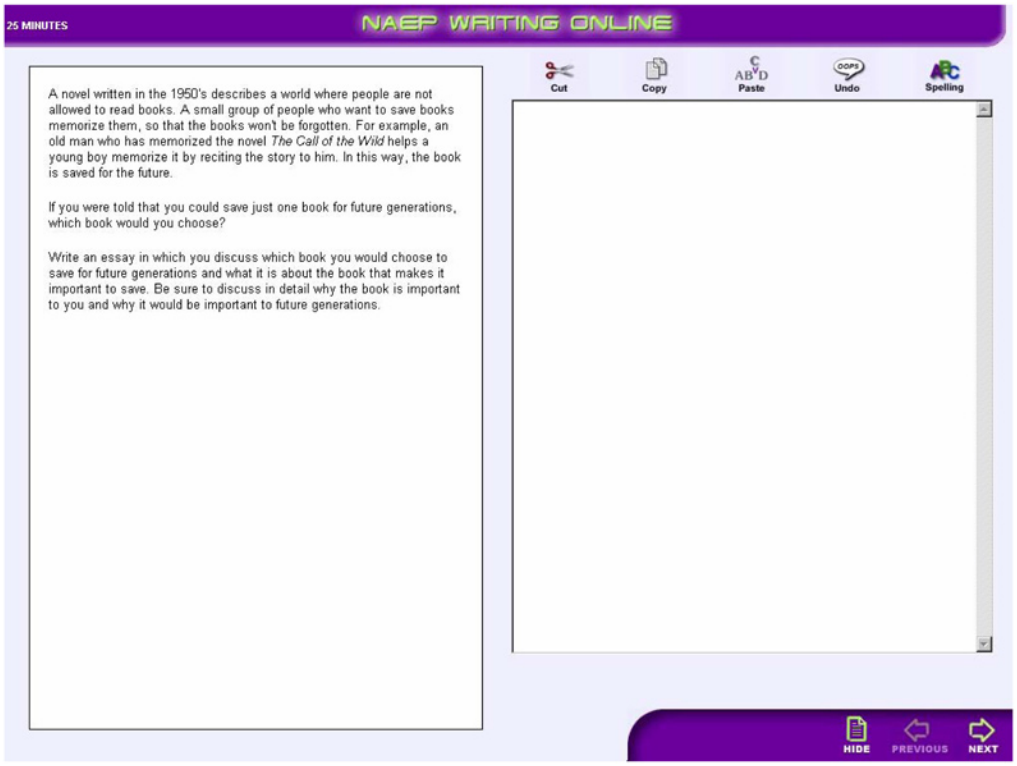
The computer-skills measure was administered to evaluate each student's facility with the computer and, specifically, word processing proficiency. The computer-skills measure presented a series of five exercises that asked students to type, insert, delete, correct, and move text. Students were also asked to type a paragraph exactly as it was shown on the screen. They were given two minutes to type the text as accurately as possible.

Two online essay tasks

As in the main NAEP writing assessment, each student was first given a brochure entitled, "Ideas for planning and reviewing your writing." Students could refer to the brochure at any point during the test, but they were specifically instructed to look at it prior to writing their responses.

Students were next shown general directions on the computer. Then they proceeded to the first WOL writing task, "Save a Book." The task was displayed on the left side of the screen, and students typed their responses in a field on the right side. The text entry area included word processing tools represented as icons on the tool bar at the top of the screen. These tools allowed students to cut, copy, and paste text; undo their last action; and check spelling. Some of these functions were also accessible through standard control-key combinations. Figure 1 (next page) shows the WOL computer interface and the first essay task.

Figure 1: The Writing Online Computer Interface



A novel written in the 1950's describes a world where people are not allowed to read books. A small group of people who want to save books memorize them, so that the books won't be forgotten. For example, an old man who has memorized the novel *The Call of the Wild* helps a young boy memorize it by reciting the story to him. In this way, the book is saved for the future.

If you were told that you could save just one book for future generations, which book would you choose?

Write an essay in which you discuss which book you would choose to save for future generations and what it is about the book that makes it important to save. Be sure to discuss in detail why the book is important to you and why it would be important to future generations.

Note: The "Save a Book" essay prompt is visible on the left and the response area is shown on the right.

Students were allowed 25 minutes for each essay task. Timing began as soon as the first task was displayed, which was consistent with the manner in which the main NAEP paper-and-pencil writing assessment was administered. If a student completed the first essay before 25 minutes elapsed, that student was able to move on to the second essay, "School Schedule." The timer then automatically reset to 25 minutes, regardless of the time used in the first essay. Students were not allowed to return to the first essay once they had moved on to the second essay. This procedure also was followed to maintain comparability with that used for the main NAEP paper-and-pencil writing administration.

Both WOL essays were drawn from the 2002 main NAEP writing assessment and administered to students in the same order as in that assessment. For "Save a Book," an informative writing task, students were asked to explain what book they would preserve through memorization if they lived in a society where reading was not allowed. Since any book could be chosen, a wide range of responses was acceptable. "School Schedule," a persuasive writing task, required students to read a short newspaper article about the sleeping habits of adults and children, and to show how those habits ought to influence school schedules. Students were able to react to the article and use the contents to frame their arguments on the topic.

Online background questions

Students were allowed 20 minutes to complete the background questions section, which consisted of 37 questions: 10 main NAEP general background questions (including race/ethnicity, parents' education level, and literacy materials available in the home), 21 questions about students' experience with computers, and 6 questions about students' instruction in writing. Background questions appeared on the screen, and students were directed to click on the bubble next to their selected response.

The following components were administered to students who took the main NAEP 2002 paper writing assessment but did not take WOL.

Two essay tasks

Each student was given a brochure entitled "Ideas for planning and reviewing your writing," which was the same brochure as used by the WOL students. Students then responded to the same two 25-minute essay questions in the same order as presented on the WOL test. If students finished before 25 minutes elapsed, they were not allowed to move ahead, but they could check over their work on that section.

Background questions

Students responded to 53 background questions, which were designed to gather information about student demographics and students' classroom writing instruction and writing experience. (Some of these background questions were also administered in WOL.)

Table 2 (next page) summarizes the instruments used in the WOL study, the two treatment conditions (paper and online), and the samples used in most study analyses.

Table 2: Instruments and Conditions

Paper Condition (N=2,878)	*Online Condition (N=1,255)*
2002 Main NAEP Assessment (January – March)	
• Paper writing test with two essays	• *Either* a paper writing test (with two essays different from the "paper condition" of N=2,878) *or* a paper reading test
• Paper background questionnaire	• Paper background questionnaire
2002 WOL Study (April – May)	
—	• Online tutorial and computer facility measure
—	• Online writing test (with the same two essays as the "paper condition" of N=2,878)
—	• Online background questionnaire

Note: The paper condition (N=2,878) consisted of different students from the online condition (N=1,255).

Procedures

Essay scoring

For the group taking the main NAEP 2002 paper writing assessment, scores for each essay were taken from data files produced as part of that assessment. (See appendix A for scoring rubrics.) In main NAEP scoring, readers grade on computer the scanned versions of students' handwritten responses. For the group taking WOL, a separate scoring session was held in which readers graded on computer students' typed responses. This WOL scoring session employed the training procedures and sample response papers used for scoring the same two essays in main NAEP. In the WOL scoring session, each of the two essays was scored by a different group of readers, which is consistent with main NAEP scoring procedures.

To evaluate reader reliability, a random sample of WOL responses was double-scored and compared to the double-scored responses of a different group of students who had taken the same two essays on paper in main NAEP. The intra-class correlations between two readers for "Save a Book" were .81 and .87 for the WOL responses and the main NAEP writing responses respectively. For "School Schedule," the comparable values were .88 for WOL and .94 in main NAEP. These results indicate that, for those responses that were double-scored, the WOL readers agreed with one another in rank ordering individuals to a slightly lesser degree than did the main NAEP readers. The discrepancy between the rater reliabilities for WOL compared with main NAEP may be due to several factors, including differences in reader groups, scoring procedures, or the modes of on-screen presentation (scanned handwritten paper images vs. typed responses).

Differences in reader agreement can impact study results to the extent that this lower agreement negatively affects the overall reliability of scores. Estimates of score reliability that incorporate reader agreement as an error component can, therefore, be helpful in evaluating this impact. Such score reliabilities can be estimated for the WOL test and the main NAEP assessment using the product-moment correlation between the two essay responses within each study group (corrected for the fact that this correlation reflects a half-length test). This correlation incorporates reader agreement as an error component because student responses in both main NAEP and WOL were assigned randomly to readers, so most students' first and second essays would have been rated by different individuals. For WOL, the corrected correlation based on the study sample of 1,255 was .77. For main NAEP, the corrected correlation based on the study sample of 2,878 was .73.[2] Thus, despite differences in reader reliability, the score reliabilities across the two samples were reasonably close to one another.

Reader scoring consistencies between modes

In main NAEP, students handwrote their essay responses, whereas in WOL students typed their responses. Several studies have found that readers award different scores to typed essays as compared with handwritten versions of the same essays. In some studies, readers have given lower scores to the typed versions (Powers & Farnum, 1997; Powers, Fowles, Farnum, & Ramsey, 1994; Russell & Tao, 2004a; Russell & Tao, 2004b), though other studies have reported either mixed or null results (Harrington, Shermis, & Rollins, 2000; MacCann, Eastment, & Pickering, 2002). To evaluate whether there was such a bias in this study, a sample of handwritten responses drawn from the main NAEP 2002 writing assessment was used. Responses were drawn separately for each essay from all students administered that question, since it was not practical to identify in NAEP files which students had been administered, and had responded to, both questions.[3] The selected responses were next keyed into the WOL online scoring system by operators instructed to transcribe the essays as faithfully as possible, including the reproduction of student errors. These transcribed responses were then rated during the WOL scoring session by randomly interspersing them with WOL responses, appearing to readers on-screen exactly as did WOL responses that had been entered by students online.

The scores for these responses were compared using a repeated-measures analysis of variance, with essay and presentation format (i.e., handwritten vs. typed) as the independent variables, essay score as the dependent variable, and repeated measures on the presentation format factor. Results showed no significant main effect for presentation format but a significant format-by-essay interaction (F,1,584=10.97, p<.05). Post-hoc tests showed that the typed transcriptions were scored lower than the handwritten originals of the same response for one essay but higher than the handwritten originals for the other essay. In both cases, the effect sizes in standard deviation units of the handwritten group were so small as to be of little practical concern (d=.05 and .07).

Practice effect

Two student samples took WOL. One sample had previously taken a NAEP writing assessment and one sample had not previously taken such an assessment. To determine whether having taken the main NAEP writing assessment affected subsequent WOL performance, the mean scores of the WOL students drawn from the main NAEP writing sample were compared to the mean scores for WOL students drawn from the main NAEP reading sample. Results showed no between-subjects main effect for WOL group (F,1,62=3.50, p>.05) and no significant interaction of WOL group with essay (F,1,62=0.01, p>.05). Because no significant difference was found between the groups, they were combined where appropriate for the analyses subsequently presented in this paper.

Results

Performance Differences Across Assessment Modes

Do eighth-grade students nationally perform differently on computer vs. paper writing tests? To address this question, three indicators were compared across delivery mode: essay score, essay length, and the frequency of valid responses.

Essay score

In this analysis, the mean scores for students taking WOL were compared with the mean scores from a different, but also reasonably representative, group of students taking the same essays in the paper-and-pencil main NAEP writing assessment. Table 3 (next page) gives the mean scores and standard deviations for each group on each essay, where scores are on a scale of 1 to 6. These statistics, as most other results presented in this paper, are weighted to estimate parameters for the population of 8th graders nationally.

Table 3: Mean scores for students responding to Writing Online (WOL) and for a different group of students responding to the same essays on paper in the main NAEP writing assessment

Essay	WOL (N=1,255)	Paper and Pencil (N=2,878)
Save a Book	3.5 (1.3)	3.6 (1.0)
School Schedule	3.5 (1.2)	3.6 (1.1)

Note: Standard deviations are in parentheses.

To test the difference between means, a repeated-measures analysis of variance (ANOVA) was conducted. Delivery mode and essay were the independent variables, and essay score was the dependent variable, with repeated measures on the essay factor. The results of this analysis did not detect a significant effect for delivery mode (F,1,62=3.39, p>.05), nor a significant interaction of delivery mode with essay (F,1,62=0.29, p>.05). (Examination of the Table 3 standard deviations does, however, suggest that the computer scores are more variable than the paper ones.) This model was run again accounting separately for gender, race/ethnicity, parents' education level, school location, eligibility for free/reduced-price school lunch, and school type in order to control for the effects of differences in the representation of these groups between the paper and computer samples. The results showed no significant mean differences for delivery mode, or for the interaction of delivery mode with essay.

Essay length

A second indicator of mode effect is essay length, which can be automatically computed once responses are in electronic form (which they already were for WOL). From the paper main NAEP writing assessment, a sample of handwritten responses had previously been transcribed to electronic form for evaluating the comparability of scoring typed vs. handwritten responses. This sample had score means closely similar to – but score variation noticeably greater than – that found for the paper main NAEP writing assessment sample responding to both essays. To adjust for this difference in variance, the transcribed paper responses were weighted to reproduce the variation found in the observed score distributions of the paper main NAEP writing assessment sample. These weights were then used in computing the mean and standard deviation of the word counts for these same transcribed handwritten responses.

Table 4 gives the mean word count and the standard deviations for each essay by delivery mode.

Table 4: Mean word count for students responding to Writing Online and for different students responding to the same essays on paper in the main NAEP writing assessment

Essay	WOL	Paper and Pencil
Save a Book	185 (103)	176 (81)
School Schedule	162 (91)	153 (72)

Note: The number of responses for "Save a Book" was 294 for paper main NAEP writing and 1,255 for WOL. The number of responses for "School Schedule" was 292 for paper main NAEP writing and 1,255 for WOL. Standard deviations are in parentheses. Word counts are not weighted to be representative of the performance of 8th graders nationally.

To test the effect of delivery on essay length, a separate ANOVA was conducted for each essay, with delivery mode the independent variable and the number of words serving as the dependent variable. Results showed that there was no effect of delivery mode on word count for "Save a Book" (F,1,815=1.93, p>.05) or for "School Schedule" (F,1,635=2.54, p>.05). When these analyses were repeated adding gender to the model, a significant mode effect was detected, but only for "School Schedule" (F,1,632=4.27, p<.05). For this essay, students taking the test on computer wrote marginally longer responses than those who took it on paper (a difference of about 9 words, or .13 standard deviations in the units of the paper group). Finally, the interaction of mode with gender was not significant for either essay. When this interaction term was subsequently removed from the ANOVA model, the mode effect for "School Schedule" remained significant.

Frequency of valid responses

A third indicator of the impact of delivery mode is the extent to which students provide valid responses to test questions. It is conceivable that response rates will be lower on computer because students with limited computer facility may fail to respond if taking an online test becomes frustrating. On the other hand, response rates could be higher for WOL if students who frequently use computers at home and school find online tests more motivating than paper examinations. Table 5 (next page) shows the percentage of students responding to each essay, where non-response included off-task, not reached, illegible, omitted, or any other missing answer.

Table 5: Percentage of students giving valid responses to Writing Online and for a different group of students responding to the same essays on paper in the main NAEP writing assessment

Essay	WOL (N=1,308)	Paper and Pencil (N=2,983)
Save a Book	98	98
School Schedule	97	99

To test statistically for differences in responding, separate logistic regressions were estimated for each essay with delivery mode as the independent variable and the dependent variable being whether or not there was a response to the essay. Results for "Save a Book" showed no significant effect for delivery mode (*F*,1,62=0.67, *p*>.05). For "School Schedule," however, delivery mode did significantly predict response rate (*F*,1,62=10.88, *p*<.05), with those taking the paper test more likely to respond to this essay than those taking WOL by about 1 percentage point. These analyses were repeated with gender as an independent variable to control for its effects. The same substantive results were obtained.

Population Group Performance

Score comparisons across delivery modes were conducted separately for gender, race/ethnicity, parents' education level, school location, eligibility for free/reduced-price school lunch (an indicator of socioeconomic status), and school type (public vs. nonpublic). Because the sample sizes for some of these groups were small, seemingly large differences may not always be statistically significant. It is not possible to distinguish for these instances whether the apparent difference is a true reflection of population performance, or alternatively, an artifact of sample selection.

Population group comparisons are reported here only for essay score. (See Horkay, Bennett, Allen, & Kaplan, 2005, for comparisons of other variables.) For each comparison, a repeated-measures ANOVA was conducted. For this analysis, the independent variables were the NAEP reporting group of interest, delivery mode, gender (if not already the reporting group of interest), and essay, with repeated measures on the essay factor. Essay score was the dependent variable. Gender was included as an independent variable even when it wasn't the reporting group of interest to control for differences between the WOL and the main NAEP writing assessment samples, which were largest on this demographic characteristic. Also included was the interaction of NAEP reporting group with delivery mode, as such an interaction would indicate that the difference in scores between modes was not the same for all categories composing a particular reporting group (e.g., for all of the parent education levels).

Gender

Table 6 presents mean scores and standard deviations for WOL and for the paper main NAEP writing assessment by gender.

Table 6: Mean scores by gender for students drawn from main NAEP who took the Writing Online test and for a different group of students responding to the same essays on paper in the main NAEP writing assessment

Essay	WOL (N=1,249)	Paper and Pencil (N=2,867)
Save a Book		
Male	3.3 (1.2)	3.4 (1.0)
Female	3.8 (1.2)	3.8 (1.1)
School Schedule		
Male	3.3 (1.3)	3.3 (1.0)
Female	3.7 (1.2)	3.8 (1.1)

Note: Students without gender designations are omitted.
Standard deviations are in parentheses.

The between-groups ANOVA results showed no effect for delivery mode (*F*,1,62=1.23, *p*>.05) as reported earlier; an expected significant main effect for gender (*F*,1,62=80.12, *p*<.05); and no significant interaction of delivery mode with gender (*F*,1,62=0.05, *p*>.05). The within-groups results showed no significant interaction of delivery mode with essay (*F*,1,62=0.73, *p*>.05), of gender with essay (*F*,1,62=1.62, *p*>.05), or of delivery mode, gender, and essay (*F*,1,62=0.35, *p*>.05). With respect to essay score, then, delivery mode does not appear to have affected either gender group more than the other.

Race/ethnicity

Table 7 gives the mean scores and standard deviations by race/ethnicity.

Table 7: Mean scores by race/ethnicity for students drawn from main NAEP who took the Writing Online test and for a different group of students responding to the same essays on paper in the main NAEP writing assessment

Essay	WOL (N=1,255)	Paper and Pencil (N=2,878))
Save a Book		
White	3.7 (1.2)	3.8 (1.0)
Black	2.9 (1.1)	3.3 (0.9)
Hispanic	3.0 (1.4)	3.2 (1.3)
Asian/Pacific Islander	3.8 (1.3)	4.0 (1.1)
Other	3.3 (1.0)	3.4 (0.9)
School Schedule		
White	3.7 (1.2)	3.7 (1.0)
Black	2.8 (1.1)	3.2 (1.1)
Hispanic	2.9 (1.2)	3.1 (1.4)
Asian/Pacific Islander	3.8 (1.3)	4.1 (1.0)
Other	3.4 (1.0)	3.4 (0.8)

Note: "Other" includes American Indian/Alaskan Native and unclassified students.
Standard deviations are in parentheses.

Because gender was included in the model and 17 students were missing gender designations, the statistical test was conducted on a slightly smaller number of students than the one used to compute the means in Table 7. Results of the ANOVA showed a significant between-groups effect for race (*F*,4,59=51.66, *p*<.05) and for gender (*F*,1,62=72.63, *p*<.05). There was no significant effect for delivery mode (*F*,1,62=1.52, *p*>.05) and no significant interaction of delivery mode with race/ethnicity (*F*,4,59=1.46, *p*>.05). The within-groups results showed no significant interaction of essay with race (*F*,4,59=1.47, *p*>.05), essay with delivery mode (*F*,1,62=0.04, *p*>.05), essay with gender (*F*,1,62=0.34, *p*>.05), or essay, delivery mode, and race/ethnicity (*F*,4,59=0.19, *p*>.05).

Parents' education level

Table 8 gives the mean scores and standard deviations by parents' education level, where that level is the higher of the levels reported by the student for his or her mother or father.

Table 8: Mean scores by parents' highest level of education for students drawn from main NAEP who took the Writing Online test and for a different group of students responding to the same essays on paper in the main NAEP writing assessment

Essay	WOL (N=1,255)	Paper and Pencil (N=2,878))
Save a Book		
High school degree or less	3.3 (1.2)	3.3 (0.9)
More than high school degree	3.6 (1.3)	3.9 (1.0)
Unavailable	3.1 (1.3)	3.0 (1.0)
School Schedule		
High school degree or less	3.2 (1.1)	3.2 (0.9)
More than high school degree	3.6 (1.3)	3.8 (1.0)
Unavailable	3.0 (1.2)	2.8 (1.2)

Note: "High school degree or less" includes students reporting parents who did not finish high school or who obtained high school degrees. "More than high school degree" includes students reporting one or more parents having some education after high school or who graduated from college. "Unavailable" includes students with missing data for this variable. Standard deviations are in parentheses.

Differences between the means were tested for the slightly smaller subset of students with gender designations (total N=4,116) than shown in Table 8. The between-groups results showed expected significant effects for parents' education level (*F*,2,61=105.83, *p*<.05) and gender (*F*,1,62=47.34, *p*<.05). There were no significant effects for delivery mode (*F*,1,62=0.02, *p*>.05) or for the interaction of delivery mode with parents' education level (*F*,2,61=2.71, *p*>.05). The within-groups results showed no significant interaction of essay with parents' education level (*F*,2,61=1.21, *p*>.05), essay with delivery mode (*F*,1,62=0.27, *p*>.05), essay with gender (*F*,1,62=0.35, *p*>.05), or essay, delivery mode, and parents' education level (*F*,2,61=0.64, *p*>.05).

School location

Table 9 gives the mean scores and standard deviations by type of school location.

Table 9: Mean scores by school location for students drawn from main NAEP who took the Writing Online test and for a different group of students responding to the same essays on paper in the main NAEP writing assessment

Essay	WOL (N=1,255)	Paper and Pencil (N=2,878))
Save a Book		
Central city	3.3 (1.4)	3.5 (1.1)
Urban fringe/large town	3.6 (1.3)	3.7 (1.2)
Rural/small town	3.7 (1.1)	3.6 (0.8)
School Schedule		
Central city	3.3 (1.4)	3.4 (1.2)
Urban fringe/large town	3.5 (1.2)	3.7 (1.1)
Rural/small town	3.6 (1.2)	3.4 (1.0)

Note: Standard deviations are in parentheses.

Here, too, the statistical tests were computed for the subset of students with gender designations. The between-groups results showed expected significant effects for school location (F,2,61=9.39, p<.05) and gender (F,1,62=44.85, p<.05). There was no significant effect for delivery mode (F,1,62=0.90, p>.05). However, the interaction of delivery mode with school location was significant (F,2,61=3.45, p<.05). The within-groups results showed no significant interaction of essay with school location (F,2,61=1.65, p>.05), essay with delivery mode (F,1,62=1.35, p>.05), essay with gender (F,1,62=0.31, p>.05), or essay, delivery mode, and school location (F,2,61=1.89, p>.05).

Post-hoc tests showed that students from urban fringe/large town locations performed significantly higher on the paper as compared to the computer test (F,1,62=5.05, p<.05).[4] The size of the effect was about .15 in the standard deviation units of the paper group, not even a "small" effect in the classification system proposed by Cohen (1988). No significant differences between modes were apparent for students from central city (F,1,62=1.55, p>.05) or from rural/small town (F,1,62=1.86, p>.05) locations.

Eligibility for free/reduced-price school lunch

Table 10 gives the mean scores and standard deviations by eligibility for free/reduced-price school lunch, an indicator of socio-economic status.

Table 10: Mean scores by student eligibility for free/reduced-price school lunch for students drawn from main NAEP who took the Writing Online test and for a different group of students responding to the same essays on paper in the main NAEP writing assessment

Essay	WOL (N=1,255)	Paper and Pencil (N=2,878))
Save a Book		
Eligible	3.1 (1.2)	3.2 (1.0)
Not eligible	3.8 (1.1)	3.8 (1.0)
Unavailable	3.4 (1.5)	3.9 (1.3)
School Schedule		
Eligible	3.1 (1.2)	3.1 (1.0)
Not eligible	3.7 (1.2)	3.7 (1.0)
Unavailable	3.2 (1.4)	3.9 (1.3)

Note: "Unavailable" includes students with missing data for this variable.
Standard deviations are in parentheses.

As in the other population group analyses, the means were tested omitting those 17 students without gender designations. The between-groups results showed expected significant effects for eligibility for free/reduced-price school lunch (F,2,61=69.26, p<.05) and gender (F,1,62=54.38, p<.05). There was also a significant effect for delivery mode (F,1,62=5.23, p<.05), but no significant interaction of delivery mode with eligibility for free/reduced-price school lunch (F,2,61=2.59, p>.05). The within-groups results showed no significant interaction of essay with eligibility for free/reduced-price school lunch (F,2,61=1.11, p>.05), essay with delivery mode (F,1,62=0.04, p>.05), essay with gender (F,1,62=0.18, p>.05), or essay, delivery mode, and eligibility for free/reduced-price school lunch (F,2,61=0.94, p>.05).

Because the effect for delivery mode was significant in the above model and the interaction of delivery mode and eligibility for free/reduced-price school lunch was not, the model was rerun without the interaction. In this new model, delivery mode was no longer significant (F,1,62=2.22, p>.05).

School type

The mean scores and standard deviations by school type are presented in table 11.

Table 11: Mean scores by school type for students drawn from main NAEP who took the Writing Online test and for a different group of students responding to the same essays on paper in the main NAEP writing assessment

Essay	WOL (N=1,255)	Paper and Pencil (N=2,878)
Save a Book		
Public	3.5 (1.2)	3.6 (1.0)
Nonpublic	3.6 (1.3)	4.0 (1.5)
School Schedule		
Public	3.5 (1.2)	3.5 (1.1)
Nonpublic	3.5 (1.4)	3.9 (1.4)

Note: Standard deviations are in parentheses.

Between-groups results for the subset of students with gender designations (total N=4,116) showed a significant effect for gender (*F*,1,62=44.69, *p*<.05) but no significant effect for school type (*F*,1,62=3.63, *p*>.05). There were no significant effects either for delivery mode (*F*,1,62=2.87, *p*>.05) or for the interaction of delivery mode with school type (*F*,1,62=2.66, *p*>.05). As to the within-groups results, there were no significant interactions of essay with school type (*F*,1,62=0.37, *p*>.05), essay with delivery mode (*F*,1,62=0.02, *p*>.05), essay with gender (*F*,1,62=0.29, *p*>.05), or essay, delivery mode, and school type (*F*,1,62=0.17, *p*>.05).

In sum, the only statistically significant interaction of population group with delivery mode detected was for one category of school location and, for that case, the effect size could be considered less than "small." This finding suggests that, in terms of mean scores, computer delivery does not generally disadvantage NAEP reporting groups.

Performance as a Function of Computer Type

Because a large number of schools did not have the particular equipment, level of Internet connectivity, or Internet software required to administer the WOL test on their own machines, NAEP staff brought laptops into schools to assess approximately 65 percent of the study participants. The laptops used in this study had smaller screens and keyboards, as well as different keyboard layouts, than those found on many school computers, the overwhelming majority of which were desktops in early 2002 when WOL was administered. These differences, combined with the fact that most students would have been more familiar with their school computers than with the NAEP laptops, may have affected writing performance in construct-irrelevant ways. The fact that tests presented on laptop and school computers might not be comparable could pose a problem for NAEP. If the performance differences were large enough, NAEP's population estimates could change simply as a function of the mix of laptops and school computers used in the assessment. Further, this mix would likely change over time as more schools were able to participate in NAEP assessments using their own web-connected machines.[5]

In this study, the assignment of students to computer type was not done at random but instead based on whether school computers and connectivity matched WOL requirements. This nonrandom assignment could have been correlated with school location, school type, or socioeconomic status, among other things, and, thereby, with writing skill. Thus, any comparison of performance between computer types must be interpreted cautiously.

To deal with this fact, two sets of analyses were conducted. The first set was quasi-experimental and was carried out with WOL students drawn from the main NAEP writing assessment. This set incorporated statistical controls to attempt to adjust for preexisting differences between the groups taking the test on different computer types. The second set involved a small experiment in which students from three WOL schools were randomly assigned to computer type. Both sets of analyses are limited. The first set had reasonably sized samples but was not experimental. The second set was experimental but had small, unrepresentative samples. If the two sets agree, however, the results should be more interpretable than results from either analysis alone.

With respect to the first set, Table 12 shows the mean scores and standard deviations for WOL students drawn from the main NAEP writing assessment sample by the type of computer on which the WOL test was taken.

Table 12: Mean scores by computer type for Writing Online students drawn from the main NAEP writing sample

Essay	NAEP laptop (N=431)	Web-connected school computer (N=256)
Save a Book	3.5 (1.2)	3.7 (1.3)
School Schedule	3.5 (1.2)	3.6 (1.3)

Note: Standard deviations are in parentheses.

These means were tested using a repeated-measures ANOVA with computer type (laptop vs. school computer), main NAEP writing performance (as a covariate), and essay as the independent variables, with repeated measures on the essay factor.[6] The dependent variable was essay score. Two students included in the above table were omitted from the analysis because of missing main NAEP data. Results of this analysis indicated that, accounting for main NAEP writing performance, there is no difference between the scores of students taking WOL on laptop vs. school computer (F,1,62=0.56, p>.05) and no interaction of computer type with essay (F,1,62=0.06, p>.05).

While the above analysis found no impact of computer type on WOL writing performance for students generally, it is fair to ask whether computer type affects certain population groups. Table 13 shows the means and standard deviations for students by gender, the only reporting group considered due to sample size limitations.

Table 13: Mean scores, by gender and computer type, for Writing Online students drawn from the main NAEP writing sample

	Male		Female	
Essay	NAEP laptop (N=224)	Web-connected school computer (N=136)	NAEP laptop (N=204)	Web-connected school computer (N=120)
Save a Book	3.4 (1.2)	3.2 (1.2)	3.6 (1.2)	4.1 (1.2)
School Schedule	3.3 (1.2)	3.2 (1.3)	3.7 (1.2)	4.0 (1.2)

Note: Standard deviations are in parentheses.

These means were tested using a repeated-measures ANOVA with computer type (laptop vs. school computer), gender, and essay as the independent variables, and main NAEP writing performance as a covariate. Repeated measures were conducted on the essay factor. The dependent variable was essay score. Four students included in the above table were omitted from the analysis due to missing data. As in the analysis above, there was no main effect for computer type (F,1,62=0.84, p>.05). There was an expected main effect for gender (F,1,62=10.66, p<.05) but, more importantly, a significant interaction of gender with computer type (F,1,62=6.38, p<.05), indicating that the difference in performance between computer types was not the same for male and female students. The within-group results showed no interaction of essay with computer type (F,1,62=0.00, p>.05), with gender (F,1,62=0.04, p>.05), or with gender and computer type (F,1,62=3.81, p>.05).

Because the difference in laptop versus school-computer performance was not the same for males and females, the above analysis was followed by conducting a repeated-measures ANOVA separately for each gender group. These ANOVAs used computer type and essay as independent variables, with repeated measures on the essay factor, and main NAEP writing performance as a covariate. The dependent variable was essay score. Accounting for main NAEP writing performance, there was no difference between the scores for male students taking WOL on laptop vs. school computer (F,1,62=0.89, p>.05), and no interaction between essay and computer type (F,1,62=1.59, p>.05). Female students, however, performed significantly higher on school computers than on the NAEP laptops (F,1,62=5.12, p<.05). According to the rule of thumb suggested by Cohen (1988), the size of the effect was small, about .39 standard deviations in the units of the school-computer group.[7] Finally, for female students, there was no interaction between essay and computer type (F,1,62=1.41, p>.05).

As noted, in addition to the above quasi-experimental analyses, a small experiment was conducted. This experiment was carried out in nine participating schools, which included three low-, three middle-, and three high-socioeconomic status (SES) institutions, based on median income as indicated by school zip-code information reported in the 1990 Census. All of the schools had the capability to administer WOL over the Internet using their own desktop computers and, as a consequence, this sample is not representative of the 8th grade population. Eighty-eight students participated (51 male and 37 female students). The selected students were randomly assigned to either a school desktop or NAEP laptop computer, and all students received the two WOL essays in the same order. The procedures for selecting students in the participating schools and for administering the test were identical to the procedures followed at all other WOL schools.

Usable data were obtained from 76 of the 88 students. The differences between the unweighted means were tested with a repeated-measures analysis of variance in which the dependent variable was essay score. The factors were computer type (laptop vs. desktop) and gender. As for the quasi-experimental analysis, the results of the ANOVA showed no significant main effect for computer type (F,1,72=2.83, p>.05) and an expected main effect for gender (F,1,72=9.40, p<.05). Unlike the quasi-experimental analysis, there was no significant effect for the interaction of gender with computer type (F,1,72=0.78, p>.05), possibly because of the small sample sizes involved.

With respect to the within-subjects effects, no significant difference was detected between essays (F,1,72=2.33, p>.05), but an essay-by-computer-type interaction was found (F,1,72=4.63, p<.05), suggesting that computer type was related to performance differently for each task. There was no interaction of essay with gender (F,1,72=2.18, p>.05), or of essay, computer type, and gender (F,1,72=0.05, p>.05). Post-hoc, one-tailed tests indicated that students performed significantly better on desktop than laptop for "Save a Book" (t,75=–2.40, p<.05), but that the computer types were not significantly different for "School Schedule" (t,75=–0.40, p>.05).

In sum, the quasi-experimental analysis and small experimental study do not give completely consistent results, though both analyses suggest that computer type may sometimes affect writing score.

Performance as a Function of Computer Experience

The last question addressed in this study relates to whether computer familiarity affects online test performance. How familiar were eighth-grade students with computers as of spring 2002? Students' responses to background questions collected in this study offer an answer. Responses suggest that most eighth-grade students had access to computers at school and home, and used them frequently. For example, the large majority of students indicated that they used a computer at home (91 percent) and that they used the computer at least to some extent to find information on the Internet for school projects or reports (97 percent). The majority also said that they used a computer outside of school at least two or three times a week (80 percent). (Only six percent of students indicated they never used a computer outside of school, and only 13 percent said they never used a computer at school.)

To what extent did students use computers for writing? Although almost all students reported using a computer to write at least to some degree, there was considerable variation: the results for all students show that 29 percent indicated using a computer to write "to a large extent," 41 percent "to a moderate extent," 22 percent "to a small extent," and 7 percent "not at all."

How did students use computers for writing? Again, there was wide variation: 32 percent reported that they "always" used a computer to write a paper from the beginning, 42 percent said they did this "sometimes," and 25 percent indicated that they "never" used a computer in this way. What the majority of students (69 percent) did report doing, however, was "always" using a computer to type final copy of a report that they wrote by hand.

Although computer familiarity can be measured in many ways, for purposes of this study, familiarity was defined as having experiential and hands-on components. Theoretically, these components should overlap but still be separable. For instance, a student may have had several years of experience with a computer but be neither fast nor accurate in typing. Furthermore, a minimal level on each component should, in theory, be present before a student can effectively take an online writing test. For example, some amount of previous computer experience might allow quicker adaptation to the test's navigational and input procedures, which in the WOL test were designed to be consistent with common software conventions. Likewise, some degree of automaticity in hands-on skill is necessary so that the student can focus on composing the substance of the essay and not on the mechanics associated with its entry.

To measure computer familiarity in the WOL study, two sets of indicators were used, one related to experience and one to hands-on skill. The first set came from the 37 self-reported background questions administered to students taking WOL. The rationale for using these questions as measures of computer familiarity is that they are routinely used in NAEP for reporting on computer access and use among school children. Additionally, similar questions have been used as indicators of computer familiarity in other major comparability studies (e.g., Taylor, Jamieson, Eignor, & Kirsch 1998). To evaluate the utility of these questions for measuring computer familiarity, various composites were created and related to WOL performance in the sample drawn from main NAEP reading.

The set of indicators selected to measure computer experience consisted of two composite variables, each created from a group of background questions. Figure 2 shows the two sets of background questions that were both substantively relevant and significantly related to WOL performance in the sample drawn from the main NAEP reading assessment. Questions

1–8 contributed to the "Extent of computer use" composite indicator, and questions 29–34 contributed to the "Computer use for writing" composite indicator.

Figure 2: Self-reported computer-familiarity questions contributing to each of two composite indicators

To what extent do you do the following on computer?
Include things you do in school and things you do outside of school.
(Choices: Not at all, Small extent, Moderate extent, Large extent)

1. **Play computer games**
2. **Write using a word processing program**
3. **Make drawings or art projects on the computer**
4. **Make tables, charts, or graphs on the computer**
5. **Look up information on a CD**
6. **Find information on the Internet for a project or report for school**
7. **Use email to communicate with others**
8. **Talk in chat groups or with other people who are logged on at the same time you are**

When you write a paper or report for school this year, how often do you do each of the following?
(Choices: Almost always, Sometimes, Never or hardly ever)

29. **Use a computer to plan your writing (for example, by making an outline, list, chart, or other kind of plan)**
30. **Use a computer from the beginning to write the paper or report (for example, use a computer to write the first draft)**
31. **Use a computer to make changes to the paper or report (for example, spell-check, cut and paste)**
32. **Use a computer to type up the final copy of the paper or report that you wrote by hand**
33. **Look for information on the Internet to include in the paper or report**
34. **Use a computer to include pictures or graphs in the paper or report.**

For each question set, a single score was created by making the response to each item dichotomous (because finer distinctions were likely to be less dependable), and then summing across the items. Thus, the responses to questions 1–8 were converted to a 0–8 scale after grouping the "Not at all" and "Small extent" categories with one another and similarly collapsing the "Moderate extent" and "Large extent" categories. Responses for questions 29–34 were converted to a 0–6 scale after grouping the "Sometimes" and "Never or hardly ever" categories together.[8]

The second set of computer familiarity indicators came from the hands-on exercises that preceded the test. Several measures were included that were intended to tap various components of computer skill related to taking an online writing test. From these measures, a subset was selected by relating the hands-on measures to WOL performance in the study sample drawn from the main NAEP reading assessment.

Three variables were theoretically meaningful and showed significant relationship to WOL performance. The variables, described in Table 14, were typing speed, typing accuracy, and editing skill. For an online writing test, some minimum level of each is helpful, if not required, for successful performance. Speed is needed to ensure that a complete response can be entered before the testing time elapses. Accuracy is important because faulty entry can obscure or change meaning. Finally, editing skill, which concerns command of basic word processing functions, can help the writer to revise text more effectively and quickly. For analysis purposes, typing speed, typing accuracy, and editing skill were combined to form a single hands-on computer skill index, with that index defined as the best linear composite from the regression of WOL score onto the three variables, where the regression was computed in the study sample drawn from the main NAEP reading assessment.[9,10]

Table 14: Components of the hands-on computer skills measure

Component	Definition	Scale Range
Typing speed	Number of words typed within two minutes from a 78-word passage presented on-screen.	0 – 78
Typing accuracy	Sum of punctuation, capitalization, spacing, omission, and insertion errors made in typing the above passage.	0 – maximum number of errors made
Editing	Number of editing tasks completed correctly, including correcting the spelling of a word, deleting a word, inserting a word, changing a word, moving a sentence.	0 – 5

To examine whether computer familiarity affects online test performance, a repeated-measures ANOVA was conducted with the students drawn from the main NAEP writing assessment who responded to both computer-administered WOL essays (N=660).[11] Because it is conducted within the WOL sample, this analysis avoids the potential effects of demographic differences between the paper and WOL samples. In this analysis, the independent variables were extent of computer use, computer use for writing, hands-on computer proficiency, main NAEP writing performance, and essay, with repeated measures on this last factor. Main NAEP writing performance was included to account for the possibility of a relationship between academic skill and computer familiarity, as when more scholastically accomplished students tend also to be more technologically proficient. The between-subjects results showed no significant effects for extent of computer use (*F*,1,62=2.65, *p*>.05) or for computer use for writing (*F*,1,62=0.64, *p*>.05). However, there was a significant effect for hands-on computer proficiency (*F*,1,62=93.40, *p*<.05). Within-subjects, there were no significant interactions of essay with extent of computer use (*F*,1,62=0.06, *p*>.05), with computer use for writing (*F*,1,62=2.20, *p*>.05), or with hands-on computer proficiency (*F*,1,62=3.86, *p*>.05). Thus, computer experience, in the form of keyboarding proficiency, does appear to play a role in WOL performance such that students with more hands-on skill score higher than those with less skill (holding constant their writing proficiency as measured by a paper writing test). Some sense of the magnitude of this role can be gained from examining the incremental variance accounted for by different variables in the model. Paper writing performance accounts for 36 percent of the variance in WOL scores. Adding the three computer familiarity variables to the model increases the variance accounted for in WOL scores by more than 11 points, to 47 percent.

These results, of course, provide only correlational evidence that computer familiarity affects WOL performance. That causal claim would be strengthened by evidence that the reverse situation does not hold. That is, computer familiarity should not add incrementally in any practically important way to the prediction of paper writing score, holding WOL score constant. This hypothesis can be tested by rerunning the ANOVA, this time with the dependent variable being main NAEP paper writing performance and the independent variables being extent of computer use, computer use for writing, hands-on computer proficiency, and the sum of the two WOL essay scores.[12] The results are the same as for the original ANOVA model: a significant effect for hands-on computer proficiency (*F*,1,62=13.74, *p*<.05) and no effect for either extent of computer use (*F*,1,62=1.29, *p*>.05) or for computer use for writing (*F*,1,62=0.26, *p*>.05). However, although significantly different from zero, the increment in variance from adding the three computer familiarity variables is less than 2 percentage points (as compared with over 11 points).

How practically important is an 11 percentage point increment in prediction? A real-world example may provide some context. Admissions test scores like those from the SAT or ACT Assessment add about 5 or 6 percentage points over high school grades in predicting first-year college GPA (Burton & Ramist, 2001, pg. 10; Noble & Sawyer, 2002, pg. 2). Conversely, high school grades add 9–10 percentage points over admissions test scores in predicting the same criterion.

Perhaps a more concrete way of expressing the impact is to use the ANOVA model to predict the scores of hypothetical students with different levels of computer familiarity. For the same paper writing performance, a student with computer familiarity one standard deviation below the mean on each of the three indicators would be predicted to get a WOL score .9 point lower on the 1–6 scale than a student having computer familiarity one standard deviation above the mean on those indicators. At levels of computer familiarity equal to –2 and +2 standard deviations, the predicted WOL scores would be almost two points apart.

Does computer familiarity matter more for one population group than another? To find out, gender was added to the original ANOVA model to see if there were significant interactions with the two self-reported familiarity variables or with the hands-on indicator. (Other population groups were not examined due to sample-size limitations.) Results showed that the main effect for hands-on computer skill was still significant, and that there was a significant interaction of this variable with essay, indicating that when gender is in the model, computer skill matters more for performance on one essay than on the other. However, none of the interactions with gender was found to be statistically significant; in other words, there were no measurable differences in the relationship between computer skill and WOL performance for male versus female students. (See Horkay, Bennett, Allen, & Kaplan, 2005 for the complete ANOVA results.)[13]

Discussion

This study investigated the comparability and fairness of scores associated with conducting a NAEP writing assessment on computer. Data were collected from samples of eighth-grade students selected to be representative of the nation. Four questions were addressed:

- Do students perform differently on computer-based versus paper-based writing assessments?
- Does test mode differentially affect the performance of NAEP reporting groups (e.g., those categorized by gender or by race/ethnicity)?
- Does performance vary as a function of the type of computer used to take the test (i.e., school computers vs. NAEP-supplied laptops)?
- Do students who are relatively unfamiliar with computers perform differently from students who are more familiar with them?

With respect to whether students taking paper-and-pencil tests performed differently than those taking computer-based writing tests, performance was measured in terms of essay score, essay length, and the frequency of valid responses. Results revealed no measurable differences between the two delivery modes on mean essay score (though computer scores generally appeared more variable than the paper ones). For the second of the two essays, there were significant differences for essay length and for the rate of valid responses but, in both instances, the differences appeared to be very small. For that second essay, about 1 percent more students responded on paper than on computer. And for that essay, the responding students wrote, on average, about 9 words (or 6%) more on computer than on paper (which could be simply an artifact of the difference in response rates).

The second study question concerned the impact of assessment mode on the performance of NAEP reporting groups. Performance on paper vs. computer versions of the same test was evaluated separately for groups categorized by gender, race/ethnicity, parents' education level, school location, eligibility for free/reduced-price school lunch, and school type. For all but one of the reporting groups, there were no significant differences between the scores of students who wrote their essays on paper and those who composed on computer. The singular exception was students from urban fringe/large town school locations, who scored higher on paper than on computer tests by a very small amount (.15 standard deviation units).

The third question was whether assignment to a NAEP laptop versus a school computer had an effect on performance. This question is important because some students may be more comfortable with the school computers on which they normally work and may perform better on them than on NAEP laptops. Results of a quasi-experimental analysis found that female students performed lower on the NAEP laptops by a small, but not inconsequential, amount (about .39 standard deviation units). A complementary experiment showed students to score better on desktop than laptop for one but not the other essay (and no interaction with gender).

The last study question addressed the impact of computer familiarity on online test performance. Hands-on skill was significantly related to online writing assessment performance: Students with greater hands-on skill achieved higher WOL scores, holding constant their performance on a paper writing test. Computer familiarity added about 11 percentage points over paper writing score to the prediction of WOL performance.

What are the implications of this study for online writing assessment? The finding that, for a given level of paper writing skill, students with more hands-on computer facility attained higher scores on WOL than students with less keyboard proficiency suggests that writing on computer may not be the same as writing on paper. Research on the use of computers in writing instruction supports this claim. In a meta-analysis of 32 studies published through 1990 covering the elementary through postsecondary levels, Bangert-Drowns (1993) found that students receiving writing instruction with a word processor improved the quality of their writing and wrote longer compositions than students receiving writing instruction with paper and pencil. From a meta-analysis of 26 additional studies conducted between 1992 and 2002 at the K–12 level, Goldberg, Russell, and Cook (2003) reported that students who used computers when learning to write not only produced written work that was of higher quality and greater length, but were more engaged and motivated in their writing. Thus, it is conceivable that, for a given level of paper writing performance, students with greater computer facility score higher on WOL because they write better on computer than on paper (relative to their less technologically-experienced peers). And, the reason they write better on computer than they do on paper may be because the computer offers them a tool that makes it possible to do so.

The complementary case may also be true. Holding paper writing proficiency constant, students with little practice writing on computer will not score as highly in an online writing test as their peers who word process routinely. And that lower relative performance will not necessarily be because the former students are less skilled writers, but because they are less skilled writers on computer.

These conclusions have implications for how NAEP writing assessments should be delivered and interpreted. No differences in mean scores were detected between the delivery modes. That finding suggests that, at the population level, the NAEP 2002 writing results would have been the same regardless of whether the assessment had been conducted with paper and pencil or on computer. However, the current study also suggests that the population estimates from either mode alone would probably be lower than the estimates resulting if students had been tested using the mode in which they wrote best. This conclusion follows logically from the fact that students with high computer facility wrote better on computer than students with lower computer facility but equal paper writing skill.

Which delivery mode is appropriate for future NAEP assessments should depend on what we wish to know about student writing proficiency. Do we want to know how well students write on paper, how well they write on computer, or how well they write in the mode of their choice? These questions are not necessarily the same. As students shift their typical mode of writing for school to computer, how well they write on paper becomes less relevant, a fact NAEP may need to address in the design of the next writing assessment.

Several limitations of this study should be noted. First, the study was restricted to a single grade and to only two essay tasks. At other grades, the findings could be different. If fourth-grade students have more limited word processing skills, or twelfth-graders more developed ones, student performance might vary much more dramatically across modes than was observed for the 8th grade participants in this study. Similarly, results could vary if questions requiring considerably longer or shorter responses were used.

Second, the samples taking the paper and computer tests diverged in relatively minor ways from their sampling frames as well as from one another. Also, the paper and computer tests were not administered at the same point in time. Divergence from the sampling frames reduces somewhat the generalizability of results to the nation. Differences between the samples themselves threaten the meaning of comparisons. This threat was dealt with through statistical control, where possible, and by conducting some analyses within the WOL sample. Differences in the times at which the paper and computer tests were administered would be a confounding factor if the writing proficiency of eighth grade students materially changed over the relatively short period between the two administrations, an eventuality we believe to be unlikely.

Third, it was not possible to offer a strong assertion about the impact of computer type on performance. In this study, females did not perform as well on NAEP-supplied laptops as on school computers, after controlling for main NAEP paper writing performance. Also, in a small experiment, students scored higher on school computers than on NAEP laptops, though no interaction with gender was found. While these results suggest that computer type may affect test performance, there are reasons to suspect that any such effect has moderated. First, today's laptops have improved considerably over the machines used in this study: The keyboards have become much more comparable in comfort and responsiveness to those of desktops; screen size and clarity also have increased remarkably. Second, students in general have become more familiar with laptop computers. These computers have, since the WOL study, become mass-market commodities and, thus, more likely to be familiar to students from a wide variety of demographic groups. Still, the possibility of disadvantage due to computer type should be investigated with respect to any future operational assessment that requires some students to test on machines unfamiliar to them.

A variation on this theme which this study did not investigate is the impact on writing test performance of differences among school computers. As school computers become the predominant delivery mechanism, variation across machines (e.g., monitor size, screen resolution, connection speed) may play a greater role in affecting performance irrelevantly. Such an effect was reported on reading assessments for the combined variation of screen resolution and monitor size (Bridgeman, Lennon, & Jackenthal, 2003). Such variation may impact writing assessment to the extent that differences in, for example, keyboard layout impact a student's ability to compose without devoting undue attention to the mechanics of text entry.

Fourth, as noted, the finding that computer familiarity affected online writing performance was based on correlational evidence. Although reversing the roles of the paper and computer test scores in the ANOVA may strengthen the causal claim, it is possible that factors other than computer familiarity were responsible for the result we observed.

Fifth, the sample sizes for NAEP reporting groups were often small. This fact reduced the power of the statistical tests for those analyses that could be conducted and made other potentially important analyses infeasible. For example, such questions as whether computer familiarity affects online test performance for particular demographic groups other than gender remain to be addressed.

Finally, differences in reader reliability across the modes were observed in this study and such differences could potentially impact results. Optimally, scoring should be done for both delivery modes at the same time by the same readers using the same procedures. For practical reasons, different groups at different times scored the online and paper responses used in this study. While these procedural differences were associated with lower levels of reader agreement for the scoring of the online responses than for the paper responses, the overall score reliabilities for the two modes did not suggest any notable divergence in scoring accuracy. Further, when WOL readers blindly scored paper responses that had been transcribed from handwritten to typed format, the total scores were not significantly different from those assigned by the original reader group. Given these facts, the lower reader reliability observed for the WOL sample does not seem likely to have affected the study conclusions in any material way.

Endnotes

1. Complete details on sample selection are given in Horkay, Bennett, Allen, & Kaplan (2005).
2. The uncorrected correlations were .63 for WOL and .57 for main NAEP. Corrections were computed using the Spearman-Brown formula (Thorndike 1982).
3. For each of these specially selected samples, the unweighted mean essay score was compared to the unweighted mean essay score from the sample of paper main NAEP students who had been administered and did respond to both essays (N=2,878). No significant mean differences were detected for Save a Book (*t*, 3170=.25, *p*>.05) or for School Schedule (*t*, 3168=.45, *p*>.05). The specially selected samples did, however, have noticeably larger standard deviations than the paper main NAEP sample.
4. The post-hoc test was a repeated-measures ANOVA done separately for each category of school location. The independent variables were delivery mode and essay, with repeated measures on the essay factor. The dependent variable was essay score.
5. School machines vary too in ways that may possibly affect performance. This naturally occurring equipment variation was not evaluated in this study.
6. Main NAEP writing performance was represented using the "plausible values" methodology as described in Allen, Carlson, and Zelenak, (1999). Essentially, for each student, five possible scores (or plausible values) are sampled from a posterior distribution predicted from item responses, item parameters, and background information. The software employed, WESVAR, uses all five plausible values in its ANOVA procedure.
7. This effect was computed using the means *unadjusted* for main NAEP writing performance because that variable appeared to have little impact on the analysis. That is, removing main NAEP writing performance from the overall (three-way) ANOVA model produced the same substantive result, and closely similar quantitative results, as including it.
8. Coefficient alpha reliabilities for the "Extent of computer use" and "Computer use for writing" scores were .55 and .65, respectively.
9. The standardized regression weights for the three index components were .52 for typing speed, .19 for editing skill, and –.10 for typing accuracy (which gets negative weight because it indicates the number of errors made). These weights give an indication of the relative importance of each component to the hands-on index.
10. The study sample drawn from the main NAEP reading assessment was used to select the hands-on variables and to derive their best linear composite. This composite was then applied in the study sample drawn from the main NAEP writing assessment. The two samples were used to avoid the potential for capitalizing on chance that would be present if the variables had been selected, their composite derived, and that composite applied all in the same sample.
11. Twenty-seven students were not included in the analysis because they did not respond to the minimum number of background questions required to form the "computer use for writing" measure, or they did not have main NAEP writing performance information.

12. This model was run with 664 students. The model was run five times, once with each plausible value as the dependent variable (representing main NAEP writing performance). The F-values from the five runs were then averaged and this average was tested. This independent treatment of plausible values is comparable to that employed for the original ANOVA model, where the plausible values were among the predictor set.
13. At first blush, it might seem sensible to rerun the model yet again, this time adding computer type to control for its effects and to identify any interactions with computer familiarity. However, the hands-on computer proficiency measure was confounded with computer type, so this model was not run.

References

Allen, N. L., Carlson, J. E., & Zelenak, C. A. (1999). *The NAEP 1996 technical report (NCES 1999–452)*. Washington, DC: National Center for Education Statistics, US Department of Education.

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.

Bridgeman, B., & Cooper, P. (1998, April). *Comparability of scores on word-processed and handwritten essays on the Graduate Management Admission Test*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2003). Effects of screen size, screen resolution, and display rate on computer-based test performance. *Applied Measurement in Education, 16*(3), 191–205.

Burton, N., & Ramist, L. *Predicting success in college: SAT studies of classes graduating since 1980* (Research Report No. 2001–2). New York: College Board. Retrieved January 31, 2006 from http://www.collegeboard.com/research/pdf/rdreport200_3919.pdf

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Computer-based testing. (2006, May 4). *Education Week, 25*(35), 52.

Goldberg, A., Russell, M., & Cook, A. (2003). The effect of computers on student writing: A meta-analysis of studies from 1992 to 2002. *Journal of Technology, Learning, and Assessment, 2*(1). Retrieved November 24, 2003, from http://www.bc.edu/research/intasc/jtla/journal/v2n1.shtml.

Harrington, S., Shermis, M. D., & Rollins, A. L. (2000). The influence of word processing on English placement test results. *Computers and Composition, 17*(2), 197–210.

Horkay, N., Bennett, R. E., Allen, N., & Kaplan, B. (2005). Online assessment in writing. In B. Sandene, N. Horkay, R. E. Bennett, N. Allen, J. Braswell, B. Kaplan, & A. Oranje (Eds.), *Online assessment in mathematics and writing: Reports from the NAEP Technology-Based Assessment Project* (NCES 2005–457). Washington, DC: National Center for Education Statistics, US Department of Education. Retrieved January 31, 2005 from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2005457

MacCann, R., Eastment, B., & Pickering, S. (2002). Responding to free response examination questions: Computer versus pen and paper. *British Journal of Educational Technology, 33*(2), 173–188.

Noble, J., & Sawyer, R. (2002). *Predicting different levels of academic success in college using high school GPA and ACT composite score* (ACT Research Report Series 2002–4). Iowa City: ACT. Retrieved January 31, 2006 from http://www.act.org/research/reports/pdf/ACT_RR2002-4.pdf

Persky, H. R., Daane, M. C., & Jin, Y. (2003). The nation's report card: Writing 2002. Washington, DC: National Center for Education Statistics, US Department of Education. Retrieved September 21, 2006 from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2003529.

Powers, D., & Farnum, M. (1997). *Effects of mode of presentation on essay scores* (RM–97–8). Princeton, NJ: Educational Testing Service.

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.

Powers, D., & Potenza, M. T. (1996). *Comparability of testing using laptop and desktop computers* (RR–96–15). Princeton, NJ: Educational Testing Service.

Russell, M. (1999). Testing on computers: A follow-up study comparing performance on computer and on paper. *Education Policy Analysis Archives,* 7(20). Retrieved June 27, 2003, from http://epaa.asu.edu/epaa/v7n20/.

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). Retrieved June 27, 2003, from http://epaa.asu.edu/epaa/v5n3.html.

Russell, M., & Plati, T. (2001). Effects of computer versus paper administration of a state-mandated writing assessment. *TCRecord*. Retrieved June 27, 2003, from http://www.tcrecord.org/Content.asp?ContentID=10709.

Russell, M., & Tao, W. (2004a). Effects of handwriting and computer-print on composition scores: A follow-up to Powers et al. *Practical Assessment, Research and Evaluation, 9*(1). Retrieved July 8, 2004, from http://pareonline.net/getvn.asp?v=9&n=1.

Russell, M., & Tao, W. (2004b). The influence of computer-print on rater scores. *Practical Assessment, Research and Evaluation, 9*(1). Retrieved July 8, 2004, from http://pareonline.net/getvn.asp?v=9&n=10.

State: Online testing helped raise scores. (2005, August 18). *eSchool News Online*. Retrieved August 18, 2005 from http://www.eschoolnews.com/news/showStoryts.cfm?ArticleID=5826.

Taylor, C., Jamieson, J., Eignor, D., & Kirsch, I. (1998). *The relationship between computer familiarity and performance on computer-based TOEFL test tasks* (Report 61). Princeton, NJ: Educational Testing Service.

Thorndike, R. L. (1982). *Applied psychometrics*. Boston: Houghton Mifflin.

Virginia Department of Education. (undated). *Virginia Standards of Learning assessment program: Update for Spring 2005*. Retrieved February 7, 2006 from http://www.doe.virginia.gov/VDOE/Technology/soltech/docs/RegionalWorkshop_Spr05.ppt.

Wolfe, E. W., Bolton, S., Feltovich, B., & Bangert, A. W. (1996). A study of word processing experience and its effects on student essay writing. *Journal of Educational Computing Research, 14*(3), 269–283.

Wolfe, E. W., Bolton, S., Feltovich, B., & Niday, D. M. (1996). The influence of student experience with word processors on the quality of essays written for a direct writing assessment. *Assessing Writing, 3*(2), 123–147.

Wolfe, E. W., & Manalo, J. R. (2004). Composition medium comparability in a direct writing assessment of non-native english speakers. *Language Learning and Technology, 8*(1), 53–65. Retrieved January 6, 2004, from http://llt.msu.edu/vol8num1/wolfe/default.html.

Appendix A: Scoring Rubrics

Informative Scoring Guide (Save a Book)

Excellent-6

- Develops and shapes information with well-chosen details across the response.
- Well organized with strong transitions.
- Sustains variety in sentence structure and exhibits good word choice.
- Errors in grammar, spelling, and punctuation are few and do not interfere with understanding.

Skillful-5

- Develops and shapes information with details in parts of the response.
- Clearly organized, but may lack some transitions and/or have occasional lapses in continuity.
- Exhibits some variety in sentence structure and some good word choices.
- Errors in grammar, spelling, and punctuation do not interfere with understanding.

Sufficient-4

- Develops information with some details.
- Organized with ideas that are generally related, but has few or no transitions.
- Exhibits control over sentence boundaries and sentence structure, but sentences and word choice may be simple and unvaried.
- Errors in grammar, spelling, and punctuation do not interfere with understanding.

Uneven-3

May be characterized by one or more of the following:

- Presents some clear information, but is list-like, undeveloped, or repetitive OR offers no more than a well-written beginning.
- Unevenly organized; the response may be disjointed.
- Exhibits uneven control over sentence boundaries and sentence structure; may have some inaccurate word choices.
- Errors in grammar, spelling, and punctuation sometimes interfere with understanding.

Insufficient-2

May be characterized by one or more of the following:

- Presents fragmented information OR may be very repetitive OR may be very undeveloped.
- Very disorganized; thoughts are tenuously connected OR the response is too brief to detect organization.
- Minimal control over sentence boundaries and sentence structure; word choice may often be inaccurate.
- Errors in grammar or usage (such as missing words or incorrect word use or word order), spelling, and punctuation interfere with understanding in much of the response.

Unsatisfactory-1

May be characterized by one or more of the following:

- Attempts to respond to prompt, but provides little or no coherent information; may only paraphrase the prompt.
- Has no apparent organization OR consists of a single statement.
- Minimal or no control over sentence boundaries and sentence structure; word choice may be inaccurate in much or all of the response.
- A multiplicity of errors in grammar or usage (such as missing words or incorrect word use or word order), spelling, and punctuation severely impedes understanding across the response.

Note: From Persky, Daane, & Jin (2003).

Persuasive Scoring Guide (School Schedule)

Excellent-6

- Takes a clear position and develops it consistently with well-chosen reasons and/or examples across the response.
- Well organized with strong transitions.
- Sustains variety in sentence structure and exhibits good word choice.
- Errors in grammar, spelling, and punctuation are few and do not interfere with understanding.

Skillful-5

- Takes a clear position and develops it with reasons and/or examples in parts of the response.
- Clearly organized, but may lack some transitions and/or have occasional lapses in continuity.
- Exhibits some variety in sentence structure and some good word choices.
- Errors in grammar, spelling, and punctuation do not interfere with understanding.

Sufficient-4

- Takes a clear position and supports it with some reasons and/or examples.
- Organized with ideas that are generally related, but there are few or no transitions.
- Exhibits control over sentence boundaries and sentence structure, but sentences and word choice may be simple and unvaried.
- Errors in grammar, spelling, and punctuation do not interfere with understanding.

Uneven-3

May be characterized by one or more of the following:

- Takes a position and offers support, but may be unclear, repetitive, list-like, or undeveloped.
- Unevenly organized; the response may be disjointed.
- Exhibits uneven control over sentence boundaries and sentence structure; may have some inaccurate word choices.
- Errors in grammar, spelling, and punctuation sometimes interfere with understanding.

Insufficient-2

May be characterized by one or more of the following:

- Takes a position, but may be very unclear, very undeveloped, or very repetitive.
- Very disorganized; thoughts are tenuously connected OR the response is too brief to detect organization.
- Minimal control over sentence boundaries and sentence structure; word choice may often be inaccurate.
- Errors in grammar or usage (such as missing words or incorrect word use or word order), spelling, and punctuation interfere with understanding in much of the response.

Unsatisfactory-1

May be characterized by one or more of the following:

- Attempts to take a position (addresses topic) but is incoherent OR takes a position but provides no support; may only paraphrase the prompt.
- Has no apparent organization OR consists of a single statement.
- Minimal or no control over sentence boundaries and sentence structure; word choice may be inaccurate in much or all of the response.
- A multiplicity of errors in grammar or usage (such as missing words or incorrect word use or word order), spelling, and punctuation severely impedes understanding across the response.

Note: From Persky, Daane, & Jin (2003).

Author Notes

All correspondence should be sent to the second author, Randy Bennett (rbennett@ets.org).

This study was funded by the National Center for Education Statistics (NCES), Institute of Education Sciences, U.S. Department of Education.

We are grateful for the contributions of Mary Daane, Doug Forer, Andreas Oranje, Hilary Persky, Mike Wagner, and Vincent Weng of ETS, Taslima Rahman of NCES, the staff of Westat, and many others without whom this project could not have been conducted.

Author Biographies

Nancy Horkay, employed by Educational Testing Service (ETS) from 1985 through 2004, was Director of Operations for the National Assessment of Educational Progress (NAEP). She also served as the project manager for the Writing Online assessment, overseeing and directing activities concerning try out, data collection, scoring, and initial report writing. During her time at ETS, Ms. Horkay managed and led operational activities for the PRAXIS program, as well as for NAEP. She retired from ETS in 2004 and is now enjoying life in South Carolina.

Randy Elliot Bennett is Distinguished Presidential Appointee in the Research & Development Division at Educational Testing Service in Princeton, New Jersey. A graduate of Teachers College, Columbia University, Dr. Bennett began his employment at ETS in 1979. Since the 1980s, he has conducted research on the applications of technology to testing, on new forms of assessment, and on the assessment of students with disabilities. Dr. Bennett's work on the use of new technology to improve assessment has included research on presenting and scoring open-ended test items on the computer, on multimedia and simulation in testing, and on generating test items automatically. For this work, he was given the ETS Senior Scientist Award in 1996 and the ETS Career Achievement Award in 2005. He is the author of many publications including "Technology and Testing" (with Fritz Drasgow and Ric Luecht) in Educational Measurement (4th Edition) and "What Does it Mean to Be a Non-profit Educational Measurement Organization in the 21st Century" (http://www.ets.org/Media/Research/pdf/Nonprofit.pdf).

Nancy Allen earned her doctoral degree in educational statistics and measurement from the University of Iowa in 1987. She spent the following year as an ETS post doctoral fellow and in 1988 began employment as an associate research scientist at the organization. In 1995, she became Director of NAEP Data Analysis and Scaling for ETS. Allen is a member of the American Educational Research Association, Division D, the National Council on Measurement in Education, the Psychometric Society, and the American Statistical Association. She has served as associate editor for the Journal of Educational and Behavioral Statistics, on the Executive Committee for the Caucus for Women in Statistics, and as president of the caucus. Allen has extensive experience with large-scale assessments, especially in areas of scaling, analysis, and the handling missing data. She left ETS in 2004 to pursue postsecondary teaching and consulting in educational measurement and statistics.

Bruce A. Kaplan holds the title of director of data analysis and interactive systems in the Research & Development Division at Educational Testing Service, in Princeton, NJ. He has been employed at ETS since 1979. He received his master's degree in economic and social statistics from the Industrial and Labor Relations School at Cornell University in 1979. He received his bachelor of science with highest of honors in applied mathematics and computer science from the State University of New York at Stony Brook in 1976. His interests include sample survey design and estimation, exploratory data analysis, regression analysis, computerization of statistical techniques, empirical Bayes techniques, and graphical displays. His experiences cover a wide range of research, computer technology, and statistical areas.

Fred Yan holds the title of research data analyst in the Center for Data Analysis Research in the Research and Development Division at Educational Testing Service, in Princeton, NJ. He has been employed at ETS since 1998. He received his bachelor's degree in Engineering Mechanics from Tianjin University, China and his master of science degree in Civil Engineering from Pennsylvania State University. He also is currently working toward completing the requirements for the master of science degree in computer science from New Jersey Institute of Technology. His experience covers statistical data analysis and application software development.

J·T·L·A

The Journal of Technology, Learning, and Assessment

www.jtla.org

Technology and Assessment Study Collaborative
Caroline A. & Peter S. Lynch School of Education, Boston College