Part II:

Online Assessment in Writing

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

In collaboration with
Mary Daane
Douglas Forer
Hilary Persky
Michael Wagner
Vincent Weng
Fred Yan
Educational Testing Service

Taslima Rahman
Project Officer
National Center for Education Statistics

THIS PAGE INTENTIONALLY LEFT BLANK.

Executive Summary

The 2002 Writing Online (WOL) study is the second of three field investigations in the Technology-Based Assessment project, which explores the use of new technology in administering the National Assessment of Educational Progress (NAEP).[1] The study addresses issues related to measurement, equity, efficiency, and operations in a computer-based writing assessment.

This report describes the results of testing a national sample of eighth-grade students on computer. The WOL study was administered to students on school computers via the World Wide Web or on NAEP laptop computers brought into schools. Both writing tasks (herein referred to as "essays") used in the WOL study were taken from the existing main NAEP writing assessment and were originally developed for paper administration.

During April and May 2002, data were collected from more than 1,300 students in about 160 schools. Student performance on WOL was compared to that of a national sample that took the main NAEP paper-and-pencil writing assessment between January and March 2002. For the samples taking WOL, background information concerning access to, use of, and attitudes toward computers was also collected. In addition, exercises designed to measure computer skills were administered. Results are considered to be statistically significant if the probability of obtaining them by chance alone does not exceed the .05 level.

Measurement

- Performance on computer versus a paper test was measured in terms of essay score, essay length, and the frequency of valid responses. Results showed no significant difference in essay scores or essay length between the two delivery modes. However, for the second of the two essays comprised in the test, delivery mode did significantly predict response rate, with roughly 1 percent more students responding to the test on paper than on computer.

Equity

- Performance on paper and computer versions of the same test was evaluated separately for the categories of gender, race/ethnicity, parents' education level, school location, eligibility for free/reduced-price school lunch, and school type. With one exception, there were no significant differences for the NAEP reporting groups examined between the scores of students who wrote their essays on paper and those who responded on computer. The exception was for students from urban fringe/large town locations, who performed higher on paper than on computer tests by about 0.15 standard deviation units.
- The effect of delivery mode on performance was also evaluated for gender groups in terms of response length and frequency of valid responses. For the second essay, males wrote significantly fewer words on paper than on computer. Also for that second essay, a significantly higher percentage of females responded on paper than on computer. The difference in percent responding was about 2 percentage points.
- The impact of assignment to a NAEP laptop versus a school computer was evaluated in two analyses. Results from the two analyses were not completely consistent. In an experimental substudy in which a small number of students were randomly assigned to computer type, those who took the test on NAEP laptops scored significantly lower than students taking the test on school computers, but for only one of the two essays. In a quasi-experimental analysis with larger sample sizes, however, only female students performed significantly lower on the NAEP laptops, but this group did so for both essays.
- To determine if computer familiarity affected online test performance, students' self-reported computer experience and hands-on measures of keyboarding skill were used to predict online writing performance, after controlling for their paper writing score. Hands-on skill was significantly related to online writing assessment performance, so that students with greater hands-on skill achieved higher WOL scores when holding constant their performance on a paper-and-pencil writing test. Computer familiarity added about 11 percentage points over paper writing score to the prediction of WOL performance.

Efficiency

- With respect to timeliness, it is anticipated that delivering assessments via computer would not have any significant short-term effect on the pilot stage of the NAEP assessment cycle, but could possibly shorten the operational stage appreciably by requiring fewer steps.

[1] The initial project in the series was the 2001 Math Online study, an investigation of the implications of delivering NAEP mathematics assessments on computer. The third project in the series is the 2003 Problem Solving in Technology-Rich Environments study, an investigation of how computers might be used to measure skills that cannot be measured in a paper test.

- Assuming similar levels of effort for current NAEP writing assessments, the costs for an online test should be similar for test development, similar or higher for assessment delivery and administration, and similar or lower for scoring.
- Results showed that the automated scoring of essay responses did not agree with the scores awarded by human readers. The automated scoring produced mean scores that were significantly higher than the mean scores awarded by human readers. Second, the automated scores agreed less frequently with the readers in level than the readers agreed with each other. Finally, the automated scores agreed less with the readers in rank order than the readers agreed with one another.

Operations

- Because the WOL delivery software supported only the Windows operating system and required broadband connections that were not available at some schools, 65 percent of students (and 59 percent of schools) were tested on laptop computers provided by NAEP administrators. The remainder were tested on school computers via the Web. Both web and laptop administrations ran very smoothly, with only minimal problems overall and almost no problems with computer hardware.

The authors believe these results have important implications for NAEP:

- Aggregated scores from writing tests taken on computer do not appear to be measurably different from ones taken on paper for the eighth-grade population as a whole, as well as for all but one of the NAEP reporting groups examined.
- Scores for *individual* students may not be comparable, however. Even after controlling for their level of paper writing skill, students with more hands-on computer facility appear to get higher scores on WOL than do students with lower levels of keyboard proficiency.
- Because scores for individuals on paper and computer writing tests do not appear to be comparable, relationships of certain demographic variables to writing proficiency may change, depending upon the mode in which that proficiency is measured.
- NAEP should expect the transition and near-term costs for conducting an electronic writing assessment to be considerable. NAEP will likely need to supplement web delivery by bringing laptop computers into some schools.
- Delivering writing assessments on computer may allow responses to be automatically scored, which could help NAEP reduce costs and speed up reporting. Although automated scores did not agree highly enough with the scores awarded by human readers to consider the two types of scoring interchangeable, this technology has been found to work effectively in some studies, is evolving rapidly, and may soon become usable by NAEP.
- Future research should address the generalizability of this study's findings to other grades and other types of essay tasks, and investigate the impact of differences in equipment configuration on NAEP population estimates. Finally, in this study, WOL readers scored student responses with lower levels of agreement than did the main NAEP readers. Future research should attempt to minimize more effectively differences in reader reliability across modes that can potentially affect the precision of scores and the meaning of results.

Foreword

The Research and Development series of reports has been initiated for the following goals:

1. To share studies and research that are developmental in nature. The results of such studies may be revised as the work continues and additional data become available.
2. To share results of studies that are, to some extent, on the cutting edge of methodological developments. Emerging analytical approaches and new computer software development often permit new, and sometimes controversial, analysis to be done. By participating in "frontier research," we hope to contribute to the resolution of issues and improved analysis.
3. To participate in discussions of emerging issues of interest to educational researchers, statisticians, and the federal statistical community in general. Such reports may document workshops and symposiums sponsored by the National Center for Education Statistics (NCES) that address methodological and analytical issues or may share and discuss issues regarding NCES practice, procedures, and standards.

The common theme in all three goals is that these reports present results or discussions that do not reach definitive conclusions at this point in time, either because the data are tentative, the methodology is new and developing, or the topic is one on which there are divergent views. Therefore, the techniques and inferences made from the data are tentative and are subject to revision. To facilitate the process of closure on the issues, we invite comment, criticism, and alternatives to what we have done. Such responses should be directed to:

Marilyn M. Seastrom
Chief Statistician
Statistical Standards Program
National Center for Education Statistics
1900 K Street NW, Suite 9000
Washington, DC 20006

Acknowledgments

The NAEP Writing Online study was part of the Technology-Based Assessment (TBA) project, a collaborative effort led by the National Center for Education Statistics (NCES) and the National Assessment Governing Board (NAGB), and carried out by Educational Testing Service (ETS) and Westat. The project was funded through NCES, in the Institute of Education Sciences of the U.S. Department of Education. We appreciate the support and guidance of Associate Commissioner of Education Statistics Peggy Carr, NAEP project director Suzanne Triplett, TBA project directors Holly Spurlock and Taslima Rahman, and NCES consultants Vonda Kiplinger and Bob Evans.

NAEP is grateful to the students and school staff who participated in the assessment, to the Westat staff who administered the assessment, and to the ETS consultants who scored the writing essays.

NAEP activities at ETS were directed by Stephen Lazer and John Mazzeo, with assistance from John Barone. The ETS management for the TBA project included Randy Bennett and Clyde Reese.

The Writing Online study was managed and coordinated by Nancy Horkay of ETS. Scoring activities were conducted at ETS with contributions from Nancy Glazer and Stephanie Wittman. Contributors to the production of the online test and tutorials included Andrew Baird, Marylou Lennon, Lou Mang, and Rob Rarich. Staff members who worked on usability testing were Holly Knott and Margaret Redman. Jill Burstein carried out the automated scoring activities with assistance from Chi Lu and Slava Andreyev.

Statistical and psychometric activities were carried out by John Willey. Kevin Bentley, Gerry Kokolis, and Katharine Pashley conducted the WOL database work.

The design and production of this report were overseen by Loretta Casalaina with assistance from Rick Hasney and Susan Mills. Ming Kuang coordinated the documentation and data checking procedures. Carmen Payton reviewed tabular presentations for consistency with NCES standards. Arlene Weiner coordinated the editorial and proofreading procedures with assistance from Patricia Hamill, Linda Myers, and Jennifer O'Bryan. The web version of this report was coordinated by Rick Hasney.

This project could not have been completed without Westat, which conducted student sampling, administration, field support, and weighting. NAEP activities at Westat were directed by Nancy Caldwell, Keith Rust, Debby Vivari, and Dianne Walsh. Westat's Writing Online study activities were managed by Dward A. Moore, Jr., Brice Hart, and Brenda Ennis. Sampling and weighting activities were managed by John Burke, and sampling and weighting statistical work was carried out by Sylvia Dohrmann and Hyunshik Lee. Weighting systems work was completed by Ngoan Vo and Phu Nguyen. Rob Dymowski developed the system used to draw student samples in each school and to report progress in the field with assistance from Sharon Hirabayashi. Brice Hart managed the day-to-day operations of Westat's technical and software systems with assistance from Fran Cohen.

Many thanks are due to the numerous reviewers, both internal and external to NCES and ETS. The comments and critical feedback of the following reviewers are reflected in the final version of this report: Tajuana Bates, Ellen Carey, Young Chun, John Clement, Mary Crovo, Aaron Douglas, Lawrence Feinberg, Ray Fields, Arnold Goldstein, Steve Gorman, Andrew Kolstad, Taslima Rahman, and Holly Spurlock.

Contents

Tables

Figures

1. Introduction

This technical report presents the methodology and results of the Writing Online (WOL) study, part of the National Assessment of Educational Progress (NAEP) Technology-Based Assessment (TBA) project. Funded by the National Center for Education Statistics (NCES), the Technology-Based Assessment project is intended to explore the use of new technology in NAEP.

The TBA project focuses on several key questions:

1. ***What are the measurement implications of using technology-based assessment in NAEP?*** Technology-based assessment may change the meaning of NAEP measures in as yet unknown ways. It may allow assessment of skills that could not be measured using paper and pencil or preclude measuring skills that could be tested by conventional means. It may permit the assessment of emerging skills, particularly those requiring students to employ new technology in learning and problem solving.
2. ***What are the implications for equity?*** If not carefully designed, technology-based assessment could inaccurately reflect the skills of some groups of students, especially those with differing degrees of access to, or skill with, computers. At the same time, it could increase participation of students with disabilities by providing additional accommodation tools. In addition, it may better reflect the skills of students who routinely use the computer to perform academic tasks like writing and composing.
3. ***What are the efficiency implications of using technology-based assessment compared with paper and pencil?*** Along with other new technologies, the Internet may afford significant time and cost savings for the delivery and scoring of large-scale assessments.
4. ***What are the operational implications of technology-based assessment?*** Moving from a paper-based program to an electronic one raises significant issues concerning school facilities, equipment functioning, administrator responsibilities, and school cooperation.

To answer these questions, the NAEP program undertook three empirical studies with students: Math Online (MOL), Writing Online (WOL), and Problem Solving in Technology-Rich Environments (TRE). These studies together address the questions above.

The WOL study focused on the issues associated with delivering existing constructed-response NAEP writing tasks on computer. The key issues were:

Measurement

- How does test mode (i.e., delivery on computer vs. delivery on paper) affect the inferences that can be drawn about students' writing skill? In particular, do students perform differently across the two modes?

Equity

- How do population groups perform and do mode effects vary across groups?
- Are students disadvantaged if they must take a writing test on a NAEP laptop instead of a school computer?
- How are students with different levels of computer experience affected by computer- vs. paper-based writing assessment?

Efficiency

- Is a technology-based writing assessment more cost-effective or less time-consuming than a paper one?
- How might technological advances like web delivery and automated essay scoring affect the cost and timeliness of assessment?

Operations

- What are the logistical challenges associated with administering a NAEP writing assessment on computer? In particular, are school facilities, equipment, software, and internet connectivity adequate? Are schools willing to cooperate with the needs of a technology-based assessment? How might NAEP use computer delivery to accommodate the needs of students with disabilities? Is the quality of data derived from an assessment delivered on computer acceptable?

2. Methodology

Study Samples

The WOL study samples were composed of nationally representative groups of eighth-grade students drawn from the main NAEP 2002 assessments, which were administered between the end of January and the beginning of March 2002.[1] 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, none of which included the essay tasks 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 main NAEP writing 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. (See appendix A for more details on the WOL sample.)

Of the 5,368 schools selected for the main NAEP 2002 writing and reading assessments, 236 were randomly selected for administration of WOL. One hundred and fifty-eight of these schools participated.[2] The weighted school response rate, which reflects the accumulated effect of main NAEP and WOL study attrition, is 67 percent. Within the 158 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 students, 1,313 participated in WOL. Reasons for nonparticipation are given in table 2-1. In addition to these nonparticipating students, five other 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 nonparticipants and misclassified individuals, the weighted student response rate reflecting both main NAEP and WOL attrition is 77 percent.[3]

Table 2-1. Reasons for student nonparticipation in Writing Online, grade 8: 2002

Reasons for nonparticipation	Number of cases 546
Absent from WOL administration	207
Absent from the NAEP administration	137
Withdrawn from school or ineligible	85
Excluded as SD or LEP[1]	65
Attempted WOL test but did not complete	29
Participated in WOL self-voicing substudy[2]	23

[1] Generally students with disabilities or limited-English-proficient students who were judged by school staff as not being able meaningfully to participate in the assessment activities without accommodation were excluded from the study.
[2] A small number of students with print-related disabilities was selected to be tested with an accommodated version of WOL.
NOTE: WOL= Writing Online. SD=Students with disabilities. LEP=Limited-English-proficient students.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Writing Online Study.

For most of the analyses conducted for this study, data were used only from those students who responded to both essay tasks. This restriction was imposed because it allows 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 permits 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.

[1] Details on sample selection are given in appendix A.
[2] One school was subsequently dropped from the analysis because, although it administered WOL, that school's students could not be matched to main NAEP data as that school did not participate in main NAEP.
[3] Analysis of nonresponse for groups with sufficient cell sizes showed that census region was significantly related to school-level nonresponse and that relative age and disability status were significantly related to student nonresponse.

Table 2-2. Numbers of students in study samples before and after excluding those who did not respond to both essays, grade 8: 2002

Study sample	Main NAEP writing students administered both paper-and-pencil essays in the same order as WOL	WOL Students		
		All students	Students drawn from main NAEP writing	Students drawn from main NAEP reading
Total	2,983	1,308	715	593
Students responding to both essays	2,878	1,255	687	568
Weighted percentage responding to both essays	98 (0.4)	96 (0.6)	97 (0.7)	95 (1.0)

NOTE: WOL=Writing Online. Standard errors are in parentheses.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Writing Online Study.

Table 2-2 shows the numbers of students before and after the elimination of those who did not respond to both essays, as well as the weighted percentages responding. In addition to the three samples described above, values are given for all students taking the WOL test (which is the sum of the other two WOL groups). As the table indicates, even after eliminating those who only responded to one essay, a very high percentage of participating students—more than 95 percent—was retained in each sample.

How representative are these samples? Table 2-3 contrasts main NAEP scores and background information for the subset of 2,878 students responding on paper to both main NAEP essays used in this study with all 118,516 students taking the main NAEP 2002 writing assessment. As table 2-3 shows, the characteristics of students in the main NAEP writing subsample were not significantly different from the corresponding characteristics of all main NAEP writing students, except for the significantly higher percentage of female students and lower percentage of male students in the subsample.

Table 2-3. Characteristics of study sample taking the main NAEP paper-and-pencil writing assessment compared with all students taking main NAEP writing, grade 8: 2002

Characteristic	Main NAEP writing students responding to both paper-and-pencil essays in the same order as WOL	All main NAEP writing students
Number of students	2,878	118,516
NAEP writing mean	156 (1.4)	153 (0.5)
	Percent of students	
Exclusion rate[1]	3 (0.4)	4 (0.2)
Gender		
Male	45 (1.5)*	50 (0.3)
Female	54 (1.6)*	50 (0.3)
Race/ethnicity		
White	65 (1.6)	65 (0.5)
Black	16 (1.5)	15 (0.4)
Hispanic	15 (1.2)	14 (0.4)
Asian/Pacific Islander	4 (0.8)	4 (0.2)
Other	1 (0.3)	2 (0.1)
Type of school		
Public	90 (0.8)	91 (0.2)
Nonpublic	10 (0.8)	9 (0.2)
Parents' highest level of education		
Less than high school	6 (0.7)	6 (0.1)
Graduated high school	17 (1.1)	17 (0.2)
Some education after high school	19 (1.5)	18 (0.2)
Graduated college	46 (1.7)	46 (0.4)
Unavailable	13 (1.1)	12 (0.2)
Student eligibility for free/reduced-price school lunch		
Eligible	30 (1.4)	31 (0.6)
Not eligible	54 (1.7)	53 (1.0)
Unavailable	15 (1.3)	16 (0.8)
Type of school location		
Central city	28 (1.2)	29 (0.6)
Urban fringe/large town	43 (1.5)	42 (0.7)
Rural/small town	29 (0.9)	29 (0.5)

* $p < .05$ for the difference between the study sample and all students administered the main NAEP assessment as computed from a *t*-test for independent samples.

[1] "Exclusion rate" is the weighted sum of the excluded students divided by the excluded plus the assessed students. For study participants, this rate is based on all students who were sampled to receive the test booklet containing the two paper-and-pencil essays given in the same order as the WOL essays.

NOTE: WOL= Writing Online. All values are weighted, except for the sample sizes. The sample size for "all main NAEP writing students" includes individuals who did not respond to either essay. "Other" category for race/ethnicity includes American Indian/Alaska Native and unclassified students. Detail may not sum to totals because of rounding. Standard errors are in parentheses.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Writing Online Study.

Table 2-4 compares characteristics of the students taking the WOL computer test and those in the main NAEP samples from which these students were drawn. The first relevant comparison is between all students taking the main NAEP *writing* assessment and the students who responded to both essays on the WOL test. The second relevant comparison is between all students taking the main NAEP *reading* assessment and the students who responded to both essays on the WOL test.

Table 2-4 indicates that the WOL sample drawn from the 2002 main NAEP writing assessment was significantly different from all main NAEP writing students on several dimensions. The study sample had a greater percentage of White students, a smaller percentage of Hispanic students, a smaller percentage of students whose parents' highest education level was unavailable, and a greater percentage of rural students than the main NAEP writing assessment as a whole. (See Appendix B for definitions of these groups.)

Similarly, the WOL sample drawn from main NAEP reading differed from the sample taking the main NAEP reading assessment. The WOL sample had greater percentages of White students, students with one or more parents having some education after high school, and rural students. The WOL sample also had smaller percentages of Asian/Pacific Islander students, students whose parents' highest education level was graduation from high school, students with parents having less than a high school education, and students whose parents' highest level of education was unavailable.

Table 2-4. Characteristics of study samples taking the Writing Online test compared with all students taking main NAEP, grade 8: 2002

Characteristic	All main NAEP writing students	WOL students drawn from main NAEP writing and responding to both essays on computer	All main NAEP reading students	WOL students drawn from main NAEP reading and responding to both essays on computer
Number of students	118,516	687	115,176	568
NAEP writing mean	153 (0.5)	157 (2.0)	†	†
NAEP reading mean	†	†	264 (0.4)	267 (1.9)
		Percent of students		
Exclusion rate	4 (0.2)	5 (1.3)[1]	5 (0.3)	4 (1.0)[1]
Gender				
Male	50 (0.3)	52 (1.8)	50 (0.3)	51 (2.5)
Female	50 (0.3)	47 (1.9)	50 (0.3)	48 (2.0)
Race/ethnicity				
White	65 (0.5)	69 (0.8)*	65 (0.5)	69 (1.0)*
Black	15 (0.4)	15 (0.6)	15 (0.4)	14 (0.7)
Hispanic	14 (0.4)	11 (0.6)*	14 (0.4)	13 (0.7)
Asian/Pacific Islander	4 (0.2)	4 (0.6)	4 (0.2)	2 (0.5)*
Other	2 (0.1)	2 (0.5)	2 (0.1)	2 (0.8)
Type of school				
Public	91 (0.2)	92 (1.1)	91 (0.2)	91 (1.0)
Nonpublic	9 (0.2)	8 (1.1)	9 (0.2)	9 (1.0)
Parents' highest level of education				
Less than high school	6 (0.1)	6 (1.1)	6 (0.2)	4 (0.9)*
Graduated high school	17 (0.2)	17 (1.7)	17 (0.2)	13 (1.2)*
Some education after high school	18 (0.2)	20 (1.6)	19 (0.3)	24 (2.0)*
Graduated college	46 (0.4)	48 (1.9)	46 (0.5)	50 (2.4)
Unavailable	12 (0.2)	10 (1.1)*	12 (0.2)	10 (1.3)*
Student eligibility for free/reduced-price school lunch				
Eligible	31 (0.6)	28 (2.5)	31 (0.6)	29 (2.9)
Not eligible	53 (1.0)	58 (3.0)	54 (1.0)	57 (3.2)
Unavailable	16 (0.8)	14 (2.6)	16 (0.9)	14 (2.5)
Type of school location				
Central city	29 (0.6)	28 (1.5)	29 (0.6)	27 (1.3)
Urban fringe/large town	42 (0.7)	38 (1.9)	42 (0.7)	39 (1.8)
Rural/small town	29 (0.5)	34 (1.8)*	29 (0.5)	35 (1.7)*

† Not applicable.

*$p < .05$ for the difference between the study sample and all students administered the relevant main NAEP assessment as computed from a *t*-test for independent samples.

[1] "Exclusion rate" is the weighted sum of the excluded students divided by the excluded plus the assessed students. This rate is based on all students who were sampled for inclusion in the study.

NOTE: WOL = Writing Online. All values are weighted, except for the sample sizes. "Other" category for race/ethnicity includes American Indian/Alaska Native and unclassified students. Detail may not sum to totals because of rounding. Standard errors are in parentheses.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Writing Online Study.

The data in table 2-4 suggest that the study samples diverge from the nationally representative main NAEP samples on one or more characteristics, depending upon the sample. How comparable are the study samples to one another on these same characteristics? Table 2-5 compares the samples responding to both essays on the WOL computer test with the main NAEP writing sample responding to the same two essays on paper. As the table indicates, the WOL computer samples significantly differ from the paper comparison sample on several characteristics. The computer samples had somewhat greater percentages of male students, White students, and students in rural/small town locations, but smaller percentages of female students and of students in urban fringe/large town locations. One of the samples also had a smaller percentage of Hispanic students, one had a smaller percentage of students who reported that at least one parent had graduated from high school, and one had a smaller percentage of students for whom the level of parents' education was unavailable. To deal with these differences, many of the study's analyses were run with gender as one of the independent variables to control for its effects, as this characteristic appeared to be associated with the largest differences between the paper and computer samples. Similarly, the main study question of whether delivery mode causes differences in mean performance was analyzed with each of the background variables from table 2-5 included in turn as an independent variable.

Table 2-5. Characteristics of study samples taking the Writing Online computer test compared with the main NAEP writing study sample responding to the same essays on paper, grade 8: 2002

Characteristic	Main NAEP writing students responding to both paper-and-pencil essays in the same order as WOL	WOL Students		
		All students responding to both essays on computer	Students drawn from main NAEP writing and responding to both essays on computer	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 (1.4)	†	157 (2.0)	†
NAEP reading mean	†	†	†	267 (1.9)
	Percent of students			
Exclusion rate[1]	3 (0.4)	5 (1.0)	5 (1.3)	4 (1.0)
Gender				
Male	45 (1.5)	52 (1.7)*	52 (1.8)*	51 (2.5)*
Female	54 (1.6)	47 (1.4)*	47 (1.9)*	48 (2.0)*
Race/ethnicity				
White	65 (1.6)	69 (0.7)*	69 (0.8)*	69 (1.0)*
Black	16 (1.5)	14 (0.5)	15 (0.6)	14 (0.7)
Hispanic	15 (1.2)	12 (0.5)	11 (0.6)*	13 (0.7)
Asian/Pacific Islander	4 (0.8)	3 (0.4)	4 (0.6)	2 (0.5)
Other	1 (0.3)	2 (0.5)	2 (0.5)	2 (0.8)
Type of school				
Public	90 (0.8)	92 (0.9)	92 (1.1)	91 (1.0)
Nonpublic	10 (0.8)	8 (0.8)	8 (1.1)	9 (1.0)
Parents' highest level of education				
Less than high school	6 (0.7)	5 (0.9)	6 (1.1)	4 (0.9)
Graduated high school	17 (1.1)	15 (1.1)	17 (1.7)	13 (1.2)*
Some education after high school	19 (1.5)	21 (1.2)	20 (1.6)	24 (2.0)
Graduated college	46 (1.7)	49 (1.7)	48 (1.9)	50 (2.4)
Unavailable	13 (1.1)	10 (0.7)*	10 (1.1)	10 (1.3)
Student eligibility for free/reduced-price school lunch				
Eligible	30 (1.4)	28 (2.4)	28 (2.5)	29 (2.9)
Not eligible	54 (1.7)	58 (2.8)	58 (3.0)	57 (3.2)
Unavailable	15 (1.3)	14 (2.3)	14 (2.6)	14 (2.5)
Type of school location				
Central city	28 (1.2)	28 (1.2)	28 (1.5)	27 (1.3)
Urban fringe/large town	43 (1.5)	38 (1.5)*	38 (1.9)*	39 (1.8)*
Rural/small town	29 (0.9)	34 (1.4)*	34 (1.8)*	35 (1.7)*

† Not applicable.

* $p < .05$ for the difference between the WOL sample and the paper comparison group as computed from a *t*-test for independent samples (e.g., between the percentage of all WOL students who were White and the percentage of main NAEP writing students responding to both paper-and-pencil essays in WOL order who were White).

[1] "Exclusion rate" is the weighted sum of the excluded students divided by the excluded plus the assessed students. For all main NAEP writing students, this rate is based on all students who were sampled to receive the test booklet containing the two paper-and-pencil essays given in the same order as the WOL essays. For WOL students, this rate is based on all students who were sampled for inclusion in the study.

NOTE: WOL = Writing Online. All values are weighted, except for the sample sizes. "Other" category for race/ethnicity includes American Indian/Alaska Native and unclassified students. Detail may not sum to totals because of rounding. Standard errors are in parentheses.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Writing Online Study.

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 sessions, students responded to 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, those students sampled for the Writing Online (WOL) study took the following components in a single session:

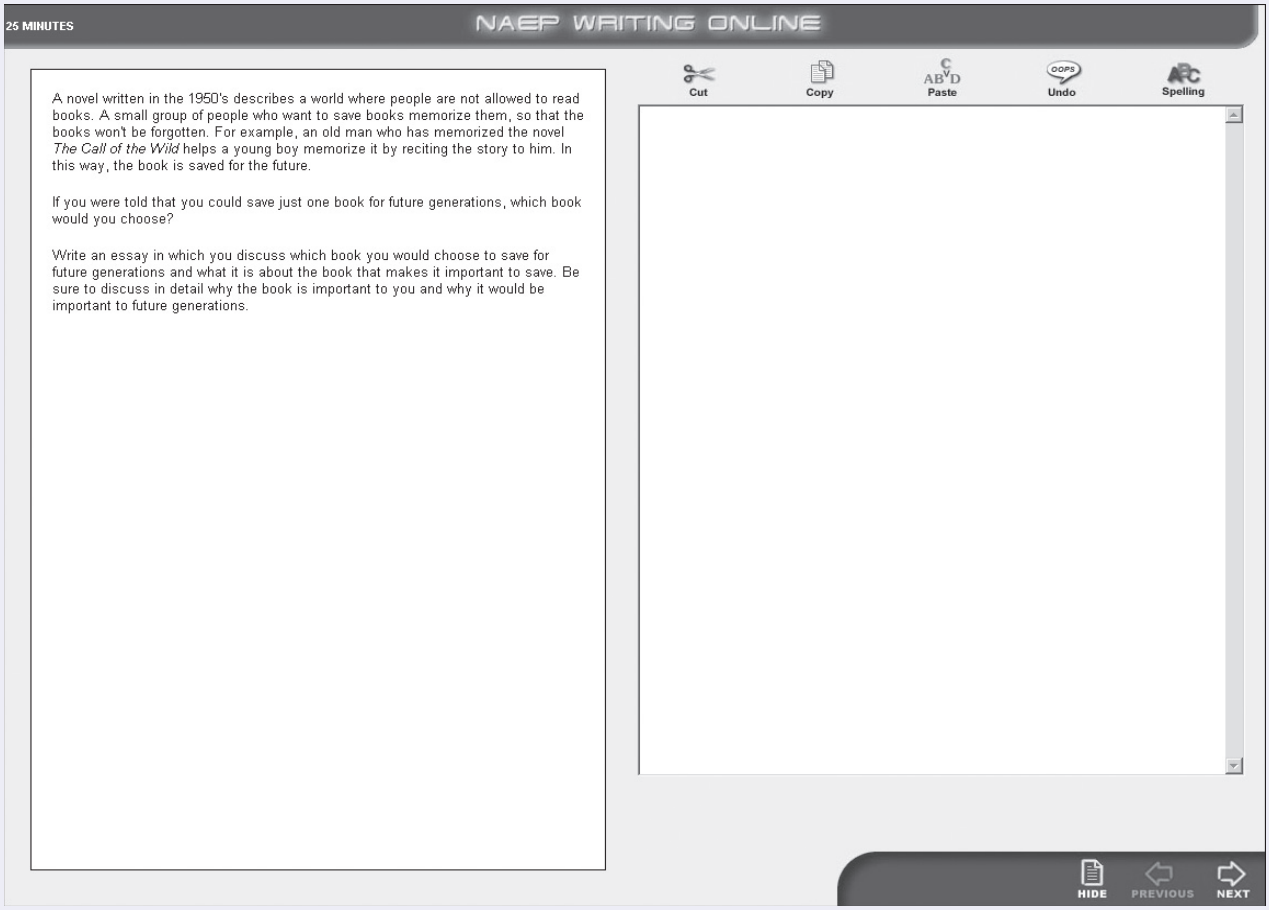
- ***Online tutorial.*** The online tutorial showed students how to use the computer to respond to the essay tasks. The tutorial provided instruction and practice in the use of the mouse and scrolling, presented information about the test interface and how to navigate from one question to the next, and described the functions of the WOL word processor (cut, copy, paste, undo, and spell-check). 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 on the NCES website (http://nces.ed.gov/nationsreportcard/studies/tbatutorial.asp#wol).
- ***Online computer skills measure.*** The computer skills measure was administered to evaluate students' facility with the computer and, specifically, word processing. 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. (See appendix C and appendix D.)
- ***Two online essays.*** 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. Figure 2-1 shows the WOL computer interface and the first essay task.

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 NAEP paper-and-pencil writing test 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 NAEP paper-and-pencil writing test administration.

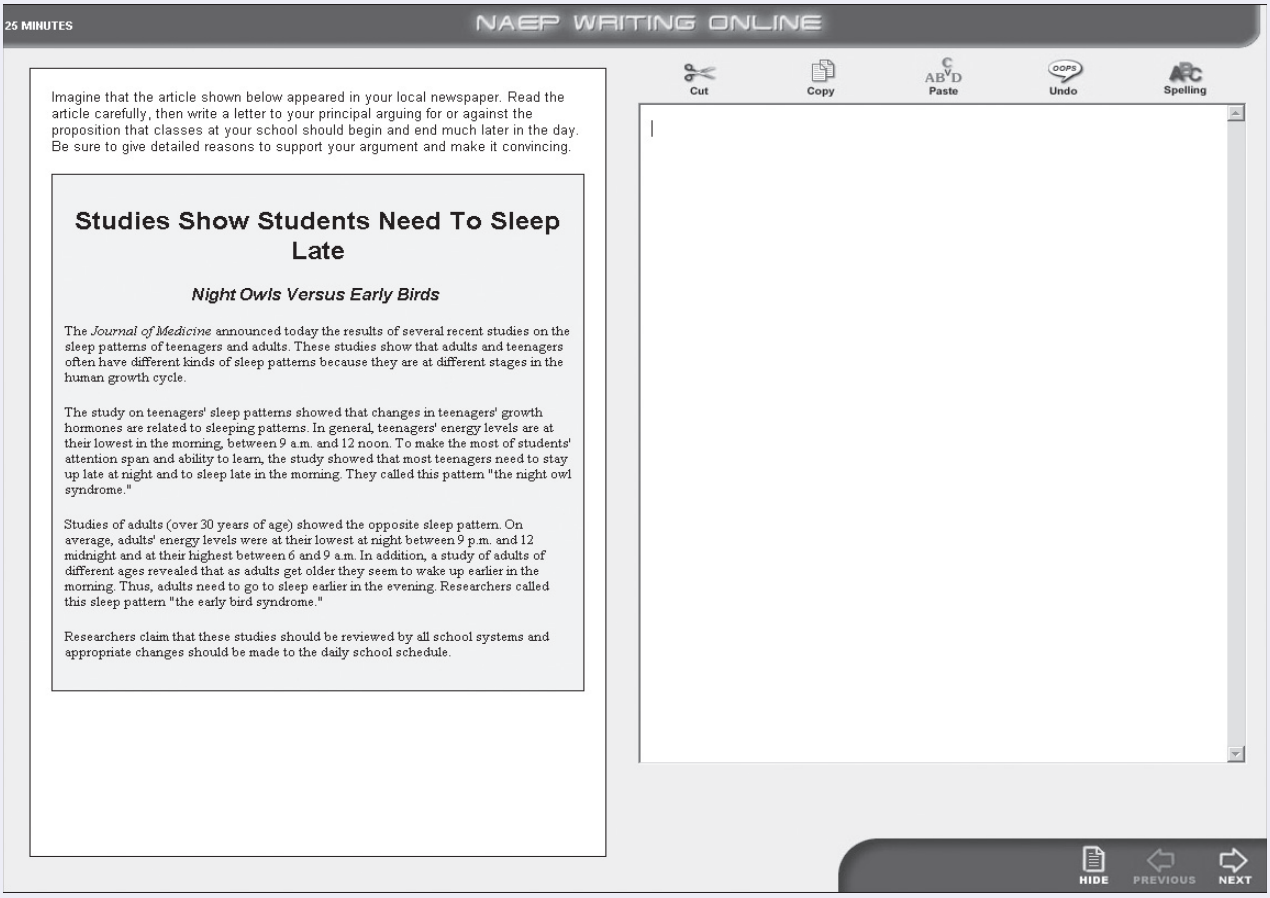
Both WOL essays were drawn from the 2002 main NAEP writing assessment and administered to students in the same order as in that assessment. For "Save a Book," an informative writing task, students were asked to explain what book they would preserve through memorization if they lived in a society where 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. Figure 2-2 shows "School Schedule."

Figure 2-1. The Writing Online computer interface showing the "Save a Book" essay, grade 8: 2002



SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Writing Online Study.

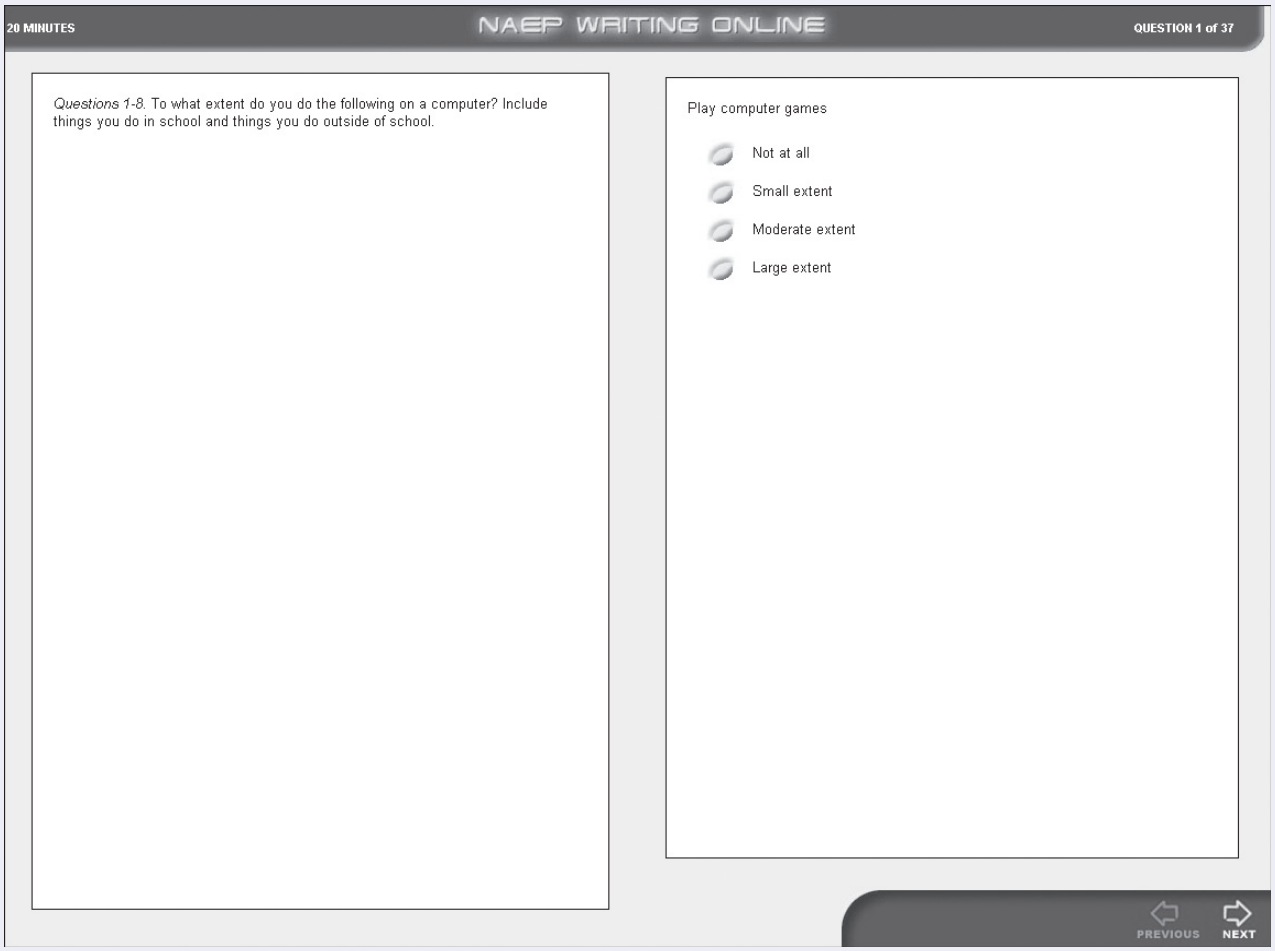
Figure 2-2. The Writing Online computer interface showing the "School Schedule" essay, grade 8: 2002



SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Writing Online Study.

- ***Online background questions.*** Students were allowed 20 minutes to complete the background questions section, which consisted of 37 questions: 10 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. (See appendix E for the specific text of the background questions.) Background questions appeared on the screen, and students were directed to click on the bubble next to their selected response. Students were able to move forward or backward throughout this section by clicking on the "Next" and "Previous" buttons. A counter in the upper-right corner of the screen indicated which question they were answering, for example, "27 of 37 questions." Figure 2-3 shows a sample background question screen.

Figure 2-3. Sample Writing Online background question screen, grade 8: 2002



SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Writing Online Study.

The following components were administered to the study participants who took the main NAEP 2002 paper writing assessment, but who did not take WOL:

- ***Two writing tasks.*** Each student was given a brochure entitled "Ideas for planning and reviewing your writing," which was the same brochure as that 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 2-6 summarizes the instruments used in the WOL study and the student samples that took each instrument.

Table 2-6. Instruments administered to each student sample, grade 8: 2002

Sample taking main NAEP writing and WOL	Sample taking main NAEP reading and WOL	Sample taking only main NAEP writing
Main NAEP administrations (January–March 2002)		
Paper test with two essays	Paper test with two blocks (9-13 items each)	Paper test with two essays
Background questions (53 items)	Background questions (29 of the 53 items administered to main NAEP writing students)	Background questions (53 items)
WOL administrations (April–May 2002)		
Online tutorial	Online tutorial	†
Online computer skills measure	Online computer skills measure	†
Online test with two essays	Online test with two essays	†
Background questions (37 items)	Background questions (37 items)	†

† Not applicable.

NOTE: WOL=Writing Online.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Writing Online Study.

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. 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 the typed versions of students' responses. This WOL scoring session used 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 writing scoring procedures. "Save a Book" was scored during one week in August 2002, and "School Schedule" was scored during one week in November 2002. Training for scoring each task was conducted by staff members who have extensive experience with scoring main NAEP writing. During the WOL scoring session, whenever useful for explication or clarification, training papers were supplemented with examples from the WOL responses to the tasks.

Reader training began with careful explanation of the anchor papers, which are tied directly to the scoring guide (see appendix F for NAEP writing scoring guides) and are intended as exemplars of each score level. Following discussion of the anchor papers, readers worked through practice sets and consensus-building sets, all designed to increase scorers' ability to score consistently and reliably, first as a group, then individually. Prior to scoring "live" WOL responses, readers took a qualifying test to determine their readiness for scoring. Once actual scoring began, readers generally worked in pairs or small groups until the trainer determined that they were maintaining a consistent level of agreement, at which time they began scoring individually. Throughout the scoring process, the trainer monitored reader agreement and intervened, if necessary, to recalibrate readers.

To evaluate reader reliability, a random sample of WOL responses was double-scored and compared to the double-scored responses of those students in the study sample who had taken the same two essays on paper in main NAEP. Table 2-7 presents the intraclass correlations between two readers for "Save a Book" and "School Schedule." As the table shows, the correlations for WOL appear lower than those for main NAEP writing, which indicates that for those responses that were double-scored, the WOL readers agreed with one another in rank ordering individuals to a lesser degree than did the main NAEP readers. Table 2-8 shows the percentage exact agreement between

two readers. The agreement percentages are accompanied by a statistic, "kappa," which corrects for the level of agreement expected by chance (Fleiss 1981). Here, the percentages appear to be lower for WOL than for main NAEP writing, suggesting that, for double-scored responses, the two WOL readers did not assign the same score to a given individual as often as 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).

The above analysis indicates that the WOL readers scored student responses with lower levels of agreement than did the main NAEP readers. Such 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.[4] Thus, despite lower levels of reader reliability, the score reliabilities across the two samples are reasonably close to one another.

Table 2-7. Intraclass correlations between two readers for Writing Online and for main NAEP writing, grade 8: 2002

Measure	Save a Book	School Schedule
WOL	.81	.88
Main NAEP writing	.87	.94

NOTE: WOL = Writing Online. For WOL, the number of students responding was 310 for "Save a Book" and 309 for "School Schedule." For main NAEP writing, the numbers were 129 and 159, respectively.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Writing Online Study.

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 generally award different scores to typed essays as compared with handwritten versions of the same essays. In most studies, readers have given lower scores to the typed versions (Powers and Farnum 1997; Powers, Fowles, Farnum, and Ramsey 1994; Russell and Tao 2004a; Russell and Tao 2004b), though other studies have reported either mixed or null results (Harrington, Shermis, and Rollins 2000; MacCann, Eastment, and Pickering 2002). To evaluate whether there was such a bias in this study, a sample of handwritten student responses from the main NAEP 2002 writing assessment was drawn separately for each essay and keyed into the WOL online scoring system. 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

Table 2-8. Percentage exact agreement between two readers for Writing Online and for main NAEP writing, grade 8: 2002

	Save a Book		School Schedule	
Measure	Percent exact agreement	Kappa	Percent exact agreement	Kappa
WOL	60	.47	63	.53
Main NAEP writing	72	.62	84	.79

NOTE: WOL = Writing Online. For WOL, the number of students responding was 310 for "Save a Book" and 309 for "School Schedule." For main NAEP writing, the numbers were 129 and 159, respectively.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Writing Online Study.

[4] The uncorrected correlations were .63 for WOL and .57 for main NAEP. Corrections were computed using the Spearman-Brown formula (Thorndike 1982).

entered by students online.[5] Table 2-9 shows the unweighted mean scores assigned to the same essays when presented to main NAEP readers in handwritten form and then to WOL readers in typed form.

Table 2-9. Unweighted means and standard deviations for the same main NAEP writing responses presented to different groups of readers in handwritten and in typed form, grade 8: 2002

Essay	Handwritten	Typed
Save a Book		
Mean	3.5	3.4
Standard Deviation	1.7	1.5
School Schedule		
Mean	3.5	3.6
Standard Deviation	1.7	1.5

NOTE: Responses were drawn from students taking the 2002 paper main NAEP writing assessment. All responses were transcribed from handwritten to typed form. The number of responses for "Save a Book" was 294, and the number for "School Schedule" was 292. The same group of students did not respond to both essays.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Writing Online Study.

These means 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 format factor. Results showed no significant difference for presentation format (F,1,584 = 0.37, $p > .05$), indicating that, summing across the two essays, the scores for the handwritten and typed formats did not differ measurably. However, there was a significant format-by-essay interaction (F,1,584 = 10.97, $p < .05$), suggesting that the size of the score difference between the formats was not the same for the two essay questions. Posthoc, dependent-samples t-tests (one-tailed) between the scores for the typed and handwritten responses showed that the typed responses were scored lower than the handwritten versions of the same essays for "Save a Book" (t, 293 = 2.05, $p < .05$), but higher than the handwritten versions for "School Schedule" (t, 291 = -2.61, $p < .05$). In both cases, the effect sizes in standard deviation units of the handwritten group were very small: .05 for "Save a Book" and .07 for "School Schedule."

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 main NAEP writing 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. Weighted means were compared using a repeated-measures analysis of variance.[6] In this analysis, the independent variables were the WOL group (reading and writing) and essay, with repeated measures on the essay factor. Essay score was the dependent measure. The analysis was run using only the 1,255 students who responded to both essays.

Table 2-10 gives the mean scores for the two groups on each essay. Results of the statistical tests showed no between-subjects main effect for the 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 performance of the groups, they were combined where appropriate for the analyses subsequently presented in this report.

Table 2-10. Mean scores for students drawn from main NAEP writing and from main NAEP reading on the Writing Online test, grade 8: 2002

Essay	WOL main NAEP writing	WOL main NAEP reading
Save a Book	3.6 (0.05)	3.5 (0.06)
School Schedule	3.5 (0.06)	3.4 (0.06)

NOTE: WOL = Writing Online. The number of students was 1,255, with 687 drawn from the main NAEP writing assessment sample and 568 from the main NAEP reading assessment sample. Standard errors are in parentheses.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Writing Online Study.

[5] Five off-topic responses were removed from this data set, as such responses are not considered in the main analyses presented later in this report.

[6] All repeated-measures ANOVAs that used sampling weights were run using WESVAR, proprietary software of Westat, which accounts for the clustered nature of NAEP samples. See appendix G for a description of the use of WESVAR.

3. Measurement Issues

This section considers how the mode of administering a writing assessment (i.e., computer vs. paper) affects the inferences that can be drawn about students' writing skill. This issue is explored by evaluating whether students perform differently across the two delivery modes:

- Do students score differently on a computer test versus a paper test?
- Do students write essays of different lengths in these two delivery modes?
- Do more students respond validly in one or the other mode?

Performance Differences Across Assessment Modes

Very few studies of the effect of mode on writing test performance have been conducted at the K-12 level. Moreover, the studies that are available generally use small, nonrepresentative samples. Even so, the results suggest that mode does have an impact on test score. For example, two studies (Russell and Haney 1997; Russell and Plati 2000) 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 computers 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.

Two studies with older students taking admissions tests also show evidence of overall mode effects. For a large group of Test of English as a Foreign Language (TOEFL®) 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 essays in each mode, students performed better on the paper than on the computer tests (Bridgeman and Cooper 1998).

Are computer and paper writing tests comparable for eighth-graders nationally? To address this question, three indicators were compared across delivery modes: essay score, essay length, and the frequency of valid responses.

Essay Score

Perhaps the most direct approach to evaluating the effect of delivery mode on performance can be provided by comparing mean scores on WOL with the mean scores from a different, but representative, group of students taking the same essays in the paper-and-pencil main NAEP writing assessment. To test the difference between means, a repeated-measures analysis of variance (ANOVA) was conducted. For this analysis, delivery mode and essay were the independent variables, and essay score was the dependent variable, with repeated measures on the essay factor. Table 3-1 gives the mean scores for each group on each essay, where scores are on a scale of 1 to 6.

Table 3-1. Mean scores for students responding to Writing Online and for a different group of students responding to the same essays on paper in the main NAEP writing assessment, grade 8: 2002

Essay	WOL	Paper and Pencil
Save a Book	3.5 (0.04)	3.6 (0.03)
School Schedule	3.5 (0.05)	3.6 (0.04)

NOTE: WOL = Writing Online. The number of students responding to both essays was 4,133, with 1,255 taking the WOL computer test and 2,878 taking the paper main NAEP writing assessment. Students were included only if they responded to both essays and, for main NAEP writing, only if the tasks were administered in the same order as those given in WOL. Standard errors are in parentheses.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Writing Online Study.

The results of this analysis did not detect a significant effect for delivery mode (F,1,62 = 3.39, $p > .05$) or a significant interaction of delivery mode with essay (F,1,62 = 0.29, $p > .05$). 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. The results, reported in the Equity Issues section, also showed no significant effect 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. From the paper main NAEP writing assessment, a random sample of handwritten responses was transcribed to electronic form for each essay task. For WOL, all responses were already in electronic form. In this analysis, the same students did not necessarily respond to both essays, and different groups took the paper and computer tests. Table 3-2 gives the unweighted mean word counts for each essay by delivery mode.

Table 3-2. Unweighted mean word count for students responding to Writing Online and for a different group of students responding to the same essays on paper in the main NAEP writing assessment, grade 8: 2002

Essay	WOL	Paper and Pencil
Save a Book	185 (2.9)	175 (6.0)
School Schedule	162 (2.6)	166 (5.4)

NOTE: WOL = Writing Online. 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 errors are in parentheses.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Writing Online Study.

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.[7] Results showed that there was no effect of delivery mode on word count for "Save a Book" (F,1,1547 = 2.34, $p > .05$) or for "School Schedule" (F,1,1545 = 0.46, $p > .05$). Thus, there were no measurable differences in the number of words written on computer as compared with paper tests. These analyses were repeated, controlling for gender. The repeated analyses, which are reported in the Equity Issues section, also showed no main effect for delivery mode.

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 3-3 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 3-3. 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, grade 8: 2002

Essay	WOL	Paper and Pencil
Save a Book	98 (0.5)	98 (0.4)
School Schedule	97 (0.5)	99 (0.2)

NOTE: WOL = Writing Online. The number of students administered both essays was 4,291, with 1,308 taking the WOL computer test and 2,983 taking the paper main NAEP writing assessment. Main NAEP writing students were included only if they were administered both essays in the same order as those given in WOL. Standard errors are in parentheses.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Writing Online Study.

To examine differences in responding more closely, 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, $p > .05$).[8] 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 and are described in the Equity Issues section.

[7] Student weights were not used because appropriate weights were not available for the sample of students whose handwritten responses had been transcribed to electronic form. The SAS generalized linear model (GLM) procedure was used to conduct this analysis.

[8] These logistic regressions were computed using WESVAR, which provides F-statistics.

4. Equity Issues

This section considers three questions:

- How do population groups perform, and do mode effects vary across groups?
- Are students disadvantaged if they must take a writing test on a NAEP laptop instead of a school computer?
- How are students with different levels of computer experience affected by computer- versus paper-based writing assessments?

Population Group Performance

To date, the performance of population groups on computer compared with paper writing tests has not been widely studied. In a small-sample study, Russell and Haney (1997) found that the differences in performance on computer versus paper writing tests were similar for male and female middle-school students. Among a large sample of prospective business school students, Bridgeman and Cooper (1998) found no interactions between delivery mode and population groups defined by gender, race/ethnicity, or whether English was their first language.

Gender

For gender, delivery mode was evaluated in terms of its effects on essay score, response length, and frequency of valid responding. (The latter two performance indicators are presented because gender was included in the model when the overall effects on these performance indicators were evaluated in the previous section.) Table 4-1 presents mean scores for WOL and for the paper main NAEP writing assessment by gender.

Table 4-1. Mean scores for students drawn from main NAEP who took the Writing Online computer test and for students responding to the same essays on paper in the main NAEP writing assessment, by gender and essay, grade 8: 2002

Gender	WOL	Paper and pencil
Save a Book		
Male	3.3 (0.05)	3.4 (0.04)
Female	3.8 (0.06)	3.8 (0.06)
School Schedule		
Male	3.3 (0.06)	3.3 (0.05)
Female	3.7 (0.06)	3.8 (0.04)

NOTE: WOL = Writing Online. The number of students responding to both essays was 4,116, with 1,249 taking the WOL computer test and 2,867 taking the paper main NAEP writing assessment. Students were included only if they responded to both essays and, for main NAEP writing, only if the tasks were administered in the same order as those given in WOL. Standard errors are in parentheses.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Writing Online Study.

To test for the presence of gender effects, a repeated-measures ANOVA was conducted with delivery mode, gender, and essay as the independent variables; essay score as the dependent variable; and repeated measures on the essay factor. The between-groups results showed no effect for delivery mode (F,1,62 = 1.23, $p > .05$), 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 one gender group more than the other.

Table 4-2 shows mean essay length by gender for students responding to WOL and for a random sample of responses to the same essay tasks drawn from students taking the paper main NAEP writing assessment. In the latter sample, the same group of students did not necessarily respond to both essays.

Table 4-2. Unweighted mean word count for students responding to Writing Online and for a different group of students responding to the same essays on paper in the main NAEP writing assessment, by gender and essay, grade 8: 2002

	Mean word count	
Gender	WOL	Paper and pencil
Save a Book		
Male	164 (3.6)	148 (8.2)
Female	209 (4.5)	204 (8.1)
School Schedule		
Male	145 (3.3)	132 (6.8)
Female	181 (3.9)	195 (7.5)

NOTE: WOL = Writing Online. The number of responses for "Save a Book" was 294 for paper main NAEP writing and 1,249 for WOL. The number of responses for "School Schedule" was 292 for paper main NAEP writing and 1,249 for WOL. The same main NAEP students did not necessarily respond to both essays. Standard errors are in parentheses.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Writing Online Study.

To test the impact on essay length, a separate unweighted ANOVA was conducted for each essay, with the number of words serving as the dependent variable.[9] The independent variables were delivery mode and gender. For the students taking the paper main NAEP writing assessment, the sample size is relatively small due to the need to key enter paper responses and the cost of doing so. As a consequence, the power of these analyses to detect differences in essay length is lower than it otherwise would be. For "Save a Book," there was a significant effect for gender, with female students producing more words than male students (F,1,1539 = 85.26, $p < .05$), but no effect of delivery mode on word count (F,1,1539 = 2.45, $p > .05$), and no significant delivery mode-by-gender interaction (F,1,1539 = 0.79, $p > .05$). For "School Schedule," there was the same significant effect of female students writing longer essays than male students (F,1,1537 = 81.81, $p < .05$), and no main effect for delivery mode (F,1,1537 = 0.46, $p > .05$). However, there was a significant delivery-mode-by-gender interaction (F,1,1537 = 5.27, $p < .05$). This interaction indicates that delivery mode affects essay length differently for male students and female students for "School Schedule." One-tailed post-hoc tests showed that, for "School Schedule," male students wrote significantly fewer words in the paper test condition than on the computer test (t, 785 = 1.77, $p < .05$), while female students showed no such difference (t, 752 = -1.59, $p > .05$). However, although male students' paper essays were about 11 percent shorter than their computer-generated ones, there was no corresponding significant difference in their mean scores across delivery modes for this essay, as described above and shown in table 4-1.

Table 4-3. 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, by gender and essay, grade 8: 2002

	Percent of students	
Gender	WOL	Paper and pencil
Save a Book		
Male	97 (0.7)	97 (0.8)
Female	99 (0.5)	99 (0.2)
School Schedule		
Male	97 (0.6)	98 (0.5)
Female	97 (0.7)	99 (0.2)

NOTE: WOL = Writing Online. The number of students administered both essays was 4,274, with 1,302 taking the WOL computer test and 2,972 taking the paper main NAEP writing assessment. Main NAEP writing students were included only if they were administered both essays in the same order as those given in WOL. Standard errors are in parentheses.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Writing Online Study.

Finally, table 4-3 shows the response rates for male and female students taking WOL compared to those taking the paper main NAEP writing assessment.

To examine differences in responding more closely, separate logistic regressions were conducted for each essay with delivery mode and gender as the independent variables. The dependent variable was whether or not there was a response to the essay. Results for "Save a Book" showed an expected significant effect for gender (F,1,62 = 21.55, $p < .05$), no main effect for delivery mode (F,1,62 = 1.97,

[9] The SAS GLM procedure was used to conduct this analysis.

$p > .05$), and no significant effect for the interaction of gender and delivery mode (F,1,62 = 2.47, $p > .05$). For "School Schedule," the gender main effect (F,1,62 = 5.53, $p < .05$) was significant, but, more importantly for the purposes of this study, so was the interaction of gender and delivery mode (F,1,62 = 8.58, $p < .05$), indicating that the difference in response rates for paper and computer was not the same for males and females. Finally, consistent with the response rate analysis reported for "School Schedule" in the Measurement Issues section, which did not include gender, there was a significant main effect for delivery mode itself (F,1,62 = 16.08, $p < .05$). Post-hoc tests showed that a significantly greater percentage of females gave valid responses to "School Schedule" on paper than on computer (F,1,62 = 17.61, $p < .05$), by about 2 percentage points.

Other NAEP Reporting Groups

Direct comparisons across modes can be made for other NAEP reporting groups. Such comparisons were made separately for 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). (A complete description of NAEP reporting groups is available in appendix B.) Because the sample sizes for some of these groups were small, differences may not always be statistically significant even if they are seemingly large. It is not possible to distinguish for these instances whether the apparent difference is a reflection of population performance, or alternatively, an artifact of sample selection.

Population group comparisons were made only for essay score. For each comparison, a repeated-measures ANOVA was conducted, similar to the analysis for gender. For this analysis, the independent variables were the NAEP reporting group of interest, delivery mode, gender, and essay, with repeated measures on the essay factor. Essay score was the dependent variable. Gender was included as an independent variable in all of the models to control for differences between the WOL and the main NAEP writing 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., all of the parent education levels). For all study samples, the ANOVA was restricted to WOL and main NAEP writing students and, in the case of main NAEP writing, to those students who were administered essays on paper given in the same order as those in WOL.

Race/ethnicity. Table 4-4 gives the mean scores by race/ethnicity. Because gender was included in the model and some students were missing gender designations, the statistical test of the means was conducted on a slightly smaller number of students (n = 4,116) than the one used to compute the means in the table (n = 4,133). Results of the ANOVA showed a significant between-groups effect for race (F,4,59

Table 4-4. Mean scores for students drawn from main NAEP who took the Writing Online test and for students responding to the same essays on paper in the main NAEP writing assessment, by race/ethnicity and essay, grade 8: 2002

Race/ethnicity	WOL	Paper and pencil
Save a Book		
White	3.7 (0.05)	3.8 (0.04)
Black	2.9 (0.10)	3.3 (0.08)
Hispanic	3.0 (0.09)	3.2 (0.12)
Asian/Pacific Islander	3.8 (0.28)	4.0 (0.18)
Other	3.3 (0.30)	3.4 (0.38)
School Schedule		
White	3.7 (0.06)	3.7 (0.03)
Black	2.8 (0.09)	3.2 (0.13)
Hispanic	2.9 (0.10)	3.1 (0.14)
Asian/Pacific Islander	3.8 (0.30)	4.1 (0.18)
Other	3.4 (0.27)	3.4 (0.18)

NOTE: WOL = Writing Online. The number of students was 4,133, with 1,255 taking the WOL computer test and 2,878 taking the paper main NAEP writing assessment. Students were included only if they responded to both essays and, for main NAEP writing, only if the tasks were administered in the same order as those in WOL. "Other" category for race/ethnicity includes American Indian/Alaska Native and unclassified students. Standard errors are in parentheses.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Writing Online Study.

= 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 4-5 gives the mean scores by parents' education level, where that level is the higher of the levels reported by the student for his or her mother or father. Differences between the means were tested for the slightly smaller subset of students with gender designations (n = 4,116). 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, $p > .05$).

Table 4-5. Mean scores for students drawn from main NAEP who took the Writing Online test and for students responding to the same essays on paper in the main NAEP writing assessment, by parents' highest level of education and essay, grade 8: 2002

Parents' highest education level	WOL	Paper and pencil
Save a Book		
High school degree or less	3.3 (0.07)	3.3 (0.07)
More than high school degree	3.6 (0.05)	3.9 (0.04)
Unavailable	3.1 (0.11)	3.0 (0.09)
School Schedule		
High school degree or less	3.2 (0.08)	3.2 (0.06)
More than high school degree	3.6 (0.06)	3.8 (0.03)
Unavailable	3.0 (0.11)	2.8 (0.09)

NOTE: WOL = Writing Online. The number of students was 4,133, with 1,255 taking the WOL computer test and 2,878 taking the paper main NAEP writing assessment. Students were included only if they responded to both essays and, for main NAEP writing, only if the tasks were administered in the same order as those in WOL. "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 errors are in parentheses.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Writing Online Study.

School location. Table 4-6 gives the mean scores by type of school location. Here, too, the statistical tests were computed for the subset of students with gender designations ($n = 4{,}116$). 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 ($F,1,62 = 5.05$, $p < .05$).[10] 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).[11] 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.

Table 4-6. Mean scores for students drawn from main NAEP who took the Writing Online test and for students responding to the same essays on paper in the main NAEP writing assessment, by school location and essay, grade 8: 2002

School location	WOL	Paper and pencil
Save a Book		
Central city	3.3 (0.09)	3.5 (0.06)
Urban fringe/large town	3.6 (0.08)	3.7 (0.05)
Rural/small town	3.7 (0.05)	3.6 (0.04)
School Schedule		
Central city	3.3 (0.09)	3.4 (0.07)
Urban fringe/large town	3.5 (0.08)	3.7 (0.06)
Rural/small town	3.6 (0.09)	3.4 (0.03)

NOTE: WOL = Writing Online. The number of students was 4,133, with 1,255 taking the WOL computer test and 2,878 taking the paper main NAEP writing assessment. Students were included only if they responded to both essays and, for main NAEP writing, only if the tasks were administered in the same order as those in WOL. Standard errors are in parentheses.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Writing Online Study.

[10] 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.

[11] Cohen (1988) suggests, as a rule of thumb, that .2 be considered a minimum for "small" effects, .5 a minimum for "medium" effects, and .8 a minimum for "large" effects.

Eligibility for free/reduced-price school lunch. Table 4-7 gives the mean scores by eligibility for free/reduced-price school lunch. As in the other population group analyses, the means were tested only for those students with gender designations ($n = 4{,}116$). 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, $p > .05$).

Table 4-7. Mean scores for students drawn from main NAEP who took the Writing Online test and for students responding to the same essays on paper in the main NAEP writing assessment, by student eligibility for free/reduced-price school lunch and essay, grade 8: 2002

Student eligibility for free/ reduced-price school lunch	WOL	Paper and pencil
Save a Book		
Eligible	3.1 (0.06)	3.2 (0.06)
Not eligible	3.8 (0.05)	3.8 (0.05)
Unavailable	3.4 (0.17)	3.9 (0.09)
School Schedule		
Eligible	3.1 (0.06)	3.1 (0.06)
Not eligible	3.7 (0.07)	3.7 (0.04)
Unavailable	3.2 (0.16)	3.9 (0.11)

NOTE: WOL = Writing Online. The number of students was 4,133, with 1,255 taking the WOL computer test and 2,878 taking the paper main NAEP writing assessment. Students were included only if they responded to both essays and, for main NAEP writing, only if the tasks were administered in the same order as those in WOL. "Unavailable" includes students with missing data for this variable. Standard errors are in parentheses.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Writing Online Study.

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, which controls for eligibility for free/reduced-price school lunch and gender, delivery mode was no longer significant (F,1,62 = 2.22, $p > .05$).

School type. The mean scores by school type are presented in table 4-8. Between-groups results for the subset of students with gender designations ($n = 4{,}116$) 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, $p > .05$).

Table 4-8. Mean scores for students drawn from main NAEP who took the Writing Online test and for students responding to the same essays on paper in the main NAEP writing assessment, by school type and essay, grade 8: 2002

School type	WOL	Paper and pencil
Save a Book		
Public	3.5 (0.04)	3.6 (0.03)
Nonpublic	3.6 (0.26)	4.0 (0.11)
School Schedule		
Public	3.5 (0.05)	3.5 (0.04)
Nonpublic	3.5 (0.23)	3.9 (0.10)

NOTE: WOL = Writing Online. The number of students was 4,133, with 1,255 taking the WOL computer test and 2,878 taking the paper main NAEP writing assessment. Students were included only if they responded to both essays and, for main NAEP writing, only if the tasks were administered in the same order as those in WOL. Standard errors are in parentheses.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Writing Online Study.

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 "small." This finding suggests that computer delivery does not generally disadvantage NAEP reporting groups. Furthermore, the fact that the delivery mode main effects were also not significant in these analyses supports the lack of performance differences found across assessment modes, as indicated earlier in this report.

Performance as a Function of Computer Type

Because a large number of schools did not have the particular equipment, connectivity, or software required to administer the WOL study, NAEP staff brought laptops into schools to administer the test. As a result, approximately 65 percent of students took the WOL test on laptop computers.

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.[12]

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 (taken on desktop). Those who took the test on laptop first showed no difference in performance between essays.

To assess the effect of computer type on writing performance, an experiment was conducted 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 population. Eighty-eight students participated (51 male and 37 female students) in the experiment.[13] The selected students were randomly assigned to either a desktop or laptop computer for the test, 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.

The essay means for students responding to the laptop and desktop administrations are shown in table 4-9.

Table 4-9. Unweighted means for students randomly assigned to take the Writing Online test on laptop and web-connected school desktop computers, grade 8: 2002

Essay	NAEP laptop	Web-connected school computer
Save a Book	3.3 (0.22)	3.9 (0.15)
School Schedule	3.4 (0.22)	3.5 (0.17)

NOTE: Only those students responding to both essays are included. The number of students responding to both essays was 76, with 31 responding on laptop and 45 on desktop. Standard errors are in parentheses.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Writing Online Study.

[12] School machines vary too in ways that may possibly affect performance. This naturally occurring equipment variation was not evaluated in this study.

[13] The total number of students used for the analysis was 76, as only students who responded to both writing tasks were included.

The differences between the unweighted student means shown in table 4-9 were tested using a repeated-measures analysis of variance in which the dependent variable was essay score.[14] The factors were computer type (laptop vs. web-connected school desktop) and gender.

The results of the ANOVA showed no significant main effect for computer type (F,1,72 = 2.83, $p > .05$). That is, across both essays, the mean score for students taking WOL on laptop computers was not significantly different from the mean score for students taking WOL on school desktops. Although there was an expected main effect for gender (F,1,72 = 9.40, $p < .05$), there was no significant effect for the interaction of gender with computer type (F,1,72 = 0.78, $p > .05$), meaning that the difference in performance between using a laptop computer and a desktop computer was the same for male and female students.

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, $p > .05$).[15] 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, $p > .05$).

Because the sample sizes in the experiment were very small and unrespresentative, the performance of students on school computers compared with NAEP laptops was also evaluated in the larger WOL sample. In contrast to this experiment, among all students taking WOL the assignment to computer type was nonrandom, based on whether school computers and connectivity matched WOL requirements. This assignment could have been correlated with school location, school type, or socioeconomic status and, thereby, with writing skill level.

Table 4-10 shows the (weighted) mean scores for WOL students drawn from the main NAEP writing sample by the type of computer on which the WOL test was taken.

Table 4-10. Mean scores, by computer type, for Writing Online students drawn from the main NAEP writing sample, grade 8: 2002

Essay	NAEP laptop	Web-connected school computer
Save a Book	3.5 (0.06)	3.7 (0.09)
School Schedule	3.5 (0.08)	3.6 (0.11)

NOTE: The number of students was 687, with 256 responding on web-connected school computers and 431 on laptop computers. Standard errors are in parentheses.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Writing Online Study.

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.[16] The dependent variable was essay score. 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, $p > .05$).[17]

While there appears to be 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 4-11 shows the means for students by gender.

[14] This analysis was conducted with the SAS GLM procedure. It was used instead of the WESVAR repeated-measures ANOVA employed elsewhere in the study because, in the absence of the need for sampling weights, the SAS GLM ANOVA is simpler to implement.

[15] The ANOVA model was rerun substituting school SES (low, medium, high) for gender with substantively the same results: no significant between-groups effect for computer type (F,1,70 = 2.59, $p > .05$), school SES (F,2,70 = 1.21, $p > .05$), or the interaction of school SES with computer type (F,2,70 = 1.43, $p > .05$). Within groups, there was no significant difference between essays (F,1,70 = 2.46, $p > .05$), a significant essay-by-computer-type interaction (F,1,70 = 4.89, $p < .05$), and no interaction of essay with SES (F,2,70 = 1.43, $p > .05$) or of essay, computer type, and SES (F,2,70 = 2.76, $p > .05$).

[16] Main NAEP writing performance was indicated by the five plausible values associated with each student, which WESVAR uses to compute the group means and variances. The sample size for this analysis was 685, with two students deleted because they were missing plausible values.

[17] When main NAEP writing performance is omitted from the model (n = 687), there is also no significant main effect for computer type (F,1,62 = 1.16, $p > .05$) and no interaction of computer type with essay (F,1,62 = 0.08, $p > .05$).

Table 4-11. Mean scores, by gender and computer type, for Writing Online students drawn from the main NAEP writing sample, grade 8: 2002

	Male		Female	
Essay	NAEP laptop	Web-connected school computer	NAEP laptop	Web-connected school computer
Save a Book	3.4 (0.09)	3.2 (0.12)	3.6 (0.10)	4.1 (0.11)
School Schedule	3.3 (0.10)	3.2 (0.12)	3.7 (0.10)	4.0 (0.11)

NOTE: The number of students was 684, with 256 responding on web-connected school computers and 428 on NAEP laptop computers. Standard errors are in parentheses.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Writing Online Study.

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. Accounting for main NAEP writing performance (n = 680), there was no difference between the scores for students taking WOL on laptop versus school computer (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, $p > .05$).[18]

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 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 there was 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 laptop computers (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. Finally, for female students, there was no interaction between essay and computer type (F,1,62 = 1.41, $p > .05$).

The preceding analysis found female students to perform better on school computers than on NAEP laptops. Do females also write longer essays on school computers? To evaluate this possibility, the same repeated-measures ANOVA as above (n = 680) with gender groups combined was executed, but with essay length instead of score as the dependent variable. Although, after accounting for main NAEP writing performance, this analysis showed a significant effect for gender (F,1,62 = 23.36, $p < .05$), there was no effect for computer type (F,1,62 = 0.01, $p > .05$) or for the interaction of gender and computer type (F,1,62 = 2.33, $p > .05$). Further, there were no significant interactions of essay with computer type (F,1,62 = 0.97, $p > .05$), with gender (F,1,62 = 0.75, $p > .05$), or with computer type and gender (F,1,62 = 1.67, $p > .05$). Thus, for any given level of writing skill, female students generate longer essays than male students, but this propensity holds regardless of computer platform.

In sum, the results comparing NAEP laptop and school computer performance are not completely consistent. In the experimental substudy, students generally scored lower on laptop than desktop for one of the two essays. This effect was not duplicated,

[18] When main NAEP writing performance is removed from the model, the same substantive results were obtained. There was no effect for laptop vs. school computer (F,1,62 = 1.58, $p > .05$), a main effect for gender (F,1,62 = 37.88, $p < .05$), an interaction of gender with computer type (F,1,62 = 10.35, $p < .05$), and no interaction of essay with computer type (F,1,62 = 0.01, $p > .05$), with gender (F,1,62 = 0.41, $p > .05$), or with gender and computer type (F,1,62 = 3.58, $p > .05$).

however, in the quasi-experimental comparison conducted in the larger WOL main NAEP writing sample. Instead, the quasi-experimental analysis showed female students performing lower on the NAEP laptops for both essays. In any case, the results do suggest that students may sometimes obtain different scores on writing tests administered on laptop versus school machines.

Performance as a Function of Computer Experience

Does familiarity with computers affect 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. 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. 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 (2000) found eighth- and tenth-grade students performed better on the computer-writing test regardless of their keyboarding speed.

Except for students from urban fringe/large town schools, the traditional NAEP reporting groups do not seem to be differentially affected by computer delivery. However, it may still be the case that computer familiarity itself 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 provide a partial answer.[19] Responses suggest that most eighth-grade students have access to computers at school and home, use computers frequently, and have positive attitudes toward them. For example, the large majority of students indicated that they use a computer at home (91 percent) and that they use 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 use a computer outside of school at least two or three times a week (80 percent). (Only six percent of students indicated they never use a computer outside of school, and only 13 percent said they never use a computer at school.) Finally, the majority of students reported that learning is more fun on the computer (85 percent), they get more done when they use a computer for schoolwork (75 percent), and they are more motivated to start schoolwork if they use the computer (71 percent).

To what extent do students use computers for writing? Although almost all students report using a computer to write at least to some degree, there is considerable variation: In rounded percentages, the results for all students show that 29 percent indicate 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 do students use computers for writing? Again, there is wide variation: 32 percent report that they "always" use a computer to write a paper from the beginning, 42 percent say they do this "sometimes," and 25 percent indicate that they "never" use a computer in this way. What the large majority of students (69 percent) report doing, however, is "always" using a computer to type final copy of a report that they wrote by hand. Appendix H gives additional response data about specific writing uses.

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

[19] The background questions used in WOL were selected from among questions previously administered in the 1998 and 2002 main NAEP writing assessments. (See appendix D for the WOL questions.) The percentages reported herein are from all students who took WOL.

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, and 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 4-1 shows the two sets of background questions that were both substantively relevant and significantly related to WOL performance in the sample drawn from main NAEP reading. 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.

For each question set, a single score was created by making the response to each question dichotomous, then summing the responses. 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.[20]

Figure 4-1. Self-reported computer-familiarity questions administered to students taking Writing Online, grade 8: 2002

To what extent do you do the following on a 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

NOTE: The responses to all questions were collapsed to a 0/1 score and the results then summed across questions within a set.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Writing Online Study.

[20] Coefficient alpha reliabilities for the "Extent of computer use" and "Computer use for writing" scores were .55 and .65, respectively.

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 sample drawn from main NAEP reading.

Three variables were theoretically meaningful and showed significant relationship to WOL performance. The variables, described in table 4-12, were typing speed, typing accuracy, and editing skill. (Summary statistics are given in appendix I.) 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 sample drawn from main NAEP reading.[21]

Table 4-13 gives the correlations among the WOL self-report computer familiarity questions, the hands-on computer skills measure, WOL performance, and the main NAEP performance for those main NAEP writing students taking WOL.[22] (Summary statistics are given in appendix H.) As the table shows, hands-on computer skill is moderately related to both WOL essays and to main NAEP writing performance. Also, hands-on computer skill is unrelated or weakly related to the self-reported computer familiarity indicators. The two types of familiarity indicators, then, seem to have little overlap with one another, suggesting that each may, in fact, be tapping relatively independent components of familiarity. Equally important, both the extent of computer use and the hands-on computer skill measure show some potential to predict online test performance.

Table 4-12. Components of the hands-on computer skills measure, grade 8: 2002

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

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Writing Online Study.

[21] The standardized regression weights for the three index components were .52 for typing speed, .19 for editing skill, and -.10 for typing accuracy. These weights give an indication of the relative importance of each component to the hands-on index.

[22] The sample drawn from main NAEP reading was used to select the hands-on variables and to derive their best linear composite. This composite was then applied in the main NAEP writing sample. 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.

Table 4-13. Correlations among Writing Online self-reported computer familiarity questions, hands-on computer skills, Writing Online scores, and main NAEP writing performance for Writing Online students drawn from the main NAEP writing assessment, grade 8: 2002

Variable	Extent of computer use	Computer use for writing	Hands-on computer skill	Save a Book	School Schedule
Computer use for writing	.40*				
Hands-on computer skill	.19*	.07			
Save a Book	.16*	.06	.48*		
School Schedule	.14*	.01	.52*	.64*	
Main NAEP writing performance	.08*	.02	.42*	.53*	.55*

* Significantly different from zero at $p < .05$.

NOTE: Sample sizes range from 679 to 687. The main NAEP writing performance is the first plausible value. Extent of computer use was scored 0-8, computer use for writing 0-6, and hands-on computer use was the best linear composite of three polytomously scored variables.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Writing Online Study.

To examine whether computer familiarity affects online test performance, a repeated-measures ANOVA was conducted with 660 students drawn from the main NAEP writing assessment who responded to both computer-administered WOL essays.[23] 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, $p > .05$), computer use for writing (F,1,62 = 2.20, $p > .05$), or hands-on computer proficiency (F,1,62 = 3.86, $p > .05$). Thus, after accounting for paper writing performance, computer experience, in the form of keyboarding proficiency, does appear to play a role in WOL performance. Some sense of the magnitude of this role can be gleaned 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 to 47 percent.[24]

[23] 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.

[24] That computer familiarity plays a significant role in WOL performance may explain why WOL score reliability was *not* lower than the paper main NAEP score reliability even though the WOL reader agreement *was* lower. The correlation between WOL essays was likely increased by the fact that the score on each essay was in part a function of each student's computer familiarity. Computer familiarity would not be expected to increase the correlations between paper main NAEP essays in the same way.

Does computer familiarity matter more for one population group than another? To find out, gender was added to the 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 from the repeated-measures ANOVA are presented in appendix J. In this model, the main effect for hands-on computer skill is still significant, and there is 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.

In sum, computer familiarity in the form of hands-on skill affects online writing test performance. The relationship is such that students with more hands-on skill score higher than those with less skill, holding constant their writing proficiency as measured by paper writing tests. Thus, while no measurable differences between computer and paper tests of writing were detected for the population as a whole, the two delivery modes are apparently not comparable for individuals.

5. Efficiency Issues

This section addresses issues concerning the efficiency of technology-based assessment. In particular:

- Is a technology-based writing assessment more cost-effective or timely than a paper one?
- How might technological advances like web delivery and automated essay scoring affect the cost and timeliness of assessment?

Relative Timeliness and Costs of Computer- vs. Paper-Based Assessment

The data presented thus far in this report speak to the measurement and equity issues around using computer delivery as an alternative to paper delivery of NAEP writing assessments. But how might a computer-delivered NAEP writing assessment compare with a paper-based assessment in terms of cost and timeliness?

Relative Timeliness of Computer vs. Paper Testing

Figure 5-1 shows the key steps in the conventional paper administration (from pilot test to operational assessment), along with the likely steps for online delivery. Also included for each step are estimated elapsed times in calendar days. The elapsed-time estimates were based on the combined judgments of two NAEP WOL test developers with considerable experience in the operational NAEP paper-testing program. Because their judgments are based on only a single online testing experience, this comparison should be regarded as suggestive. For the pilot stage, the estimated number of calendar days needed would be similar for paper delivery (217 days) and for computer delivery (206 days). For the operational stage, however, the estimates are about 30 percent shorter for computer delivery (109 days) than for paper delivery (156 days). The primary reason for this difference is that fewer steps are expected to be required in the computer delivery process.

Figure 5-1. Key steps in NAEP paper vs. computer writing test delivery, with estimated elapsed times

Pilot test			
Paper delivery		**Computer delivery**	
Step	**Estimated elapsed time in days**	**Step**	**Estimated elapsed time in days**
Total	**217**	**Total**	**206**
• Draft items created on paper, reviewed, and revised by NAEP staff	30	• Draft items created on paper, reviewed, and revised by NAEP staff	30
• NCES review of items	10	• NCES paper review of items	10
• Subject-area committee review of items	3	• Initial version of items produced online	5
• Items revised by NAEP staff	5	• Subject-area committee review of items online via World Wide Web (WWW)	7
• Items reviewed by state education officials	5	• Items revised by NAEP staff	5
• Subject-area committee review of items	3	• Items reviewed by state education officials online via WWW	7
• Clearance package sent to NAGB/NCES for review	5	• Subject-area committee review of items online via WWW	7
• Comments received from NCES/NAGB	10	• NAGB/NCES review items online via WWW for clearance	10
• Items revised as necessary and assembled into pilot blocks	5	• Items revised online as necessary and assembled into blocks	5
• Camera-ready blocks produced and sent to be printed	10	• Items formatted for online delivery	10
• Bluelines (printer proofs) of test booklets produced	15	• Test administered online or on NAEP laptops	35
• Test booklets printed, spiraled, bundled, and shipped to administrators	17	• Student data transferred from laptops (where used) to NAEP database. School computer data delivered directly to scoring contractor	10
• Test administered	35	• Training samples selected for scoring	15
• Test booklets returned to scoring contractor for scanning	10	• Student responses used to refine automated scoring algorithms for those items to be scored by machine	20
• Training samples selected for scoring	15	• Items either automatically scored or scored online by trained NAEP raters	10
• Selection of training samples reviewed at committee meeting	4	• Scores sent to NAEP database	10
• Scanned handwritten responses scored online by trained NAEP raters	15	• Data sent to analysis contractor	10
• Scores sent to NAEP database	10		
• Data sent to analysis contractor	10		

See notes at end of figure. ▶

Figure 5-1. Key steps in NAEP paper vs. computer writing test delivery, with estimated elapsed times—Continued

Operational Assessment			
Paper delivery		**Computer delivery**	
Step	**Estimated elapsed time in days**	**Step**	**Estimated elapsed time in days**
Total	**156**	**Total**	**109**
• Final test items selected and revised as necessary	7	• Final test items selected and revised as necessary	7
• Subject-area committee review of final versions of items	4	• Subject-area committee review of final versions of items online via World-Wide Web (WWW)	5
• Items revised by NAEP staff	3	• Items revised by NAEP staff	3
• Clearance package sent to NAGB/NCES for review	5	• NAGB/NCES review items online via WWW for clearance	10
• Comments received from NCES/NAGB	10	• Items revised online as necessary	3
• Items revised as necessary	3	• Test administered online or on NAEP laptops	35
• Camera-ready blocks produced and sent to be printed	5	• Student data transferred from laptops (where used) to NAEP database. School computer data delivered directly to scoring contractor	10
• Bluelines (printer proofs) of test booklets produced	10	• Training samples selected for scoring	8
• Test booklets printed, spiraled, bundled, and shipped to administrators	17	• Items either automatically scored or scored online by trained NAEP raters	8
• Test administered	35	• Scores sent to NAEP database	10
• Test booklets returned to scoring contractor for scanning	10	• Data sent to analysis contractor	10
• Training samples selected for scoring	8		
• Selection of training samples reviewed at committee meeting	4		
• Scanned responses scored online by trained NAEP raters	15		
• Scores sent to NAEP database	10		
• Data sent to analysis contractor	10		

NOTE: Time estimates assume a 40-item pilot test and a 20-item operational test. Elapsed times do not represent levels of effort. NAGB = National Assessment Governing Board. NCES = National Center for Education Statistics.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Writing Online Study.

Relative Costs of Computer vs. Paper Testing

This section looks at the comparative costs of item and software development, delivery and administration, and scoring for the two testing modes.

Relative costs of item and software development. The cost of creating new items for online delivery of writing assessments should be similar to costs for paper delivery, but will depend somewhat on whether the requisite online tools exist in the delivery software. Commercial web-delivery systems generally have the necessary templates for item authoring, tutorials to show students how to respond, and the associated tools for word processing. For writing tests, the screen will usually consist of an area that displays the essay task, a response area into which text can be typed, and associated tools. In the software used for WOL, these tools included cut, copy, paste, undo, spell-checker, and hide task (to increase the size of the response area).

Relative costs of assessment delivery and administration. Delivery and administration costs for an online assessment, which are not needed in a paper assessment, include licenses or development fees for the testing software; central hosting of that software, the item bank, and the student-response database; lease or rental of laptops for schools that cannot participate using their own computer equipment; copying of test software and item banks to the laptops and removal of student data from them; shipping of laptops; and telephone technical support for field administrators.

Some of these delivery and administration costs will be quite variable. In particular, laptop costs will depend on student sample sizes, number of schools participating, and the number of school computers that can be used. The number of available school computers, will, in turn, depend on the ability of the delivery software to accommodate a wide range of configurations (e.g., PC and Macintosh, broadband and dial-up, Internet Explorer and Netscape). Such a range, however, could also impact standardization in ways that materially affect assessment performance. How machine variation affects performance is not yet well known.

As implemented in WOL, fewer students per session were tested online than in paper-and-pencil sessions. (NAEP paper administrations routinely assess groups of about 30 students at a time.) This difference was largely a function of server capacity and the need to minimize burden on the field administrators. In an operational assessment, NAEP would use a production delivery system with greater server capacity and would expect administrators to handle larger groups comfortably. Assessing groups of 30 students online may be possible in schools that can devote a room of certifiable computers to the assessment. In those cases where a school cannot, the group size will range from five students (the number of laptops an administrator can comfortably transport) to that amount plus the number of machines the school can supply. On average, this number may still be fewer than the amount NAEP currently tests on paper (perhaps by one-half). That differential will diminish, however, as computers become less expensive.

While the delivery and administration costs of online assessment can be considerable, these expenses can be offset to a degree by eliminating some of the high-cost factors of paper delivery, such as test book printing, packing, shipping, and tracking the return of test materials. In addition, the expense associated with occasional last-minute changes to the assessment would be reduced. Changes to test instruments, spiraling designs, or sampling plans would otherwise need to be made by reprinting, reassembling, or repackaging test materials.

Relative costs of scoring.
The cost of scoring computerized writing assessments should not differ from current NAEP scoring expenses, so long as human readers are used to evaluate essay responses. However, if automated scoring can be used along with, or instead of, human readers, large cost savings may be achievable.

Automated essay scoring has been used operationally in several testing programs for scoring essay responses. These programs include the Graduate Management Admission Test, in which a computer-generated score is used in conjunction with the score of a human reader, and the College Board's Writeplacer and ACT's COMPASS e-Write, where the computer is the only grader.

For automated scoring to be implemented in NAEP, one-time investments might need to be made in existing operational systems to allow for efficiently training the grading software, integrating scores, and back-reading papers. Also, automated scoring may be of only limited value at the pilot stage, as opposed to the operational stage, of a writing assessment. For pilot tests, the sample sizes are smaller than for operational assessments and the cost for human scoring is, therefore, relatively low. Furthermore, since items may be dropped after pilot testing, any effort and cost expended on training automated systems to score specific items might not carry over to the operational stage. However, to the extent that scoring systems do not need to be trained for specific items, this may not be a limitation.

In a NAEP writing assessment, automated scoring would offer the greatest increase in cost-effectiveness for new items delivered to large samples of students and for trend items to be used in multiple (computer-delivered) assessments taken across years. Currently, substantial staff preparation, training, and scoring time are devoted in each assessment cycle to maintaining trend. These "trend validation" procedures are implemented to ensure that readers score items with the same accuracy and standards as in previous years. A significant benefit to automated scoring would be the elimination of score drift or change in agreement from one year to the next.

Figure 5-2 summarizes the relative costs for NAEP of computer versus paper assessment. Assuming writing items similar to those currently used in NAEP, the costs for an online writing assessment should be similar for test development, similar or higher for test delivery and administration, and similar or lower for scoring.

Figure 5-2. Relative costs for NAEP of computer vs. paper writing assessment

Process	Relative Cost	Comment
Item and Software Development		
Developing writing tasks (essays)	Similar	Commercial delivery systems generally have item templates, tutorial segments, essay presentation and answer formats, and supplementary text-processing tools.
Test Delivery and Administration		
Delivering test to schools	Similar or higher than paper	Includes cost of licensing or developing delivery software and hosting software, item bank, and student response database. Also includes cost of leasing laptops, loading software, shipping, and removing student data. Computer delivery eliminates costs of printing, packaging, shipping, and returning test booklets. Overall cost difference depends greatly on sample size and number of laptops required.
Preparing for and administering test	Similar or higher than paper	More time may be required for initial contacts with schools and for certifying computers, although that need should decrease over time.
Providing telephone technical support	Similar	Help desk is routinely used for paper assessments at similar staffing level.
Changing items, spiral designs, and sampling plans	Lower than paper	Eliminates need to reprint, repackage, or reassemble test materials.
Scoring		
Automatically scoring items	Lower than paper	So long as student samples are large or scoring includes trend items.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Writing Online Study.

Automated Scoring: E-rater®

A technological advance that could help NAEP increase efficiency once it begins delivering writing assessments online is the automated scoring of responses. By reducing or eliminating the need for human readers, automated scoring could reduce scoring costs while increasing the speed with which NAEP analyses can be completed.

To investigate the feasibility of automated scoring for a NAEP writing assessment, all WOL essays were scored using e-rater®, a computer essay scoring system developed by Educational Testing Service (ETS). The version of e-rater® used for this study, 2.0, is a recently released upgrade to the program used for production scoring of Graduate Management Admission Test® (GMAT®) essays by ETS. For the GMAT®, each essay was scored by both e-rater® and one human reader. If there was a discrepancy of more than one point on the 1–6 score scale, a second human reader was assigned to resolve the discrepancy.

The scoring process implemented by e-rater® 2.0 involves several steps (Burstein, Chodorow and Leacock in press; Burstein 2003). First, a training sample of essay responses for a given question is selected, where human judges have already scored each response. Next, e-rater® extracts values for a fixed set of 12 features from these essays. (See figure 5-3 for a list of features.) Third, the weights for 11 of these features are determined through multiple regression to optimally predict the human scores. (The weight for the last feature, essay length, is set judgmentally so as not to overemphasize the influence of this feature on score computation.) Fourth, this regression model is cross-validated by using it to predict human scores for a new sample of responses to the same essay question. Finally, if the model is judged to be acceptable, it is used to score the remainder of the essay responses.

For potential use in NAEP writing assessments, a relevant question is whether e-rater® scores are comparable to, or exchangeable with, those of human readers. In psychometric terms, scores from two assessments are considered comparable when they have approximately the same distribution and rank order. In scoring NAEP writing tasks, the program strives for comparability between readers, that is, which particular reader scored the responses should not matter because the end result should be approximately the same.

There have been many studies of the extent to which automated scoring programs like e-rater® produce scores comparable to those rendered by human readers. Keith (2003, pp. 154, 158, 161) summarized results from studies suggesting, for example, that the scores produced by such systems correlate as highly with the scores assigned by a human reader as two human readers' scores correlate with one another. To date, however, no studies could be found using middle-school students responding to essay prompts like those used in main NAEP.

Figure 5-3. Writing features extracted by e-rater®, grouped by logical dimensions

Dimension	Feature
Grammar, usage, mechanics, and style	1. Ratio of grammar errors to the total number of words
	2. Ratio of mechanics errors to the total number of words
	3. Ratio of usage errors to the total number of words
	4. Ratio of style errors (repetitious words, passive sentences, very long sentences, very short sentences) to the total number of words
Organization and development	5. The number of "discourse" elements detected in the essay (i.e., background, thesis, main ideas, supporting ideas, conclusion)
	6. The average length of each element as a proportion of total number of words in the essay
Topical analysis	7. Similarity of the essay's content to other previously scored essays in the top score category
	8. The score category containing essays whose words are most similar to the target essay
Word complexity	9. Word repetition (ratio of different content words to total number of words)
	10. Vocabulary difficulty (based on word frequency)
	11. Average word length
Essay length	12. Total number of words

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Writing Online Study.

In the main NAEP writing assessment, one human reader is assigned to score each response. Then, a sample of the responses are independently scored by a second human reader to estimate the degree to which the scores from different readers are, in fact, interchangeable. This standard double-scoring by human readers was also implemented for the WOL study. For a subsample of responses that were double-scored by human readers, table 5-1 gives the unweighted means and standard deviations of the scores assigned by each of them, as well as by e-rater®. Results for this and all subsequent analyses employ cross-validation samples; that is, the samples of essays are different from the ones used to train e-rater® and, therefore, do not include every double-scored response.

These means were tested using an analysis of variance with reader and essay as the independent variables, and repeated measures on the rater factor (but not on the essay factor because a different random sample of papers was double-scored for each task). Score was the dependent variable. No significant differences were detected between the two essay means (F,1,500 = 3.21, $p > .05$) or in the interaction of readers and questions (F,2,1000 = 1.22, $p > .05$). However, the means assigned by the three "raters" did differ significantly (F,2,1000 = 26.92, $p < .05$). One-tailed, post-hoc, dependent-sample t tests showed that the two human-reader means did not differ significantly from one another (t, 501 = -0.97, $p > .05$), but that the e-rater® mean was significantly higher than the mean of the first reader (t, 501 = -6.29, $p < .05$) as well as higher than the mean for the second reader (t, 501 = -5.59, $p < .05$).[25] In effect-size terms, the differences between e-rater® and the first and second human reader were 0.25 and 0.19 standard deviations, respectively (in the units of each human reader).

In addition to differences in mean scores between automated and human raters, the two methods may also order individuals differently. To investigate whether scores were similarly ordered, the intra-class correlation between e-rater® scores and the scores assigned by the human readers was computed for each essay (see table 5-2). For "Save a Book," the two human readers' scores correlated significantly more highly with one another than the e-rater® scores correlated with the first reader (t, 258 = 4.38, $p < .05$) or than e-rater® correlated with the second reader

Table 5-1. Unweighted means and standard deviations for essay scores, by human readers and e-rater®, grade 8: 2002

Essay	First reader	Second reader	E-rater®
Save a Book			
Mean	3.4	3.5	3.7
Standard Deviation	1.2	1.1	1.0
School Schedule			
Mean	3.4	3.3	3.5
Standard Deviation	1.1	1.1	1.1

NOTE: The number of responses was 261 for "Save a Book" and 241 for "School Schedule."

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Writing Online Study.

[25] An analysis of variance was also run using responses for which there was only a single human score. This analysis used 797 students with responses to the two writing tasks and who had not been included in the data set used to train e-rater®. For this analysis, the rater factor had only two levels, e-rater® and the first human reader, and there were repeated measures on both the essay and the rater factors. A significant difference was detected for essay (F,1,796 = 32.68, $p < .05$) and for reader (F,1,796 = 68.15, $p < .05$), as well as for the interaction between the two (F,1,796 = 9.00, $p < .05$), indicating that the size of the difference between e-rater® and the human reader was not the same for the two essays. For each essay, however, the e-rater® mean score was higher than the human reader's mean score.

Table 5-2. Unweighted intraclass correlations for essay scores, by human readers and e-rater®, grade 8: 2002

Variable pair	Save a Book	School Schedule
First reader with second reader	.79	.84
First reader with e-rater®	.66	.66
Second reader with e-rater®	.67	.67

NOTE: The number of students responding was 261 for "Save a Book" and 241 for "School Schedule."

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Writing Online Study.

(t, 258 = 4.00, $p < .05$). The same result was obtained for "School Schedule," where the correlations between readers significantly exceeded the e-rater® correlation with the first reader (t, 238 = 6.53, $p < .05$) and with the second reader (t, 238 = 6.20, $p < .05$).

Table 5-3 shows the unweighted percentage exact agreement between the e-rater® and human reader scores and between two human reader scores. Differences in agreement among pairs of raters were tested with repeated-measures ANOVA. The independent variables were essays and pairs of raters, with repeated measures on the latter factor. The dependent variable was whether or not the members of a pair (e.g., first reader and second reader) agreed with one another exactly. Between-groups results showed no significant difference between the agreement levels for the two essays (F,1,500 = 0.00, $p > .05$). Of more relevance to the comparability of automated and human scoring, however, were the within-group results. These results showed a significant effect for rater pairs (F,2,1000 = 4.97, $p < .05$), but no interaction of essays and rater pairs (F,2,1000 = 0.15, $p > .05$). Thus, these results suggest that, across essays, some combinations of raters agreed more highly with one another than did other combinations. Post-hoc, dependent-sample t tests (one-tailed) indicated that agreement of e-rater® with the first reader was not significantly different from its agreement with the second reader (t, 501 = -0.47, $p > .05$). However, the agreement of e-rater® with the first reader was lower than the first reader's agreement with the second reader (t, 501 = 2.85, $p < .05$). Likewise, agreement of e-rater® with the second reader was lower than the agreement between the two human readers (t, 501 = 2.38, $p < .05$).

Table 5-3. Unweighted percentage exact agreement between e-rater® and human readers and between two human readers, grade 8: 2002

	Save a Book		School Schedule	
Variable pair	Percent exact agreement	Kappa	Percent exact agreement	Kappa
First reader with second reader	61	.48	62	.49
First reader with e-rater®	54	.38	53	.37
Second reader with e-rater®	55	.38	55	.40

NOTE: The number of students responding was 261 for "Save a Book" and 241 for "School Schedule."

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Writing Online Study.

Tables 5-4 and 5-5 show for "Save a Book" the exact agreement between e-rater® and the two human readers, respectively, for each of six score levels. Table 5-6 shows the comparable agreement between two human readers. The far right-hand column of each table gives the percentage exact agreement for each level. For each score level, figure 5-4 shows the difference between the percentage agreement achieved by the human readers and the mean percentage agreement between e-rater® and the humans. Note that, as has been found in studies with earlier versions of e-rater® (e.g., Burstein, Kukich, Wolff, Lu, and Chodorow 1998), the scoring program's agreement with human readers appears in this sample to be considerably higher at the middle score levels (i.e., 3, 4, 5) than at the extremes (i.e., 1, 2, 6).

Table 5-4. Unweighted score distributions and percentage exact agreement between e-rater® and first human reader at each of six score levels for "Save a Book," grade 8: 2002

First human reader score level	e-rater® score level						Percent exact agreement
	1	2	3	4	5	6	
1	**5**	3	10	0	0	0	28
2	1	**4**	17	7	0	0	14
3	0	6	**49**	19	11	0	58
4	0	0	14	**54**	12	2	66
5	0	0	0	10	**29**	2	71
6	0	0	0	0	5	**1**	17

NOTE: The number of students responding was 261 for "Save a Book."

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Writing Online Study.

Table 5-5. Unweighted score distributions and percentage exact agreement between e-rater® and second human reader at each of six score levels for "Save a Book," grade 8: 2002

Second human reader score level	e-rater® score level						Percent exact agreement
	1	2	3	4	5	6	
1	**5**	3	9	0	0	0	29
2	0	**7**	14	3	1	0	28
3	1	3	**49**	20	5	1	62
4	0	0	17	**51**	19	1	58
5	0	0	1	15	**30**	2	63
6	0	0	0	1	2	**1**	25

NOTE: The number of students responding was 261 for "Save a Book."

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Writing Online Study.

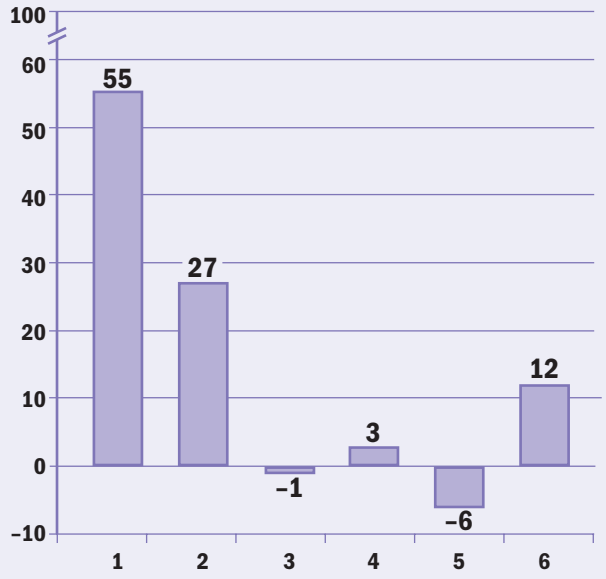
Table 5-6. Unweighted score distributions and percentage exact agreement between two human readers at each of six score levels for "Save a Book," grade 8: 2002

First human reader score level	Second human reader score level						Percent exact agreement
	1	2	3	4	5	6	
1	**15**	3	0	0	0	0	83
2	2	**14**	12	1	0	0	48
3	0	8	**50**	19	8	0	59
4	0	0	15	**53**	13	1	65
5	0	0	2	13	**25**	1	61
6	0	0	0	2	2	**2**	33

NOTE: The number of students responding was 261 for "Save a Book."

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Writing Online Study.

Figure 5-4. Unweighted difference between mean of e-rater® percentage exact agreements with two human readers and percentage exact agreement of two human readers with one another at each of six score levels for "Save a Book," grade 8: 2002



NOTE: The number of students responding was 261 for "Save a Book." Positive differences indicate that the human readers agree with one another to a greater degree than e-rater® agrees with the human readers.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Writing Online Study.

Tables 5-7, 5-8, and 5-9 show the same statistics for "School Schedule," with the same general result: The exact agreement of e-rater® relative to human readers is higher in this sample for the middle scores than at the extremes, as figure 5-5 illustrates.

Table 5-7. Unweighted score distributions and percentage exact agreement between e-rater® and first human reader at each of six score levels for "School Schedule," grade 8: 2002

First human reader score level	E-rater® score level						Percent exact agreement
	1	2	3	4	5	6	
1	**5**	4	4	2	1	0	31
2	3	**13**	11	3	0	0	43
3	1	12	**50**	20	3	2	57
4	0	1	11	**44**	14	2	61
5	0	0	1	13	**11**	5	37
6	0	0	0	1	0	**4**	80

NOTE: The number of students responding was 241 for "School Schedule."

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Writing Online Study.

Table 5-8. Unweighted score distributions and percentage exact agreement between e-rater® and second human reader at each of six score levels for "School Schedule," grade 8: 2002

Second human reader score level	E-rater® score level						Percent exact agreement
	1	2	3	4	5	6	
1	**6**	3	3	2	1	0	40
2	2	**17**	12	4	1	0	47
3	1	10	**49**	18	3	2	59
4	0	0	13	**46**	14	2	61
5	0	0	0	12	**9**	4	36
6	0	0	0	1	1	**5**	71

NOTE: The number of students responding was 241 for "School Schedule."

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Writing Online Study.

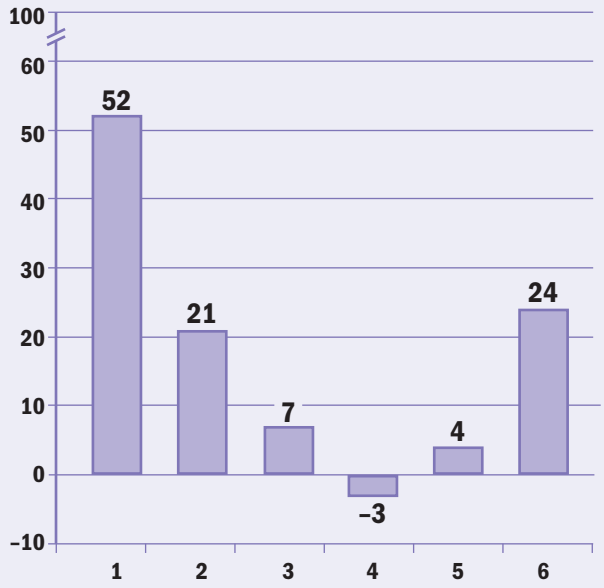
Table 5-9. Unweighted score distributions and percentage exact agreement between two human readers at each of six score levels for "School Schedule," grade 8: 2002

First human reader score level	Second human reader score level						Percent exact agreement
	1	2	3	4	5	6	
1	**14**	1	1	0	0	0	88
2	1	**20**	9	0	0	0	67
3	0	14	**57**	17	0	0	65
4	0	1	16	**42**	13	0	58
5	0	0	0	16	**12**	2	40
6	0	0	0	0	0	**5**	100

NOTE: The number of students responding was 261 for "Save a Book."

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Writing Online Study.

Figure 5-5. Unweighted difference between mean of e-rater® percentage exact agreements with two human readers and percentage exact agreement of two human readers with one another at each of six score levels for "School Schedule," grade 8: 2002



NOTE: The number of students responding was 241 for "School Schedule." Positive differences indicate that the human readers agree with one another to a greater degree than e-rater® agrees with the human readers.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Writing Online Study.

6. Operational Issues

This section reports on the logistical challenges associated with administering a NAEP writing assessment on computer. In particular, the discussion considers whether school facilities, equipment, software, and internet connectivity; administrator effectiveness; school cooperation; and data quality are sufficient to conduct NAEP assessments electronically. Westat, the NAEP sampling and data collection contractor, sampled and recruited both the NAEP and WOL schools and also administered all instruments. Westat contributed much of the information for this section of the report (Westat 2002).

Recruiting Schools

The sample of schools for WOL was drawn from among the schools selected for the main NAEP 2002 reading and writing assessments. Thus, it was not possible to identify the WOL schools until the main NAEP selection had been finalized. The WOL sampling was completed in early January 2002, and school recruiting began in February 2002.

Letters were sent to NAEP state coordinators and state test directors on February 18, 2002, informing them about the WOL sample selection. On February 26, 2002, letters were sent to superintendents of districts that included selected schools. After sending an initial mailing about upcoming NAEP assessments, a special letter that focused on the WOL project was sent to principals. Because of the need for computer delivery, Westat engaged in more telephone interaction with school administrators and school technology staff than for the typical NAEP assessment.

Westat reported that it was initially somewhat difficult to recruit schools to participate for WOL, due mainly to the late nature of the contacts and conflicts with other testing already scheduled for the eighth grade. Factors that helped gain cooperation from the schools were the need for only about 10 students per school to complete the online test (about 20 students fewer than the usual NAEP assessment); no need for teacher or school questionnaires to be completed, which was a reduction in burden from the main NAEP assessment; and the offer of a $200 honorarium for participating in the WOL study.

Training Field Administrators

A WOL training session for 26 field administrators and one field manager was held at Westat's headquarters in Rockville, Maryland on March 26–28, 2002. The presentations focused on the technical issues associated with certifying school computers and troubleshooting problems, as well as on administering WOL. Most of the WOL field administrators had previous experience administering either Math Online, the first of the NAEP technology-based assessment projects, or the spring 2001 WOL pretest.

Preparing for the Administrations

Westat supervisors conducted preliminary phone calls with schools to determine the type of computers available (IBM-compatible versus Macintosh), whether the school had an internet connection that could be used for WOL, and what type of internet connection was available. Based on the answers to these questions, the supervisors determined how much time was needed to certify the school computers, or if they would need to use the NAEP laptop computers.

Westat staff visited each school approximately two weeks prior to its test date, as is routine for NAEP assessments. During these pre-administration visits, supervisors worked with school personnel to draw the student sample, establish locations and times for the administration, and make any other necessary arrangements. Westat also worked with the school or district computer technician to certify the school's computers for the study (or to arrange space for laptops if they were to be used). The procedure, repeated on each school computer, involved the technician logging on to the computer and the supervisor accessing a special NAEP website. A program run from this website remotely evaluated the school computer hardware and software to determine if the computer met the WOL specifications, or, if it did not, indicated what needed to be done for the system to be certified. In some instances, the technician was simply able to modify a setting to allow the computer to be used.

Because school and district technicians generally were disappointed when their PCs failed to certify, many spent much time and effort attempting to remedy problems. Occasionally, the administrator arrived on the day of the test to find that upgrades to systems had been made in the interim and that the school PCs now could be certified. Even in those schools in which the computers met the WOL specifications, the administrators re-certified the computers before beginning the test to ensure that the settings had not been changed between the original certification and the day of the administration.[26]

The primary reason for PCs failing WOL certification was slow data transmission: Many schools were unable to meet the standard required to efficiently administer the test. Other reasons for failing certification included insufficient memory or available hard drive space to download the Macromedia Flash and Java software components needed to run the test.

The technical specifications required by the web-based delivery system for the study are shown in figure 6-1. Because this system was developed for research use, it supported only computers that use Microsoft® Windows. For an operational assessment, NAEP would employ a commercial delivery system. Such systems typically accommodate both Windows and Macintosh computers, thereby accounting for the vast majority of internet machines found in schools.

The system used in this study delivered the test from a server via the Internet. However, the system also could be run from a stand-alone laptop computer. In that configuration, the server software resided on the laptop hard drive and presented information to the machine's browser as if there were an active internet connection.

Figure 6-1. Technical specifications for school computers used to deliver the Writing Online test, grade 8: 2002

Feature	Requirement
Computer type	IBM (or compatible) personal computer
Processor type	Pentium or higher
Processor speed	266 MHz or faster
Screen resolution	800 x 600 resolution minimum
Screen colors	65,536 (16 bit) colors minimum
Random access memory	32MB or greater for Windows 95 or 98; 64MB for other Windows operating systems
Data transmission	Dedicated (non-dial-up) connection with 200 kilobits per second minimum
Web browser	Microsoft® Internet Explorer Version 5.0 or higher
Hard drive	10 MB free disk space minimum
Macromedia Flash Player™ software	Version 5.0 or higher. If not available, downloaded from Web during certification process
Java Virtual Machine™ software	If not available, downloaded from Web during certification process

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Writing Online Study.

[26] A minitest was developed to ensure that computers had adequately rapid data transmission and capacity to administer the assessment efficiently, and also to determine that the appropriate software had been downloaded.

Conducting the Administrations

When some or all of a school's computers could not be used to deliver WOL, the Westat administrator brought up to five laptop computers into the school to use for testing.[27] Table 6-1 summarizes the method of WOL delivery. As shown, the majority of students were tested on NAEP laptop computers. In many cases, this was because schools had only Macintosh computers available, which were not supported by the WOL software, or the school's internet connectivity was not sufficiently robust to support the WOL administration.

In most cases, WOL was conducted in a similar way to an individualized administration. After the administrator logged a student on to the computer, the student was given a one-page handout of directions to read silently, and then moved through the tutorial and the test at his or her own pace. As students completed the WOL session, they were dismissed. This procedure allowed more students to be tested in a shorter period of time, as some students finished more quickly than others, and new students could then be logged on immediately.

Although some computer-based testing programs have had problems with security, Westat administrators did not report any such concerns. This may have been due in part to the small numbers of students tested at any given time, which allowed for close monitoring, and to NAEP not being perceived as a high-stakes test. In addition, security precautions were taken in the design and delivery of WOL. These included logging onto the test delivery website with an administrator ID and password, and logging students on with specific ID numbers. At the conclusion of the testing session, Westat administrators routinely cleared each computer's cache, which might have retained copies of items, and deleted the browser history, which would have retained the secure delivery site's web address. Further precautions would be taken in an operational NAEP assessment, which would employ commercial, rather than research grade, test delivery software. Commercial software typically incorporates security mechanisms that prevent students from temporarily exiting the test to use other programs or files, and that automatically clear the computer of any residual test content once the assessment has ended.

Table 6-1. Percentage distribution of students and schools, by computer configuration, used to deliver the Writing Online test, grade 8: 2002

Computer configuration	Percent of students	Percent of schools
NAEP laptop	65	59
Internet	35	27
Both	†	14

† Not applicable.
NOTE: Detail may not sum to totals because of rounding. The number of students who participated in the study was 1,308 and the number of schools was 157.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Writing Online Study.

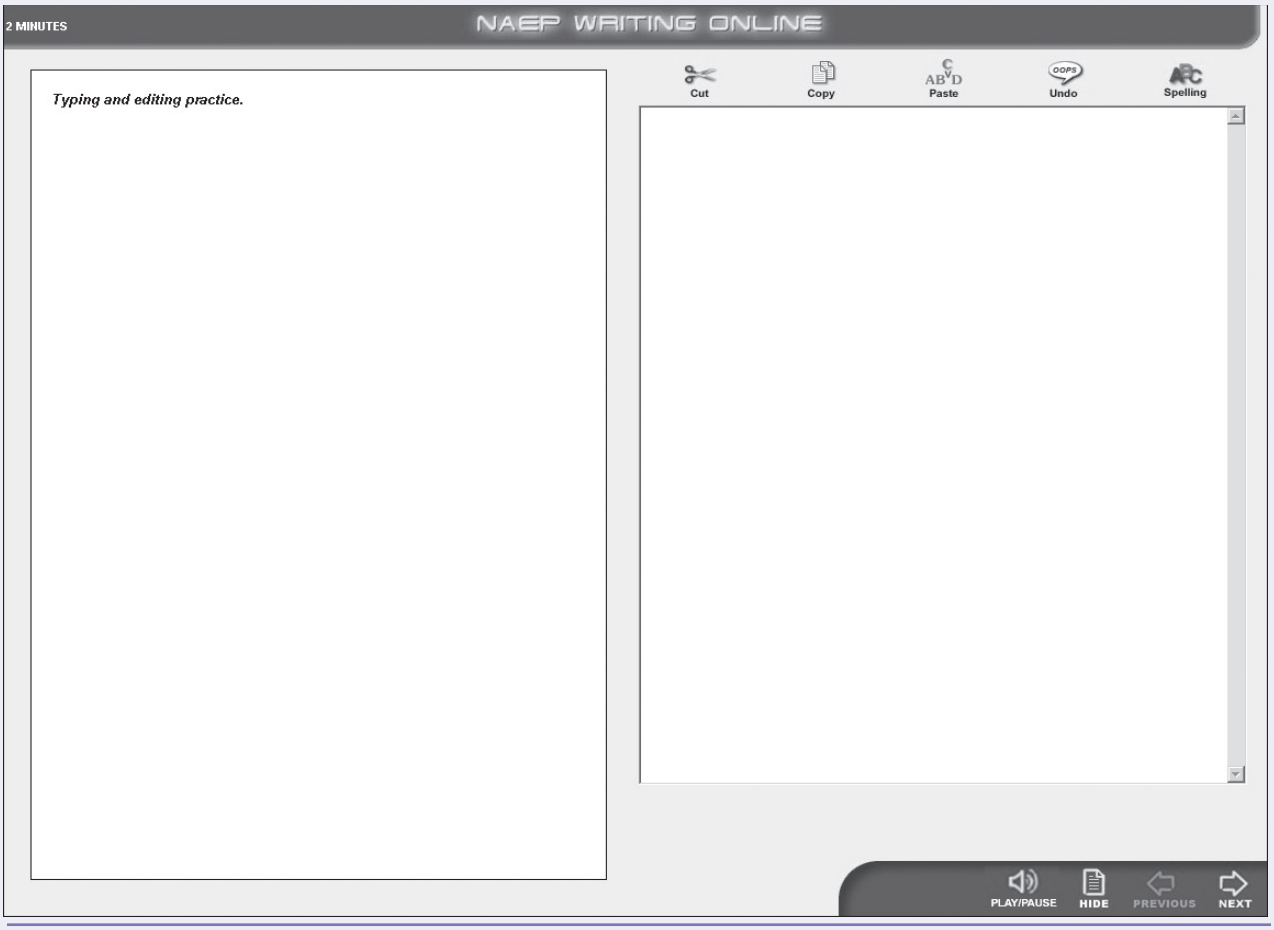
[27] The laptops used were Toshiba 1800 S203 notebook computers with a Windows 2000 operating system, 14 GB hard drive, 256 MB memory, external Microsoft mouse, and a Xircom Realport network card installed.

Accommodations for Students With Disabilities: WOL Voicing

As a preliminary step in studying how technology might be used to assist students with disabilities, a voicing version of WOL was developed, which presented selected components of WOL aloud through digitally recorded speech. Seventy-four students (female, 41 percent; White, 57 percent) participated in a preliminary voicing study, separate from the main WOL study. For the voicing study, field administrators were instructed to select students who had Individualized Education Programs (IEPs) that required "read aloud" accommodations, students with a print-related learning disability who might benefit from having directions read to them, or low-vision students who could be tested with the available accommodations.

WOL directions and the two essay tasks were the only voiced components. The voicing of the text was activated whenever a student clicked onto one of the directions or task screens. Figure 6-2 shows a sample screen from the voicing form of WOL. Once the voicing started on a given screen, clicking on the Play/Pause button in the lower-right corner of the screen paused the voice recording at that point. When the text for a given screen had been read completely, clicking on the Play/Pause button began the voicing of the text for that screen over again.

Figure 6-2. A sample Writing Online voicing screen, grade 8: 2002



SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Writing Online Study.

All WOL voicing tests were administered via NAEP laptop, with headphones attached. Field administrators were asked to report any difficulties students had with either the headphones or adjusting the volume. Only six field staff reported any difficulty with the headphones, which largely related to students' complaints about the headphones being uncomfortable for their ears.

Some field staff found the voicing test burdensome to administer because they still needed to read the tutorial and other unvoiced portions to the students. Field administrators reported that most of the students who used the voicing version thought the prerecorded audio was helpful. These students were especially enthusiastic about their ability to control volume and to repeat passages. Although a majority of the students thought the accommodations were "adequate," some expressed their disappointment that all sections of the test, particularly the tutorial, were not available with voicing. When asked whether they would prefer more complete voicing to a human reader, just over one half of the students said they would prefer the voicing. The most common reasons were because they "wouldn't waste someone's time reading," would find it "clearer/more understandable," and it would allow repetition of the voicing sections.

Table 6-2 presents the unweighted means for performance on the voicing version of the WOL test. Because the sample is neither large nor representative, the data should be regarded as descriptive only.

The correlation of scores between the two essays on the voicing test was .70 ($n = 66$). As a reference, the comparable value for the total group of students taking WOL was .63 ($n = 1{,}255$).

Table 6-2. Unweighted means for students with disabilities taking the voicing version of Writing Online, by essay and demographic group, grade 8: 2002

	Save a Book		School Schedule	
	Mean	Standard deviation	Mean	Standard deviation
Total	2.3	1.0	1.9	1.0
Gender				
Male	2.1	0.9	1.8	0.9
Female	2.4	1.1	2.1	1.1
Race/Ethnicity				
White	2.5	1.0	2.3	1.1
All other races	1.9	0.8	1.5	0.6

NOTE: The number of students responding was 70 for "Save a Book" and 69 for "School Schedule."

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Writing Online Study.

Equipment Performance

Overall, the WOL administrations ran smoothly. However, some minor difficulties did occur. The Westat Help Desk logged 80 requests for assistance from the field administrators. As indicated in table 6-3, many of these calls were unrelated to the WOL test itself (e.g., 32 percent concerned "other issues," and 9 percent problems with administrator computers). The most common test-related calls concerned difficulties with either laptops or PC certification. The Westat Help Desk also received 16 calls from staff at participating schools, most of which were requests for general information about the study or questions regarding the administration date and procedures.

Very few hardware-related problems were reported, and none of the laptops experienced a failure serious enough to require replacement. The WOL software functioned extremely well, and only two software updates were distributed during the field period. The first update was sent to correct a problem with accepting booklet IDs during the login process, and the other to eliminate a dialog box labeled "Done initializing applet" from appearing. In both cases, the updates were handled by mailing a computer diskette to the administrators, who were instructed to apply the update to each of their WOL laptops. These updates were performed with little difficulty, and the Help Desk was able to assist with the few problems that did arise.

More notable is the fact that few instances were reported of computers locking up, which did occur with some frequency in the 2001 Math Online study (Sandene et al. 2005, Part I). Table 6-4 summarizes the most common technical difficulties reported by Westat administrators, most of which were resolved on-site by the administrators themselves.

Table 6-3. Percentage distribution of calls reported to the Westat help desk, by reason for call, grade 8: 2002

Reason for call	Percent of calls
Total	100
Laptop problems during administration	23
PC certification difficulties	19
Software problems	9
Administrator computer problems	9
Administration procedures	8
Other (including problems with school control system, e-mail, data transmission, and data transfer)	34

NOTE: Administrator computers were not used for testing students, but were used by the Westat administrators to maintain field records and to transmit data to the Westat home office. The Westat help desk received a total of 80 calls. Detail may not sum to total because of rounding.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Writing Online Study.

Table 6-4 Percentage distribution of technical problems reported by the Westat administrators, grade 8: 2002

Technical problem	Percent of calls
Total	100
Computer(s) freezing	15
Slow computer(s)/connection	13
Invalid ID	13
Data lost	10
Error message	6
Spell check	4
Other (including problems with mouse, tab keys, highlighting on screen, and one-time-only situations)	40

NOTE: Detail may not sum to total because of rounding. The number of technical problems reported was 124.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Writing Online Study.

Initially, NAEP administrators had some concerns about the security and transporting of the NAEP laptops. To ensure protection of the laptops, each administrator packed his or her supply in a single, locked suitcase on wheels. This made transporting the laptops from school to school relatively safe, although size made the case somewhat cumbersome to maneuver.

Student and School Staff Reactions

NAEP administrators informally obtained feedback from students regarding their reaction to the test. Although numbers of comments are reported, these should be taken as descriptive only (Westat 2002). Generally, administrators reported far more positive reactions than negative ones from students. When asked what they liked most about WOL, 722 comments were received compared with 417 comments received regarding what students liked least. The most common positive responses were the following: liked using the computer format (185), liked typing (68), test was easy (66), liked writing (42), liked using the laptop (32), and it was fun (30). The most common negative responses were the following: time limit/too short/too long (78), did not like writing (34), did not like typing (33), and did not like essay portion (28).

Students were also asked if they thought they write better on computer or paper. Of 929 responses, the overwhelming majority (76 percent) reported that they write better on the computer, while 21 percent indicated that they write better on paper. (The remaining three percent of students reported that they write equally well on computer and paper.) Those students who reported that they write better on the computer gave reasons such as the following: typing is faster (119), editing is easier (107), editing tools are useful (102), neatness is improved (83), typing is easier (65), and writing by hand cramps their hands (35). Students who reported that they write better on paper gave reasons such as the following: writing is faster (43), not a proficient typist (29), easier to express ideas (26), and not comfortable using the computer (26).

NAEP administrators also informally asked school staff for their reactions to the WOL administration. Of the 124 school staff comments received, 96 were positive, 2 negative, and 26 mixed or neutral. The most frequent positive comment was that the WOL administrators were very supportive of the school staff. Field staff also received comments about how smoothly the administration went and how eagerly and diligently the students participated. It should be noted that, per NAEP security policy, school staff did not actually view the content of the WOL test, so there were no comments about the test itself.

Data Quality

Because of technical problems, some students were prevented from working through the tutorials or the test questions without interruption. These problems included school internet connections that were occasionally dropped, and laptops that sometimes froze during administration. In these cases, administrators attempted to restart students where they had stopped. If this procedure was unsuccessful, students had to begin writing their responses again. Only eighteen cases, or about 1 percent of the 1,308 students administered the WOL test, experienced interruptions. This percentage was greatly reduced from that of the Math Online study (Sandene et al. 2005, Part I) conducted the previous year. In that study, 15 percent of the fourth-grade students and 11 percent of the eighth-grade students had their tests interrupted. The decline in incidence of interrupted sessions was due in large part to better functioning of the laptops used for the WOL test.

To help insure the integrity of the WOL data, when laptop computers were used in schools, the administrators were trained to back up the record files. A program on each of the WOL laptops allowed the administrators to quickly copy all of a day's data onto a diskette. After backing up the data onto the diskette, the data were copied onto the administrator's laptop, and then transmitted to Westat, as an additional safeguard. Files copied directly from the laptops were returned to NAEP at the end of the WOL study for data analysis.

7. Summary and Conclusions

The Writing Online study addressed measurement, equity, efficiency, and operational issues associated with conducting a NAEP writing assessment on computer. Data were collected from samples of eighth-grade students in approximately 160 schools throughout the United States.

The primary measurement question was 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 in essay score or essay length. However, for the second of the two essays, delivery mode significantly predicted the rate of valid responses. Approximately 1 percent more students responded to the second essay when it was delivered on paper rather than on computer.

With respect to equity, the study addressed three issues. The first equity issue 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 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-group categories examined, 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 about 0.15 standard deviation units.

In addition to its impact on scores, the effect of delivery mode on performance was evaluated for gender groups in terms of response length and frequency of valid responses. For the second essay, males wrote significantly fewer words on paper than on computer. Also for that second essay, a signficantly higher percentage of females responded on paper than on computer. The difference was about 2 percent.

The second equity issue 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 they normally work on and would perform better on them than on NAEP laptops. To address this question, a small experiment was conducted in which students were randomly assigned to take the WOL test on NAEP laptops or on school computers. In addition, analyses were done in the larger WOL sample, contrasting the performance of students who had been nonrandomly assigned to the two computer types but controlling for performance on the paper main NAEP writing assessment. Results from the two analyses were not completely consistent. In the experimental substudy, students scored lower on laptop than desktop but for only one of the two essays. In the quasi-experimental analysis, however, only female students performed lower on the NAEP laptops, but this group did so for both essays. In any case, the results do suggest that students may sometimes obtain different scores on writing tests administered on laptop versus school computers.

The last equity question concerned the impact of computer familiarity on online test performance. Students' responses to background questions suggest that the overwhelming majority had access to computers at home (91 percent) and used a computer to write at least to some degree (93 percent), although there was considerable variation on the extent of this type of computer use. To determine if this variation in computer familiarity affected WOL performance, self-reported computer experience and hands-on measures of keyboarding skill were used to predict online writing performance after controlling for paper writing score. This analysis showed that hands-on skill was significantly related to online writing assessment performance, so that students with greater hands-on skill achieved higher WOL scores, even when 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.

In addition to measurement and equity issues, the study considered questions related to efficiency. Here, the relative costs and timeliness of different assessment delivery modes were analyzed, as was the feasibility of one technological innovation, automated scoring. With respect to timeliness, it is anticipated that moving assessments to computer would not have any significant short-term effect on the pilot stage of the NAEP assessment cycle, but could possibly shorten the operational stage considerably by requiring fewer steps. The costs for an online assessment should be similar for assessment development, similar or higher for assessment delivery and administration, and similar or lower for scoring. Among the key cost drivers for assessment delivery are student sample sizes, the number of schools participating, how many students need to be assessed on NAEP laptops, and the number of students per school who can be assessed simultaneously. A considerable increase in program costs would result, for instance, from assessing a large sample in small groups, primarily on NAEP laptop computers.

Although human readers scored all student responses, the e-rater® automated scoring technology also was used to score all responses. Results showed that the automated scoring did not agree with the scores awarded by human readers. The automated scoring produced mean scores that were significantly higher than the mean scores awarded by human readers. Human scores also correlated significantly more highly with one another than with the automated scoring. Finally, the two human readers assigned the same score to papers with significantly greater frequency than the automated grader assigned the same score as either human reader.

The last set of issues considered in this study concerned field operations. At pre-administration visits, field staff worked with school personnel to determine whether local hardware and connectivity were sufficient to support internet delivery. If not, administrators brought in NAEP laptop computers, which were used for testing 65 percent of the students. The two principal reasons for laptop use were that schools had only Macintosh equipment, which was not supported by the WOL web-delivery system, or that school internet connectivity was not robust enough to administer the test. While administrations ran very smoothly overall, technical problems did cause a small number of interruptions. Even so, reactions from students and school staff to electronic test delivery were more often positive than negative.

8. Implications for NAEP

The study authors believe these results have important implications for NAEP. The main study finding was that the scores from writing tests taken by eighth-graders on computer are generally not different from ones taken on paper, at least at the level of aggregated group results.

Several important caveats, however, must be considered along with this claim of score comparability. First, although the NAEP reporting groups examined generally showed no significant differences between performance on paper and computer tests, these findings should be confirmed with larger samples before concluding that the two delivery modes are interchangeable for population groups. Second, under some conditions, comparability appears to be affected by whether the test is taken on a NAEP laptop or on a school computer. Also, even though measurable differences were not detected for group scores, the scores for individuals do appear to be affected by delivery mode. For a given level of paper writing skill, students with more hands-on computer facility appear to get higher scores on WOL than do students with less keyboard proficiency. Whether this score boost is an irrelevant one is not entirely clear.

A score advantage for students with keyboard proficiency was also found in the Math Online study (Sandene et al. 2005, Part I). In that case, a strong argument could be made for attributing the score boost to factors unrelated to mathematics skill. That is, students with higher levels of keyboard proficiency scored better on the online math test than did students with less keyboarding skill because the latter group would have had more trouble entering their answers, especially on constructed-response questions that called for more intensive computer interaction. Likewise, those with high keyboard proficiency did not have greater command of mathematics, just better command of the computer. This argument rests largely on the fact that the Math Online test did not include mathematically related tools (such as spreadsheets) that might have allowed the more intensive computer users to show mathematical proficiencies that could not be expressed on a paper test.

WOL, however, presents a more complex situation. In contrast to Math Online, the Writing Online study included a construct-relevant writing tool, the word processor. 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 use computers when learning to write not only produce written work that is of higher quality and greater length, but are 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 peers). And, they write better on computer than they do on paper because the computer offers them a tool that makes it possible to do so.

The complementary interpretation also holds. Holding paper writing proficiency constant, students with little practice writing on computer will not score as high 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 measurement and equity results have implications for how NAEP writing assessments should be interpreted. This study implies that, at the population level, 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 study also suggests that the population estimates from *either* mode alone are probably lower than the performance that would have resulted if students could have been tested using the mode in which they wrote best. This situation 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.

A second implication for interpretation is that the relationships of certain demographic variables to writing proficiency might have been different if that proficiency had been measured on computer. This would have likely been the case for any demographic variable related to computer familiarity, with the magnitude of the difference being a function of the strength of the relationship between familiarity and that demographic characteristic.

With respect to efficiency, the implications of this study for back-end processing are not completely clear. In this study, automated scores did not agree with scores assigned by human readers as highly as did scores between human readers. However, the operational scores from a pair of human readers may not be a sufficient validation criterion. Ideally, scores taken across a greater number of readers grading under less pressured conditions, in combination with other measures of writing skill, would provide a more sound comparative standard. Additionally, it is not clear how much lower levels of reader agreement would affect NAEP. Even if automated scoring were less accurate, it would be important to know the impact of that accuracy loss on NAEP population estimates. If the loss were small enough, the use of automated scoring could have little negative impact on results but considerable effect in lowering costs and faster reporting. Further, the writing component scores and diagnostics that are now available in some scoring programs could add to the type of information that NAEP provides. More research will be required to address these issues.

NAEP should expect the costs for conducting an electronic writing assessment to be considerable. A primary reason for high costs is that the school technology infrastructure is not yet developed enough to support national delivery via the Web directly to school computers. Thus, NAEP will need to supplement web delivery by bringing laptop computers into schools, though undoubtedly not to the same extent as in this study because school technology is being improved continually. In the longer term, however, cost issues may be overshadowed by considerations of validity and credibility. As students do more of their writing on computer, NAEP may find it difficult to defend the assessment of that skill on paper.

Future research on the delivery of electronic writing assessment in NAEP might address several issues. First, this 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 eighth-grade participants in this study. Similarly, results could vary if questions requiring considerably longer or shorter responses were used.

Second, future research should investigate the impact of differences in equipment configuration on NAEP population estimates. This study found some differences in performance as a function of whether a student used a NAEP laptop or a school computer to take the writing test. As school computers become the predominant delivery mechanism, variation across computers (e.g., monitor size, screen resolution, connection speed) may play a greater role in affecting performance irrelevantly. Such an effect has already been reported for differences in screen resolution and monitor size on reading assessments (Bridgeman, Lennon, and Jackenthal 2003). Such variation may impact writing assessment to the extent that differences in keyboard layout impact a student's ability to compose without devoting undue attention to the mechanics of text entry.

Finally, future studies should control as well as possible for differences in reader reliability across the modes because such differences can potentially invalidate 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 the current 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 of response did not suggest any notable divergence in score 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 may not have affected the study conclusions in any substantial manner.

NAEP's history has been one of leadership and innovation, and NAEP continues this tradition by looking at what is promising and what is problematic about technology-based assessment. A third Technology-Based Assessment study of problem solving in technology-rich environments will add to our understanding of how computers may help improve NAEP and educational assessment generally.

References

Bangert-Drowns, R.L. (1993). The Word Processor as an Instructional Tool: A Meta-Analysis of Word Processing in Writing Instruction. *Review of Educational Research, 63*(1): 69–93.

Bridgeman, B., and 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., and 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.

Burstein, J. (2003). The E-rater® Scoring Engine: Automated Essay Scoring With Natural Language Processing. In M.D. Shermis and J.C. Burstein (Eds.), *Automated Essay Scoring: A Cross-Disciplinary Perspective.* Mahwah, NJ: Lawrence Erlbaum Associates.

Burstein, J., Chodorow, M., and Leacock, C. (in press). Criterion℠ Online Essay Evaluation: Automated Evaluation of Student Essays for Writing Instruction. *AI Magazine.*

Burstein, J., Kukich, K., Wolff, S., Lu, C., and Chodorow, M. (1998, April). *Computer Analysis of Essays.* Paper presented at the annual meeting of the National Council for Measurement in Education, San Diego.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences.* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Fleiss, J.L. (1981). *Statistical Methods for Rates and Proportions* (2nd ed.). New York: John Wiley & Sons.

Goldberg, A., Russell, M., and 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., and Rollins, A.L. (2000). The Influence of Word Processing on English Placement Test results. *Computers and Composition, 17*(2): 197–210.

Keith, T.Z. (2003). Validity of Automated Essay Scoring Systems. In M.D. Shermis and J.C. Burstein (Eds.), *Automated Essay Scoring: A Cross-Disciplinary Perspective.* Mahwah, NJ: Erlbaum.

MacCann, R., Eastment, B., and Pickering, S. (2002). Responding to Free Response Examination Questions: Computer Versus Pen and Paper. *British Journal of Educational Technology, 33*(2): 173–188.

Powers, D., and 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., and 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., and 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., and 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., and Plati, T. (2000). *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., and 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., and 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.

Sandene, B., Horkay, N., Bennett, R., Allen, N., Braswell, J., Kaplan, B., and Oranje, A. (2005). *Online Assessment in Mathematics and Writing: Reports From the NAEP Technology-Based Assessment Project, Research and Development Series.* U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Taylor, C., Jamieson, J., Eignor, D., and 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.

Westat. (2002). *NAEP TBA 2002 Operations Report.* Unpublished report, Rockville, MD: Author.

Wolfe, E.W., Bolton, S., Feltovich, B., and 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., and 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., and 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.