



U.S. Department of Education
Institute of Education Sciences
NCES 2005-457

The Nation's Report Card™

Online Assessment in Mathematics and Writing:

Reports From the
NAEP Technology-Based Assessment Project,
Research and Development Series



The National Assessment of Educational Progress

What is The Nation's Report Card™?

THE NATION'S REPORT CARD™, the National Assessment of Educational Progress (NAEP), is a nationally representative and continuing assessment of what America's students know and can do in various subject areas. Since 1969, assessments have been conducted periodically in reading, mathematics, science, writing, history, geography, and other fields. By making objective information on student performance available to policymakers at the national, state, and local levels, NAEP is an integral part of our nation's evaluation of the condition and progress of education. Only information related to academic achievement is collected under this program. NAEP guarantees the privacy of individual students and their families.

NAEP is a congressionally mandated project of the National Center for Education Statistics within the Institute of Education Sciences of the U.S. Department of Education. The Commissioner of Education Statistics is responsible, by law, for carrying out the NAEP project through competitive awards to qualified organizations.

In 1988, Congress established the National Assessment Governing Board (NAGB) to oversee and set policy for NAEP. The Board is responsible for: selecting the subject areas to be assessed; setting appropriate student achievement levels; developing assessment objectives and test specifications; developing a process for the review of the assessment; designing the assessment methodology; developing guidelines for reporting and disseminating NAEP results; developing standards and procedures for interstate, regional, and national comparisons; determining the appropriateness of all assessment items and ensuring the assessment items are free from bias and are secular, neutral, and nonideological; taking actions to improve the form, content, use, and reporting of results of the National Assessment; and planning and executing the initial public release of NAEP reports.

The National Assessment Governing Board

Darvin M. Winick, Chair

President
Winick & Associates
Dickinson, Texas

Sheila M. Ford, Vice Chair

Principal
Horace Mann Elementary
School
Washington, D.C.

Francie Alexander

Chief Academic Officer,
Scholastic, Inc.
Senior Vice President,
Scholastic Education
New York, New York

David J. Alukonis

Chairman
Hudson School Board
Hudson, New Hampshire

Amanda P. Avallone

Assistant Principal &
Eighth-Grade Teacher
Summit Middle School
Boulder, Colorado

Honorable Jeb Bush

Governor of Florida
Tallahassee, Florida

Barbara Byrd-Bennett

Chief Executive Officer
Cleveland Municipal
School District
Cleveland, Ohio

Carl A. Cohn

Clinical Professor
Rossier School of Education
University of Southern
California
Los Angeles, California

Shirley V. Dickson

Educational Consultant
Laguna Niguel, California

John Q. Easton

Executive Director
Consortium on Chicago
School Research
Chicago, Illinois

Honorable Dwight Evans

Member
Pennsylvania House of
Representatives
Philadelphia, Pennsylvania

David W. Gordon

Sacramento County
Superintendent of Schools
Sacramento County Office
of Education
Sacramento, California

Kathi M. King

Twelfth-Grade Teacher
Messalonskee High School
Oakland, Maine

Honorable Keith King

Member
Colorado House of
Representatives
Colorado Springs,
Colorado

Kim Kozbial-Hess

Fourth-Grade Teacher
Fall-Meyer Elementary
School
Toledo, Ohio

Andrew C. Porter

Professor
Leadership Policy and
Organizations
Vanderbilt University
Nashville, Tennessee

Luis A. Ramos

Community Relations
Manager
PPL Susquehanna
Berwick, Pennsylvania

Mark D. Reckase

Professor
Measurement and
Quantitative Methods
Michigan State University
East Lansing, Michigan

John H. Stevens

Executive Director
Texas Business and
Education Coalition
Austin, Texas

Mary Frances Taymans, SND

Executive Director
National Catholic
Educational Association
Washington, D.C.

Oscar A. Troncoso

Principal
Socorro High School
Socorro Independent School
District
El Paso, Texas

Honorable Thomas J. Vilsack

Governor of Iowa
Des Moines, Iowa

Michael E. Ward

Former State Superintendent
of Public Instruction
North Carolina Public Schools
Jackson, Mississippi

Eileen L. Weiser

Member, State Board
of Education
Michigan Department
of Education
Lansing, Michigan

Grover J. Whitehurst (Ex officio)

Director
Institute of Education
Sciences
U.S. Department of
Education
Washington, D.C.

Charles E. Smith

Executive Director
NAGB
Washington, D.C.



U.S. Department of Education
Institute of Education Sciences
NCES 2005-457

Online Assessment in Mathematics and Writing:

Reports From the
NAEP Technology-Based Assessment Project,
Research and Development Series

August 2005

Brent Sandene
Nancy Horkay
Randy Elliot Bennett
Nancy Allen
James Braswell
Bruce Kaplan
Andreas Oranje

Educational Testing Service

In collaboration with

Mary Daane
Douglas Forer
Claudia Leacock
Youn-Hee Lim
Hilary Persky
Dennis Quardt
Fred Schaefer
Michael Wagner
Vincent Weng
Fred Yan
April Zenisky

Educational Testing Service

Taslima Rahman
Holly Spurlock
Project Officers
National Center for Education Statistics

U.S. Department of Education

Margaret Spellings
Secretary

Institute of Education Sciences

Grover J. Whitehurst
Director

National Center for Education Statistics

Grover J. Whitehurst
Acting Commissioner

The National Center for Education Statistics (NCES) is the primary federal entity for collecting, analyzing, and reporting data related to education in the United States and other nations. It fulfills a congressional mandate to collect, collate, analyze, and report full and complete statistics on the condition of education in the United States; conduct and publish reports and specialized analyses of the meaning and significance of such statistics; assist state and local education agencies in improving their statistical systems; and review and report on education activities in foreign countries.

NCES activities are designed to address high-priority education data needs; provide consistent, reliable, complete, and accurate indicators of education status and trends; and report timely, useful, and high-quality data to the U.S. Department of Education, the Congress, the states, other education policymakers, practitioners, data users, and the general public. Unless specifically noted, all information contained herein is in the public domain.

We strive to make our products available in a variety of formats and in language that is appropriate to a variety of audiences. You, as our customer, are the best judge of our success in communicating information effectively. If you have any comments or suggestions about this or any other NCES product or report, we would like to hear from you. Please direct your comments to

National Center for Education Statistics
Institute of Education Sciences
U.S. Department of Education
1990 K Street NW
Washington, DC 20006-5651

August 2005

The NCES World Wide Web Home Page address is <http://nces.ed.gov>.

The NCES World Wide Web Electronic Catalog is <http://nces.ed.gov/pubsearch>.

Suggested Citation

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* (NCES 2005-457). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

For ordering information on this report, write to

U.S. Department of Education
ED Pubs
P.O. Box 1398
Jessup, MD 20794-1398

or call toll free 1-877-4ED-Pubs or order online at <http://www.edpubs.org>.

Content Contacts

Taslima Rahman, 202-502-7316, Taslima.Rahman@ed.gov
Holly Spurlock, 202-502-7458, Holly.Spurlock@ed.gov

The work upon which this publication is based was performed for
the National Center for Education Statistics, Institute of Education Sciences,
by Educational Testing Service and Westat.

Online Assessment in Mathematics and Writing:

Reports From the NAEP Technology-Based Assessment Project, Research and Development Series

This publication presents the reports from two studies, Math Online (MOL) and Writing Online (WOL), 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?*
2. *What are the implications for equity?*
3. *What are the efficiency implications of using technology-based assessment compared with paper and pencil?*
4. *What are the operational implications of technology-based assessment?*

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.

This publication is organized into two parts. Part I contains the report from the Math Online study. Part II contains the report from the Writing Online study. Each report is paginated separately. The results from the TRE study will be found in a separate, subsequent report.

THIS PAGE INTENTIONALLY LEFT BLANK.

Part I:

Online Assessment in Mathematics

Brent Sandene
Randy Elliot Bennett
James Braswell
Andreas Oranje
Educational Testing Service

In collaboration with
Mary Daane
Douglas Forer
Claudia Leacock
Youn-Hee Lim
Dennis Quardt
Fred Schaefer
Michael Wagner
April Zenisky
Educational Testing Service

Taslina Rahman
Holly Spurlock
Project Officers
**National Center for
Education Statistics**

THIS PAGE INTENTIONALLY LEFT BLANK.

Executive Summary

The Math Online (MOL) study is one of three field investigations in the National Assessment of Educational Progress (NAEP) Technology-Based Assessment Project, which explores the use of new technology in administering NAEP. The MOL study addresses issues related to measurement, equity, efficiency, and operations in online mathematics assessment. The other two studies focus on the use of computers in assessing writing and problem solving.

In the MOL study, data were collected in spring 2001 from more than 100 schools at each of two grade levels. Over 1,000 students at grade 4 and 1,000 at grade 8 took a test on a computer via the World Wide Web or on laptop computers taken into schools. At both grades 4 and 8, the study collected background data concerning students' access to computers, use of them, and attitudes toward them. In addition, students were administered hands-on exercises designed to measure input skill.

Over 2,700 students at grade 8 took comparable paper-and-pencil tests. The students taking paper-and-pencil tests were assigned randomly to one of three forms. One paper-and-pencil form, which presented identical items to the grade 8 computer-based test, provides the main comparisons for the effect of computer delivery vs. paper delivery. The other two paper-and-pencil forms were used to study psychometric questions related to the automatic generation of test items.

A priori and empirical analyses were performed to explore the implications of technology-based assessment for measurement, equity, efficiency, and operations. A review of findings in these categories follows.

Measurement

- In general, eighth-grade NAEP mathematics items appear suitable for computer delivery. Content review of the questions from the 2000 mathematics assessment suggested that most questions could be computer-delivered with no or only moderate difficulty.
- At grade 8, mean scale scores on the computerized test were about 4 points lower than on the paper version, a statistically significant difference.
- At the item level, there was a mean difficulty difference of .05 on the proportion-correct scale between the computer and paper tests, meaning that on average 5 percent more students responded to the items correctly on paper than on computer. Also, on average, the differences appeared to be larger for constructed-response items than for multiple-choice questions.

Equity

- At grade 8, no significant difference in performance on the computer test vs. the paper test was detected for the NAEP reporting groups examined (gender, race/ethnicity, parents' education level, region of the country, school location, and school type), except for students reporting that at least one parent graduated from college. These students performed better on paper than on computer tests.
- Background data suggest that the majority of fourth- and eighth-grade students have some familiarity with using a computer. For example, 85 percent of fourth-graders and 88 percent of eighth-graders reported that they use a computer at home.
- Use of computers by students at school also appears to be common. Eighty-six percent of fourth-graders and 80 percent of eighth-graders reported using a computer at school.
- To explore the possibility that, for some students, lack of computer familiarity impeded online test performance, both self-reported and hands-on indicators of computer familiarity were used to predict online test performance. At both grades, results suggested that performance on computer-delivered mathematics tests depended in part on how familiar a student was with computers.

Efficiency

- On the basis of a content analysis, about three-quarters of the items used on the NAEP 2000 mathematics assessment appear amenable to automatic generation. Geometry and Spatial Sense was the only framework content area for which the majority of the items could not be automatically generated.
- The degree to which the item-parameter estimates from one automatically generated item could be used for related automatically generated items was also investigated. Results suggested that, while the item-parameter estimates varied more than would be expected from chance alone, this added variation would have no statistically significant impact on NAEP scale scores.
- Eight of the nine constructed-response items included in the computer test at each grade were scored automatically. For both grades, the automated scores for the items requiring simple numeric entry or short text responses generally agreed as highly with the grades assigned by two human raters as the raters agreed with each other. Questions requiring more extended text entry were scored automatically, with less agreement with the grades assigned by two human raters.
- Based on an analysis of typical test development cycles, it is estimated that moving NAEP assessments to the computer would not have any significant short-term effect on the pilot stage of the NAEP development cycle but could possibly shorten the operational stage somewhat by requiring fewer steps.

Operations

- Although most tests were administered via laptop computers brought into schools by NAEP administrators (80 percent of students at fourth grade and 62 percent at eighth grade), a portion of schools tested some or all of their students via the Web (25 percent of the schools at grade 4 and 46 percent of schools at grade 8).
- Most administrations went smoothly, but technical problems caused some tests to be interrupted. Interrupted test sessions were associated with lower test scores by a statistically significant, but small, amount.
- Perhaps due in part to experiencing more frequent technical problems, eighth-grade students taking tests on NAEP laptops scored significantly lower than those taking tests on school computers, thereby contributing to the lack of comparability found between computer and paper tests.

Implications of Findings

The authors believe that these findings have several implications for NAEP:

- Most NAEP mathematics items could be computer delivered, arguably improving the measurement of some content areas specified by the mathematics framework. At the same time, conventional delivery may be needed for other items, especially those that require the manipulation of a real (as opposed to a simulated) physical object.
- Although the computerized test was somewhat more difficult than its paper counterpart for the population as a whole, it may be possible in future assessments to put tests given in the two modes on the same scale by administering a subset of common items in each mode to different randomly assigned groups of students.
- Even though most students reported some familiarity with technology, differences in computer proficiency may introduce irrelevant variance into performance on NAEP mathematics test items presented on computer, particularly on tests containing constructed-response items. For the near term, NAEP should be particularly thoughtful about delivering computer mathematics tests, especially when they include constructed-response items or where students have limited experience with technology.
- In the not-too-distant future, constructed-response mathematics tests may be feasible as keyboarding skills become pervasive, improved computer interfaces offer simpler means of interaction, and designers become more proficient in their renditions of open-ended items. When that occurs, automated scoring may help reduce NAEP's costs, increase speed of reporting, and improve scoring consistency across trend years.
- Automatic item generation might help to increase NAEP's efficiency, security, and depth of content coverage. Item variants could offer the opportunity to cover framework content areas more comprehensively, permit generation of precalibrated replacements for questions that have been disclosed, and allow the creation of item blocks as the assessment is administered.
- NAEP should expect the transition and near-term operating costs for electronic assessment to be substantial. However, the program may still need to deliver some assessments via computer despite higher cost. As students do more of their academic work on computers, NAEP may find it increasingly hard to justify documenting their achievement with paper tests.
- For the foreseeable future, occasional equipment problems and difficulties with internet connectivity are likely to cause interruptions in testing for some students or for some schools. Options for dealing with these events include discarding the data and reducing the representativeness of samples, retaining the data and possibly introducing bias into results, or conducting make-up sessions that could add considerable expense for NAEP.
- School technology infrastructures may not yet be advanced enough for national assessments to be delivered exclusively via the Web to school computers. However, if assessment blocks are initially composed solely of multiple-choice items and short constructed-response items, with more complex constructed-response questions left for paper blocks, web delivery may be possible for most schools.
- Future research should examine several factors related to irrelevant variation in online test scores. These factors include the impact of using laptop vs. school computers, the effectiveness of methods that attempt to compensate for differences in the operating characteristics of school machines, the effect of test interruptions on performance and comparability, the impact of constructed-response questions requiring different degrees of keyboard activity, the extent to which repeated exposure to tutorials and online practice tests might reduce variation in performance due to computer familiarity, and the impact of typed vs. handwritten responses on human grading.

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 Math 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 of Associate Commissioner of Education Statistics Peggy Carr, NAEP project directors Suzanne Triplett and Steven Gorman, TBA project directors Holly Spurlock and Taslima Rahman, and NCES consultants Vonda Kiplinger and Bob Evans. Without their consistent and strong backing, the Math Online study could not have been completed.

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 constructed-response items.

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, James Braswell, and the late Charlotte Solomon. Beth Durkin, Christine O’Sullivan, and Clyde Reese also participated in earlier stages.

The Math Online study was coordinated by Brent Sandene of ETS. Test booklet production, printing, distribution, scoring, and processing activities were conducted by ETS with contributions from Mary Anne Dorofee, Girish Jerath, Jeff Haberstroh, Pat Meier, and Catherine Shaughnessy. Contributors to the production of the online test and tutorials included Andrew Baird, Douglas Forer, Marylou Lennon, Lou Mang, Debbie Pisacreta, and Rob Rarich. Staff members who worked on usability testing included Holly Knott and Margaret Redman. Douglas Forer and Michael Wagner designed and implemented the delivery of the online test and the system for online scoring. Dennis Quardt and Claudia Leacock carried out the automated scoring.

Statistical and psychometric activities were overseen by Catherine McClellan and Andreas Oranje. Analysis was directed by Andreas Oranje and carried out by Fred Schaefer, Bruce Kaplan, Steve Isham, and Youn-Hee Lim. April Zenisky, of the University of Massachusetts, also contributed to analysis and writing. Database work was managed by Katharine Pashley, with assistance from Gerry Kokolis.

The design and production of this report were overseen by Loretta Casalaina, with contributions from Carmen Payton, Joseph Kolodey, and Rick Hasney. 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 and Linda Myers. 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. Westat’s activities were managed by Dianne Walsh, Dward A. Moore, Jr., Brice Hart, David Goldberg, and Brenda Ennis, with the assistance of Nia Davis. Sampling and weighting were conducted by Louis Rizzo and Tom Krenzke. Weighting systems work was completed by Bill Wall. Lonnie Broadnax assisted with student sampling software. Rob Dymowski developed the system used to draw student samples in each school and to report progress in the field. The day-to-day management of Westat’s technical and software systems was accomplished by Brice Hart with assistance from Fran Cohen and Karen Dennis.

Thanks are due to many reviewers both internal and external to NCES, ETS, and Westat, including Jim Carlson, Bob Evans, Steven Gorman, Vonda Kiplinger, Andrew Kolstad, Stephen Lazer, Anthony Lutkus, John Mazzeo, Michael Planty, and Marilyn Seastrom.

Contents

Executive Summary	vii
Foreword	x
Acknowledgments	xi
1. Introduction	1
2. Methodology	2
Study Sample	2
Instruments	3
Procedure	5
Constructed-Response Scoring	5
Scaling and Proficiency Estimation	5
3. Measurement Issues	6
Suitability of the Modes for Assessing NAEP 2000 Framework Content Areas	6
Ease of Measuring Existing Framework Content Areas on Computer	6
Framework Content Areas That Might Be Measured Better With Computer	7
Performance Differences Across Test Modes	8
Analysis of Item Difficulty for Eighth Grade	11
Analysis of Item Discrimination for Eighth Grade	16
4. Equity Issues	18
Population Group Performance	18
Performance as a Function of Computer Experience	18
5. Efficiency Issues	24
Automatic Item Generation	24
Empirical Analysis	25
A Priori Analysis	27
Automated Scoring	31
Scoring by Pattern and Feature Matching	32
Scoring Using Natural Language Processing	33
Procedure and Data Analysis Method	33
Cross-Validation Results	34
Relative Costs and Timeliness of Computer vs. Paper-Based Assessment	36
Relative Timeliness of Computer vs. Paper Testing	36
Relative Costs of Computer vs. Paper Testing	38
Relative Costs of Item and Software Development	38
Relative Costs of Test Delivery and Administration	38
Relative Cost of Scoring	39
6. Operational Issues	41
Recruiting Schools	41
Training Field Administrators	41
Preparing for the Administration	41
Conducting the Administrations	43
Student and School Reactions	44
Data Quality	44
7. Summary and Conclusions	46
8. Implications for NAEP	48
References	51
Appendix A: Inter-Rater Reliability	53
Appendix B: Ease of Assessing Existing NAEP Framework Content Areas on Computer	54
Appendix C: Students Omitting, Not Reaching, and Giving Off-Task Responses	57
Appendix D: Test Mode by Population Group Contrasts	58
Appendix E: Self-Reported Computer Experience	61
Appendix F: Student Mathematics Performance on Computer-Based Test and Paper-and-Pencil Test by Self-Reported Computer Experience	66

Tables and Figures

List of Tables

Table 2-1. Percentage of study participants, by gender and race/ethnicity, grades 4 and 8: 2001	3
Table 2-2. Instruments administered to each student sample, grades 4 and 8: 2001	4
Table 3-1. IRT b parameter estimates for items presented on computer and on paper, by extent of change required for computer and size of mode difference, grade 8: 2001	12
Table 3-2. Proportion-correct ($p+$) values for items presented on computer and on paper, by extent of change required for computer and size of mode difference, grade 8: 2001	13
Table 3-3. Comparison of IRT b parameter estimates for the MOL test to parameter estimates from three paper forms, grade 8: 2001	14
Table 3-4. IRT a parameter estimates for items presented on computer and on paper, by extent of change required for computer and size of mode difference, grade 8: 2001	16
Table 3-5. Comparison of IRT a parameter estimates for the MOL test to parameter estimates from three paper forms, grade 8: 2001	17
Table 4-1. Components of the input-skill measure, grades 4 and 8: 2001	21
Table 4-2. Coefficient alpha values for computer familiarity measures, grades 4 and 8: 2001	21
Table 4-3. Sample correlations among computer familiarity measures and with mathematics performance, grades 4 and 8: 2001	21
Table 4-4. Regression results for the effect of input skill and computer experience on computer mathematics test raw score, controlling for paper mathematics proficiency, grade 4: 2001	22
Table 4-5. Regression results for the effect of input skill and computer experience on computer mathematics test raw score, controlling for paper mathematics proficiency, grade 8: 2001	23
Table 5-1. IRT b parameter estimates for isomorphic vs. identical items for the three paper forms, grade 8: 2001	26
Table 5-2. Mean scores from scalings in which item parameters were and were not constrained to be equal across paper forms, grade 8: 2001	27
Table 5-3. Percentages of items from the NAEP 2000 mathematics assessment, by feasibility of automatic item generation, grade 8: 2001	28
Table 5-4. Percentage exact agreement between human judges and between automated grader and each human judge for the pattern-and-feature-matching method, grades 4 and 8: 2001	34
Table 5-5. Resolution of scoring disagreements between automated grader and either or both human scores for the pattern-and-feature-matching method, grades 4 and 8: 2001	35
Table 5-6. Percentage exact agreement between human judges and between c-rater™ and each human judge, grades 4 and 8: 2001	35
Table 5-7. Resolution of scoring disagreement between machine and either or both human scores for c-rater™, grades 4 and 8: 2001	35
Table 6-1. Primary reasons some school PCs failed certification for online testing, grades 4 and 8: 2001	42
Table 6-2. Number and percentage of students and schools, by method of computer-based test delivery, grades 4 and 8: 2001	43
Table 6-3. Percentage of performance problems, by cause reported to the Westat Help Desk, grades 4 and 8: 2001	43
Table 6-4. Mean MOL scale scores for students with and students without fragmented test-session records, grades 4 and 8: 2001	45
Table 6-5. Mean MOL scale scores for students testing on school computers and NAEP laptops, grades 4 and 8: 2001	45
Table A-1. Inter-rater reliability for constructed-response items, grade 4: 2001	53
Table A-2. Inter-rater reliability for constructed-response items, grade 8: 2001	53
Table B-1. Percentage of NAEP items, by framework content area and ease of implementation for computer delivery, grade 8: 2001	54
Table B-2. Percentage of NAEP mathematics items, by format and ease of implementation for computer delivery, grade 8: 2001	54
Table C-1. Mean percentages of students omitting, not reaching, and giving off-task responses for the MOL and paper tests, grade 8: 2001	57

Table E-1.	Percentage of students who report computer or Internet use at home, grade 4: 2001	61
Table E-2.	Percentage of students who report using a computer in and out of school, by frequency levels, grade 4: 2001	61
Table E-3.	Percentage of students who report using a computer for various purposes, grade 4: 2001	62
Table E-4.	Percentage of students who report using a computer for mathematics, by frequency level, grade 4: 2001	62
Table E-5.	Percentage of students agreeing with a positive statement about computer use, grade 4: 2001	63
Table E-6.	Percentage of students who report computer or Internet use at home, grade 8: 2001	63
Table E-7.	Percentage of students who report using a computer in and out of school, by frequency levels, grade 8: 2001	63
Table E-8.	Percentage of students who report using a computer for various purposes, grade 8: 2001	64
Table E-9.	Percentage of students who report using a computer for mathematics, by frequency level, grade 8: 2001	65
Table E-10.	Percentage of students agreeing with a positive statement about computer use, grade 8: 2001	65
Table F-1.	Mean scale scores and standard errors, by frequency of general computer use in and out of school, grade 8: 2001	66
Table F-2.	Mean scale scores and standard errors, by technology in the home, grade 8: 2001	66
Table F-3.	Mean scale scores and standard errors, by frequency of specific computer use, grade 8: 2001	67

List of Figures

Figure 3-1.	Overview of test items, grade 8: 2001	10
Figure 3-2.	Comparison of IRT b parameter estimates for items presented on computer and on paper, grade 8: 2001	11
Figure 3-3.	Comparison of IRT b parameter estimates for the MOL test vs. three paper forms, grade 8: 2001	15
Figure 3-4.	Comparison of IRT b parameter estimates for three paper forms, grade 8: 2001	15
Figure 5-1.	Pair-wise comparisons of IRT b parameter estimates for 14 isomorphs on three paper forms, grade 8: 2001	25
Figure 5-2.	Pair-wise comparisons of IRT b parameter estimates for 11 identical items on three paper forms, grade 8: 2001	26
Figure 5-3.	An item suitable for automatic generation that would require relatively limited effort for model creation, grade 8: 2001	29
Figure 5-4.	An item suitable for automatic generation that would require substantial effort for model creation, grade 8: 2001	30
Figure 5-5.	An item not suitable for automatic generation, grade 8: 2001	30
Figure 5-6.	Item for which the student must provide an answer and an explanation, grade 4: 2001	31
Figure 5-7.	Item for which the student must provide only an answer, grade 4: 2001	32
Figure 5-8.	Key steps in NAEP paper vs. computer test delivery, with estimated elapsed times	37
Figure 5-9.	Relative costs for NAEP of computer vs. paper assessment	40
Figure 6-1.	Technical specifications for school computers	42
Figure B-1.	A NAEP item measuring the geometry and spatial sense content area that requires a drawn response, grade 8: 2001	55
Figure B-2.	A NAEP item assessing the measurement content area that requires paper stimulus materials, grade 8: 2001	56
Figure D-1.	Mean scale score for MOL and P&P, by gender, grade 8: 2001	58
Figure D-2.	Mean scale score for MOL and P&P, by race/ethnicity grade 8: 2001	58
Figure D-3.	Mean scale score for MOL and P&P, by parents' education level, grade 8: 2001	59
Figure D-4.	Mean scale score for MOL and P&P, by region of country, grade 8: 2001	59
Figure D-5.	Mean scale score for MOL and P&P, by school location, grade 8: 2001	60
Figure D-6.	Mean scale score for MOL and P&P, by school type, grade 8: 2001	60

1. Introduction

This technical report presents the methodology and results of the Math Online (MOL) study, part of the National Assessment of Educational Progress (NAEP) Technology-Based Assessment Project. Funded by the National Center for Education Statistics (NCES), the Technology-Based Assessment (TBA) Project is intended to explore the use of new technology in NAEP. There are many possibilities for introducing new technologies to NAEP, in specific NAEP processes (e.g., item creation, test delivery), in specific content domains, and in specific assessment activities (e.g., the Main NAEP assessment vs. a special study). NAEP has historically been known for both rigorous and innovative methods. Since it is used to compare the progress of groups of students across time and to compare the progress of particular populations (those defined, for example, by gender, race/ethnicity, and school location), it is essential to NAEP's mission to preserve the comparability of assessments.

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 our 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 allow us to assess 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 computers. At the same time, it could increase participation of students with disabilities. In addition, it may better reflect the skills of students who routinely use the computer to perform academic tasks like writing.

3. What are the efficiency implications of using technology-based assessment compared with paper and pencil?

The Internet is facilitating a revolution in how companies do business. Along with other new technologies, the Internet may afford significant time and cost savings for large-scale assessments too.

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 MOL study focused on the issues associated with translating existing multiple-choice and constructed-response mathematics items from paper-and-pencil to computer delivery. The issues were:

- Measurement issues

How does test mode (i.e., presentation on computer vs. presentation on paper) affect the inferences that can be drawn about students' mathematics skill?

How do the modes compare with respect to the framework content areas that can be tested?

Do students perform differently across modes?

- Equity issues

How do population groups perform, and do mode effects vary across groups?¹

How are students with different levels of computer experience affected by technology vs. paper-based mathematics assessment? In particular, does a lack of computer familiarity appear to affect online test performance negatively?

- Efficiency issues

Is a technology-based mathematics assessment more cost-effective or timely than a paper one?

How might technological advances like automatic item generation and automated scoring affect the cost and timeliness of assessment?

- Operational issues

What are the logistical challenges associated with administering a NAEP mathematics assessment on a computer?

Are school facilities, equipment, software, and internet connectivity adequate?

Are schools willing to cooperate with the needs of a technology-based assessment?

Is the quality of data derived from an assessment delivered on computer acceptable?

¹ Issues related to students with disabilities were not addressed in this study.

2. Methodology

Study Sample

The target population for the MOL study consisted of fourth- and eighth-grade students enrolled in public and private elementary and secondary schools. The target sample sizes were 1,000 fourth-grade students and 1,000 eighth-grade students for online testing, and 2,750 eighth-grade students for paper-and-pencil testing. (A paper-and-pencil sample was not included at the fourth-grade level due to resource constraints.)

The sample, designed by Westat, was a full multi-stage, probability-based sample. In the first stage, the primary sampling units (PSUs) were counties or groups of counties. Because the MOL study did not require the same large sample sizes as a NAEP assessment, a subset of 52 PSUs was sampled from the 94 PSUs selected for the NAEP history and geography assessments (Lapp, Grigg, and Tay-Lim 2002; Weiss, Lütkus, Hildebrant, and Johnson 2002). To increase the chance of getting a representative subset, the sampling was done to include the 10 largest PSUs, half of the 12 smallest PSUs, and half of the remaining 72 PSUs.

In the second stage, schools were the sampling units. For fourth grade, elementary schools were sampled, and for eighth grade, middle and secondary schools were sampled. For each grade level, schools were chosen (without replacement) across all PSUs from a sorted list, with probabilities proportional to size.² The samples were designed to oversample large schools and schools with more than 10 percent Black students or 10 percent Hispanic students.

In the third stage, schools for the eighth-grade sample were assigned to testing conditions, with 110 schools to deliver both online and paper-and-pencil tests, and two schools to administer only paper-and-pencil examinations. Because it would be costly to transport computers to a school to test only a few students, the assignment of schools to conditions differed by school size. *Large* schools were assigned to administer tests in both delivery modes. *Small* schools, on the other hand, were assigned to be either all paper-and-pencil or both online and paper-and-pencil. Finally, the *smallest* schools were assigned

to be either all online or all paper-and-pencil, so that when a school was assigned to the online group, all of its selected students were tested on computer.³

In the fourth stage, students were selected. In the fourth-grade schools, 10 students were selected from each sampled school with equal probability and assigned to take the online test. (When the school had fewer than 10 eligible students, all eligible students were included.) In the 110 eighth-grade schools selected to administer both testing conditions, the students were assigned randomly to the online or paper-and-pencil forms.⁴ For all 112 eighth-grade schools, students in the paper-and-pencil condition were assigned randomly to one of three parallel forms.

Students were tested in April and May 2001. At grade 4, some 126 of 138 sampled schools (92 percent) and 1,094 of 1,255 sampled students (88 percent) were eligible and willing to participate in the study.⁵ Of these 1,094 examinees, 58 were not able to take the test because of technology problems, bringing the tested sample to 1,036. On average, 8 fourth-grade students per school were assessed. At grade 8, 110 of 129 sampled schools (87 percent) participated in the online condition and 108 of 131 schools (83 percent) took part in the paper condition. Schools participating in the online condition contributed 1,072 of 1,297 sampled students (84 percent). Of these 1,072 students, 56 were nonrespondents because of technology problems, reducing the tested sample to 1,016 participants. Schools administering the three paper test forms contributed the following numbers of students: 954 of 1,680 (83 percent), 926 of 1,652 (83 percent), and 906 of 1,628 (83 percent). On average, 9 eighth-grade students per school were assessed online and 26 were tested on paper.

Students who were judged by standard NAEP exclusion criteria as not being able to participate meaningfully in the testing activities without accommodations were excluded. At grade 4, 99 of the 1,255 sampled students were excluded. At grade 8, 94 of the 1,297 sampled students were excluded from online testing and 229 of 3,522 sampled students were excused from paper testing. These exclusion

² For fourth grade, the sorted list contained 25,184 elementary schools. For eighth grade, the list contained 14,836 secondary schools.

³ To avoid having to test too few students online at any given school, the following decision rules were used. For the smallest schools (between 1 and 11 grade-eligible students), a school was selected as all paper-and-pencil with 33/45 probability, and all online with 12/45 probability. For small schools (between 12 and 23 grade-eligible students), a school was selected as all paper-and-pencil with 21/45 probability, and half online, half paper-and-pencil with 24/45 probability.

⁴ This assignment was made with probabilities of 12/45 and 33/45, respectively, to ensure that roughly equal numbers of students were allocated to the computer and to each of the three paper forms.

⁵ Percentages of schools and students are weighted and may differ substantially from raw percentages.

rates, of between 6 percent and 8 percent, are similar to those for unaccommodated samples tested in the recent NAEP assessments in history and geography (Lapp, Grigg, and Tay-Lim 2002; Weiss, Lutkus, Hildebrant, and Johnson 2002).

Table 2-1 displays information about gender and race/ethnicity for the fourth-grade and eighth-grade samples assessed. Values in this table and throughout the report are weighted to make the results representative of the national fourth- and eighth-grade populations.

Table 2-1. Percentage of study participants, by gender and race/ethnicity, grades 4 and 8: 2001

	Grade 4 (n = 1,036)	Grade 8 (n = 3,802)
Gender		
Male	48 (1.7)	50 (1.0)
Female	52 (1.7)	50 (1.0)
Race/Ethnicity		
White	64 (0.5)	66 (0.3)
Black	14 (0.5)	14 (0.3)
Hispanic	17 (0.3)	14 (0.2)
Asian/Pacific Islander	3 (0.4)	4 (0.2)
American Indian/Alaska Native	2 (0.4)	1 (0.2)

NOTE: Standard errors of the percentages appear in parentheses. Detail may not sum to totals because of rounding.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

Instruments

All students took

- a *paper-and-pencil block of questions*, administered first. The paper-and-pencil block contained items from the NAEP 2000 mathematics assessment: 10 multiple-choice items for grade 4 and 20 multiple-choice items for grade 8. The block was used for scaling (described on p. 5) and as a covariate in selected analyses (p. 22).
- a *background questionnaire* to gather information about demographics and computer experience, presented last. The background questionnaire for

grade 4 contained 24 background questions with a 20-minute time limit, and that for grade 8 contained 30 questions with a 20-minute time limit.

After the initial paper-based block, students taking the *computer-based* test (hereinafter referred to as MOL) received

- an *online tutorial* in how to use the computer to complete the test. The online tutorial included instruction and practice in clicking on choices, clicking to shade or darken regions, moving back and forth between screens, correcting errors, and typing answers and explanations. The tutorial also had embedded tasks to provide a measure of the student's computer skill. The tutorial was split into two portions: a basic portion that preceded the test and a calculator portion that preceded the third test section. The tutorials can be viewed on the NCES web site (<http://nces.ed.gov/nationsreportcard/studies/tbatutorial.asp#mol>).
 - *online mathematics questions*, drawn from the existing NAEP item inventory and presented in three sections. Students were given paper to use for scratch work in answering these questions. The MOL fourth-grade test was based on an experimental NAEP administration conducted in 2000 that was composed of a "market basket" of questions intended to broadly represent the NAEP mathematics framework.⁶ The MOL version of this test included 32 questions: 22 multiple choice; 9 short constructed-response, which required such actions as entering a number or clicking on line segments to form a figure; and 1 extended constructed-response, which asked the student to provide an answer and enter an explanation. There were roughly 10 questions in each section, and the time allowed per section was either 15 or 20 minutes, depending on the number of constructed-response questions in that section. The third section permitted use of a four-function calculator that was on screen throughout the section.
- In the eighth-grade online test, there were 26 questions: 16 multiple choice, 8 short constructed response, and 2 extended constructed response. The time allowed for each section was 15 minutes and the number of questions per section was 10, 9, and 7, respectively. The third section permitted use of a scientific calculator available on screen throughout.

⁶ Twenty-eight of the market basket items were included in MOL. An additional item that appeared in the market basket as a single polytomously scored constructed-response question was broken into three dichotomously scored multiple-choice items. In addition, one item that did not appear in the market basket was used in MOL.

A paper version of the online test was administered only at grade 8. Funding limitations prevented concurrent collection of a paper sample at grade 4.

After the initial paper block, the students taking the eighth-grade *paper* tests took one of three forms: P&P (paper-and-pencil), Form A, or Form B. P&P contained exactly the same three sections of 26 mathematics questions as the online test, with the same time limits. Forms A and B contained 11 of the items that appeared on P&P. For each of the remaining 15 items on P&P, a variant was created, one for Form A and one for Form B. Each variant was designed to be mathematically identical to, but superficially different from, its P&P counterpart.

These variants were intended to investigate psychometric questions related to the computer generation of items discussed later in this report. For each of the paper-and-pencil test forms, the third section permitted the use of a scientific calculator provided by NAEP administrators.

Table 2-2 provides an overview of the instruments and student samples. Performance on the initial paper block provides a convenient mechanism for checking the equivalence of the grade 8 samples. For this grade, the raw-score means were 12.4, 12.3, 12.3, and 12.5 for MOL and the three paper samples, respectively.

Table 2-2. Instruments administered to each student sample, grades 4 and 8: 2001

Grade 4		Grade 8		
MOL (n = 1,036)	MOL (n = 1,016)	P&P (n = 954)	Form A (n = 926)	Form B (n = 906)
Initial paper block (10 items)	Initial paper block (20 items)	Initial paper block (20 items)	Initial paper block (20 items)	Initial paper block (20 items)
Online tutorial	Online tutorial	†	†	†
Online test (32 items) with embedded calculator tutorial	Online test (26 items) with embedded calculator tutorial	Paper test (P&P) (26 items)	Paper test (Form A) (26 items)	Paper test (Form B) (26 items)
Background questions (24 items)	Background questions (30 items)	Background questions (30 items)	Background questions (30 items)	Background questions (30 items)

† Not applicable.

NOTE: MOL=Math Online. P&P=Paper and Pencil. Forms A and B were paper forms constructed to be parallel to P&P. One item was removed from the analysis of the grade 8 tests due to poor scaling properties in the calibration step.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

Procedure

Constructed-Response Scoring

The test administered at each grade contained 10 constructed-response questions. A team of trained raters scored responses to these items. Raters used the rubrics and sample answers that had been developed for the items from NAEP paper assessments. Where needed, supplemental training responses were printed from the online versions of the items.

At grade 8, a single team of individuals led by a trainer scored both the online and the paper responses to each item. Responses written in test booklets were scored on paper; those completed on computer were presented to raters for scoring on computer. At grade 4, all student responses were scored by raters on computer.

A random sample of approximately 25 percent of the responses was double-scored to compute inter-rater reliability. For grade 4, exact agreement levels ranged from 87 percent to 98 percent for MOL. This range compares favorably to the agreement range of 88 percent to 100 percent for the earlier scoring of the experimental market basket form (NCS Pearson n.d.).

For grade 8, exact agreement ranged from 80 percent to 99 percent for P&P and 84 percent to 98 percent for MOL. Agreement levels within items for both grades can be found in appendix A.

Scaling and Proficiency Estimation

To scale items and estimate examinee proficiencies, the study used essentially the same multi-step process employed for NAEP assessments (see Allen, Donoghue, and Schoeps 2001, for complete details on these NAEP technical procedures). This process included calibration, conditioning, imputation, and transformation. Departures from the procedures typically used for NAEP assessments are noted, as appropriate.

The calibration process employed item response theory (IRT). IRT is a statistical method for relating item responses to estimates of student proficiency. For grade 4, calibration entailed estimating item parameters simultaneously for the initial paper-and-pencil block and MOL (42 items in total). For grade 8, the item parameters for the initial paper block, MOL, and the three paper forms were estimated together (45 questions in all). (One item and its variants on Forms A and B were omitted from the analysis because they introduced difficulties in obtaining a satisfactory scaling solution.) This univariate calibration step was repeated with several model variations for use in different analyses.⁷ For example, to facilitate the study of total-score mode effects, the calibration was conducted with item parameters constrained to be equal across MOL and the P&P form. For item-level comparisons, however, the calibration was conducted with parameters permitted to vary across the two testing modes. For such calibrations, the initial paper block items were constrained to be equal across examinee groups, thereby defining a common scale on which the MOL and the paper forms could be compared.

While the calibration step differed for the two grades, the conditioning, imputation, and transformation steps were the same at both levels. In conditioning, a univariate total score distribution on an arbitrary scale was predicted for each student based on demographic information, the item parameters estimated in the calibration step, and item responses to the MOL or paper test. This conditioning was done separately for the grade 8 sample taking MOL and for each of the samples taking the paper forms.

Next, for each student, five plausible values were sampled from the appropriate total-score (posterior) distribution. Finally, these plausible values were transformed to a scale with a mean of 200, a standard deviation of 30, and a range from 0 to 400.⁸

⁷ Although the main NAEP mathematics framework and assessment contain five subscales, the sample size and scope of this study only allowed a subset of the instrument to be used. Therefore, a multivariate calibration could not be obtained psychometrically or substantively. In addition, the high correlation between mathematics subscales in main NAEP supports the validity of a univariate calibration for this study.

⁸ This scale was chosen to avoid confusion with the NAEP mathematics scale, which is multidimensional and may measure a somewhat different construct.

3. Measurement Issues

Many studies have investigated the comparability of paper and computer tests for adults (e.g., Bridgeman 1998; Schaeffer, Bridgeman, Golub-Smith, Lewis, Potenza, and Steffen 1998; Schaeffer, Steffen, Golub-Smith, Mills, and Durso 1995). Mead and Drasgow (1993) reported a meta-analysis of studies that estimated the correlation between testing modes after correcting for unreliability. Across 159 estimates derived from tests in a variety of skill domains, they found the correlation for timed power tests, such as those used in achievement domains, to be .97, suggesting score equivalence, but the correlation for speeded measures, like clerical tests, to be .72. Further, for the timed power tests, the standardized mean difference between modes was .03, indicating that computerized tests were harder than paper versions, but only trivially so.⁹

At the elementary and secondary school level, the data are far more limited. Among the studies with large samples are those sponsored by the Oregon Department of Education and the North Carolina Department of Public Instruction. Choi and Tinkler (2002) assessed approximately 800 Oregon students in third and tenth grades with multiple-choice reading and mathematics items delivered on paper and by computer. They discovered that items presented on computer were generally more difficult than items presented on paper, but that this difference was more apparent for third-grade than for tenth-grade students, and more apparent for reading than for mathematics tests. For the North Carolina Department of Public Instruction, Coon, McLeod, and Thissen (2002) evaluated third-grade students in reading and fifth-grade students in mathematics, with roughly 1,300 students in each grade taking paper test forms and 400 students taking the same test forms on computer. All items were multiple-choice. Results indicated that for both subjects scale scores were higher for paper than for the online examinations.

Similar findings are emerging from the few, relatively small, studies that have been done with constructed-response items. These studies suggest that scores from free-response writing tests, and possibly from open-ended mathematics tests, may

differ across delivery mode (e.g., Russell and Haney 1997; Russell 1999; Russell and Plati 2001; Wolfe, Bolton, Feltovich, and Niday 1996).

This section considers how the mode of administering the mathematics assessment (i.e., on computer vs. on paper) affects the inferences that can be drawn about students' mathematics skill. In particular, two questions are addressed:

- How do the modes compare with respect to the framework content areas that can be tested?
- Do students perform differently across modes?

Suitability of the Modes for Assessing NAEP 2000 Framework Content Areas

In principle, test mode can make a difference in what can be measured. Paper presentation may allow some skills to be assessed that computer delivery does not, and vice versa. Since NAEP is a framework-governed assessment and the existing mathematics frameworks were developed with paper delivery in mind, this discussion focuses primarily on content areas that are already easily tested on paper but might be difficult to assess on computer.

Ease of Measuring Existing Framework Content Areas on Computer

To investigate the feasibility of using an online assessment to cover an entire NAEP mathematics framework, two ETS test developers and two technology staff members analyzed qualitatively each of the 160 items used in the NAEP 2000 eighth-grade assessment in terms of their potential for computer-based delivery. Each staff member reviewed each item independently. In their review, staff members considered suitability for on-screen presentation and general compatibility with the technology used for delivering Math Online items, as well as content-based issues. Staff members rated items as easy, moderately difficult, or difficult to implement online.¹⁰ Disagreements among judges over the suitability of individual items were resolved by using the more restrictive judgment. This strategy was employed to ensure a relatively conservative result.

The results of this analysis suggest that approximately 86 percent of the items from the NAEP 2000

⁹ The standardized mean difference is the difference between the means of the paper and computer groups divided by the within-groups standard deviation. A rule of thumb suggested by Cohen (1988) is to consider .2 as the minimum for "small" differences, .5 the minimum for "medium" differences, and .8 the minimum for "large" differences.

¹⁰ Most items rated as moderately difficult or difficult could be included in an operational online assessment; however, the development costs and potential problems associated with delivering such items online might argue for administering those items in paper form.

grade 8 mathematics assessment could be implemented for computer delivery with no or only moderate difficulty.¹¹ Of the five content areas specified by the framework, items from the Number Sense, Properties, and Operations area, the Algebra and Functions area, and the Data Analysis, Statistics, and Probability area appeared generally easier to implement than those from the Measurement and the Geometry and Spatial Sense areas (see appendix B). Measurement items judged difficult to implement included ones requiring use of rulers or protractors, where part of the intent of the framework is to determine how effectively the student can manipulate these tools in solving problems. Geometry items judged hard to deliver involved the manipulation of three-dimensional objects (e.g., arranging cut-out shapes to form a specified geometric shape), or required students to create detailed drawings as part of problem solving.

In addition to framework content areas, different response formats might be more or less difficult to implement in computer-based testing than in paper-based testing. Analysis suggested that NAEP constructed-response items were less often appropriate for computer delivery than multiple-choice items. These items tended to cluster in the Measurement, the Geometry, and the Data Analysis framework content areas. The constructed-response items judged difficult to implement included those that required tools such as rulers or protractors, manipulatives (such as cut-out shapes), and detailed drawings.

Based on their review, staff members concluded that certain kinds of items are currently less likely to work well for computer delivery in NAEP. It should be emphasized that these conclusions may not apply to other testing programs. Further, as the tools for creating and delivering computerized tests become more sophisticated, such items may work effectively in electronic tests.

Items less likely to be appropriate for an online NAEP assessment included those that

- are multipart or that would require more than a screen (e.g., because they have graphics needing a large amount of space);
- are intended in part to determine how effectively the student can manipulate some physical tool (e.g., a ruler or protractor);
- require the student to create drawings, enter a lengthy amount of text, or produce mathematical

formulas, each of which can be done on computer but not with equal facility by all students;

- require extended tutorials or lengthy item-specific directions for responding;
- require paper stimulus materials; or
- assume a screen resolution that is the same across all student computers as, for example, would be required if an on-screen object was to be measured and the delivery system was not able to control monitor resolution.

See appendix B for examples.

Framework Content Areas That Might Be Measured Better With Computer

While some current NAEP framework content areas may pose challenges to computer delivery, there are aspects of other content areas that arguably could be better measured on computer. Consider, for example, the Data Analysis, Statistics, and Probability area. Data analysis involves collecting, organizing, summarizing, and interpreting data. To assess these skills, NAEP has typically presented students with questions requiring the manipulation of very small data sets. Questions have revolved around the most common statistics (e.g., the mean, median, and mode). The large data sets found in the real world are not used because of constraints on the length of time a student can be tested and because those data sets would be impossible to analyze with the standard calculator provided by NAEP. Computer delivery, however, affords the opportunity to assess data analysis skills more authentically by making it possible to ask students to manage and manipulate reasonably large data sets.

Although this report focuses on assessment of fourth- and eighth-grade mathematics, a particularly good example of how better measurement might be achieved through computer presentation is found at twelfth grade. The new NAEP 2005 mathematics framework calls for twelfth-graders to

- calculate, interpret, or use mean, median, mode, range, interquartile range, or standard deviation;
- compare two or more data sets using mean, median, mode, range, interquartile range, or standard deviation describing the same characteristics for two populations or subsets of the same population; and
- estimate the probability of simple or compound events in familiar or unfamiliar contexts.

¹¹ These results should apply to the 2005 mathematics assessment framework to the extent that this new framework overlaps with the old one.

These subtopics might be more effectively assessed in computer-based testing than in paper-based testing. The computer could provide a small collection of utilities that the student could call up to carry out routine manipulations, e.g., “sort,” “sum,” “count,” and “find” functions. For the triplet (23, 13, 17), “sort” would produce 13, 17, 23, “sum” would produce 53, “count” would give the number of members in the set, in this case 3, and “find (2)” would find the second value in the sorted list, 17.

Students could be presented with a data set or partial set containing an unknown number of members. The first few members of the set, S , could be given as 342, 409, 153, etc. One item might ask the student to find for S the range of values, the mean value, and the median value. The student could then apply the “count S ,” “sort S ,” “sum S ,” and “find $S(n)$ ” utilities to get, for example:

Count S : 1287 (set S has 1287 members)

Sort S : 103, 105, 105, 106 ... 542, 543, 555
(the greatest value in S is 555)

Sum S : 415,701
(the sum of the values in S is 415,701)

Find $S(644)$: 299 (the median in S is 299)

Using this information, the student can also derive the range, $555-103 = 452$, and the mean, $415,701/1287 = 323$.

The measurement of NAEP mathematics content areas might be improved through computer delivery in other ways. Some framework subtopics require students to locate points on a number line, plot points on a coordinate grid, graph linear and nonlinear equations, or classify figures according to their properties. However, when answered on paper, it can be difficult to score such constructed responses reliably. For example, if a student is asked to mark the location of $2/3$ on a number line, is the response close enough to receive credit? With paper delivery, scoring is currently done by human judges

working on computer with the screen images of students’ responses. Templates are generally available for questions like the one above and these templates do make scoring more reliable. However, the templates must be accurately applied and some judgment is often required. In addition, when hundreds of papers are scored, errors do occur. In the case of automated scoring, a tolerance can be established for allowable deviations from the correct answer. Arguably, a score could be assigned by the machine with higher reliability than could be achieved through human grading.

Performance Differences Across Test Modes

Given that a framework content area can feasibly be measured on computer, it is still important to investigate whether computer presentation affects students’ scores, and whether it affects subgroups of the population differently. If such differences are found, the scores on computer-based assessment are not equivalent to scores on traditional paper-based assessment. This section reports analysis of student performance in the online and the paper-based tests. It focuses on the eighth grade, since computer-based and paper-based tests were administered simultaneously to independent representative samples of eighth-graders.

The most direct method of detecting performance differences is to compare the eighth-grade mean scale scores for MOL and the paper form using the same items (P&P). For this analysis, mean scores were generated from a scaling in which the item parameters for each mode were constrained to be equal, thereby forcing mode differences into the total scores. For MOL, the mean eighth-grade scale score was 198, whereas for P&P it was 202. This difference is statistically significant ($t = -2.26$, $p < .05$).¹² In terms of practical importance, the difference of .14 standard deviation units is less than the .2 minimum for “small” effects suggested by Cohen (1988).¹³

¹² One possible cause of these differences is the extent to which students omit, don’t reach, or give off-task responses more frequently in one versus the other mode. Table C-1 gives the mean percentages of eighth-grade students not responding in each of these three ways. In general, the percentages were so small as to be of limited consequence.

¹³ The effect size is given in the standard deviation units of the total-score scale, which is 30 points.

To explore the impact of test mode on eighth-grade performance in more detail, an analysis of the difficulty of each item and how well it discriminated between higher- and lower-performing students was performed. Comparisons were made of the estimated item parameters across paper and computer delivery. Figure 3-1 gives a description of each item, its NAEP framework content area, its format, how much the item was changed in rendering for computer, and whether it was entirely text-based or included a table or graphic. Three item formats were used: multiple-choice (MC), short constructed-response (SCR), and extended constructed-response (ECR). SCR questions were scored on either a 2- or 3-point scale, while ECRs were scored on a 5-point scale. Twenty-one of the items were changed only minimally for computer presentation. Four differed more in their computer format from the originals in paper format.

For each of the eighth-grade items, IRT a (discrimination) and b (difficulty) parameters were estimated as part of scaling, using the examinee response data from the two administration modes. Proportion-correct ($p+$) values were also computed. Two-tailed z -tests for independent samples were conducted to determine whether the item's IRT difficulty and discrimination estimates differed significantly when the item was presented on computer vs. when it was presented on paper.¹⁴

¹⁴ To compute the difference between item parameters, the standard errors produced by Parscale were used to compute a pooled standard error: $SE_p = \sqrt{SE_1^2 + SE_2^2}$. Next, a test statistic was computed: $Z = \frac{\theta_1 - \theta_2}{SE_p}$ where θ_1 and θ_2 are item parameter estimates for MOL and P&P. The distribution of this statistic is approximated by a normal distribution. This assumption seems justified given that the item parameters were estimated based on the total sample within each mode, resulting in a relatively large number of degrees of freedom. At the 0.05 level (two-sided), this statistic has confidence interval bounds of -1.96 and +1.96. This statistic assumes examinees were drawn from a simple random sample and does not take into account the clustered nature of the sample used in this study.

Figure 3-1. Overview of test items, grade 8: 2001

Item	Description	Framework content area	Format	Changes required for computer rendering	Stimulus type
1	Choose the numerical expression that best represents the area of a given rectangle	Measurement	MC	Minimal	Graphic (Picture, Rectangle)
2	Mark the place on number line to show the location of a given fraction	Number sense	2 pt SCR	Minimal	Graphic (Number Line)
3	Extend a pattern of numbers and provide the rule used to find the answer	Algebra & functions	3 pt SCR	Minimal	Text-based
4	Given objects that balance on a scale, identify equivalent weights between objects	Algebra & functions	MC	Minimal	Graphic (Symbols and figure)
5	Identify the best estimate of floor area	Measurement	MC	Minimal	Text-based
6	Compute the effect of an incremental increase of a variable in a mathematical expression	Algebra & functions	MC	Minimal	Text-based
7	Given the sum of three numbers, answer a question related to the relationship between the smallest and largest number; explain this answer	Number sense	2 pt SCR	Minimal	Text-based
8	Given certain angle measures related to a triangle, determine the angle measure of a specified angle in the triangle	Geometry and SS	MC	Minimal	Graphic (Angle meas. of triangle)
10	Describe the speed of a cyclist at various points in time, given a graph of time vs. distance	Data analysis, S & P	5 pt ECR	Considerable	Graphic (Graph of speed)
11	Estimate the difference between two weights	Number sense	MC	Minimal	Text-based
12	Given a table of data, apply the concept of a pictograph to represent one piece of data in the table	Data analysis, S & P	MC	Minimal	Graphic (Symbol, table)
13	Apply the concept of symmetry to visualize the result of folding a marked strip of paper	Geometry	3 pt SCR	Minimal	Graphic (Picture, click-on)
14	Identify a point on a grid that is the fourth vertex of a rectangle, given the location of the other three vertices	Algebra & functions	MC	Minimal	Graphic (Cartesian coordinates)
15	Determine the value of a point on a number line	Algebra & functions	2 pt SCR	Considerable	Graphic (Number Line)
16	Determine the value of a point on a number line	Algebra & functions	2 pt SCR	Considerable	Graphic (Number Line)
17	Determine the value of a point on a number line	Algebra & functions	2 pt SCR	Considerable	Graphic (Number Line)
18	Identify a geometric figure to illustrate a logical argument	Geometry and SS	MC	Minimal	Graphic (Picture answer choices)
19	Evaluate the appropriateness of a sampling design and explain the answer	Data analysis, S & P	3 pt SCR	Minimal	Text-based
20	Compute the total product cost, given unit pricing	Number sense	MC	Minimal	Text-based
21	Given deposits and debits in a checkbook, determine the final balance	Number sense	MC	Minimal	Graphic (Table)
22	Select the best graphical representation of an inequality	Algebra & functions	MC	Minimal	Graphic (Picture choices, shapes)
23	Demonstrate an understanding of scientific notation	Number sense	MC	Minimal	Graphic (Picture of calculator)
24	Given the formula, convert a temperature between °F and °C	Algebra & functions	MC	Minimal	Text-based
25	Given the formula, compute the volume of a figure	Measurement	MC	Minimal	Text-based
26	Given a diagram showing a detour and a car with a partially full tank of gas, determine whether the car will make it to a gas station shown on the map before running out of gas.	Number sense	5 pt ECR	Minimal	Graphic (Map)

NOTE: Item 9 was dropped from analysis because it introduced scaling difficulties. SCR=short constructed-response. ECR=extended constructed-response. MC=multiple choice.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

Analysis of Item Difficulty for Eighth Grade

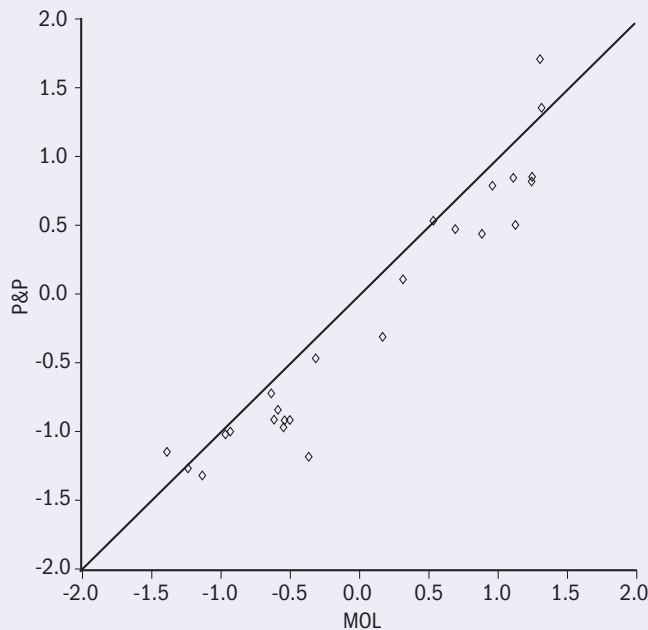
The IRT b parameter positions the item on the ability scale at the point where the probability of a correct response is .5 (after adjusting for guessing in multiple-choice items). The parameter is commonly estimated to range from -2.0 to 2.0. Items with higher b values are more difficult.

Figure 3-2 presents the scatter plot of the IRT b values for the 25 paper-administered items against

the b values for the same 25 MOL items. Two results stand out. First, the relationship of the estimated parameters to one another is almost identical across modes: the product-moment correlation is .96. Second, the preponderance of items falls on the MOL side of the identity line, suggesting that items presented on computer were more difficult than the same items on paper.

Table 3-1 shows the IRT b parameter estimates for

Figure 3-2. Comparison of IRT b parameter estimates for items presented on computer and on paper, grade 8: 2001



NOTE: MOL=Math Online. P&P=Paper and Pencil.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

each item, along with the z -test for statistical significance of the difficulty differences. This test was performed only for the 20 dichotomously scored items because standard errors for the polytomous items could not be reliably estimated. The items needing minimal change for presentation on computer appear in the upper section of the table. Within that category, and within the list of items that needed greater change, the items are listed by the size of the difference in the b parameter estimates

(computer minus P&P). As the table indicates, 8 of the 20 items were significantly different, with all 8 more difficult on computer than on paper. Taken across all 25 items, the mean of the differences was equal to .22 logits (range = -.25 to .81). Because positive and negative differences can cancel each other out, the mean of the absolute values of the differences was also calculated. This equaled .28 logits.¹⁵

¹⁵ All 25 items were included in computing the mean differences to give an item-level representation of the mode effect already detected for the mean scale scores. These scale scores incorporate all items whether or not the items show significant differences across delivery modes.

Table 3-1. IRT *b* parameter estimates for items presented on computer and on paper, by extent of change required for computer and size of mode difference, grade 8: 2001

Item and format	Estimated <i>b</i>		Difference (computer minus paper)	z value
	Computer	Paper		
Items needing minimal change to render on computer				
11 MC	-1.38 (.131)	-1.14 (.128)	-.25	-1.34
8 MC	1.32 (.090)	1.37 (.117)	-.05	-0.35
1 MC	.54 (.090)	.55 (.083)	-.01	-0.06
18 MC	-1.23 (.136)	-1.25 (.110)	.02	0.11
12 MC	-.96 (.123)	-1.01 (.124)	.05	0.27
3 SCR	-.93 (**)	-.99 (**)	.06	***
2 SCR	-.63 (.078)	-.72 (.079)	.09	0.78
5 MC	-.31 (.193)	-.46 (.187)	.15	0.55
23 MC	.97 (.070)	.80 (.086)	.17	1.52
20 MC	-1.13 (.140)	-1.31 (.149)	.18	0.88
14 MC	.32 (.064)	.13 (.068)	.20	2.13 *
22 MC	.70 (.104)	.49 (.082)	.21	1.59
4 MC	-.58 (.127)	-.84 (.131)	.26	1.40
25 MC	1.12 (.117)	.85 (.073)	.27	1.94
6 MC	1.26 (.081)	.87 (.065)	.39	3.73 *
19 SCR	-.50 (**)	-.90 (**)	.41	***
24 MC	1.25 (.078)	.83 (.077)	.41	3.75 *
7 SCR	.89 (.075)	.46 (.054)	.43	4.69 *
21 MC	.18 (.107)	-.30 (.109)	.48	3.12 *
26 ECR	1.14 (**)	.52 (**)	.62	***
13 SCR	-.36 (**)	-1.16 (**)	.81	***
Items needing considerable change to render on computer				
10 ECR	1.31 (**)	1.73 (**)	-.42	***
17 SCR	-.60 (.040)	-.90 (.044)	.30	5.00 *
15 SCR	-.53 (.040)	-.91 (.052)	.38	5.84 *
16 SCR	-.54 (.040)	-.96 (.051)	.41	6.39 *

* $p < .05$.

** Standard errors from Parscale for polytomous constructed-response item parameters could not be estimated reliably.

*** z-value could not be calculated because a reliable standard error could not be estimated.

NOTE: MC= multiple choice. SCR=short constructed-response.

ECR=extended constructed-response. Standard errors of the estimated *b* parameters appear in parentheses. For polytomous items, the estimated *b* is the item location following the parameterization of Muraki (1990).

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

As the bottom section of table 3-1 indicates, three of the four items requiring considerable change for computer rendering were significantly more difficult than their paper counterparts. (The difference for the fourth item, which was less difficult on computer, could not be tested.) Taken across all four items, the mean differences for the changed vs. unchanged items were .17 vs. .23, respectively, and the mean absolute differences were .38 vs. .26.

The three considerably changed items that were significantly harder on computer were implemented quite differently compared with their paper renderings. For each of these items (numbers 15–17), the general task was to determine the value of a point on a number line. On the paper test, the examinee needed to write a value on the number line in the space provided. On the computer test, the student first had to choose the appropriate answer template (a whole number, decimal, fraction, or mixed number), and then type the answer into that template.¹⁶

As the table suggests, change in presentation was related to response format: the questions needing considerable change were all constructed response. Classifying the data by item format also suggests an impact on difficulty. On average, the discrepancies were about twice as large for constructed-response questions as for multiple-choice items: the mean difference for constructed-response was .31 vs. .16 for multiple-choice, and the mean absolute differences were .39 and .20, respectively.

Finally, items were classified by whether or not a calculator was present. (Recall that a scientific calculator was made available for section three of P&P, and an online scientific calculator was available for that same section in MOL.) Since the calculator was only present for items in the final section of the test, it should be noted that this comparison confounds position with difficulty. The mean difference between paper and computer presentation for the seven calculator-present items was .33 and the mean absolute difference was also .33. For the 18 items where the calculator was not available, the comparable figures were .18 and .26, suggesting the possibility that the presence of a calculator might increase mode differences somewhat.

¹⁶ Templates were used to avoid the ambiguity that can result from typing fractions and mixed numbers in an unstructured horizontal text box. For example, $22/3$ could be intended as either $2\frac{2}{3}$ or as $\frac{22}{3}$.

Table 3-2 presents the difficulty results in the $p+$ (proportion-correct) metric. In this metric, values range from 0 to 1.00. For example, a value of zero indicates that all students answered the item incorrectly, while a value of 1.00 indicates that all students answered the item correctly. (For the $p+$ results, no significance test was conducted, since significance had already been tested using the more theoretically sound IRT metric. Where only the median difference values are given, the median absolute difference was identical except for sign.) Over all items, the median of the difficulty differences was $-.05$ (range = $-.17$ to $.02$). The median difference for the items needing considerable change was $-.08$ and the median difference for the items needing minimal change was $-.04$. With regard to item format, the median difference for the short- and extended-constructed-response items was $-.08$, whereas the comparable value for multiple-choice items was $-.03$. Finally, for calculator items, the median difference was $-.05$ and the median absolute difference $.05$, whereas for the other items the comparable figures were $-.03$ and $.04$, respectively. Thus, in general, the $p+$ results are consistent with the differences in the b parameter estimates described above.

In addition to the paper form that contained items identical to those used on computer, two other paper forms were administered. Eleven of the items analyzed in these two paper forms also appeared on the base form (P&P). The remaining 14 items were generated to be mathematically identical to but superficially different from their base-form counterparts (e.g., the story problem context might vary although the operations performed to solve the problem were the same).

Table 3-2. Proportion-correct ($p+$) values for items presented on computer and on paper, by extent of change required for computer and size of mode difference, grade 8: 2001

Item and format	$p+$		Difference (computer minus paper)
	Computer	Paper	
Items needing minimal change to render on computer			
13 SCR	.58 (.013)	.76 (.016)	-.17
26 ECR	.18 (.013)	.34 (.016)	-.16
4 MC	.70 (.018)	.79 (.018)	-.09
21 MC	.56 (.016)	.63 (.020)	-.08
19 SCR	.62 (.014)	.70 (.016)	-.08
7 SCR	.31 (.019)	.38 (.017)	-.08
24 MC	.31 (.015)	.37 (.016)	-.06
14 MC	.53 (.020)	.58 (.020)	-.06
22 MC	.45 (.012)	.50 (.013)	-.05
23 MC	.34 (.017)	.39 (.020)	-.05
6 MC	.34 (.016)	.37 (.013)	-.04
20 MC	.80 (.016)	.83 (.017)	-.03
8 MC	.30 (.018)	.33 (.019)	-.03
5 MC	.66 (.016)	.69 (.014)	-.03
12 MC	.76 (.014)	.78 (.015)	-.02
18 MC	.83 (.011)	.85 (.013)	-.02
2 SCR	.63 (.016)	.65 (.015)	-.02
1 MC	.49 (.026)	.51 (.020)	-.02
25 MC	.44 (.024)	.44 (.025)	#
3 SCR	.69 (.014)	.69 (.014)	#
11 MC	.85 (.014)	.83 (.013)	.02
Items needing considerable change to render on computer			
16 SCR	.68 (.017)	.77 (.016)	-.09
15 SCR	.67 (.013)	.76 (.017)	-.09
17 SCR	.69 (.016)	.77 (.015)	-.08
10 ECR	.20 (.008)	.18 (.009)	.02

The estimate rounds to zero.

NOTE: ECR=extended constructed-response. MC=multiple choice.

SCR=short constructed-response. Standard errors of the estimated $p+$ values appear in parentheses. For polytomous items, $p+$ was computed as a category-weighted mean; for example, if there were three response categories, the sum of the responses in the first category was multiplied by 0, the sum in the second by 0.5, and the sum in the last by 1.0.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

Table 3-3 gives the correlations between the IRT b parameter estimates, the mean difference between parameter estimates, and the mean absolute differences. These statistics are given between the computer-based test and the three paper forms, as well as among the three paper forms.¹⁷ As the table shows, comparing the computer-based test to the two other paper forms produces essentially the same result as comparing it to the P&P base form. That is, consistent difficulty differences are apparent. (The mean differences range from .22 to .27 logits and the mean absolute differences from .28 to .29 logits.) Moreover, comparing the paper forms among themselves produces lower mean difficulty differences (mean differences from -.04 to .02 logits and mean absolute differences from .11 to .14 logits). Finally, as indicated by their correlations, the relationship between the parameter estimates is essentially the same within and across test mode.

Figure 3-3 shows the scatter plot of the IRT b parameter estimates for the computer test in comparison to the parameter estimates on each of the three paper forms. Thus, each b parameter estimate for the computer test is compared to three IRT b parameter estimates, each generated from a closely parallel paper form administered to a comparable sample. All three paired comparisons are presented on the same plot without individually identifying the forms to emphasize the overall contrast between computer and paper performance as opposed to any variation among the forms. This combined plot shows the same trend toward greater difficulty on the computer-presented test vs. the paper forms that is found in contrasting the computer test to P&P alone.¹⁸

Table 3-3. Comparison of IRT b parameter estimates for the MOL test to parameter estimates from three paper forms, grade 8: 2001

Comparison	Mean difference between parameter estimates	Mean absolute difference between parameter estimates	Correlation between parameter estimates
MOL vs. P&P	.22	.28	.96
MOL vs. Form A	.25	.28	.96
MOL vs. Form B	.27	.29	.96
P&P vs. Form A	.02	.14	.98
Form A vs. Form B	.02	.11	.99
Form B vs. P&P	-.04	.14	.98

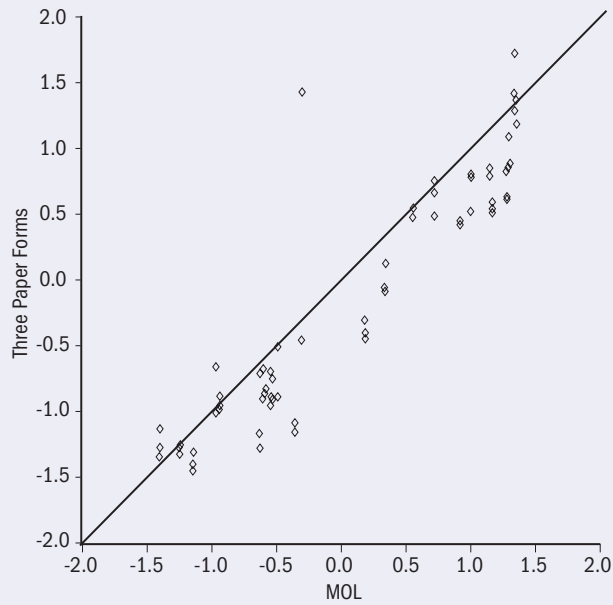
NOTE: MOL=Math Online. P&P=Paper and Pencil. Forms A and B were paper forms constructed to be parallel to P&P.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

Figure 3-4 shows the IRT b parameter estimates for all pairs of the three paper forms. Thus, this plot compares each P&P item parameter estimate to its counterpart on Form A, each Form A item parameter estimate to its Form B counterpart, and each Form B item parameter estimate to P&P. The 75 possible points of comparison in figure 3-4 are more clearly clustered around the identity line than the points in figure 3-3, further evidence that the difference in difficulty apparent in figure 3-2 is indeed an effect of the mode of presentation and not just variation due to the examinee sample.

¹⁷ The IRT b parameter estimate for one multiple-choice item included on paper Form A diverged dramatically from the b parameter estimate for the original version of the item included on MOL and P&P, as well as from the b parameter estimate for the variant of the item included on Form B. Examination of the items and data showed that the first response choice for the Form A item was a plausible but incorrect answer that attracted many examinees. For the version of the item found on MOL and P&P, and for the variant on Form B, however, the correct answer appeared *before* any other plausible answer option, making these two versions considerably easier than the Form A variant. As a consequence, this variant was removed from all Form A comparisons shown in table 3-3, along with its counterpart item in each comparison. Comparisons of MOL with P&P, MOL with Form B, and Form B with P&P were not affected.

¹⁸ This plot shows, as an outlying data point, the divergence of IRT b parameter estimates for the item described in the preceding footnote. The outlying data point represents the difference between the b parameter estimate for the item presented on Form A and the variant of that item included on MOL. Similar divergences with the estimates for the other two variants of the item presented on P&P and on Form B can be seen in figure 3-4 and in figure 5-1 as a pair of outlying data points (one point for the comparison with each of the other paper forms).

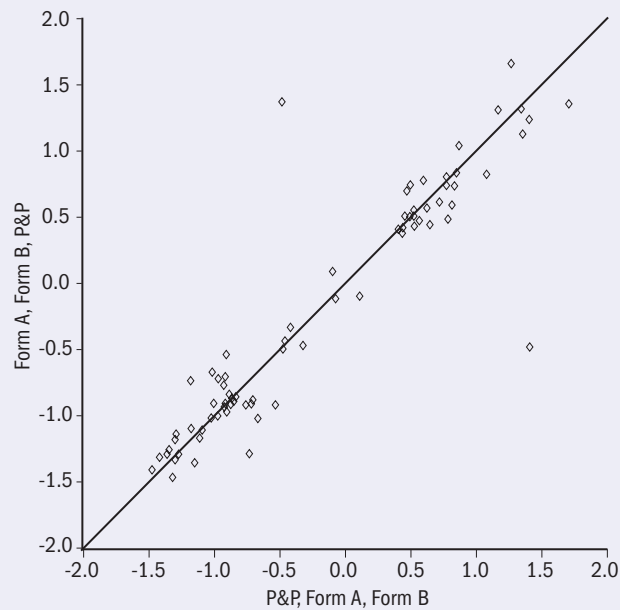
Figure 3-3. Comparison of IRT b parameter estimates for the MOL test vs. three paper forms, grade 8: 2001



NOTE: MOL=Math Online.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

Figure 3-4. Comparison of IRT b parameter estimates for three paper forms, grade 8: 2001



NOTE: P&P=Paper and Pencil. Forms A and B were paper forms constructed to be parallel to P&P.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

Analysis of Item Discrimination for Eighth Grade

The IRT a parameter describes the discrimination of an item, and is commonly considered to be the analog of the classical item-total correlation. Strictly speaking, it is an estimate of the slope of the item characteristic curve at the inflection point (the b value). Items with lower a values do not differentiate between examinees at particular points on the ability scale as well as items with higher values.

Table 3-4 gives the discrimination estimates for each item in computer- and paper-based administration, the difference between the estimates, and the results of the significance tests. As in the comparison of IRT b parameter estimates, only the differences for the 20 dichotomously scored questions were tested for significance. As the table indicates, 16 of the 20 items showed no difference in discrimination between modes. Of the four items with differences, all had lower discrimination in the computer-based test. Across all 25 items, the mean of the discrimination differences was $-.04$ and the mean of the absolute differences was $.13$, suggesting minimal effects. Also, the parameter estimates were highly related across modes ($r = .86$), though not as highly as the difficulty estimates.

Items needing considerable change for computer presentation did not differ much from items needing minimal change in their power to discriminate as measured by IRT a parameter estimates. The mean difference for the changed items was $.11$ and for the unchanged items $-.07$. The mean absolute differences were $.16$ versus $.13$.

Table 3-4. IRT a parameter estimates for items presented on computer and on paper, by extent of change required for computer and size of mode difference, grade 8: 2001

Item and format	Estimated a		Difference (computer minus paper)	z value
	Computer	Paper		
Items needing minimal change to render on computer				
25 MC	.86(.155)	1.35 (.192)	-.50	-2.01 *
22 MC	.83(.112)	1.13 (.141)	-.30	-1.68
18 MC	.87(.085)	1.17 (.114)	-.30	-2.08 *
4 MC	.74(.071)	.98 (.103)	-.25	-1.97 *
7 SCR	.70(.056)	.88 (.063)	-.18	-2.12 *
1 MC	1.01 (.130)	1.16 (.152)	-.16	-0.79
6 MC	1.22 (.192)	1.31 (.162)	-.10	-0.38
12 MC	.76(.069)	.84 (.080)	-.09	-0.81
11 MC	.92(.090)	.99 (.098)	-.06	-0.48
5 MC	.58(.073)	.63 (.078)	-.05	-0.46
2 SCR	.62(.048)	.66 (.052)	-.04	-0.57
13 SCR	.39(**)	.43 (**)	-.04	***
20 MC	.79(.076)	.81 (.082)	-.02	-0.22
19 SCR	.47(**)	.49 (**)	-.02	***
14 MC	1.37 (.147)	1.39 (.149)	-.01	-0.07
3 SCR	.42(**)	.42 (**)	#	***
26 ECR	.78(**)	.77 (**)	.01	***
21 MC	.88(.103)	.80 (.078)	.09	0.67
8 MC	1.05 (.169)	.91 (.171)	.14	0.56
24 MC	1.19 (.179)	1.03 (.130)	.17	0.76
23 MC	1.13 (.139)	.93 (.120)	.20	1.09
Items needing considerable change to render on computer				
17 SCR	1.49 (.095)	1.60 (.111)	-.11	-0.77
16 SCR	1.44 (.091)	1.32 (.092)	.12	0.89
15 SCR	1.45 (.091)	1.27 (.088)	.17	1.38
10 ECR	.61 (**)	.36 (**)	.25	***

The estimate rounds to zero.

* $p < .05$.

** Standard errors from Parscale for polytomous constructed-response item parameters could not be reliably estimated.

*** z-value could not be calculated because a reliable standard error could not be estimated.

NOTE: MC=multiple choice. SCR=short constructed-response.

ECR=extended constructed-response. Standard errors of the estimated a parameters appear in parentheses.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

Table 3-5 compares the IRT a parameter estimates across modes and within the paper forms.¹⁹ As the table shows, there is relatively little variation in the mean discrimination-parameter-estimate differences comparing each of the paper tests to the computer-based test (-.17 to -.04 for the mean differences and .13 to .23 for the absolute differences). Additionally, the differences between the computer presentation and each of the paper forms are very similar in magnitude to the differences between pairs of paper forms (whose mean differences range from -.13 to .07 and mean absolute differences from .13 to .22). The correlation between the parameter estimates does vary considerably, though it is not clear that this variation is much greater across modes than within modes.

Table 3-5. Comparison of IRT a parameter estimates for the MOL test to parameter estimates from three paper forms, grade 8: 2001

Comparison	Mean difference between parameter estimates	Mean absolute difference between parameter estimates	Correlation between parameter estimates
MOL vs. P&P	-.04	.13	.86
MOL vs. Form A	-.17	.23	.49
MOL vs. Form B	-.11	.19	.82
P&P vs. Form A	-.13	.21	.71
Form A vs. Form B	.07	.22	.68
Form B vs. P&P	.06	.13	.91

NOTE: MOL=Math Online. P&P=Paper and Pencil. Forms A and B were paper forms constructed to be parallel to P&P.
 SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

¹⁹ All items were included in this analysis. The Form A variant excluded from the table 3-3 difficulty analyses was included here because it functioned similarly to its counterparts in terms of item discrimination.

4. Equity Issues

This section considers two basic questions:

- How do population groups perform and do mode effects vary across groups?
- How are students with different levels of computer experience affected by technology vs. paper-based mathematics assessment? In particular, does a lack of computer familiarity appear to have a negative impact on online test performance?

Population Group Performance

Perhaps the most comprehensive study of the comparability of delivery modes for population groups is that of Gallagher, Bridgeman, and Cahalan (2000), who addressed the issue with large samples of examinees taking a variety of admissions and licensure tests. The tests were the Graduate Record Examinations (GRE[®]) General Test, Graduate Management Admission Test (GMAT[®]), SAT I: Reasoning Test, Praxis: Professional Assessment for Beginning Teachers, and Test of English as a Foreign Language (TOEFL[®]). These investigators discovered that delivery mode consistently changed the size of the differences between focal- and reference-group performance for some groups on both verbal and mathematical tests, but only by small amounts. Of particular interest to the current study is that for Black students and Hispanic students the difference in mathematical performance relative to White students was smaller on computer-based tests than on paper tests. From one mode to the other, the difference in performance between groups changed by up to .24 standard deviation units, depending upon the test. Also, the difference on mathematical tests between White female students and White male students was smaller on the paper versions than on the online editions. This difference changed as a function of delivery mode by up to .12 standard deviations, again depending upon the particular test.

At the school level, only one study with reasonably large samples was identified. Coon, McLeod, and Thissen (2002) evaluated third-graders in reading and fifth-graders in mathematics, using two forms of each test and delivering each form on computer and on paper to a different student group. Their analysis included an examination of the interaction of delivery mode with gender and with ethnicity. The researchers found a significant delivery-mode by ethnic-group interaction for one (but not both) of the mathematics forms, indicating the possibility that mode differences varied among population groups.

To investigate whether traditional NAEP population groups were differentially affected by computer

presentation, eighth-graders' performance on the computer-presented test was compared directly with performance on the paper form (P&P). Comparisons were made by gender, race/ethnicity, parents' education level, region of the country, school location, and school type (see appendix D).²⁰

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 true reflection of the population performance or, alternatively, an artifact of sample selection. For the groups examined, only one statistically significant difference was detected: Students reporting that at least one of their parents graduated from college performed better on P&P than a comparable group taking the same test on computer ($t = -2.73, p < .05$). For this group, the difference in mean scores was 6 points, or an effect of .21 standard deviation units, which would be characterized as "small" in Cohen's (1988) classification.

Performance as a Function of Computer Experience

While the demographic groups examined do not, in general, seem to be differentially affected by computer delivery, students who differ in their familiarity with computers might be affected. Very few recent studies of the role of computer familiarity in online test performance exist, especially at the school level. The recency of the study is important because the student population at all levels is rapidly developing basic computer proficiency. One of the more recent large-scale studies, conducted with TOEFL[®] examinees, found no meaningful relationship between computer familiarity and online performance on a multiple-choice test after controlling for language skill and after examinees had completed the online tutorial (Taylor, Jamieson, Eignor, and Kirsch 1998). However, several smaller-scale studies conducted with younger students have found that computer experience may interact with delivery mode on constructed-response writing tests (e.g., Russell and Haney 1997; Russell 1999; Russell and Plati 2001; Wolfe, Bolton, Feltovich, and Niday 1996). In addition, one study found that, compared to a paper test, taking a constructed-response mathematics test on computer had a negative effect, which moderated as keyboarding skill increased (Russell 1999).

If computer familiarity affects online test performance, a central question relates to how familiar fourth- and eighth-grade students actually are with

²⁰ Comparisons were made within each demographic variable using *t*-tests between MOL and P&P, correcting for chance via the false discovery rate (FDR) procedure.

computers. The current study addressed this question by looking at students' responses to background questions selected from those used in the NAEP 2001 history and geography assessments. Responses to these questions suggested that most fourth-grade students had access to computers at school and home, and used computers frequently (see appendix E). For example, the large majority of students indicated that they use a computer at home (85 percent) and that they use it to access the Internet (69 percent). In addition, the majority said that they used a computer at school (74 percent) or outside school (66 percent) at least once a week. (Only four percent said they never or hardly ever used a computer at either of these locations.) At least half of the students reported using a computer to play games, write, make pictures or drawings, look up information on a CD, and look up information on the Internet. The large majority reported using a computer at school for mathematics at least once a week (74 percent). Students split evenly in their attitudes about doing homework on the computer and about productivity, but most students reported that learning is more fun on the computer (77 percent vs. 21 percent).

The results for eighth-graders give a similar picture. The overwhelming majority indicated they use a computer at home (88 percent) and that they use it to access the Internet (79 percent). In addition, the majority said that at least once a week, they used a computer at school (55 percent) and used a computer elsewhere (83 percent). (Two percent said they never or hardly ever used a computer at either of those locations.) More than half of the group reported employing a computer to find information on the Internet for school (94 percent) or personal use (88 percent), to play games (90 percent), to write (87 percent), to look up information on a CD (81 percent), to communicate via e-mail (81 percent), to chat (76 percent), to make drawings (72 percent), or to make tables, charts, or graphs (59 percent).²¹ Finally, more than half agreed or strongly agreed with statements that using computers was more motivating for starting schoolwork, was more fun for learning, and helped get more schoolwork done.

To determine whether familiarity with computers affects online test performance, the relationship between computer familiarity and performance in the MOL test was examined. These analyses were conducted only for the overall populations of fourth-

and eighth-grade students, as questions of the impact of computer familiarity on test performance for population groups were beyond the scope of the study.

Computer familiarity can be measured in many ways. For purposes of this study, familiarity was conceived as having three components: computer experience, input accuracy, and input speed. Theoretically, these components should overlap but still be separable. A student may have had several years of experience with a computer but be neither fast nor accurate in typing. Similarly, a student may be a rapid but sloppy typist. In any event, a minimal level on each component should, in theory, be present before a student can effectively take an online test, especially one that includes constructed-response questions. For example, some amount of previous computer experience might allow quicker adaptation to the test's navigational and input procedures, which in the MOL test were designed to follow common software conventions. Likewise, input accuracy should be necessary for the student's intended answer to be recorded correctly. Finally, reasonable speed is required because the MOL test gives students a limited time for completion; time lost to input that is accurate but slow might introduce irrelevant variance into test performance. In fact, such an effect for speed in online mathematics test performance has been found in at least one previous comparability study (Russell 1999).

To measure the first component of familiarity, computer experience, a scale was created based on students' responses to computer-related background questions.²² The rationale for using background questions as a measure of experience was two-fold. First, these questions are the type that NAEP has used to document the extent and type of computer use among students. Second, very similar background questions have been used in other comparability studies as surrogates for computer proficiency (e.g., Taylor, Jamieson, Eignor, and Kirsch 1998).

Questions were selected for inclusion in the scale based on expert judgment. The score was the simple sum of the responses to each question, ranging from 0–20 for the fourth-grade instrument and 0–40 for the eighth-grade measure. While other question-aggregation rules are possible, this scheme was judged reasonable given research suggesting that different aggregation rules often produce similar results (Stanley and Wang 1970).

²¹ These figures were computed from table E-8 by summing the percentages of students who reported use to a large, moderate, and small extent.

²² Appendix F presents MOL vs. P&P performance for students by response to most of these questions.

For fourth grade, the questions and the number of response categories for each were:

- How often do you use a computer at school? (5)
- How often do you use a computer outside of school? (5)
- Is there a computer at home that you use? (2)
- Do you use the Internet at home? (2)
- Do you ever use a computer to do any of the following?
 - Play computer games (2)
 - Write reports, letters, stories, or anything else on the computer (2)
 - Make pictures or drawings on the computer (2)
 - Make tables, charts, or graphs on the computer (2)
 - Look up information on a CD (2)
 - Look up information on the Internet (2)
 - Send e-mail or talk in chat groups (2)
- When you do mathematics in school, how often do you do each of the following?
 - Use a computer (4)

For eighth grade, the composite consisted of questions covering essentially the same content and included the following:

- How often do you use a computer at school? (5)
- How often do you use a computer outside of school? (5)
- Is there a computer at home that you use? (2)
- Do you use the Internet at home? (2)
- To what extent do you do the following on a computer?
 - Play computer games (4)
 - Write using a word processing program (4)
 - Make drawings or art projects on the computer (4)
 - Make tables, charts, or graphs on the computer (4)
 - Look up information on a CD (4)
 - Find information on the Internet for a school project or report (4)
 - Find information on the Internet for personal use (4)
 - Use e-mail to communicate with others (4)
 - Talk in chat groups or with other people who are logged on at the same time you are (4)
- When you do mathematics in school, how often do you do each of the following?
 - Use a computer (4)

The second and third components of computer familiarity, input accuracy and input speed, were measured using tasks embedded in the MOL tutorials (available at <http://nces.ed.gov/nationsreportcard/studies/tbatutorial.asp#mol>).

The evidentiary basis for these tasks was content validity. Coming from the MOL tutorial, the tasks were essentially the same mechanical ones that students needed to perform in taking the MOL test.

Table 4-1 shows the tasks included in the accuracy and speed measures. For fourth grade, the accuracy scale range was 0–15 and the speed scale range was 0–16. For eighth grade, the comparable ranges were 0–17 and 0–22, respectively.

Variable	Number of score levels	
	Grade 4	Grade 8
Accuracy		
Typing and editing		
Accuracy typing a brief given passage	3	3
Accuracy inserting a word	3	3
Accuracy changing a word	3	3
Navigating the test		
Accuracy pointing and clicking with mouse	3	3
Accuracy scrolling	3	3
Accuracy clicking on “Next” icon	3	3
Accuracy clicking on “Previous” icon	3	3
Entering responses		
Accuracy filling in a mixed number	–	3
Using the calculator		
Accuracy in performing a given operation	2	2
Speed		
Typing and editing		
Time to type brief passage	3	3
Time to insert word	3	3
Time to change word	3	4
Navigating the test		
Time to point and click	3	4
Time to scroll	3	3
Time to click on “Next”	3	4
Time to click on “Previous”	3	3
Entering responses		
Time to fill in mixed number	–	3
Using the calculator		
Total time to complete the calculator tutorial	3	4
– Not applicable. Eighth grade only.		
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.		

Table 4-2 shows the internal consistency reliabilities for the computer familiarity measures.

	Computer experience	Input accuracy	Input speed
Grade 4	.62	.55	.58
Grade 8	.78	.48	.72
NOTE: All values are unweighted.			
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.			

Table 4-3 gives the sample correlations among the measures and with mathematics performance.

As the tables show, the three computer-familiarity measures have limited, but acceptable, reliabilities for research purposes, and their correlations with one another are generally quite a bit lower than the limit imposed by those values. Finally, for the hands-on measures, the correlations in these samples with MOL test performance are larger than their relationships with one another. Thus, empirically, the measures generally appear to be functioning as intended.

	Initial paper mathematics block	MOL test	Computer experience	Input accuracy
Grade 4				
MOL test	.57			
Computer experience	.13	.19		
Input accuracy	.31	.46	.12	
Input speed	.25	.32	.19	.13
Grade 8				
MOL test	.72			
Computer experience	.13	.21		
Input accuracy	.35	.39	.12	
Input speed	.44	.54	.31	.26
NOTE: All values are unweighted. The initial paper mathematics block contained 10 items for fourth grade and 20 items for eighth grade.				
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.				

To explore the relationship between computer familiarity and performance in the computer-based test, an ordinary least-squares multiple regression was executed. The goal of this analysis was to determine if, in the overall student population, computer familiarity predicted performance on the computer-based test after controlling for mathematics skill measured on paper. The independent variables were self-reported computer experience, input accuracy, input speed, and number-right raw score on the initial paper mathematics block, which served as a covariate. The dependent variable was the sum of the dichotomously scored and polytomously scored MOL test items. The three computer-experience variables were used because they are logically and empirically related to taking a mathematics test on computer, and not highly correlated with one another. Population-group variables were not included because the relevant difference among these

groups is in mathematics skill, which was controlled in the regression by including the initial paper block. Finally, because it is restricted to the group that took the computer test, this analysis avoids any confounding due to uncontrolled differences between the paper and computer groups (e.g., in the scoring of constructed responses).

Table 4-4 presents the results of the regression for fourth grade. Only the main effects model is presented because adding the two- and three-way interactions among the computer familiarity indicators did not add significantly to the prediction of MOL performance ($F_{4,914} = 0.64, p > .05$). After controlling for mathematics proficiency on the paper-based block, each of the three components—self-reported computer experience, input accuracy, and input speed—significantly added to the prediction of mathematics score on the computer-based test. Some sense of the magnitude of the effect can be gleaned from examining the incremental variance accounted for by different variables in the model. The initial paper block accounted for 33 percent of the variance in MOL scores. Adding the computer familiarity variables to the model increased the variance accounted for in MOL scores to 45 percent.

Table 4-4. Regression results for the effect of input skill and computer experience on computer mathematics test raw score, controlling for paper mathematics proficiency, grade 4: 2001

Variable	Estimated regression coefficient	Standard error
Intercept	-14.75	1.926
Initial paper block (covariate)	1.79 *	0.131
Input accuracy	1.23 *	0.096
Input speed	.37 *	0.073
Computer experience	.12 *	0.039

* $p < .05$, two-tailed t -test (df -range 26 to 35, t -range 3.12 to 13.63).

NOTE: The number of students included in the analysis was 1,034. A jackknife replicate weight standard error procedure was used to compute the standard errors (see: Allen, Donoghue, and Schoeps 2001).

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

Table 4-5 gives the regression results for the eighth grade. Again, only the main effects model is given because the interactions were not significant ($F_{4,539} = 0.73, p > .05$). After controlling for mathematics proficiency on the paper-based block, input accuracy and input speed significantly added to the prediction of MOL score; self-reported computer experience did not add significantly. In terms of the size of the effect, the initial paper block accounted for 49 percent of the variance in MOL scores. Adding the computer familiarity variables to the model increased the variance accounted for in MOL scores to 57 percent.

Thus, the regression results for both grades suggest that computer familiarity plays a role in online mathematics test performance. That role is such that the more familiar a student is with computers—and particularly the more efficiently he or she can manipulate the keyboard and mouse—the better that student will score. This influence would seem to be an unwanted one; it affects online performance independently of mathematics skill and suggests that some students may score better on mathematics tests like MOL simply because they are more facile with computers.

Table 4-5. Regression results for the effect of input skill and computer experience on computer mathematics test raw score, controlling for paper mathematics proficiency, grade 8: 2001

Variable	Estimated regression coefficient	Standard error
Intercept	-15.78	2.327
Initial paper block (covariate)	.87 *	0.136
Input accuracy	.67 *	0.131
Input speed	.37 *	0.067
Computer experience	.05	0.025

* $p < .05$, two-tailed t -test (df -range 3 to 12, t -range 1.86 to 6.36).

NOTE: The number of students included in the analysis was 1,011. A jackknife replicate weight standard error procedure was used to compute the standard errors (see: Allen, Donoghue, and Schoeps 2001).

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

5. Efficiency Issues

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

- How might two particular technological advances, “automatic item generation” and “automated scoring,” affect the cost and timeliness of assessment?
- Is a technology-based mathematics assessment in general more cost-effective or timely than a paper one?

First, the feasibility of automated item generation is discussed and then automated scoring. Finally, the probable cost-effectiveness of technology versus traditional paper-based methods in the context of the National Assessment of Educational Progress (NAEP) is explored.

Automatic Item Generation

Automatic item generation rests on two premises. The first premise is that a class of test items can be described in enough detail for a computer to generate instances of that class. The second is that enough can be known about the determinants of item difficulty so that each of the generated instances does not have to be individually calibrated.

The description the computer uses to generate instances of a class is called an item “model” and the instances are called “variants.” Computer-generated variants can be inexpensively created in large numbers. To the degree that large numbers could be employed effectively, computer generation of items would increase efficiency considerably.

A testing program like NAEP could, in principle, use computer-generated variants to increase depth of content coverage. In NAEP mathematics assessments, coverage of some subtopics specified by the framework is based on only a few items. For example, the subtopic, “Apply basic properties of operations” might be covered at grade 4 by a few items testing the four basic operations. The inference that policymakers and other NAEP users wish to derive, however, is not whether the nation’s fourth-grade students can perform those operations for this sparse

sample of instances but, rather, whether they can use those operations throughout the class of items those few instances represent. Expanding the number of items used to assess each subtopic can arguably support stronger inferences about what students know and can do at a finer level than current NAEP assessments.

Is it possible to generate test items automatically? It has been repeatedly demonstrated that a class of items can be described in sufficient detail for a computer to generate variants. Irvine and Kyllonen (2002) give several illustrations. In addition, for several years ETS has used a software tool, the Mathematics Test Creation Assistant (Singley and Bennett 2002), for limited item generation in selected testing programs.

Beyond feasibility, is automatic item generation efficient? If an item model can be calibrated and that calibration somehow imputed to the variants it produces, it will not be necessary to calibrate each variant individually. This calibration can be accomplished by basing the model on an empirically calibrated item and then constraining the model so that it, ideally, produces variants that diverge little in substance and psychometric properties from the original “parent” question. Variants that preserve the underlying problem structure are termed “isomorphs.” Because the variants created by a model are not only isomorphs of one another, but also isomorphs of the parent item, the model’s parameters may, in theory, be imputed from those of the parent.

A second calibration method is to pretest a sample of variants from the item model and use that information to establish model parameters. The psychometric methods for such calibration are beyond the scope of this report, but see Glas and van der Linden (2001), or Johnson and Sinharay (2002), for applications of hierarchical methods, and Bejar, Lawless, Morley, Wagner, Bennett, and Revuelta (2002), for use of the expected response function.

Empirical Analysis

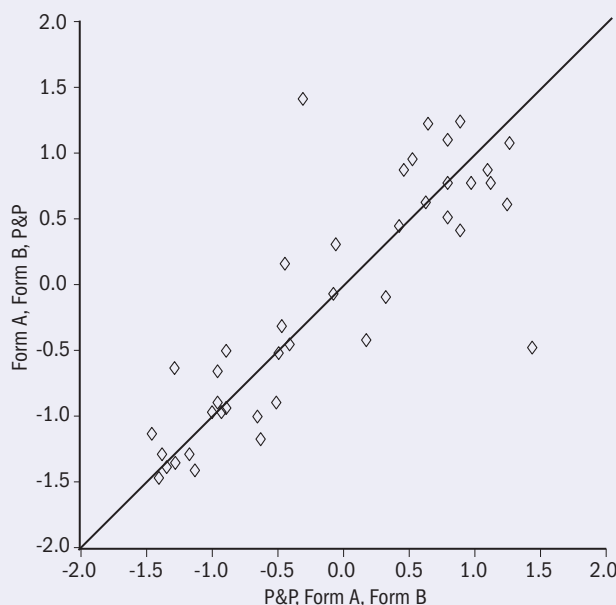
In this study, models were created for eighth-grade items using the Mathematics Test Creation Assistant (Singley and Bennett 2002). Each model resembles a test item in which elements of the stimulus, stem, and response options are treated as variables. Both linguistic and mathematical elements can be manipulated in this way. Also included in the model are constraints that govern how the values of a linguistic or numeric element may vary.

Models were created for 15 of the 26 items in the eighth grade P&P form, including both multiple-choice and constructed-response questions. Isomorphs were then generated and reviewed by staff members trained to recognize and remove instances that might inappropriately disadvantage one or another demographic group. Next, for each item, one isomorph was selected at random to be included in Form A and one to be included in Form B. Each isomorph occupied the same position as its counterpart across the three paper forms. Within forms, automatically generated items appeared in each of the three sections.

All three paper forms were administered to randomly parallel student samples at the eighth grade: 954 students for P&P, 926 students for Form A, and 906 students for Form B. The three test forms were scaled using the 20-item common paper test as an anchor. The item parameters across each form were unconstrained. This scaling makes it possible to examine differences in item difficulty parameter estimates across forms, both for the 11 items common to the 3 forms and for the 14 sets of isomorphs. (One set was dropped from the analysis because of scaling difficulties.)

Figure 5-1 shows the IRT b values for each set of 14 isomorphs on the three paper forms. Each isomorph on the P&P base form appears twice on the plot, once in comparison to its sibling on Form A and once in comparison to its sibling on Form B. The parameter estimate comparisons between Forms A and B appear as well, making for 42 pair-wise comparisons in all. Figure 5-2 shows the comparable plot for the 11 items that were identical on all three forms. As the plots suggest, there is variation in both sets of parameter estimates.

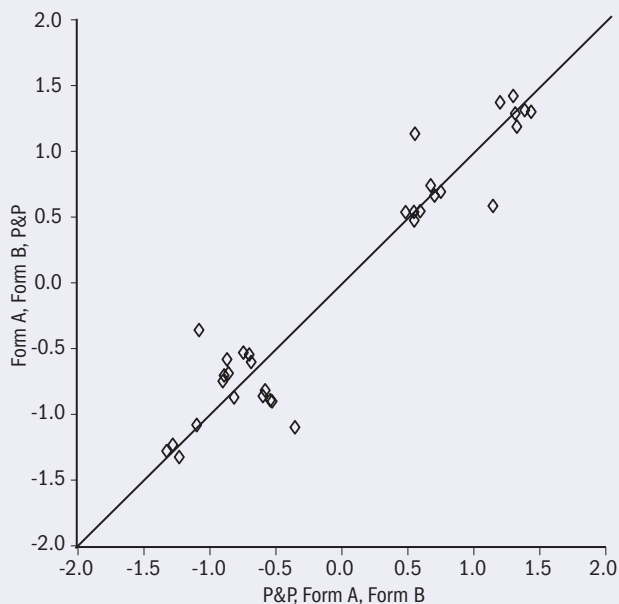
Figure 5-1. Pair-wise comparisons of IRT b parameter estimates for 14 isomorphs on three paper forms, grade 8: 2001



NOTE: P&P=Paper and Pencil. Forms A and B were paper forms constructed to be parallel to P&P.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

Figure 5-2. Pair-wise comparisons of IRT b parameter estimates for 11 identical items on three paper forms, grade 8: 2001



NOTE: P&P=Paper and Pencil. Forms A and B were paper forms constructed to be parallel to P&P.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

Table 5-1 gives the mean differences, the mean absolute differences, and the correlations between the parameter estimates. Each statistic is computed on only a small number of items, so the values should be taken as suggestive only. Consistent with the patterns shown in the plots, the parameter estimates for the isomorphs seem somewhat more variable than the ones for the identical items. This effect is clearest in the absolute differences.

Table 5-1. IRT b parameter estimates for isomorphic vs. identical items for the three paper forms, grade 8: 2001

Test form	Mean difference between parameter estimates	Mean absolute difference between parameter estimates	Correlation between parameter estimates
Isomorphic items			
P&P vs. Form A	.10	.41	.80
Form A vs. Form B	.23	.25	.85
P&P vs. Form B	.34	.35	.98
Identical items			
P&P vs. Form A	.25	.25	.97
Form A vs. Form B	-.07	.10	1.00
P&P vs. Form B	-.18	.22	.97

NOTE: P&P=Paper and Pencil. Forms A and B were paper forms constructed to be parallel to P&P. The analysis for each form included 14 isomorphic items and 11 identical items.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

Although the parameter estimates for the isomorphs seem somewhat more variable than those for the identical items, of central importance is how much that variability affects population estimates. Table 5-2 addresses this question by comparing the mean scores from two scalings. In the first scaling, the item parameters were constrained to be equal across the three paper forms, an assumption that would hold true if the variants behaved like identical items. In the second scaling, the items were free to vary, as if each form were composed of completely different items, a theoretically better-fitting model. The mean scores for a form will diverge across these two scalings to the extent that the isomorphs do not function similarly. Table 5-2 gives the means. Significant differences between the means from the two scalings were not detected for any form (t range = 0.16 to -0.39 , $p > .05$). Further, in the scaling in which the parameters were constrained to be equal across the three paper forms, no significant difference was found between the means for any pair of forms (t range = 0.20 to -1.30 , $p > .05$). Overall, this lack of variation implies that the parameter fluctuation due to the isomorphs had little impact. These results are consistent with those from simulation studies, which have shown that significant amounts of variability in item parameters can be tolerated without affecting NAEP population estimates (Dresher and Hombo 2001; Hombo and Dresher 2001).

A Priori Analysis

Although the empirical results for automatic item generation are positive, this technology certainly has limits. For example, item generation in NAEP may not be well suited to classes that

- do not have a sizable number of meaningful variants,
- employ stylized or complex graphics, or
- generate constructed-response variants requiring changes in the scoring rubric that human readers might find difficult to apply.

At the same time, many item classes typically used in NAEP are well suited for this technology. Examples include

- pure computation items;
- story problems for which the underlying mathematics can be applied to a variety of real-world situations; and
- items based on relatively simple figures, graphs, or tables whose elements can be meaningfully varied.

In order to assess the feasibility of automatic-item-generation technology for NAEP mathematics assessments, two ETS test development and two technology staff members each independently examined the items administered in the eighth-grade NAEP 2000 mathematics assessment. They examined each item to determine if a model could be created from it to generate a class containing multiple variants. Items were categorized as feasible for automatic generation or not, either because the existing generation technology was not capable of modeling the content or because the item class itself was not broad enough to support more than a few potential variants. If an item was considered feasible, it was also classified as to whether it required relatively limited effort for model creation or more substantial effort, primarily because it would entail the manipulation of such nontextual components as figures or multimedia stimuli. When there were disagreements among judges about classification, the more restrictive judgment was used.

On four of the five content areas in the mathematics framework, most of the items were judged suitable for automatic generation (see table 5-3).

Table 5-2. Mean scores from scalings in which item parameters were and were not constrained to be equal across paper forms, grade 8: 2001

Test form	Item parameters constrained to be equal across paper forms	Item parameters unconstrained across paper forms	t value
P&P	199 (1.4)	199 (1.4)	0.16
Form A	199 (1.1)	200 (1.1)	-0.39
Form B	201 (1.3)	201 (1.3)	0.16

NOTE: P&P=Paper and Pencil. Forms A and B were paper forms constructed to be parallel to P&P. Standard errors of the scale scores appear in parentheses.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

Table 5-3. Percentages of items from the NAEP 2000 mathematics assessment, by feasibility of automatic item generation, grade 8: 2001

Framework content area	Percent not feasible for automatic generation	Percent feasible for automatic generation	
		Requires relatively limited effort to model	Requires substantial effort to model
Total (160 items)	28	51	22
Number Sense, Properties, and Operations (43 items)	23	65	12
Measurement (22 items)	27	64	9
Geometry and Spatial Sense (32 items)	63	6	31
Data Analysis, Statistics, and Probability (24 items)	17	46	38
Algebra and Functions (39 items)	10	67	23

NOTE: Detail may not sum to totals because of rounding.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

Taken over all content areas, 73 percent of items appeared feasible to generate, regardless of the effort needed for model creation. The only framework content area for which the majority of items probably cannot be automatically generated was Geometry and Spatial Sense. Even for this category, however, 37 percent of items appeared suitable. If computer generation is restricted to those items needing only limited effort, then about half of NAEP items (51 percent) still appear feasible to model.

Figures 5-3 to 5-5 are released NAEP mathematics items that illustrate each of these classifications. Figure 5-3 shows a good candidate for automatic generation. This item, a grade 8 item from the Data Analysis, Statistics, and Probability content area, comes from a large class of probability problems

that, in its most general form, centers on drawing objects of different kinds from a container. In this particular item, the “objects” are boys and girls and the “container” is the mathematics class. Thus, the variable parts of the item include not only the numeric mix of the objects in the container but the type of object and type of container. A model written to generate such items would specify the acceptable values for each of these variables, making sure to hold as constant as possible the difficulty of the mathematical operations and the familiarity of the context. The multiple-choice options would be specified as algebraic constraints, such as option $A = (x - y)/(x + y)$, option $B = y/(x + y)$, and so forth, which the generation software would use to create the appropriate numeric fractions.

Figure 5-3. An item suitable for automatic generation that would require relatively limited effort for model creation, grade 8: 2001

18. There are 15 girls and 11 boys in a mathematics class. If a student is selected at random to run an errand, what is the probability that a boy will be selected?

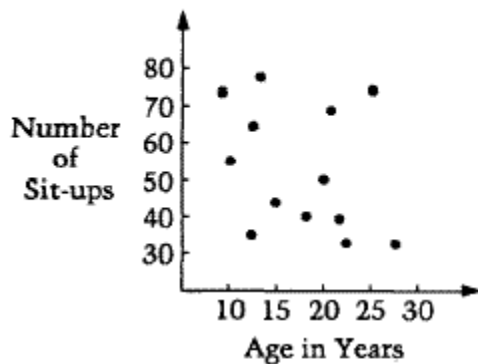
- A) $\frac{4}{26}$
- B) $\frac{11}{26}$
- C) $\frac{15}{26}$
- D) $\frac{11}{15}$
- E) $\frac{15}{11}$

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

Figure 5-4 shows an eighth-grade question that would require substantial effort to model because of the nature of its figural stimulus. The question, which is intended to assess the Data Analysis, Statistics, and Probability framework content area, is from a large class covering the interpretation of bivariate scatter plots. An item model to generate instances from this class would vary the two quantities being

plotted by changing the text of the item, the labels on the graphs, the points plotted, and the response options. Again, the test developer creating the model would need to take special care to make as invariant as possible the familiarity of the context created by the two variables chosen, the shape of the plot, and the cognitive operations posed by the question and response options.

Figure 5-4. An item suitable for automatic generation that would require substantial effort for model creation, grade 8: 2001



3. In the graph above, each dot shows the number of sit-ups and the corresponding age for one of 13 people. According to this graph, what is the median number of sit-ups for these 13 people?

- A) 15
- B) 20
- C) 45
- D) 50
- E) 55

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

Finally, figure 5-5 shows an item probably *not* well suited for automatic generation. This item assesses the Geometry and Spatial Sense framework content

area at grade 8. The number of potential variants in this problem class appears too small to make modeling worthwhile.

Figure 5-5. An item not suitable for automatic generation, grade 8: 2001

5. Which of the following figures has two circular bases?

- A) A pyramid
- B) A sphere
- C) A cube
- D) A cylinder
- E) A cone

Did you use the calculator on this question?

- Yes No

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

Automated Scoring

Another application of technology that could help NAEP increase efficiency is the automated scoring of constructed-response items. By reducing the need for human judges, automated scoring could potentially increase the speed with which NAEP analyses can be completed and also reduce the cost of scoring.

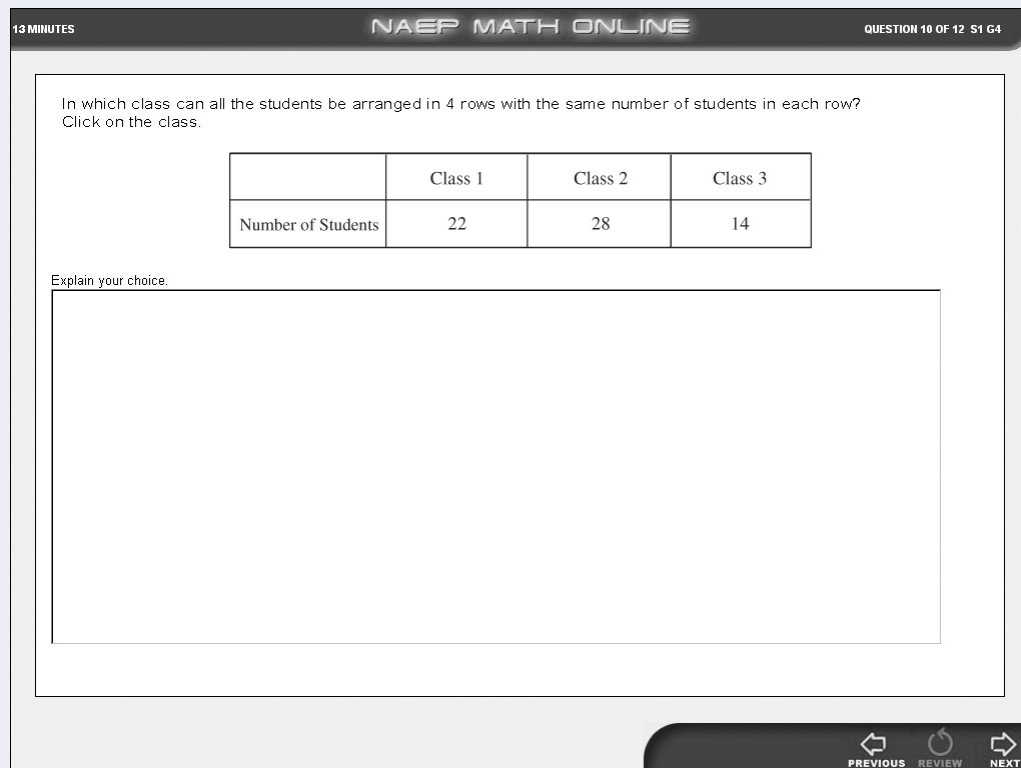
To investigate the feasibility of automated scoring for mathematics, nine of the ten constructed-response items from the fourth- and eighth-grade computer-based mathematics tests were selected. (One item from each grade was considered too complex for efficient development of scoring algorithms.)

The selected items included ones for which students were asked to give both an answer and an

explanation, and those for which they provided only an answer. Figures 5-6 and 5-7 show examples (but not the actual questions used in the test, which are still in active use).

The answers students gave to items that did not require explanations were either numeric or simple text responses (e.g., “30” or “thirty”). In contrast, the answers students gave to the items requiring explanations were usually more elaborated text, consisting of phrases or sentences. These two kinds of responses differ substantially in the scoring technology they require. Consequently, two different approaches were applied to the items, depending on the complexity of the natural language they evoked: pattern- and feature-matching for numeric and simple text responses and natural language processing for elaborated text responses.

Figure 5-6. Item for which the student must provide an answer and an explanation, grade 4: 2001



13 MINUTES

NAEP MATH ONLINE

QUESTION 10 OF 12 S1 G4

In which class can all the students be arranged in 4 rows with the same number of students in each row?
Click on the class.

	Class 1	Class 2	Class 3
Number of Students	22	28	14

Explain your choice.

PREVIOUS REVIEW NEXT

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

Figure 5-7. Item for which the student must provide only an answer, grade 4: 2001

The screenshot shows the NAEP Math Online interface. At the top, it displays "15 MINUTES", "NAEP MATH ONLINE", and "QUESTION 7 OF 10 S3 G4". On the left, there is a calculator with a display showing "0" and buttons for "+/-", \sqrt{x} , %, \div , 7, 8, 9, \times , 4, 5, 6, -, 1, 2, 3, +, CLR, 0, ., and =. On the right, the question text reads: "The band members have a goal to sell 650 candy bars. They have sold 235 candy bars so far. How many more candy bars do they have to sell to reach their goal?" Below the text is an "Answer:" label followed by a text input field. At the bottom right, there are three navigation buttons: "PREVIOUS", "REVIEW", and "NEXT".

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

Scoring by Pattern and Feature Matching

Eight of the nine grade 4 items and five of the nine grade 8 items were scored using pattern and feature-matching. For most questions in this class, a unique answer key was written. Responses were classified as “text” or “numeric.” A response was considered numeric if all characters were one of {0123456789+/-}. If one or more characters in the response was not from this set, the response was classified as text (e.g., “3 and one half” would be considered text).

The scoring of text responses consisted of comparing the response with a list, specialized to each item, of common responses and common misspellings. No natural language processing was applied to text responses that were not found in any of the lists (i.e., only an exact match of the response to the list was used).

For each item there were at least two lists:

- phrases recognized as correct (full credit)
- phrases recognized as incorrect (no credit)

For some items, there was a third list for partial-credit responses.

When the student response was not found in any list, a code of “unscorable” was assigned. In an operational assessment, a human judge would resolve such responses. Once resolved, the response would be added to the appropriate list so that if another student submitted the same answer, the automated system could grade it without assistance.

The scoring of a numeric response consisted of verifying that all of the characters were numeric and, if so, assigning a score. The logic used to assign a score was not just a simple match. A general-purpose automated scoring program for mathematics, created by ETS, was used for each item. This program determined whether the response conformed to a set of rules based on the rubric for the item. Partial credit can be assigned for breaking some rules, but not others.

As an example of such a rule set, consider an item that asks the student to find two whole numbers, each greater than a specific whole number, that have a specified whole-number product. In order to receive full credit, the response would need to satisfy each of the rules below:

- Does the response contain exactly two numbers?
- Is the first number a whole number?
- Is the second number a whole number?
- Is the first number greater than the specified number?
- Is the second number greater than the specified number?
- Is the product of the two numbers as specified?

This method could directly score all but one of the computer-based mathematics items to which it was applied. For this item, a program was written to filter the data into a format acceptable to the general-purpose engine. The item was unique in that it provided the student with the option of entering text to describe a particular geometric figure or of using the mouse to draw the figure. To process figures drawn with the mouse, the line segments generated by students were automatically analyzed to see if they approximated a straight line. Segments were then connected to form a figure. This figure was next rotated to the horizontal. Finally, the general-purpose engine processed the figure to see if it matched the required shape.

Scoring Using Natural Language Processing

The program used to score responses containing elaborated text is called *c-rater*[™] (Leacock and Chodorow 2003). *C-rater*[™] is designed to score short-answer responses by matching concepts in a student's answer to the concepts that represent a correct, partially correct, or incorrect response. In effect, it is a system that recognizes paraphrases. To recognize paraphrases, *c-rater*[™] breaks down the response's predicate-argument structure to distin-

guish syntactic variety (e.g., active versus passive sentences), and morphologically analyzes each word to recognize, for example, that different forms of the same word (e.g., add, adding, and addition) represent a single concept. The program then resolves pronoun references when words (e.g., it, he, or she) are used to refer to the previous sentence, or to the question. *C-rater*[™] also recognizes synonyms and similar words (e.g., that "minus" is similar to "subtract").

C-rater[™] matches responses against a set of model answers, which is called the "gold standard." The gold standard consists of one or more grammatical English sentences that ideally represent a comprehensive set of possible correct answers. *C-rater*[™] breaks each of these answers into an underlying representation and then matches student responses against them in turn. The scoring guide that human judges use to score an item is not by itself sufficient for deriving the gold standard because the guide does not always anticipate the range of correct or partially correct answers that students produce. Therefore, correct but unusual solutions provided by a student may not be recognized successfully until such responses are explicitly added to the gold standard.

Procedure and Data Analysis Method

The development of automated scoring keys for the computer-based mathematics test began with an analysis of scoring guides and sample responses used to train human graders for scoring paper-and-pencil questions. (Training papers for NAEP mathematics items are chosen to provide a range of correct and incorrect responses to help readers understand how to grade in a reliable manner.) Next, for each item, a sample of 500 single-scored student responses was selected to develop and test the initial algorithms. After these 500 responses were processed, the automated scores were compared with those assigned by the human raters. This comparison offered the opportunity to revise the scoring programs. Adjustments to the pattern-and-feature scoring were made, but no adjustments were made to the gold standards of *c-rater*[™].

For cross-validation, a new sample of approximately 250 responses was scored without knowledge of the scores that had previously been assigned to each response by the two human judges. A NAEP test-development staff member subsequently resolved all discrepancies between the automated and human scores.

Cross-Validation Results

Tables 5-4 and 5-5 show the results for the pattern-and-feature-matching method. The agreement percentages are accompanied by a statistic, “kappa,” which corrects for the level of agreement expected by chance (Fleiss 1981). Such levels are considerable, given the fact that most constructed-response items on the computer-based mathematics test were scored on 2- or 3-point scales.

As noted, the questions in this group generally called for numerical and single-word answers. In some cases, the algorithm was unable to process particular responses (e.g., because they could not be found on the list either of correct or of incorrect answers). As table 5-4 indicates, for grade 8, every response was scorable; for grade 4, almost every

response for six of the eight questions was scorable. For two questions (number 5 and number 14), only 80 percent and 91 percent of the responses, respectively, were scorable automatically. For the scorable responses, automated grading tended to match closely the human judgments for all items, except for item number 5. This question, described previously, allowed the student to draw a figure using the mouse. However, even for this question, the difference between human-human and automated-human agreement levels was relatively small, from 5–7 percentage points. More important, as indicated in table 5-5, when the machine score disagreed with either or both human scores, the resolution was overwhelmingly in favor of the automated score for 7 out of 8 items. The single exception was for the “drawing” item (number 5).

Table 5-4. Percentage exact agreement between human judges and between automated grader and each human judge for the pattern-and-feature-matching method, grades 4 and 8: 2001

Item	Number of responses	Percent scored by automated method	Percentage exact agreement		
			Reader 1 vs. Reader 2	Automated grader vs. Reader 1	Automated grader vs. Reader 2
Grade 4					
5	263	80	96	89	91
14	257	91	98	99	98
15	256	96	91	96	94
21	254	100	95	95	98
24	258	98	98	100	98
26	257	100	98	100	99
29	256	98	97	99	98
31	254	98	99	100	100
Grade 8					
2	249	100	98	99	100
13	251	100	98	99	99
15	247	100	98	99	98
16	245	100	98	99	99
17	247	100	98	99	99

NOTE: Kappa was .75 or higher, a strong level of agreement, for all comparisons.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

Table 5-5. Resolution of scoring disagreements between automated grader and either or both human scores for the pattern-and-feature-matching method, grades 4 and 8: 2001

Item	Number scored by automated method	Number of disagreements	Percent of disagreements resolved in favor of the automated score
Grade 4			
5	211	24	42
14	234	6	83
15	246	24	96
21	254	16	94
24	253	5	100
26	257	4	100
29	250	7	100
31	250	2	100
Grade 8			
2	249	4	100
13	251	5	100
15	247	6	100
16	245	5	100
17	247	4	100

NOTE: A disagreement was recorded when the machine score differed from one or both human scores.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

In contrast to the questions scored with the pattern-and-feature-matching method, those scored with c-rater™ called upon the examinee to enter more text. Table 5-6 provides machine-judge and inter-judge exact-agreement results for c-rater™, which assigned a score to all responses. The results indicate that for one of the five items, c-rater™ closely agreed with the score awarded by the human readers. For the other four items, agreement with c-rater™ was somewhat lower, differing by between 2 and 13 percentage points from the inter-judge levels.

Table 5-6. Percentage exact agreement between human judges and between c-rater™ and each human judge, grades 4 and 8: 2001

Item	Number of responses	Percentage exact agreement		
		Reader 1 vs. Reader 2	Automated grader vs. Reader 1	Automated grader vs. Reader 2
Grade 4				
10	253	94 *	83	81
Grade 8				
3	253	92 *	91 *	90 *
7	249	91 *	80	81
19	250	90 *	83	81
26	245	87 *	85	85

* Kappa was .75 or higher, indicating strong agreement. For all other items, kappa was between .40 and .74, indicating moderate agreement. SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

Table 5-7 shows that when c-rater™ disagreed with one or both human scores, the resolutions favored the human graders by wide margins in three cases and c-rater™ by a small margin in two other instances.

Table 5-7. Resolution of scoring disagreement between machine and either or both human scores for c-rater™, grades 4 and 8: 2001

Item	Number of responses	Number of disagreements	Percent of disagreements resolved in favor of the automated score
Grade 4			
10	253	54	26
Grade 8			
3	253	34	53
7	249	59	29
19	250	57	30
26	245	48	52

NOTE: A disagreement was recorded when the machine score differed from one or both human scores.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

Analysis of the resolved disagreements showed that the primary problem, especially with lower scores, was the inability of the program to allow for spelling mistakes. The version of c-rater™ used in this study recognized, for example, the word “subtracting” but not the misspelling “subtrackting.” Based on this finding, the following misspellings of “subtract” or “subtracted,” which appeared in student responses, were appended to the c-rater™ dictionary for future use:

subtract subctact subdtract subctacted subctacted
substract subctacd subctct subctacted sbctacted
subrtract subtrctct subtracat subctacked subctacted
subtrstct subctacted

Unfortunately, c-rater™ will still be confounded by keyboarding errors such as “add umber” and “ode nuber,” which some students used to mean “odd number.” These misspellings will confound c-rater™ because “add,” “ode,” and “umber” are all English words, and c-rater™ currently will attempt to correct only words not found in a dictionary (e.g., “nuber”).

In general, c-rater™ will not recognize creative or unusual responses if those responses do not appear in the training set used to create the gold standard. Making sure that the training sets are large and diverse in the responses they contain should help minimize this unwanted result.

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

The data presented above suggest that automated scoring and automatic item generation hold promise for NAEP. Both technologies, of course, presume computer delivery. But how might a computer-delivered NAEP assessment, in and of itself, compare with a paper one in terms of timeliness and cost?

Relative Timeliness of Computer vs. Paper Testing

Figure 5-8 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 MOL 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 (165 days) and for computer delivery (160 days). For the operational stage, however, the estimates are about 15 percent shorter for computer delivery (106 days) than for paper (144 days). The primary reason for this difference is that fewer steps are expected to be required in the computer delivery process.

Figure 5-8. Key steps in NAEP paper vs. computer test delivery, with estimated elapsed times

Paper delivery		Computer delivery	
Pilot test			
Step	Estimated elapsed time in days	Step	Estimated elapsed time in days
• Draft items created on paper, reviewed, and revised internally	28	• Draft items created on paper, reviewed, and revised internally	28
• Items reviewed/revised at committee meeting	4	• Initial version of items produced online	13
• Camera-ready items produced for clearance package	15	• Committee review of items online via World Wide Web (WWW)	7
• Clearance package sent to NAGB/NCES	4	• Items revised	13
• Items reviewed, comments received from NAGB/NCES	9	• NAGB/NCES review and clearance via WWW	15
• Final versions of items produced, sent to be published	13	• Final versions of items available on WWW	10
• Sample versions of test booklets produced	13	• Test administered	16
• Test booklets printed and shipped to administrators	15	• Student data transferred from laptops (where used) to NAEP database	10
• Test administered	15	• Student responses used to refine automated scoring algorithms for those constructed-response items to be scored by machine	18
• Test booklets sent to scoring contractor for scanning	8	• Items either automatically scored or evaluated online by NAEP raters	10
• Training samples selected for scoring	13	• Scores entered directly into NAEP database	10
• Scanned responses scored on computer by NAEP raters	8	• Data sent to contractor for analysis	<u>10</u>
• Scores sent to NAEP database	10		160
• Data sent to contractor for analysis	<u>10</u>		
	165		
Operational assessment			
Step	Estimated elapsed time in days	Step	Estimated elapsed time in days
• Final test items selected and revised	14	• Final test items selected and revised	14
• Committee reviews final versions of items	4	• Committee reviews final versions of items via WWW	4
• Camera-ready test forms developed, sent to NAGB/NCES for clearance	14	• Final versions of items developed	9
• Items reviewed, comments received from NAGB/NCES	9	• NAGB/NCES review and clearance via WWW	15
• Final versions of items produced, sent to be published	13	• Test administered	13
• Sample versions of test booklets produced	13	• Student data transferred from laptops (where used) to NAEP database	10
• Test booklets printed and shipped to administrators	15	• Student samples collected for training	13
• Test administered	13	• Scoring completed automatically or responses evaluated on computer by NAEP raters	8
• Test booklets sent to scoring contractor for scanning	8	• Scores entered into NAEP database	10
• Training samples selected	13	• Data sent to contractor for analysis	<u>10</u>
• Scanned responses scored on computer by NAEP raters	8		106
• Scores entered into NAEP database	10		
• Data sent to contractor for analysis	<u>10</u>		
	144		

NOTE: Time estimates assume a 100-item test with 75 percent multiple-choice items and 25 percent short constructed-response items. Elapsed times do not represent level of effort.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math 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 depends primarily on the item format and whether an authoring template, an examinee tutorial, a response type for that format, and supplementary tools (e.g., on-screen calculator) exist in the delivery software. For multiple-choice items, development costs should generally be comparable for either delivery mode. Commercial web-delivery systems have the templates to allow item authoring, the tutorials to show examinees how to answer, the response types to display the items and give students an entry mechanism, and the associated tools for any additional processing that students may need to perform. Further, the trend in test development is toward item authoring and display systems built around Extensible Markup Language (XML). In such systems, static multiple-choice items can be written and entered in the same way regardless of whether they are destined for online or paper delivery. Thus, the development costs for online tests comprising multiple-choice items should be indistinguishable from items destined for paper delivery.

For static constructed-response item formats, like essays or short answers, the development costs for online tests should also be closely similar to paper. Again, commercial web-delivery systems will generally have the necessary authoring templates, examinee tutorial segments, response types, and tools. For essay items, the response type will consist of a screen area that displays the prompt text, an answer box into which text can be typed, and one or more associated functions. In the experimental system used for NAEP writing research, these functions include copy, delete, insert, and hide prompt (to increase the size of the response area). A spelling checker is available as an associated tool.

Commercial systems also typically include more interactive response types. Some of these response types, like drag-and-drop and hot-spot items, are analogous to the matching and marking tasks that are currently used on paper tests. Writing the item, entering the text, and creating and entering any graphical components should also be no more time-consuming than the processes involved with conventional question creation.

Costs may be higher for online delivery in other cases. One case is when the template, tutorial, response type, and tools exist but the content development itself is labor-intensive. Such may be the case for multimedia items that require the creation of animations, editing of existing audio or video, or the recording of new audio or video. These activities can be very costly compared to simpler computer-delivered item types or to paper types intended to measure similar skills. However, if the target skill can be measured only by dynamic presentation, then the development of online items may be no more expensive than creating the same content for delivery by cassette recorder.

A second situation in which development may be more expensive is when the authoring templates, tutorials, response types, or tools needed for the envisioned item do not exist. For example, development committees might request items that ask the examinee to manipulate large data sets using such canned statistical functions as mean, median, standard deviation, and range. This new response type could certainly be built using existing components. The extant screen frame that presents the item stem and the response area that allows selection of a multiple-choice option or entry of a number could be reused. But ways to display the data set and apply the statistical tools might have to be designed, programmed, and evaluated for usability with students. A template for creating new items of this class would need to be invented so developers could easily insert new data sets. Finally, content describing how to use the statistical tools would have to be added to the examinee tutorial.

This discussion is not to suggest that such effort would be wasted. If the item type is able to measure an important framework content area in a way that could not be done through conventional methods, the investment would be justified. Once developed, these components would be added to the delivery system, making creation of new “large data set” items a relatively straightforward task.

Relative costs of test delivery and administration. The NAEP mathematics assessment is a “trend” assessment that, in addition to employing new items, regularly reuses questions from previous years in an effort to measure change. This trend measurement is conducted over relatively short times, with new trend lines begun periodically. To avoid an impact on trend, it would be safest to use computer-based testing only for presenting newly developed items.

In past assessments, such items have been integrated with trend items. Since switching between paper and online delivery might also affect trend, computer-presented items are probably best restricted to their own sections and administered to samples of students taking the larger assessment. Alternatively, one could wait until a new trend line has begun and plan for the appropriate portions of that assessment to be delivered online.

Delivery and administration costs for an online assessment include licenses 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 equipment; copying of test software and item banks to the laptops and removal of student data from them; shipping of laptops; field administrators' salaries; and telephone technical support for these individuals.

Some of these delivery and administration costs will be quite variable. In particular, laptop costs will depend on examinee sample size and the number of school machines that can be used. The number of school machines 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 reduce standardization in ways that materially affect test performance. How machine variation affects performance is not well known.

The MOL field test showed that the staff employed to administer paper NAEP assessments could successfully carry out an online examination. They were able to manage pre-assessment contacts with schools, help school staff certify that local machines were capable of delivering the assessment, and conduct the assessment. In the process, they also were able to solve routine technical problems (e.g., reestablishing connections to the MOL server in the middle of a test). They were challenged, as even more technically skilled staff would be, when more serious computer difficulties occurred. The implication for an operational NAEP assessment, however, is that the use of well-tested delivery systems would probably be more advantageous than the use of more costly, technically skilled administrators.

Compared to a pencil-and-paper administration, online testing requires slightly more staff time for telephoning schools to plan the assessment and

more pre-assessment time on site to certify computers. As school technology improves and delivery systems support a greater range of configurations, the need for preadministration planning should decrease.

As implemented in MOL, fewer students per session were tested online than in the paper sittings. This difference was a function of server capacity and of the need to keep the burden on the field administrators low for this first national study. In an assessment, NAEP would use a production delivery system with greater server capacity and would expect administrators to handle larger groups comfortably. NAEP paper administrations routinely assess groups of 30 students. Assessing groups of 30 students online may be possible in schools that can devote a laboratory of certifiable machines to the assessment. In those cases where a school cannot, the group size will range from five (the number of laptops an administrator can 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 tests on paper (perhaps by half). That differential will diminish as the technologies used for assessment become smaller and cheaper (e.g., personal digital assistants).

While the additional delivery and administration expenses of electronic assessment are considerable, they are partly balanced by eliminating some of the larger costs of paper delivery, including the printing and shipping of test booklets and the purchase and shipping of calculators. In addition, the expense associated with last-minute changes should be reduced. Changes to instruments, to spiraling designs, or to sampling plans would otherwise need to be made by reprinting or reassembling materials.

Relative cost of scoring. The cost of scoring computerized tests should not differ from current NAEP processes so long as human judges are used to evaluate constructed responses. However, if automated scoring can be used instead of human judges, a large cost savings may be achievable. Currently, in NAEP mathematics it costs roughly as much per student to score constructed-response items manually as to print, ship, perform receipt control, and track assessment booklets. For automated scoring to be implemented, though, 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.

At the pilot-test stage of an assessment, as opposed to the operational stage, automated scoring may be of only limited value. For pilot tests, the sample sizes involved are small and the cost for human scoring is relatively low. Furthermore, items are sometimes dropped after pilot testing, so any effort put into training automated systems for specific items would not carry over to the operational stage.

In the operational stage of a NAEP 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 raters grade items with the same accuracy and standards as in previous years. A significant benefit to automated grading would be that there should be no score drift or change in agreement from one year to the next.

Figure 5-9 summarizes the relative costs for NAEP of computer vs. paper assessment. Assuming an assessment of 100–120 newly developed NAEP mathematics items with no more than limited interactivity, the costs for an online assessment should be similar for test development, similar or higher for test delivery and administration, and similar or lower for scoring.

Figure 5-9. Relative costs for NAEP of computer vs. paper assessment

Process	Relative cost	Comment
Item and software development		
Creating static multiple-choice (MC) or constructed-response (CR) items	Similar	Commercial delivery systems will have item templates, tutorial segments, response presentation and answer formats, and supplementary tools.
Creating MC or simple CR items with limited interactivity (e.g., drag and drop)	Similar	Commercial delivery systems will have item templates, tutorial segments, response presentation and answer formats, and supplementary tools.
Creating multimedia items	Higher than static paper items	Commercial systems may or may not have needed authoring or delivery components. Cost of creating audio, video, or animation usually high but probably similar to that for audiocassette or videocassette delivery.
Creating new item types	Higher than paper	Item templates, tutorial segments, response presentation and answer formats, and supplementary tools will need to be created and tested for usability.
Test delivery and administration		
Delivering test to schools	Similar or higher than paper	Includes cost of licensing 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 and shipping test booklets, and purchasing and shipping calculators. Overall cost difference depends greatly on size of examinee sample and on number of laptops required.
Preparing for and administering test	Higher than paper	More time required for initial contacts with schools and for certifying computers.
Providing telephone technical support	Similar	Help desk routinely used for paper assessments at similar staffing level.
Changing items, spiral designs, and sampling plans	Lower than paper	Eliminates need to reprint or reassemble materials.
Scoring		
Automatically scoring items	Lower than paper	As long as examinee 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), 2001 Math Online Study.

6. Operational Issues

This section reports on the logistical challenges associated with administering a NAEP mathematics survey 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 data collection contractor, supplied much of the information for this section of the report (Ennis, Hart, and Moore 2001). Westat sampled and recruited schools, and administered all instruments.

Recruiting Schools

Westat began recruiting for the spring 2001 data collection in fall 2000. After sending an initial mailing about upcoming NAEP assessments, Westat sent a special letter to principals that focused on the MOL project. 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 study.

Westat reported that most of the schools contacted were interested in participating. Factors that helped gain cooperation were a principal's interest in technology and the need for only ten students per school to complete the online test (about 20 fewer students than the usual NAEP survey). Additionally, for some school officials, the fact that the study did not require collection of teacher, special-needs student, or school questionnaires helped reduce concerns about burden.

Training Field Administrators

A two-and-one-half day training session was held at Westat's headquarters in Maryland on March 26–28, 2001. The presentations focused on the technical issues associated with readying school computers and trouble-shooting problems, as well as on administering MOL.

Preparing for the Administration

Westat staff visited each school approximately two weeks prior to its test date, as is routine for NAEP assessments. For MOL, the staff member's goal was to arrange for testing 10 students, either simultaneously or split into morning and afternoon sessions.

During the visit, the staff member worked with school personnel to draw the sample, establish locations and times for the administration, and make any other necessary arrangements. Scheduling computer labs for testing often proved challenging because that space was generally used throughout the day. In order to accommodate MOL, schools often had to cancel computer lab classes.

In addition to the above activities, the Westat administrator met with the school technology representative to determine whether the sessions would be delivered via the Internet, by laptop, or a combination of the two. To make this decision, each school computer that was potentially available for the testing had to be checked against the technical specifications for MOL. This certification was conducted by asking school staff to log onto an ETS web site from each computer. Through this process, each computer was evaluated for the required characteristics. On the day of the administration, many Westat staff performed portions of the procedure again to ensure that speed of Internet transmission was adequate to allow the test to be conducted properly at that time.

The technical specifications, shown in figure 6-1, were dictated by the web-based testing system ETS uses to study the potential of the Internet for large-scale assessment. Because it was developed for experimental use, this system supports only Windows machines. 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.

When the test is administered via the Internet, the ETS system delivers one item at a time to the browser residing on the school computer. In an alternative configuration, the system can be used in the same way on a laptop that is not connected to the Internet. In that case, the server software resides on the laptop hard drive and presents items to the machine's browser as if there were an active Internet connection. When some or all of a school's computers could not be used to deliver MOL, Westat brought a maximum of five laptops into the building.

Figure 6-1. Technical specifications for school computers

Feature	Requirement
Computer type	Personal computer
Screen resolution	Capable of 800 x 600 resolution
Screen colors	Capable of 256 colors
Processor type	Pentium or higher
Processor speed	166 MHz or faster
Random access memory	At least 32 MB
Internet bandwidth	At least 128 kilobits per second
Web browser	Microsoft Internet Explorer Version 5.0 or higher
Browser cookies	Enabled
Hard drive	Required
CD-Rom drive	Required
Macromedia Flash software	Version 5.0 or higher available for download from Web
Java Virtual Machine software	Available for download from Web

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

School staff attempted to certify 868 personal computers.²³ Of this number, 704 machines (81 percent) ultimately passed. Table 6-1 summarizes the primary reasons school PCs did not pass.

Table 6-1. Primary reasons some school PCs failed certification for online testing, grades 4 and 8: 2001

Reason for Failure	Number of PCs
Throughput less than 128 kilobits per second	83
Screen resolution capability less than 800 x 600	41
Central processing unit less than 166 megahertz	19
Java not installed on computer	12
Flash plug-in not installed on computer	6
Random access memory less than 32 megabytes	3

NOTE: A PC could fail for more than one reason, but only the primary reason is given.
 SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

Westat found school and district technicians to be helpful, but variable in interest, skills, and availability. Most frequently, a school-based technician worked with the Westat staff member to deal with computer-related issues. District technicians were often consulted by telephone to assist with specific problems. In addition to assisting with certification, technicians sometimes needed to reset screen resolution, disable firewalls, or download plug-ins.

During the preadministration visit, most Westat staff asked that the school technician also be present at the beginning of the test to troubleshoot any difficulties, and again at the end to restore any configuration changes to their original settings. In most instances, Westat staff were successful in securing this assistance and, in many cases, the technician was present throughout the entire session.

In some schools, the technician was also appointed to serve as the NAEP coordinator. Westat expressed frustration with this arrangement, since many technicians lacked the authority, time, and skills needed for arranging the administrations.

²³ Not included in this figure are a small number of computers that were not able to run the certification process because school system firewalls or filters prevented it. Macintosh computers also are not included. Schools with such computers were automatically designated for laptop delivery.

Conducting the Administrations

Table 6-2 summarizes the method of MOL test delivery. At grade 4, the overwhelming majority of students and schools completed the test on laptops not connected to the Internet. At grade 8, the methods were more balanced: 38 percent of students used Internet-connected school computers and 46 percent of schools tested some or all of their students that way.

Table 6-2. Number and percentage of students and schools, by method of computer-based test delivery, grades 4 and 8: 2001

Students			
Number	Percent tested on NAEP laptops	Percent tested on school computers	
Grade 4			
1,036	80	20	
Grade 8			
1,013	62	38	
Schools			
Number	Percent with laptop delivery only	Percent with delivery by school computers only	Percent with both laptop and school computer delivery
Grade 4			
124	75	17	8
Grade 8			
109	53	29	17

NOTE: Detail may not sum to totals because of rounding.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

Westat reported that some students especially enjoyed completing the test on a laptop. A small number of students accustomed to using desktop PCs or Macintosh computers needed a few minutes to adjust (e.g., to the keyboards), but no significant problems were reported. Westat staff noted only minor problems securing suitable space to set up. Occasionally, classroom lighting made it difficult to read the laptop screens clearly; administrators dealt with this problem by moving the laptops, tilting the screens, or adjusting the contrast settings.

Some performance problems did occur. The Westat Help Desk logged 141 requests for assistance. As indicated in table 6-3, the single most common source was the laptops. Laptop problems had two causes: (1) hardware malfunctioning and (2) a time-out setting in the test delivery software. These problems were resolved by replacing computers with newer models and by increasing the time-out limit.

Table 6-3. Percentage of performance problems, by cause reported to the Westat Help Desk, grades 4 and 8: 2001

Category	Percent of calls
Certifying school computers	9
School-computer problems during assessment	16
Laptop problems during assessment	37
Administrator computer problems	18
Other	20

NOTE: Administrator computers were not used for testing.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

In addition to their performance problems, Westat administrators found the laptops cumbersome. Although they were packed in a single suitcase on wheels, it was difficult to get that case in and out of cars and up and down stairs. In addition, setup and breakdown were time consuming, and assessing a maximum of five students at a time was less efficient than traditional administrations.

As noted, the study design called for all students to complete a paper-and-pencil test before taking MOL. Westat staff found the combination of paper and computer activities problematic because of the difficulties posed by distributing materials and managing space.

Two administration methods were employed for the computer-based sessions. The first, used at all grade 4 and some grade 8 sessions, was akin to a group administration: all students started the test at the same time and waited until all were finished before being dismissed. For the second, each student began the test when she or he arrived and left as soon as she or he had finished. Westat administrators preferred this option for eighth-grade students because it freed the staff to log students on as soon as they arrived instead of having to wait for all students to be present.

Although some computer-based testing programs have had security problems, Westat administrators did not report any such concerns. It may be that no security concerns were reported in part because the number of students tested in each session was small enough to monitor carefully and because the test was not perceived as having high stakes. In addition to monitoring, other security precautions were taken in the design and delivery of the MOL test. For instance, access to the test was obtained by locating the proper web site and logging on with an administrator ID and password. Also, at the conclusion of the testing session, Westat administrators routinely cleared each machine's Internet cache, which might have retained copies of item displays, and deleted the browser history, which would have retained the delivery site's web address. Commercial test-delivery software typically incorporates additional security mechanisms, such as limiting keyboard functions that may facilitate item theft, preventing students from temporarily exiting the test to use other programs or files, and clearing the computer's hard drive of any residual test content when the test has ended.

Student and School Reactions

Westat administrators informally obtained feedback from students and school staff (Ennis, Hart, and Moore 2001). Staff reported student feedback from 88 of the 126 grade 4 schools. Administrators reported far more positive responses by students than negative ones and were in agreement that student behavior during the computer sessions was much better than in the paper administrations. The most common reasons students gave for liking the test were that it was fun, that they liked using the computer more than paper and pencil, that they liked using the calculator on the computer, and that it was easy. The most common reasons students gave for not liking the test were that the mathematics was too hard, that they had problems with typing, that they had problems with the computer (e.g., laptops freezing), and that the test was too long.

Westat administrators also informally asked school staff for their reaction to the test. Of the 92 school staff who offered comments, 75 were positive, and the rest were negative, mixed, or neutral.

At grade 8, Westat staff received student feedback from 63 of the 110 schools. The most common reasons students gave for liking the test were that they liked using the computer more than paper and pencil, it

was fun, and it was easier. The most common reasons for not liking it were difficulty using the on-screen calculator, difficulty typing, and that the mathematics was hard. (The online calculator was a scientific one similar to that provided to students completing a conventional grade 8 NAEP mathematics test.)

Westat administrators received reactions from 73 school staff. There were 61 positive responses, and the remainder were negative, mixed, or neutral comments.

Data Quality

Because of technical problems, some sampled students were unable to take the online test. At the fourth grade, 58 students fell into this category. At the eighth grade, 56 sampled students were nonrespondents because of problems with the online test.

In addition to the technology failures noted above, some students were prevented from working through the tutorials and the test questions without interruption. These problems included school Internet connections that were occasionally dropped and NAEP laptops that sometimes froze during administration. In such cases, test administrators attempted to restart students where they had stopped or, if this was unsuccessful, from the beginning of the test. Regardless of where students restarted, an additional test-session record was created. After all tests had been completed, ETS technical staff resolved these multiple records. Approximately 15 percent of the fourth-grade and 11 percent of the eighth-grade records needed to be reconstructed in this way.

An interruption could potentially affect performance in either negative or positive ways. Being interrupted could have negative consequences by reducing motivation or generating frustration that would translate into poorer performance than the student might otherwise achieve. Positive consequences would result if an interruption provided a needed break or even a small amount of extra time. Extra time could accrue because the test would sometimes be restarted from the beginning, allowing students the opportunity to answer more quickly items they had already considered, giving them more time than they would otherwise have had for subsequent items. Even if the test were not restarted from the beginning, some extra time might also be provided, as the student would be brought back to the last completed question.

Table 6-4 shows the mean scale scores for students with and without fragmented test-session records.

Table 6-4. Mean MOL scale scores for students with and students without fragmented test-session records, grades 4 and 8: 2001

	Students with fragmented records	Students without fragmented records
Grade 4	193 (3.0)	201 (1.2)
Grade 8	192 (3.4)	199 (1.6)

NOTE: MOL=Math Online. Standard errors of the scale scores appear in parentheses.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

To evaluate whether the technical problems that necessitated restarting might have affected student performance, MOL score was regressed on test-session status (fragmented vs. nonfragmented), controlling for performance on the initial paper block. This regression produced a significant effect for session status for the fourth grade ($F, 1,35=15.66, p < .01$) and for the eighth grade ($F, 1,35=12.43, p < .01$). However, the impact on scores appears to be minimal. For eighth grade, which was the main focus of the analyses in this report, the effect's magnitude can be estimated by using the regression to predict what the MOL scores of students with fragmented records would have been had their sessions not been interrupted. When the MOL mean for the total eighth-grade group is recalculated using predicted scores for students with fragmented records and the actual scores of those with nonfragmented records, the sample mean increases marginally from 198 to 199.

In addition to technical problems, a second factor that could have affected study results was that the NAEP laptop machines on which most students took MOL would have been less familiar than their school computers. Table 6-5 shows the scale-score means for students taking MOL on school computers and NAEP laptops.

Table 6-5. Mean MOL scale scores for students testing on school computers and NAEP laptops, grades 4 and 8: 2001

	Students on school computers	Students on NAEP laptops
Grade 4	200 (2.1)	200 (1.4)
Grade 8	202 (2.3)	195 (1.7)

NOTE: MOL=Math Online. Standard errors of the scale scores appear in parentheses.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

To determine whether computer type might have affected student performance, MOL score was regressed onto computer type (school computer vs. NAEP laptop), with score on the initial paper block serving as a covariate. For fourth grade, computer type was not related to MOL score after controlling for performance on the paper block ($F, 1,35=3.52, p > .05$). At eighth grade, however, computer type was a significant predictor ($F, 1,35=82.54, p < .00$). An estimate of the effect of computer type can be gained by using the regression to predict what the MOL scores of students who took the test on laptop would have been had they taken it on desktop. This estimate needs to be regarded cautiously, however, because there may be other factors correlated with taking the test on laptop that would affect performance *regardless* of computer type (e.g., level of computer familiarity). When the MOL mean for the total eighth-grade group was recalculated using predicted scores for students taking the test on laptop and the actual scores of those administered the test on desktop, the sample mean increased from 198 to 200. This increase in mean score likely overlaps with that of the increase predicted for students with fragmented records, as close to half of those students took their tests on laptop computers. In any event, at eighth grade, it seems that somewhat greater comparability between the computer and paper tests might have resulted from administering a larger proportion of the tests on school computers.

7. Summary and Conclusions

The Math Online study addressed measurement, equity, efficiency, and operational issues associated with conducting a NAEP mathematics assessment via computer. Data were collected from samples of fourth- and eighth-grade students in more than 100 schools at each grade level throughout the United States.

The study considered measurement issues related to how delivery mode might affect what can be measured and how students perform. An analysis of items used on the NAEP 2000 eighth-grade mathematics assessment suggested that most questions could be delivered electronically. Items from the Number Sense, the Data Analysis, and the Algebra and Functions content areas were generally judged easier to implement than those from the Measurement and Geometry content areas. The specific characteristics of items felt to be less amenable to computer delivery included ones that require more than a single screen; that are intended to determine how effectively a student can manipulate a physical tool (e.g., a protractor); that ask the student to create a drawing, enter extended text, or produce formulae; that require a lengthy tutorial or directions; that are accompanied by paper stimuli; or that presume constant size of graphics (when delivery software doesn't control screen resolution).

With respect to performance, the mean scale score for eighth-graders taking the computer test was 4 points lower than for a randomly parallel group taking the paper version of the same 25-item measure. At the item-parameter level, although the IRT difficulty estimates for the two modes were almost perfectly correlated, the item difficulties for the computer test were generally greater (by .22 logits on the IRT scale and .05 points on the proportion-correct scale).

The study also considered the impact of test mode on equity. In grade 8, performance of selected NAEP reporting groups was evaluated to see whether their scores differed on paper vs. computer versions of the same test. Separate comparisons were made by gender, race/ethnicity, parents' education level, region of the country, school location, and school type. Results showed that, for the NAEP reporting groups examined, performance generally was not differentially affected by electronic vs. paper delivery.

In addition to effects on the examined NAEP reporting groups, the study investigated the impact of computer familiarity on test performance. Students' responses to background questions suggested that the overwhelming majority used computers at home and at school.

To determine if lack of computer familiarity affected online test performance, hands-on measures of input accuracy and input speed and a measure of self-reported computer experience were used to predict online test performance. After controlling for performance on a paper mathematics test, self-reported computer experience, input speed, and input accuracy predicted MOL score for fourth-grade students. For eighth-grade students, input speed and input accuracy were the significant predictors. This finding suggests that computer familiarity may distort the measurement of mathematics achievement when tests are administered online to students who lack basic technology skills.

In addition to measurement and equity issues, the study considered questions related to efficiency. Here, the relative costs and timeliness of different test delivery modes were analyzed, as were the feasibility of two technological innovations, automated item generation and automated scoring. With respect to timeliness, it is anticipated that moving tests to computer would not have any significant effect on the pilot stage of the NAEP development cycle, but could possibly speed up the operational stage somewhat by requiring fewer steps. The costs for an online assessment should be similar for test development, similar or higher for test delivery and administration, and similar or lower for scoring, if one assumes an assessment of 100–120 newly developed NAEP mathematics items with no more than limited interactivity. Among the key cost drivers are examinee sample size, the number of items, how many students would need to be assessed on laptops, and the number of students per school that can test simultaneously. A very considerable increase in costs would result, for instance, from assessing a large sample in small groups primarily on laptop machines.

One potential cost-saving technology is automatic item generation. This technology rests on two assumptions: that classes of items can be described in sufficient detail to allow computer generation of instances and that enough is known about the determinants of difficulty to reduce the need for calibrating each instance individually. For the study, general descriptions, or models, were created for 15 NAEP items and instances, or variants, of each item were generated by computer. Three different versions of each item were administered to randomly parallel student samples in paper-and-pencil format, along with 11 items that were identical across samples. Results suggested that, on average, the item parameter estimates for each instance changed somewhat more from one sample to the next than did the parameter estimates for the identical items. However, this added variation had no significant impact on NAEP scale scores. This result implies that variants could be automatically generated, a subset empirically calibrated, and parameters for the remaining variants imputed without affecting the quality of NAEP population estimates.

Overall, about three quarters of the items used on the NAEP 2000 mathematics assessment appear amenable to automatic item generation. The only framework content area for which the majority of items could probably not be automatically generated was Geometry and Spatial Sense, for which some four in ten items appeared suitable. In general, the more suitable items for automatic generation were pure computation, story problems where the underlying mathematics could be applied to a variety of contexts, and figural questions with simple graphical or tabular elements that could be meaningfully varied.

Although human raters scored all constructed-response items, automated scoring technology was also employed to score eight of the nine fourth-grade items and eight of the nine eighth-grade items. These questions either required simple numerical or text responses, or more extended textual responses. Automated scoring of the items requiring simple responses was highly successful. For the items at grade 8, automated scoring agreed with the judgments of human readers to the same degree

as human readers agreed with each other. For the items at grade 4, a small percentage of the simple responses could not be graded automatically (i.e., less than 10 percent for all but one item). Of those responses that could be scored, the machine's grades were interchangeable with human scores for seven of the eight items. For the five questions requiring extended text responses, all answers were scored but, in most cases, at agreement levels somewhat lower than those of human judges. The primary cause of the disagreements was the machine's tendency to treat correct responses that were misspelled as incorrect, a shortcoming that can be addressed by including common misspellings in the automated scoring key or including a spell-check before an answer is submitted.

The last set of issues concerned field operations. At preadministration visits, field staff worked with school personnel to determine if local computers could be used for the test and, if not, made plans to use NAEP laptop machines. Most students were tested on laptops: 80 percent at grade 4 and 62 percent at grade 8. The principal reasons for laptop use were that schools employed Macintosh equipment, which was not supported by the ETS research web-delivery system, or that their Internet connection speeds were not fast enough for this system. While web delivery worked well, taking the test on laptop computer was associated with lower performance in eighth grade than taking the test on a Web-connected school computer, after controlling for score on the initial paper mathematics test. This lower performance may have, in part, been due to technical problems that affected the functioning of the NAEP laptops. Technical problems also occasionally occurred on school computers, manifested primarily in lost Internet connections. Both laptop failures and occasional Internet connection difficulties caused some examinations to be interrupted. Interruptions were associated with marginally lower performance and may be one small component of the noncomparability of computer and paper tests detected in this study. Equipment problems aside, reaction from students and school staff to electronic delivery was overwhelmingly positive at both grade levels.

8. Implications for NAEP

The authors believe that these results have several implications for NAEP. First, most NAEP mathematics items could be computer delivered, arguably improving the measurement of some framework content areas. At the same time, conventional delivery may be needed for other items, especially those that require the manipulation of a real (as opposed to a simulated) physical object.

Second, although the computer test was harder than its paper counterpart, this effect generally did not differentially impact the NAEP reporting groups examined. For instance, there was no statistically significant indication that taking a test on computer disadvantaged students of any particular gender or race/ethnicity. Because the sample sizes were small, however, this finding should be subjected to further research. Also, because socioeconomic status (SES) was not one of the population groups investigated, future research might address whether computer delivery negatively affects any SES group. In the absence of differential impact, it may be possible that the paper and computer mathematics tests can be equated to remove mode effects (as would be necessary if the scores from different modes were to be aggregated or compared from one year to the next).

Third, even though almost all students claimed some familiarity with computers, the data suggest that lack of computer proficiency may introduce irrelevant variance into NAEP online mathematics test performance. This result is similar to that found by Russell (1999). For mathematics, his study included only constructed-response items given to some 200 eighth-grade students in Massachusetts. Russell found that, compared to taking a test on paper, taking the test on computer had a negative effect, which lessened as keyboarding skill increased. What causes the effects found in these two studies? One possible contributing factor is the presence of constructed-response items which, depending upon the response requirements, can demand computer skill. In the Russell study, all items required the student to generate at least a sentence of text. When asked what problems they had taking the mathematics test online, 30 percent of the students in that study indicated difficulty typing.

In the present investigation, constructed-response items appeared to shift in difficulty more than multiple-choice items when presented on computer.

Constructed-response items also needed to be adapted more than multiple-choice items in order to be rendered on computer. These results suggest that, in moving paper mathematics items to computer, it may sometimes be harder to hold difficulty constant for constructed-response than for multiple-choice questions. This transition may introduce the need for computer skill in responding, may make it impossible for students to show their work in alternative ways (e.g., diagrammatically), or may otherwise change the nature of the task.

Also associated with item format is the potential for a presentation effect in scoring, as has been found for writing assessments. Several studies have noted that human readers grade the same essays differently depending upon whether they were handwritten or typed (Powers, Fowles, Farnum, and Ramsey 1994; Powers and Farnum 1997). Handwritten answers tended to receive higher grades than typed responses, possibly because handwritten answers look less finished, thus encouraging readers to be more tolerant of minor errors. These studies considered only essay tests, so it is unclear if the same effect would occur for NAEP mathematics items. For NAEP mathematics items, scoring emphasizes content rather than the way that content is communicated. In addition, the responses in the present study involved much less text than in essay examinations. On the eighth-grade test, five of the ten constructed-response items required only simple numeric entry or clicking on hot spots, while the remaining five questions entailed explanations of no more than a few sentences. Further research might examine whether the MOL mode effect is partly due to reader bias by transcribing a sample of responses from each mode to the other, and having different readers grade subsets of the transcribed and original versions blindly.

The presentation and response characteristics of the constructed-response format may not, of course, be the only cause of mode differences. In the present study, several multiple-choice items also showed significant difficulty shifts. This finding is consistent with that of two other studies conducted with reasonably large samples of school-age students, both of which found scores on computer-delivered multiple-choice mathematics tests to be lower than those for the paper-and-pencil versions (Choi and Tinkler 2002; Coon, McLeod, and Thissen 2002).

Further research might attempt to untangle the relationship between response format and online performance by randomly assigning students at different grade levels (or at different degrees of computer familiarity) to high-keyboard-intensive constructed-response items, low-keyboard-intensive constructed-response items, and multiple-choice items presented in each delivery mode. In addition, varying students' exposure to tutorials and online practice tests might be tried. Repeated practice in advance of the testing session may be enough to ameliorate at least some types of mode effect. (However, this practice would need to be accomplished in ways that would not create additional burden on participating schools.) For the near term, then, students' computer proficiency should remain a concern with respect to online delivery of NAEP mathematics assessments, especially when the measures include constructed-response questions, or when students have limited computer experience.

Student access to and use of computers is growing rapidly (National Center for Education Statistics 2002; U.S. Department of Commerce 2002). Further, computer use among minority-group students is approaching the use rates for the majority, due to the presence of machines in school (U.S. Department of Commerce). As students become more experienced with technology, and as computer interfaces improve, any mode effects associated with computer familiarity are likely to disappear, even for constructed-response tests.

The fourth implication of this study for NAEP is that, when constructed-response tests are deemed desirable, automated scoring may help reduce costs and possibly speed up reporting. The use of these techniques fits nicely into the NAEP operational process. The algorithms needed to score particular items can be trained with pretest data, then checked with an initial sample of responses from the assessment before production grading commences. During production grading, back-reading by human judges can occur to check the accuracy of machine scores.

A fifth implication is that, in addition to automated scoring, automatic item generation might increase NAEP's efficiency. One or more item models could be written for each particular framework subtopic. Each model could be calibrated by generating a small sample of variants and pretesting them. One of two operational delivery options could then be used. For paper assessment, additional variants would be generated from each model, with

each variant assigned to a different block, thereby providing greater coverage of each framework subtopic. For a computer-delivered assessment, variants could be generated on the fly, so that rather than being preassembled, item blocks would be created in the field as the assessment was administered. For future assessments, new variants could be generated from the same set of calibrated item models.

The sixth implication is that NAEP should expect the transition and operating costs for electronic assessment to be substantial. These costs are more likely to be recovered in the long rather than the short term. All the same, NAEP may need to move some assessments to computer delivery regardless of higher cost. As students do more of their academic work on computer, documenting that learning in a medium different from the one they routinely employ will become increasingly unjustifiable (Bennett 2002). That is, for those areas in which computers have become standard tools for doing intellectual work (e.g., in writing, information search), NAEP may have no choice but to assess the associated proficiencies online.

The seventh implication is that the technology infrastructure is not yet developed enough to support national delivery via the Web directly to school computers. In this study, Web delivery was supplemented by bringing laptop computers into schools, giving most tests on these machines. Perhaps because of technical problems, unfamiliar or more cramped keyboards, or smaller screens, NAEP laptops were associated with somewhat lower scores for eighth-graders than were school computers. However, the need for NAEP to bring laptops into schools will certainly not be as great for future NAEP assessments. First, the technical requirements for using school machines can be considerably lower if the assessment blocks assigned to computer delivery initially employ only multiple-choice and simple constructed-response items. Additionally, school technology is being improved continually, especially as states move components of their assessment systems to online delivery. At least a dozen states are piloting such delivery or actively implementing operational tests (Bennett 2002; Olson 2003). Finally, laptop screens and keyboards have improved considerably since MOL was administered in 2001, so that detrimental effects apparently due to taking a test on these computers may disappear.

As school machines become the predominant delivery mechanism, variation across machines (e.g., monitor size, screen resolution, connection speed) may play a greater role in introducing irrelevant variance. Such an effect has already been reported for differences in screen resolution and monitor size on reading tests (Bridgeman, Lennon, and Jackenthal 2003). Various means exist to control such variation, including manipulating resolution through the delivery software or, in the case of connection speed, downloading the entire test before the session commences. Consequently, it may be possible to keep irrelevant effects within tolerable limits. NAEP's delivery systems should consider the use of similar controls. In addition, research might evaluate the controls' effectiveness.

The final study implication is that there occasionally will be equipment failures that interrupt assessment for some students, regardless of what equipment is used. NAEP can deal with these events by discarding the affected data, retaining it, or returning to schools to conduct make-up sessions. Future research might investigate the nature and magnitude of the bias that might be introduced by retaining, as compared to discarding, the affected data.

NAEP's history has been one of leadership and innovation. NAEP has continued this tradition by conducting one of the first studies of the comparability of computer versus paper assessment using a nationally representative sample of school-age students. This study gives a glimpse of what is promising and what is problematic about electronic delivery. Follow-up projects on NAEP writing and problem solving in technology environments will add to the understanding of how computers will, and will not, help improve NAEP and educational assessment generally.

References

- Allen, N.L., Donoghue, J.R., and Schoeps, T.L. (2001). *The NAEP 1998 Technical Report* (NCES 2001–509). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics.
- Bejar, I.I., Lawless, R.R., Morley, M.E., Wagner, M.E., Bennett, R.E., and Revuelta, J. (2002). *A Feasibility Study of On-the-Fly Adaptive Testing* (Research Rep. No. 02–23). Princeton, NJ: Educational Testing Service.
- Bennett, R.E. (2002) Inexorable and Inevitable: The Continuing Story of Technology and Assessment. *Journal of Technology, Learning, and Assessment*, 1(1). Retrieved August 23, 2002, from <http://www.bc.edu/research/intasc/jtla/journal/v1n1.shtml>.
- Bridgeman, B. (1998). Fairness in Computer-Based Testing: What We Know and What We Need to Know. In *New Directions in Assessment for Higher Education: Fairness, Access, Multiculturalism, and Equity* (GRE®, FAME Report Series, Vol. 2), pp. 4–10. Retrieved April 19, 2002, from <ftp://ftp.ets.org/pub/gre/241343.pdf>.
- 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.
- Choi, S.W., and Tinkler, T. (2002, April). *Evaluating Comparability of Paper-and-Pencil and Computer-Based Assessment in a K–12 Setting*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Coon, C., McLeod, L., and Thissen, D. (2002). *NCCATS Update: Comparability Results of Paper and Computer Forms of the North Carolina End-of-Grade Tests* (RTI Project No. 08486.001). Raleigh, NC: North Carolina Department of Public Instruction.
- Dresher, A.R., and Hombo, C.M. (2001, April). *A Simulation Study of the Impact of Automatic Item Generation on Item Parameter and Ability Estimation*. Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.
- Ennis, B., Hart, B., and Moore, D. (2001). *NAEP TBA 2001 Operations Report*. Unpublished report. Rockville, MD: Westat.
- Fleiss, J.L. (1981). *Statistical Methods for Rates and Proportions* (2nd ed.). New York: John Wiley & Sons.
- Gallagher, A., Bridgeman, B., and Cahalan, C. (2000). *The Effect of Computer-Based Tests on Racial/Ethnic, Gender, and Language Groups* (ETS Research Rep. No. 00-8). Retrieved August 5, 2003, from ftp://ftp.ets.org/pub/gre/gre_96-21p.pdf.
- Glas, C.A.W., and van der Linden, W.J. (2001). *Modeling Variability in Item Parameters in Educational Measurement* (OMD Rep. No. 01-11). Newtown, PA: Law School Admission Council.
- Hombo, C.M., and Dresher, A.R. (2001, April). *A Simulation Study of the Impact of Automatic Item Generation Under NAEP-Like Data Conditions*. Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.
- Irvine, S., and Kyllonen, P. (Eds.). (2002). *Item Generation for Test Development*. Hillsdale, NJ: Erlbaum.
- Johnson, M.S., and Sinharay, S. (2002, April). *A Hierarchical Model for Item Model Calibration*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.
- Lapp, M.S., Grigg, W.S., and Tay-Lim, B.S. (2002). *The Nation's Report Card: U.S. History 2001*. Washington, DC: National Center for Education Statistics, Office of Educational Research and Improvement, U.S. Department of Education.
- Leacock, C., and Chodorow, M. (2003). C-rater™: Automated Scoring of Short-Answer Questions. *Computers and the Humanities*, 37(4): 389–405.

- Mead, A. D., and Drasgow, F. (1993). Equivalence of Computerized and Paper-and-Pencil Cognitive Ability Tests: A Meta-Analysis. *Psychological Bulletin*, *114*(3): 449–458.
- Muraki, E. (1990). Fitting a Polytomous Item Response Model to Likert-Type Data. *Applied Psychological Measurement*, *14*(1): 59-71.
- National Center for Education Statistics. (2002). *Internet Access in U.S. Public Schools and Classrooms: 1994–2001* (NCES 2002-018). Retrieved October 2, 2002, from <http://nces.ed.gov/pubs2002/2002018.pdf>.
- NCS Pearson. (undated). *National Assessment of Educational Progress: 2000 Report of Processing and Professional Scoring Activities, Main and State NAEP*. Minneapolis, MN: Author.
- Olson, L. (2003, May 8). Legal Twists, Digital Turns: Computerized Testing Feels the Impact of “No Child Left Behind.” *Education Week*, *22*(35), pp. 11–14, 16.
- 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.
- Russell, M. (1999). Testing on Computers: A Follow-Up Study Comparing Performance on Computer and on Paper. *Education Policy Analysis Archives*, *7*(20). Retrieved August 5, 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*. Retrieved August 5, 2003, from <http://epaa.asu.edu/epaa/v5n3.html>.
- Russell, M., and Plati, T. (2001). Effects of Computer Versus Paper Administration of a State-Mandated Writing Assessment. Retrieved April 19, 2002, from <http://www.tcrecord.org/Content.asp?ContentID=10709>.
- Schaeffer, G.A., Bridgeman, B., Golub-Smith, M.L., Lewis, C., Potenza, M.T., and Steffen, M. (1998). *Comparability of Paper-and-Pencil and Computer Adaptive Test Scores on the GRE® General Test* (Research Rep. No. 98-38). Princeton, NJ: Educational Testing Service.
- Schaeffer, G. A., Steffen, M., Golub-Smith, M. L., Mills, C. N., and Durso, R. (1995). *The Introduction and Comparability of the Computer-Adaptive GRE® General Test* (Research Rep. No. 95-20). Princeton, NJ: Educational Testing Service.
- Singley, M. K., and Bennett, R.E. (2002). Item Generation and Beyond: Applications of Schema Theory to Mathematics Assessment. In S. Irvine and P. Kyllonen (Eds.), *Item Generation for Test Development* (pp. 361–384). Hillsdale, NJ: Erlbaum.
- Stanley, J.C., and Wang, M.D. (1970). Weighting Test Items and Test-Item Options, an Overview of the Analytical and Empirical Literature. *Educational and Psychological Measurement*, *20*: 21–35.
- Taylor, C., Jamieson, J., Eignor, D., and Kirsch, I. (1998). *The Relationship Between Computer Familiarity and Performance on Computer-Based TOEFL® Test Tasks* (Report No. 61). Princeton, NJ: Educational Testing Service.
- U.S. Department of Commerce. (2002). *A Nation Online: How Americans Are Expanding Their Use of the Internet*. Retrieved April 19, 2002, from http://www.ntia.doc.gov/ntiahome/dn/nationonline_020502.htm.
- Weiss, A.R., Lutkus, A.D., Hildebrant, B.S., and Johnson, M.S. (2002). *The Nation’s Report Card: Geography 2001* (NCES 2002–484). Washington, DC: National Center for Education Statistics, Office of Educational Research and Improvement, U.S. Department of Education.
- Wolfe, E.W., Bolton, S., Feltoich, 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.

Appendix A

Inter-Rater Reliability

This appendix presents data on inter-rater reliability for constructed-response items on the 2001 mathematics online test (MOL) and for similar items on the pencil and paper (P&P) test.

Table A-1. Inter-rater reliability for constructed-response items, grade 4: 2001

Item	Percentage exact agreement for market basket form	Percentage exact agreement for MOL
5	99	93
10	97	93
14	99	96
15	99	90
21	100	94
22	88	87
24	98	97
26	98	98
29	99	96
31	98	98

NOTE: MOL=Math Online. The number of students responding ranged from 234 to 265. Item 22 was scored on a 5-point scale. All other items were scored on 2- or 3-point scales.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2000 Mathematics Assessment; 2001 Math Online Study.

Table A-2. Inter-rater reliability for constructed-response items, grade 8: 2001

Item	Percentage exact agreement for paper-based test	Percentage exact agreement for MOL
2	99	98
3	95	92
7	93	91
10	80	84
13	99	98
15	97	98
16	99	98
17	99	98
19	94	90
26	85	85

NOTE: MOL=Math Online. The number of students responding ranged from 239 to 254 on the paper test; from 249 to 253 on MOL Items. Items 10 and 26 were scored on a 5-point scale. All other items were scored on 2- or 3-point scales.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

Appendix B

Ease of Assessing Existing NAEP Framework Content Areas on Computer

This appendix presents results of the a priori analysis to determine which content areas of the NAEP mathematics framework are easily assessed in computer-based testing and which are not. It also presents examples of released NAEP mathematics items not easily rendered on computer.

Table B-1. Percentage of NAEP items, by framework content area and ease of implementation for computer delivery, grade 8: 2001

Framework content area	Percent of items		
	Easy to implement	Moderately difficult to implement	Difficult to implement
Number sense, properties, and operations (43 items)	95	5	#
Measurement (22 items)	64	5	32
Geometry and spatial sense (32 items)	53	9	38
Data analysis, statistics and probability (24 items)	75	21	4
Algebra and functions (39 items)	77	18	5

The estimate rounds to zero.
 NOTE: Detail may not sum to totals because of rounding.
 SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

Table B-2. Percentage of NAEP mathematics items, by format and ease of implementation for computer delivery, grade 8: 2001

Item format	Percent of items		
	Easy to implement	Moderately difficult to implement	Difficult to implement
Standard multiple-choice (100 items)	95	1	4
Constructed-response (60 items)	38	32	30

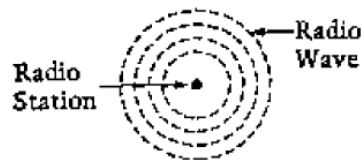
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

Figure B-1. A NAEP item measuring the geometry and spatial sense content area that requires a drawn response, grade 8: 2001

This question requires you to show your work and explain your reasoning. You may use drawings, words, and numbers in your explanation. Your answer should be clear enough so that another person could read it and understand your thinking. It is important that you show all your work.

13. Radio station KMAT in Math City is 200 miles from radio station KGEO in Geometry City. Highway 7, a straight road, connects the two cities.

KMAT broadcasts can be received up to 150 miles in all directions from the station and KGEO broadcasts can be received up to 125 miles in all directions. Radio waves travel from each radio station through the air, as represented below.



On the next page, draw a diagram that shows the following.

- Highway 7
- The location of the two radio stations
- The part of Highway 7 where both radio stations can be received

Be sure to label the distances along the highway and the length in miles of the part of the highway where both stations can be received.

NOTE: This item is shown in an onscreen version taken from the NAEP database of publicly released questions available on the Web (<http://nces.ed.gov/nationsreportcard/itmrls/>).

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

Figure B-2. A NAEP item assessing the measurement content area that requires paper stimulus materials, grade 8: 2001

With this test booklet, you will receive a packet of 6 pieces: 2 each of shape *N*, shape *P*, and shape *Q*. You will use these pieces in answering some of the questions. You can turn the pieces in any way or flip them over. You may use drawings to help explain your answers.

5. Bob, Carmen, and Tyler were comparing the areas of *N* and *P*. Bob said that *N* and *P* have the same area. Carmen said that the area of *N* is larger. Tyler said that the area of *P* is larger.

Who was correct? _____

Use words or pictures (or both) to explain why.

NOTE: This item is shown in an onscreen version taken from the NAEP database of publicly released questions available on the Web (<http://nces.ed.gov/nationsreportcard/itmrls/>).

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

Appendix C

Students Omitting, Not Reaching, and Giving Off-Task Responses

This appendix presents data on the rate at which students omitted, did not reach, or gave off-task responses to constructed-response items on the 2001 mathematics online test (MOL) and to similar items on the paper-based test.

Table C-1. Mean percentages of students omitting, not reaching, and giving off-task responses for the MOL and paper tests, grade 8: 2001

Test section	Mean percent of students			
	Omitting an item	Not reaching items	Giving off-task answer to a dichotomous CR item	Giving off-task answer to polytomous CR item
MOL				
1	1.1	0.8	0.1	0.3
2	1.5	0.4	#	0.1
3	1.5	1.7	–	2.1
Paper and pencil				
1	1.1	0.5	#	0.5
2	1.1	0.2	0.9	0.4
3	0.6	0.7	–	1.4

The estimate rounds to zero.

– Not available. No dichotomous CR items were included in this section.

NOTE: MOL=Math Online. CR=constructed-response. Each figure is the percentage of students omitting, not reaching, or giving an off-task response to an item, as the case may be, averaged over all items.

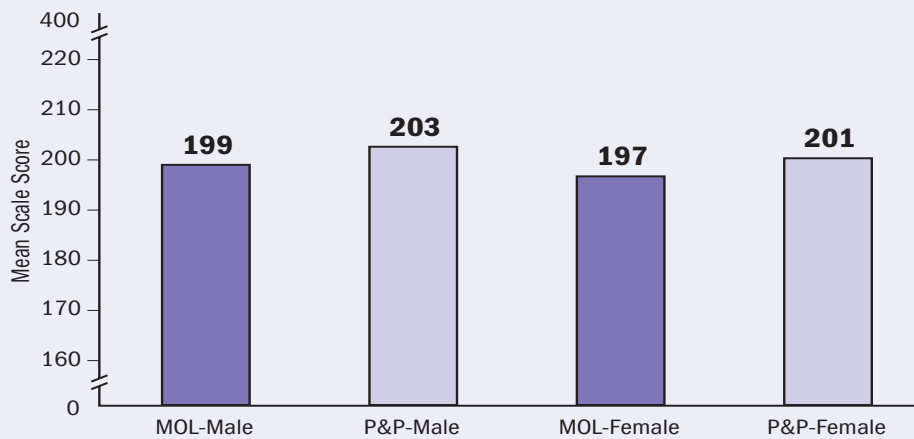
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

Appendix D

Test Mode by Population Group Contrasts

This appendix presents data on the performance of NAEP reporting groups on the 2001 mathematics online test (MOL) and on the paper-based form.

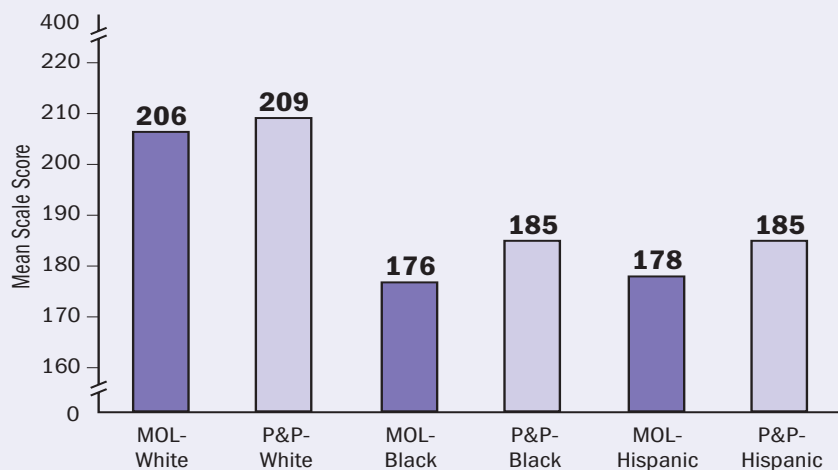
Figure D-1. Mean scale score for MOL and P&P, by gender, grade 8: 2001



NOTE: MOL=Math Online. P&P=Paper and Pencil. Average MOL scores are reported on 0–400 scale.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

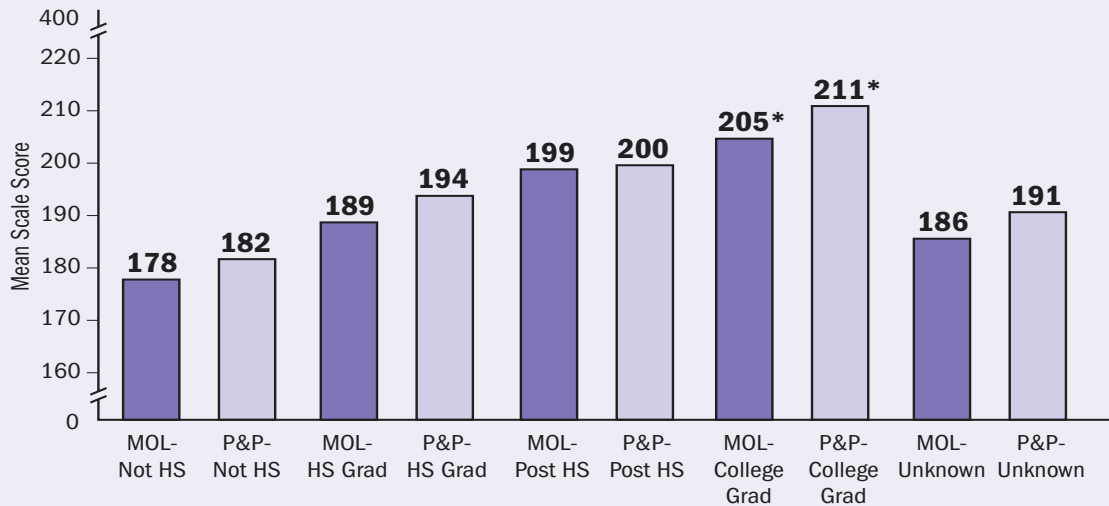
Figure D-2. Mean scale score for MOL and P&P, by race/ethnicity, grade 8: 2001



NOTE: MOL=Math Online. P&P=Paper and Pencil. Sample sizes for other racial/ethnic groups were too small to analyze statistically. Average MOL scores are reported on 0–400 scale.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

Figure D-3. Mean scale score for MOL and P&P, by parents' education level, grade 8: 2001

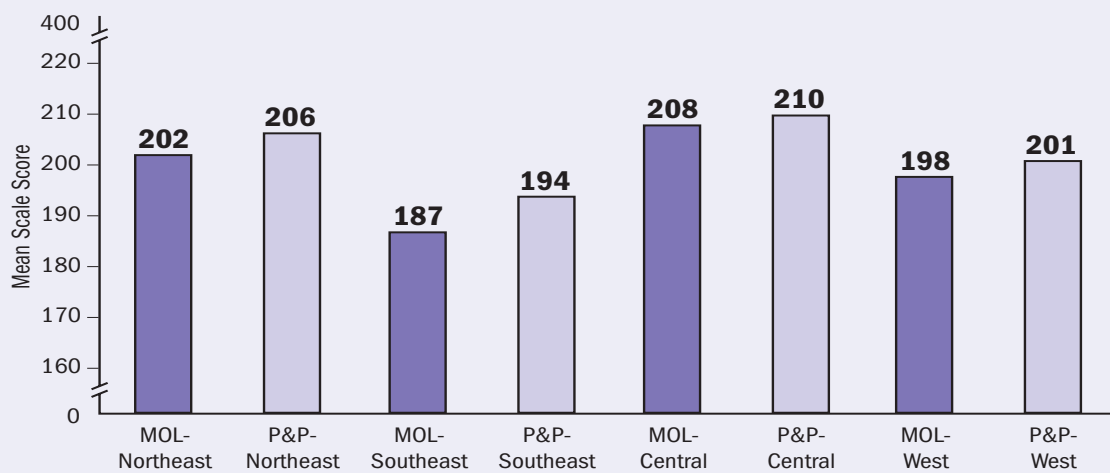


* MOL and P&P values differ significantly, $p < .05$, for students reporting parent graduated from college.

NOTE: MOL=Math Online. P&P=Paper and Pencil. Average MOL scores are reported on 0-400 scale.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

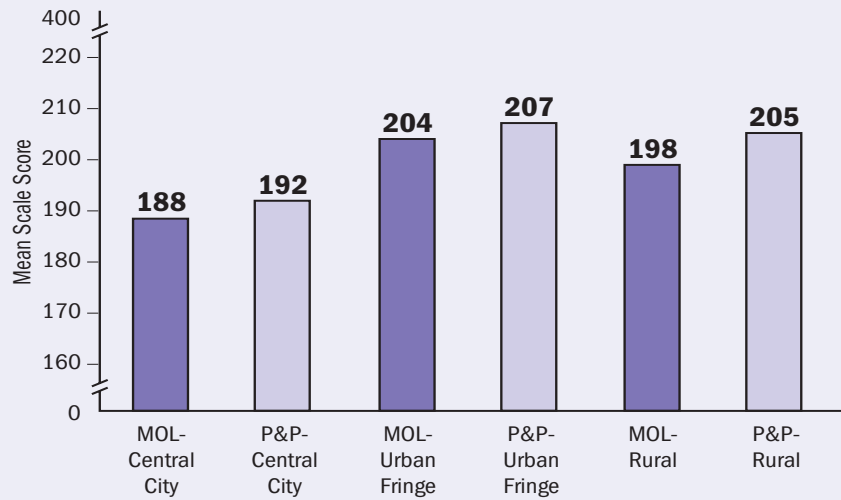
Figure D-4. Mean scale score for MOL and P&P, by region of country, grade 8: 2001



NOTE: MOL=Math Online. P&P=Paper and Pencil. Average MOL scores are reported on 0-400 scale.

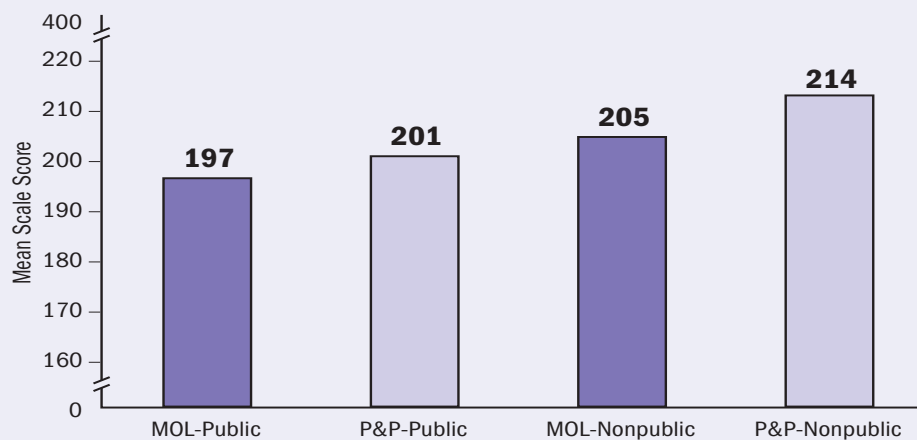
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

Figure D-5. Mean scale score for MOL and P&P, by school location, grade 8: 2001



NOTE: MOL=Math Online. P&P=Paper and Pencil. Average MOL scores are reported on 0-400 scale.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

Figure D-6. Mean scale score for MOL and P&P, by school type, grade 8: 2001



NOTE: MOL=Math Online. P&P=Paper and Pencil. Average MOL scores are reported on 0-400 scale.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

Appendix E

Self-Reported Computer Experience

This appendix presents data on students' responses to questions about their access to and use of computers.

Table E-1. Percentage of students who report computer or Internet use at home, grade 4: 2001

Item	Yes	No
Is there a computer at home that you use?	85 (1.4)	15 (1.4)
Do you use the Internet at home?	69 (1.8)	31 (1.8)

NOTE: The number of students responding ranged from 1,028 to 1,031. The standard errors of the percentages appear in parentheses. Detail may not sum to totals because of rounding.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

Table E-2. Percentage of students who report using a computer in and out of school, by frequency levels, grade 4: 2001

Item	Every day	Two or three times a week	About once a week	Once every few weeks	Never or hardly ever
How often do you use a computer at school? Include use anywhere in the school at any time of the day.	13 (1.4)	28 (2.7)	33 (2.5)	12 (1.2)	14 (2.1)
How often do you use a computer outside of school?	28 (1.5)	24 (1.5)	14 (1.1)	10 (1.1)	24 (1.4)

NOTE: The number of students responding ranged from 1,025 to 1,029. The standard errors of the percentages appear in parentheses. Detail may not sum to totals because of rounding.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

Table E-3. Percentage of students who report using a computer for various purposes, grade 4: 2001

Do you ever use a computer to do any of the following?	Yes	No
Play computer games	88 (1.0)	12 (1.0)
Write reports, letters, stories, or anything else on the computer	75 (1.5)	25 (1.5)
Make pictures or drawings on the computer	75 (1.6)	25 (1.6)
Make tables, charts, or graphs on the computer	37 (2.5)	63 (2.5)
Look up information on a CD	50 (1.8)	50 (1.8)
Look up information on the Internet	80 (1.4)	20 (1.4)
Send e-mail or talk in chat groups	47 (2.2)	53 (2.2)

NOTE: The number of students responding ranged from 1,017 to 1,032. The standard errors of the percentages appear in parentheses. Detail may not sum to totals because of rounding.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

Table E-4. Percentage of students who report using a computer for mathematics, by frequency level, grade 4: 2001

When you do mathematics in school, how often do you do each of the following?	Almost every day	Once or twice a week	Once or twice a month	Never or hardly ever
Use a computer	37 (1.9)	38 (2.0)	8 (0.8)	18 (1.4)

NOTE: The number of students responding was 1,023. The standard errors of the percentages appear in parentheses. Detail may not sum to totals because of rounding.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

Table E-5. Percentage of students agreeing with a positive statement about computer use, grade 4: 2001

Which of the following statements about using a computer are true for you?

	True	False	I never use a computer
I like doing homework more when I use a computer.	42 (1.7)	49 (2.0)	9 (1.3)
I have more fun learning when I use the computer.	77 (1.7)	21 (1.9)	3 (0.6)
I get more done when I use a computer for schoolwork.	50 (2.0)	44 (1.9)	6 (0.6)

NOTE: Detail may not sum to totals because of rounding. The number of students responding ranged from 1,026 to 1,032. The standard errors of the percentages appear in parentheses.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

Table E-6. Percentage of students who report computer or Internet use at home, grade 8: 2001

Item	Yes	No
Is there a computer at home that you use?	88 (0.7)	12 (0.7)
Do you use the Internet at home?	79 (1.1)	21 (1.1)

NOTE: The number of students responding ranged from 3,419 to 3,403. The standard errors of the percentages appear in parentheses. Detail may not sum to totals because of rounding.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

Table E-7. Percentage of students who report using a computer in and out of school, by frequency levels, grade 8: 2001

Item	Every day	Two or three times a week	About once a week	Once every few weeks	Never or hardly ever
How often do you use a computer at school? Include use anywhere in the school and at any time of the day.	16 (1.3)	21 (1.5)	18 (1.5)	24 (1.7)	20 (1.3)
How often do you use a computer outside of school?	52 (1.4)	24 (0.7)	7 (0.4)	8 (0.5)	9 (0.7)

NOTE: The number of students responding ranged from 3,777 to 3,779. The standard errors of the percentages appear in parentheses. Detail may not sum to totals because of rounding.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

Table E-8. Percentage of students who report using a computer for various purposes, grade 8: 2001

Item	Not at all	Small extent	Moderate extent	Large extent
Play computer games	10 (0.5)	42 (0.9)	33 (0.8)	15 (0.7)
Write using a word processing program	13 (0.9)	30 (1.0)	35 (1.1)	22 (1.0)
Make drawings or art projects on the computer	29 (1.3)	43 (0.9)	19 (0.7)	10 (0.6)
Make tables, charts, or graphs on the computer	41 (1.2)	39 (1.1)	15 (0.8)	5 (0.4)
Look up information on a CD	19 (1.0)	32 (1.0)	30 (1.0)	19 (0.7)
Find information on the Internet for a school project or report	6 (0.5)	16 (0.8)	34 (0.7)	44 (1.2)
Find information on the Internet for personal use	11 (0.8)	21 (0.7)	26 (1.0)	41 (1.1)
Use e-mail to communicate with others	19 (1.2)	17 (0.8)	20 (0.6)	44 (1.3)
Talk in chat groups with other people who are logged on at the same time you are	24 (1.1)	20 (0.8)	19 (0.8)	37 (1.2)

NOTE: The number of students responding ranged from 3,765 to 3,775. The standard errors of the percentages appear in parentheses. Detail may not sum to totals because of rounding.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

Table E-9. Percentage of students who report using a computer for mathematics, by frequency level, grade 8: 2001

When you do mathematics in school, how often do you do each of the following?

	Almost every day	Once or twice a week	Once or twice a month	Never or hardly ever
Use a computer	26 (1.2)	16 (1.0)	13 (0.8)	46 (1.5)

NOTE: The number of students responding was 3,739. The standard errors of the percentages appear in parentheses. Detail may not sum to totals because of rounding.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

Table E-10. Percentage of students agreeing with a positive statement about computer use, grade 8: 2001

Please indicate the extent to which you AGREE or DISAGREE with the following statements.

Item	Strongly agree	Agree	Disagree	Strongly disagree	I never use a computer
I am more motivated to get started doing my schoolwork when I use a computer.	17 (0.7)	47 (0.8)	25 (0.7)	6 (0.5)	5 (0.4)
I have more fun learning when I use the computer.	33 (1.1)	45 (0.9)	16 (0.8)	4 (0.3)	3 (0.3)
I get more done when I use a computer for schoolwork.	29 (0.9)	40 (0.9)	22 (0.7)	5 (0.4)	4 (0.4)

NOTE: The number of students responding ranged from 3,762 to 3,766. The standard errors of the percentages appear in parentheses. Detail may not sum to totals because of rounding.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

Appendix F

Student Mathematics Performance on Computer-Based Test and Paper-and-Pencil Test by Self-Reported Computer Experience

This appendix compares student performance on the 2001 mathematics online test (MOL) and the paper-based test for groups of students reporting different levels of computer access or use.

Table F-1. Mean scale scores and standard errors, by frequency of general computer use in and out of school, grade 8: 2001

Item	Test mode	Every day	Two or three times a week	About once a week	Once every few weeks	Never or hardly ever
How often do you use a computer at school?	MOL	199 (2.8)	199 (3.3)	200 (3.2)	202 (2.7)	190 (2.5)
	P&P	204 (3.2)	200 (2.2)	203 (1.7)	207 (2.3)	197 (1.9)
How often do you use a computer outside of school?	MOL	205 (2.0)	198 (2.5)	193 (4.3)	186 (2.9)	172 (3.6)
	P&P	208 (1.7)	201 (2.5)	203 (3.8)	190 (3.2)	179 (2.8)

NOTE: MOL=Math Online. P&P=Paper and Pencil. The standard errors of the mean scale scores appear in parentheses.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

Table F-2. Mean scale scores and standard errors, by technology in the home, grade 8: 2001

Item	Test mode	Yes	No
Is there a computer at home that you use?	MOL	201 (1.3) *	174 (4.0)
	P&P	205 (1.4) *	183 (3.4)
Do you use the Internet at home?	MOL	203 (1.2)	179 (3.4)
	P&P	206 (1.3)	189 (3.1)

* Values differ significantly for the contrast between MOL and P&P, $p < .05$.
NOTE: MOL=Math Online. P&P=Paper and Pencil. The standard errors of the mean scale scores appear in parentheses.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

Table F-3. Mean scale scores and standard errors, by frequency of specific computer use, grade 8: 2001

Item	Test mode	Not at all	Small extent	Moderate extent	Large extent
Play computer games	MOL	187 (4.5)	197 (2.0)	203 (2.2)	199 (2.9)
	P&P	192 (3.0)	202 (1.5)	205 (2.0)	203 (2.9)
Write using a word processing program	MOL	177 (3.5) *	190 (1.9) *	204 (1.7)	208 (2.5)
	P&P	188 (2.5) *	198 (1.9) *	205 (1.8)	213 (2.1)
Make drawings or art projects on the computer	MOL	199 (2.3)	200 (2.1)	198 (3.4)	186 (4.7)
	P&P	203 (2.0)	205 (1.8)	198 (2.3)	194 (4.0)
Make tables, charts, or graphs on the computer	MOL	192 (2.3)	204 (1.7)	201 (3.3)	‡
	P&P	198 (1.4)	206 (1.8)	204 (3.1)	‡
Look up information on a CD	MOL	192 (2.4)	201 (2.1)	202 (2.2)	193 (2.5)
	P&P	199 (2.8)	204 (1.7)	206 (1.7)	197 (2.7)
Find information on the Internet for a school project or report	MOL	‡	190 (3.3)	201 (2.0)	201 (1.8) *
	P&P	188 (3.7)	196 (2.3)	203 (1.8)	207 (2.0) *
Find information on the Internet for personal use	MOL	180 (3.4) *	196 (3.1)	202 (2.9)	202 (1.8)
	P&P	193 (2.8) *	200 (2.5)	204 (1.9)	205 (1.8)
Use e-mail to communicate with others	MOL	186 (2.9)	194 (4.1)	204 (2.2)	202 (1.7)
	P&P	191 (2.2)	203 (2.8)	210 (2.2)	203 (1.6)
Talk in chat groups or with other people who are logged on at the same time you are	MOL	193 (2.9)	196 (2.9)	201 (2.7)	201 (2.0)
	P&P	197 (1.9)	202 (2.7)	205 (2.3)	204 (1.7)

‡ Reporting standards not met. Sample size is insufficient to permit a reliable estimate.

*Values differ significantly for the contrast between MOL and P&P, $p < .05$.

NOTE: MOL=Math Online. P&P=Paper and Pencil. The standard errors of the mean scale scores appear in parentheses.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

THIS PAGE INTENTIONALLY LEFT BLANK.

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

Taslina 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.

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 Kolsat, Taslima Rahman, and Holly Spurlock.

Contents

Executive Summary	iii
Foreword	v
Acknowledgments	vi
1. Introduction	1
2. Methodology	2
Study Samples	2
Instruments	9
Procedures	13
Essay Scoring	13
Practice Effect	15
3. Measurement Issues	16
Performance Differences Across Assessment Modes	16
Essay Score	16
Essay Length	17
Frequency of Valid Responses	17
4. Equity Issues	18
Population Group Performance	18
Gender	18
Other NAEP Reporting Groups	20
Race/ethnicity	20
Parents' education level	21
School location	22
Eligibility for free/reduced-price school lunch	23
School type	23
Performance as a Function of Computer Type	24
Performance as a Function of Computer Experience	27
5. Efficiency Issues	32
Relative Timeliness and Costs of Computer- vs. Paper-Based Assessment	32
Relative Timeliness of Computer vs. Paper Testing	32
Relative Costs of Computer vs. Paper Testing	35
Automated Scoring: E-rater®	37
6. Operational Issues	45
Recruiting Schools	45
Training Field Administrators	45
Preparing for the Administrations	45
Conducting the Administrations	47
Accommodations for Students With Disabilities: WOL Voicing	48
Equipment Performance	50
Student and School Staff Reactions	52
Data Quality	52
7. Summary and Conclusions	53
8. Implications for NAEP	55
References	57
Appendix A. Sample Selection	59
Appendix B. Understanding NAEP Reporting Groups	61
Appendix C. Writing Online Hands-On Editing Tasks	63
Appendix D. Writing Online Speed and Accuracy Tasks	69
Appendix E. Background Questions Administered in Writing Online	71
Appendix F. NAEP Grade 8 Writing Scoring Guides	74
Appendix G. Statistical Procedures	76
Appendix H. Percentage of Writing Online Students Who Report Using a Computer for Different Specific Writing Purposes	77
Appendix I. Summary Statistics for Computer Familiarity Measures	78
Appendix J. Analysis of Variance Results Relating Computer Familiarity and Gender to Writing Online Performance	79

Tables

Table 2-1.	Reasons for student nonparticipation in Writing Online, grade 8: 2002	2
Table 2-2.	Numbers of students in study samples before and after excluding those who did not respond to both essays, grade 8: 2002	3
Table 2-3.	Characteristics of study sample taking the main NAEP paper-and-pencil writing assessment compared with all students taking the main NAEP writing, grade 8: 2002	4
Table 2-4.	Characteristics of study samples taking the Writing Online test compared with all students taking the main NAEP, grade 8: 2002	6
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	8
Table 2-6.	Instruments administered to each student sample, grade 8: 2002	13
Table 2-7.	Intraclass correlations between two readers for Writing Online and for the main NAEP writing, grade 8: 2002	14
Table 2-8.	Percentage exact agreement between two readers for Writing Online and for the main NAEP writing, grade 8: 2002	14
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	15
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	15
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	16
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	17
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	17
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	18
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	19
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	19
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	20
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	21
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	22
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	23
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	23

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	24
Table 4-10.	Mean scores, by computer type, for Writing Online students drawn from the main NAEP writing sample, grade 8: 2002	25
Table 4-11.	Mean scores, by gender and computer type, for Writing Online students drawn from the main NAEP writing sample, grade 8: 2002	26
Table 4-12.	Components of the hands-on computer skills measure, grade 8: 2002	29
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	30
Table 5-1.	Unweighted means and standard deviations for essay scores, by human readers and e-rater®, grade 8: 2002	39
Table 5-2.	Unweighted intraclass correlations for essay scores, by human readers and e-rater®, grade 8: 2002	40
Table 5-3.	Unweighted percentage exact agreement between e-rater® and human readers and between two human readers, grade 8: 2002	40
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	41
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	41
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	42
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	43
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	43
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	44
Table 6-1.	Percentage distribution of students and schools, by computer configuration, used to deliver the Writing Online test, grade 8: 2002	47
Table 6-2.	Unweighted means for students with disabilities taking the voicing version of Writing Online, by essay and demographic group, grade 8: 2002	49
Table 6-3.	Percentage distribution of calls reported to the Westat help desk, by reason for call, grade 8: 2002	50
Table 6-4.	Percentage distribution of technical problems reported by the Westat administrators, grade 8: 2002	51
Table H-1.	Percentage of Writing Online students who report using a computer for different specific writing purposes, grade 8: 2002	77
Table I-1.	Summary statistics for components of the hands-on computer skills measure, grade 8: 2002	78
Table I-2.	Summary statistics for computer familiarity measures, grade 8: 2002	78
Table J-1.	Results of repeated-measures analysis of variance testing the effects of gender and of self-reported and hands-on computer familiarity variables on Writing Online performance, controlling for main NAEP writing performance, grade 8: 2002	79

Figures

Figure 2-1. The Writing Online computer interface showing the “Save a Book” essay, grade 8: 2002.	10
Figure 2-2. The Writing Online computer interface showing the “School Schedule” essay, grade 8: 2002	11
Figure 2-3. Sample Writing Online background question screen, grade 8: 2002	12
Figure 4-1. Self-reported computer-familiarity questions administered to students taking Writing Online, grade 8: 2002	28
Figure 5-1. Key steps in NAEP paper vs. computer writing test delivery, with estimated elapsed times	33
Figure 5-2. Relative costs for NAEP of computer vs. paper writing assessment	37
Figure 5-3. Writing features extracted by e-rater®, grouped by logical dimensions.	38
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	42
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	44
Figure 6-1. Technical specifications for school computers used to deliver the Writing Online test, grade 8: 2002	46
Figure 6-2. A sample Writing Online voicing screen, grade 8: 2002.	48
Figure C-1. Writing Online hands-on editing tasks, screen 1, grade 8: 2002	63
Figure C-2. Writing Online hands-on editing tasks, screen 2, grade 8: 2002	64
Figure C-3. Writing Online hands-on editing tasks, screen 3, grade 8: 2002	65
Figure C-4. Writing Online hands-on editing tasks, screen 4, grade 8: 2002	66
Figure C-5. Writing Online hands-on editing tasks, screen 5, grade 8: 2002	67
Figure C-6. Writing Online hands-on editing tasks, screen 6, grade 8: 2002	68
Figure D-1. Writing Online speed and accuracy tasks, screen 1, grade 8: 2002.	69
Figure D-2. Writing Online speed and accuracy tasks, screen 2, grade 8: 2002.	70

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

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.

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

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 rate¹	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 rate¹	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:

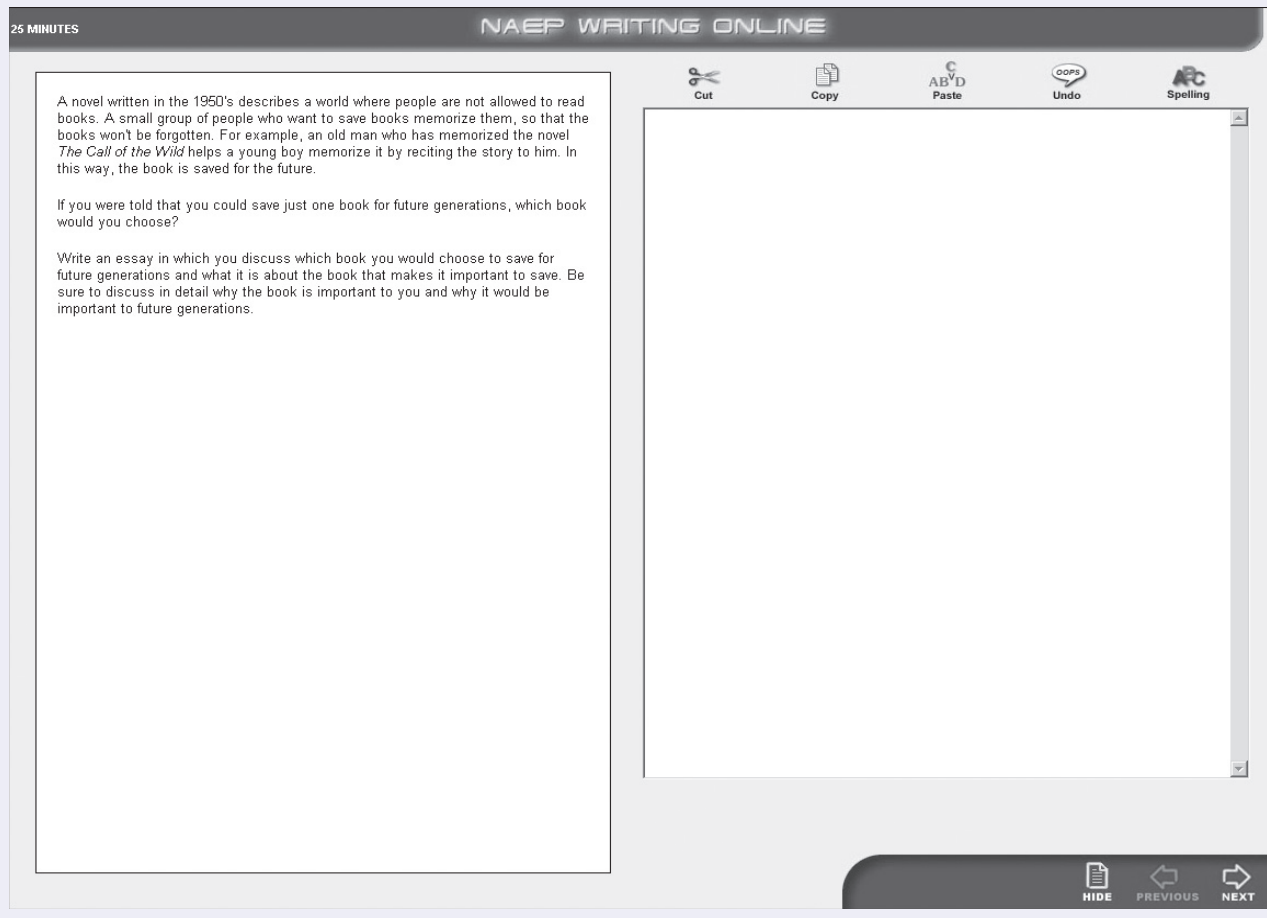
- **Online tutorial.** The online tutorial showed students how to use the computer to respond to the essay tasks. The tutorial provided instruction and practice in the use of the mouse and scrolling, presented information about the test interface and how to navigate from one question to the next, and 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.

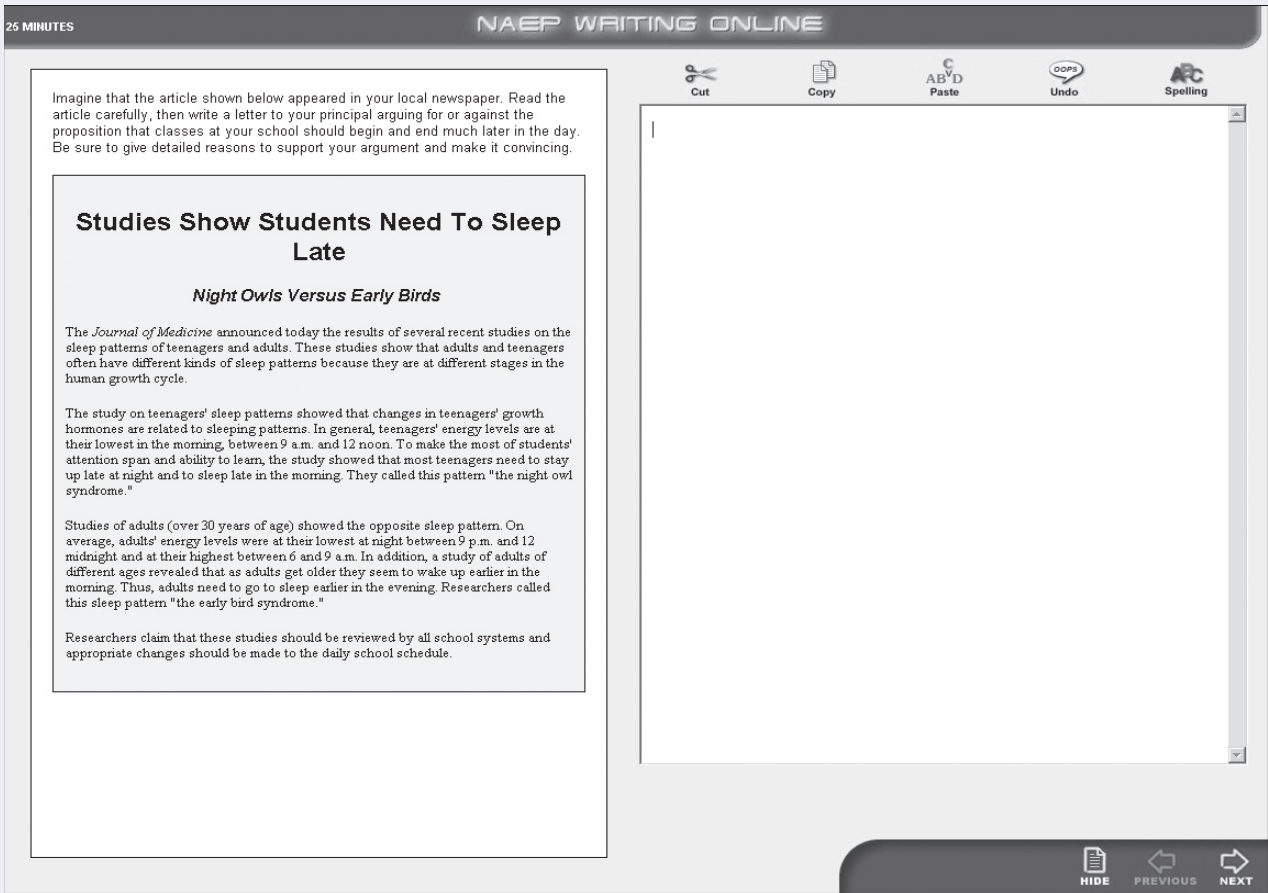
Both WOL essays were drawn from the 2002 main NAEP writing assessment and administered to students in the same order as in that assessment. For "Save a Book," an informative writing task, students were asked to explain what book they would preserve through memorization if they lived in a society where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

20 MINUTES

NAEP WRITING ONLINE

QUESTION 1 of 37

Questions 1-8. To what extent do you do the following on a computer? Include things you do in school and things you do outside of school.

Play computer games

- Not at all
- Small extent
- Moderate extent
- Large extent

PREVIOUS NEXT

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

Measure	Save a Book		School Schedule	
	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, Feltoovich, 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

Gender	Mean word count	
	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 stu-

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

dents 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

Gender	Percent of students	
	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 ($F_{1,62} = 1.97$,

$p > .05$), 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 unrepresentative, 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

Essay	Male		Female	
	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, Feltoovich, and Bangert (1996) and Wolfe, Bolton, Feltoovich, 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 on-line 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 op-

erational 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 read-

ers 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

Variable pair	Save a Book		School Schedule	
	Percent exact agreement	Kappa	Percent exact agreement	Kappa
	First reader with second reader	61	.48	62
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.

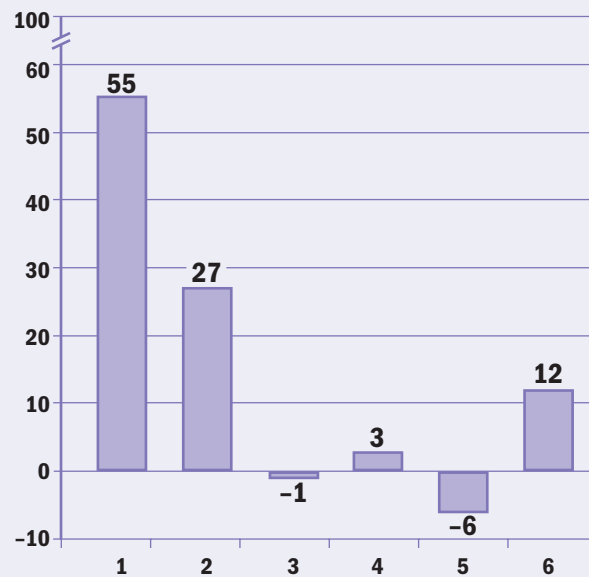
Table 5-6. Unweighted score distributions and percentage exact agreement between two human readers at each of six score levels for “Save a Book,” grade 8: 2002

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 read-

ers 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.

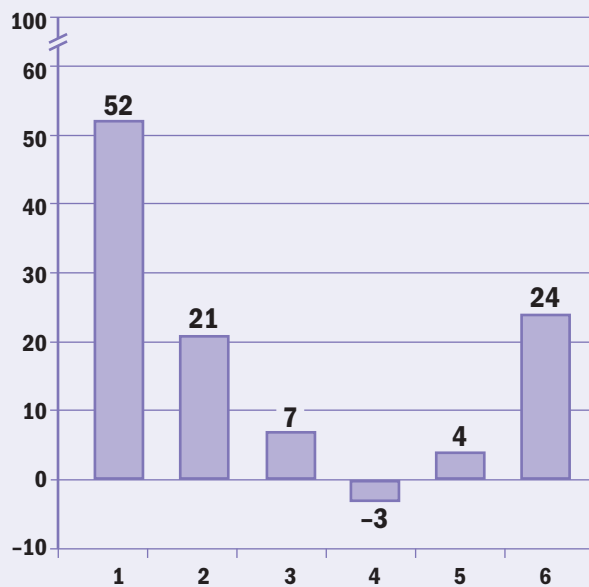
Table 5-9. Unweighted score distributions and percentage exact agreement between two human readers at each of six score levels for “School Schedule,” grade 8: 2002

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.

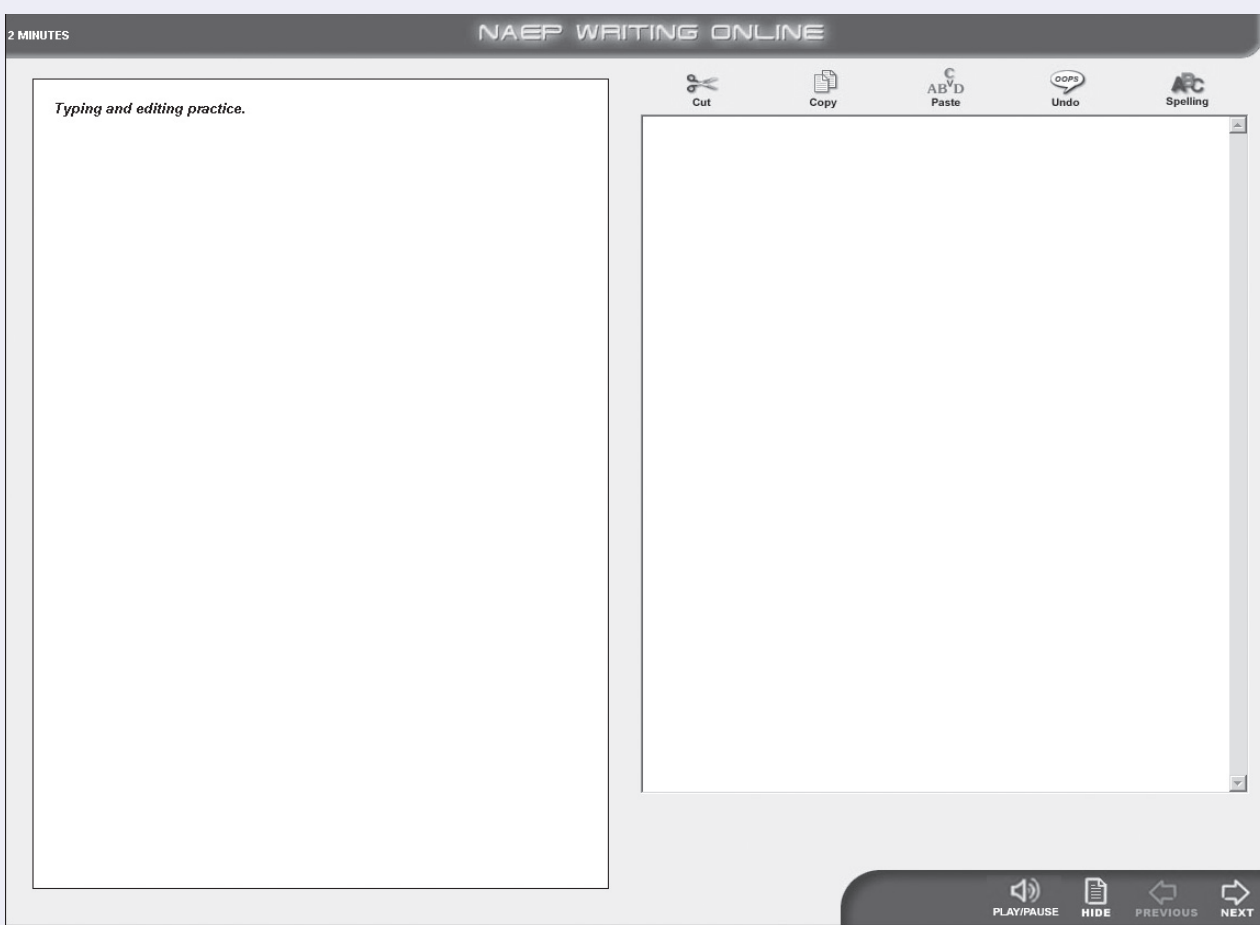
Accommodations for Students With Disabilities: WOL Voicing

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 significantly 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). CriterionSM 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>.

Appendix A. Sample Selection

The WOL study design called for a nationally representative sample of 1,400 eighth-graders to take the computer test. These students were selected from among those taking certain booklets administered as part of the main NAEP 2002 writing or reading assessments. The selection procedures for WOL involved multi-stage, multi-phase sampling of schools and students.

Sample Selection for Main NAEP 2002 Assessment

The grade 8 main NAEP 2002 assessment tested public and private school students. Samples were selected based on a two-stage design: (1) selection of schools and (2) selection of students within schools. The first-stage sample of schools was selected with probability proportional to a measure of size based on estimated enrollment at grade 8. Each participating school provided a list of eighth-graders from which a systematic sample of students was drawn. Depending on the school's size, one or more sessions of 60 students were sampled. Half of the selected students were assigned a reading assessment booklet and the remainder were assigned a writing booklet.

The public and private school sample designs differed with respect to sample size requirements and stratification. For public schools, representative samples were drawn within each state and the District of Columbia, as well as from separate lists of Bureau of Indian Affairs (BIA) schools and Department of Defense Domestic Dependent Elementary and Secondary Schools (DDESS). Each sample was designed to produce aggregate estimates with approximately equal precision. The target sample in each state was 6,300 grade 8 students. With a general target of 60 sampled students per school, roughly 100 participating schools were needed per state. Special procedures to reduce overall burden were used for states with many small schools, and for states having small numbers of grade-eligible schools.

Prior to sample selection, public schools were hierarchically stratified by district status,¹ urbanization, and race/ethnicity. Within the race/ethnicity strata, schools were sorted by state achievement data for states where it was available. Where state achievement data were not available, schools were sorted by median household income of the zip code area where the school was located. Achievement data were supplied by the states themselves. Median income data were obtained from the 1990 Census. Other stratification

variables were obtained from the National Center for Education Statistics' Common Core of Data (CCD).

For private schools, target student sample sizes were set for four separate reporting groups: Roman Catholic (6,000 students), Lutheran (1,500 students), Conservative Christian (1,500 students), and Other Private (3,000 students). Within these reporting groups, the private schools were implicitly stratified by census division, urbanization, and percent Black/Hispanic/American Indian. Implicit strata were collapsed extensively to ensure that the expected number of schools within each implicit stratum was reasonably large.²

Participation in state NAEP was not mandatory in 2002. Since the aggregate of the individual state samples was planned to be used as the public school sample for the national study, some provision needed to be made to ensure representation from a state even if that state declined to participate in state NAEP. Subsamples of schools were drawn from the state samples to use for the national sample under these circumstances. These subsamples were drawn for each and every state to cover all contingencies. As such, they provided a suitable starting point for selecting the public school portion of the WOL sample.


The process for drawing a national subsample for use in NAEP involved computing appropriate school probabilities of selection using a national target sample size assigned proportionally to each jurisdiction (as if no state NAEP samples had been drawn) and then dividing these probabilities by the full-sample or private-school NAEP probabilities to obtain conditional probabilities of selection for subsampling. School samples were drawn using the conditional probabilities. The resultant unconditional probabilities of selection for the subsample of schools are equal to the appropriate values for a stand-alone national sample. The target sample size for the main NAEP 2000 assessments was 35,500 assessed students at grade 8.

Sample Selection for the Writing Online (WOL) Study

The target student sample size for WOL was 1,400 eighth-graders. Even though considerably fewer than 60 students were selected from each school for the WOL study, further school subsampling was required. To increase operational efficiency, nationally subsampled schools were grouped into 167 geographic clusters, each containing at least 5 eligible sampled

¹ Districts with more than 20 percent of their state's students were in a separate stratum.

² In *explicit* stratification, the population is divided into strata and a separate sample is chosen from each stratum. In *implicit* stratification, the population is first sorted by a chosen characteristic. Next, the sample is selected from this sorted list using a random starting point and a fixed sampling interval.



schools. (A cluster could be an individual county if it met the minimum size requirement, or two or more adjacent counties.) From the 626 counties with at least one eligible eighth-grade school, 167 geographic clusters were defined and 48 were selected with probability proportional to the number of eligible schools. One of the 48 was selected with certainty because of its large size. In each of the remaining 47 sampled clusters, 5 schools were selected with equal probability. In the one certainty cluster, schools were also subsampled with equal probability, at a rate equal to the product of the cluster probability and within-cluster probability for noncertainty clusters.

The WOL study design targeted students who had been assessed in NAEP using any one of 10 specific writing assessment booklets or 9 specific reading booklets, which together comprise slightly less than 23 percent of NAEP-assessed students. Since the booklets are assigned randomly, the set of students assessed using these booklets constitutes a valid random sample of students capable of taking the NAEP assessment. In most schools, all such students were recruited to participate in WOL. Usually, this produced a caseload of about 10 students per school. In a very small number of schools where the sample size was larger than was operationally practical, targeted students were subsampled with equal probability.

Appendix B. Understanding NAEP Reporting Groups

NAEP results are provided for groups of students defined by shared characteristics—gender, race/ethnicity, parental education, region of the country, type of school, school’s type of location, and eligibility for free/reduced-price school lunch. Based on participation rate criteria, results are reported for subpopulations only when sufficient numbers of students and adequate school representation are present. The minimum requirement is at least 62 students in a particular subgroup from at least five primary sampling units (PSUs).¹ However, the data for all students, regardless of whether their subgroup was reported separately, were included in computing overall results. Definitions of the subpopulations are presented below.

Gender

Results are reported separately for male students and female students.

Race/Ethnicity

In all NAEP assessments, data about student race/ethnicity are collected from two sources: school records and student self-reports. Prior to 2002, NAEP used students’ self-reported race as the primary race/ethnicity reporting variable. As of 2002, the race/ethnicity variable presented in NAEP reports is based on the race reported by the school. When school-recorded information is missing, student-reported data are used to determine race/ethnicity. The mutually exclusive racial/ethnic categories are White, Black, Hispanic, Asian/Pacific Islander, American Indian (including Alaska Native), and Other. Information based on student self-reported race/ethnicity is available on the NAEP Data Tool (<http://nces.ed.gov/nationsreportcard/naepdata/>).

Parental Education

Eighth-graders were asked the following two questions, the responses to which were combined to derive the parental education variable.

How far in school did your mother go?

- She did not finish high school.
- She graduated from high school.
- She had some education after high school.
- She graduated from college.
- I don’t know.

How far in school did your father go?

- He did not finish high school.
- He graduated from high school.
- He had some education after high school.
- He graduated from college.
- I don’t know.

The information was combined into one parental education reporting variable in the following way: If a student indicated the extent of education for only one parent, that level was included in the data. If a student indicated the extent of education for both parents, the higher of the two levels was included in the data. If a student responded “I don’t know” for both parents, or responded “I don’t know” for one parent and did not respond for the other, the parental education level was classified as “I don’t know.” If the student did not respond for either parent, the student was recorded as having provided no response.

Region of the Country

Results by region were not included in the main NAEP 2002 writing assessment (except for the Southeast) because response adjustments for non-participating states cut across region. As a consequence, region was also not included among the examined population groups for the WOL study.

Type of School

Results are reported by the type of school that the student attends—public or nonpublic. Nonpublic schools include Catholic and other private schools.² Because they are funded by federal authorities (not state/local governments), Bureau of Indian Affairs (BIA) schools and Department of Defense Domestic Dependent Elementary and Secondary Schools (DDESS) are not included in either the public or nonpublic categories; they are included in the overall national results.

Type of Location

Results from the 2003 assessment are reported for students attending schools in three mutually exclusive location types: central city, urban fringe/large town, and rural/small town.

¹ For the NAEP national assessments prior to 2002, a PSU is a selected geographic region (a county, group of counties, or metropolitan statistical area). Since 2002, the first-stage sampling units are schools (public and nonpublic) in the selection of the combined sample. Further details about the procedure for determining minimum sample size will appear in the technical documentation section of the NAEP website (<http://nces.ed.gov/nationsreportcard/>).

² A more detailed breakdown of nonpublic school results is available on the NAEP website (<http://nces.ed.gov/nationsreportcard/naepdata/>).

Central city: Following standard definitions established by the Federal Office of Management and Budget, the U.S. Census Bureau (see <http://www.census.gov/>) defines “central city” as the largest city of a Metropolitan Statistical Area (MSA) or a Consolidated Metropolitan Statistical Area (CMSA). Typically, an MSA contains a city with a population of at least 50,000 and includes its adjacent areas. An MSA becomes a CMSA if it meets the requirements to qualify as a metropolitan statistical area, has a population of 1,000,000 or more, its component parts are recognized as primary metropolitan statistical areas, and local opinion favors the designation. In the NCES Common Core of Data (CCD) locale codes are assigned to schools. For the definition of central city used in this report, two locale codes of the survey are combined. The definition of each school’s type of location is determined by the size of the place where the school is located and whether or not it is in an MSA or CMSA. School locale codes are assigned by the U.S. Census Bureau. For the definition of central city, NAEP reporting uses data from two CCD locale codes: large city (a central city of an MSA or CMSA with the city having a population greater than or equal to 25,000) and midsize city (a central city of an MSA or CMSA having a population less than 25,000). Central city is a geographical term and is not synonymous with “inner city.”

Urban fringe/large town: The urban fringe category includes any incorporated place, census designated place, or nonplace territory within a CMSA or MSA of a large or mid-sized city and defined as urban by the U.S. Census Bureau, but which does not qualify as a central city. A large town is defined as a place outside a CMSA or MSA with a population greater than or equal to 25,000.

Rural/small town: Rural includes all places and areas with populations of less than 2,500 that are classified as rural by the U.S. Census Bureau. A small town is defined as a place outside a CMSA or MSA with a population of less than 25,000, but greater than or equal to 2,500. Results for each type of location are only compared across years 2000 and after. This is due to new methods used by NCES to identify the type of location assigned to each school in the CCD. The new methods were put into place by NCES in order to improve the quality of the assignments, and they take into account more information about the exact physical location of the school. The variable was revised in NAEP beginning with the 2000 assessments.

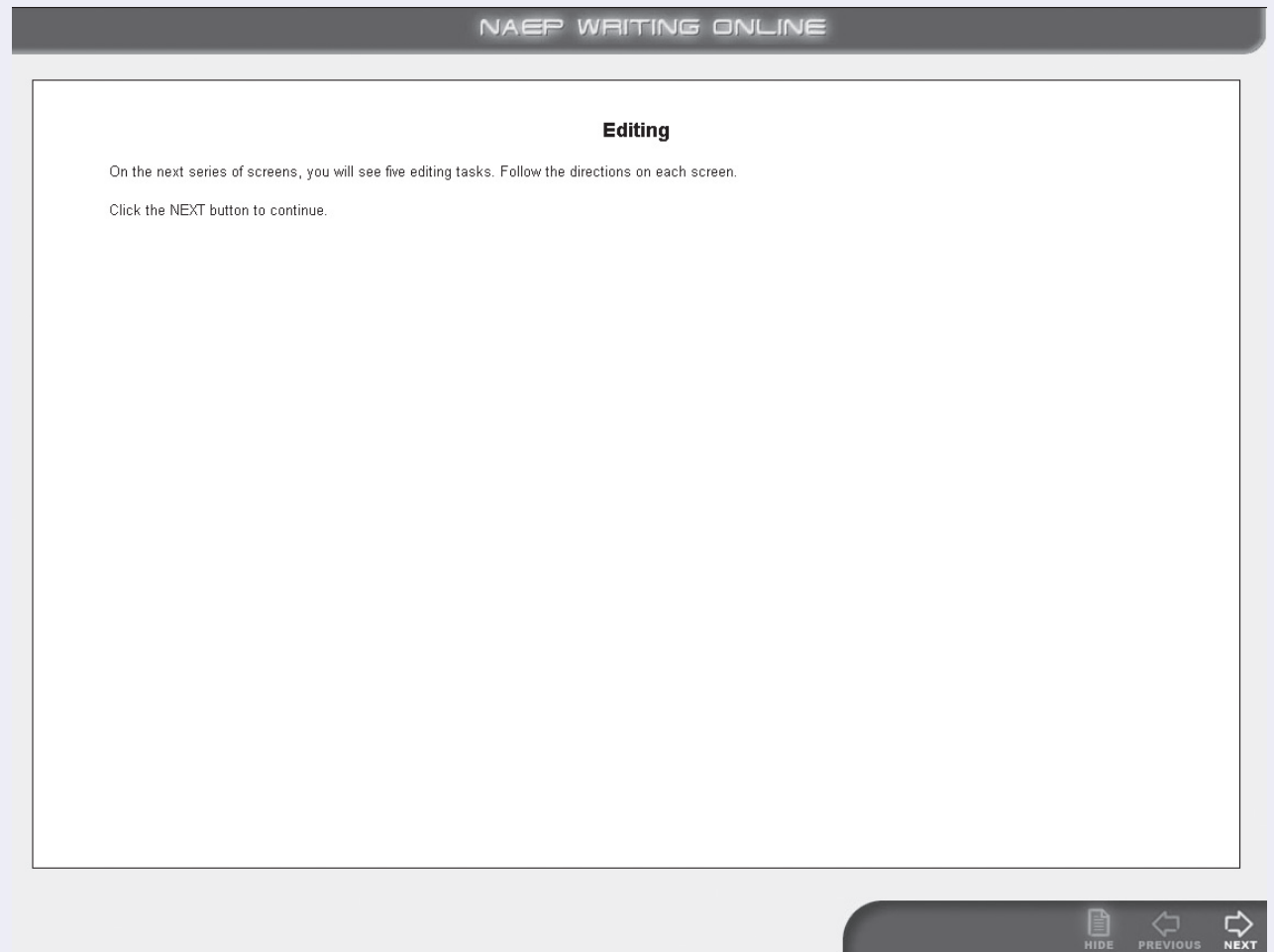
Eligibility for Free/Reduced-Price School Lunch

As part of the Department of Agriculture’s National School Lunch Program, schools can receive cash subsidies and donated commodities in turn for offering free or reduced-price lunches to eligible children. Based on available school records, students were classified as either currently eligible for free/reduced-price school lunch or not eligible. Eligibility for the program is determined by students’ family income in relation to the federally established poverty level. Free lunch qualification is set at 130 percent of the poverty level, and reduced-price lunch qualification is set between 130 and 185 percent of the poverty level. Additional information on eligibility may be found at the Department of Agriculture website (<http://www.fns.usda.gov/cnd/lunch/>). The classification applies only to the school year when the assessment was administered (i.e., the 2002–2003 school year) and is not based on eligibility in previous years. If school records were not available, the student’s information was recorded as “Unavailable.” If the school did not participate in the program, all students in that school were classified as “Unavailable.”

Appendix C. Writing Online Hands-On Editing Tasks

This appendix presents screen shots of the tasks used to measure students' online editing skills.

Figure C-1. Writing Online hands-on editing tasks, screen 1, 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 C-2. Writing Online hands-on editing tasks, screen 2, grade 8: 2002

MINUTE **NAEP WRITING ONLINE**

Editing

Next, make the following change to the bolded words and phrases in the paragraph on the right according to the directions below. You will have 1 minute.

- Delete the word "major".

When planning a vacation, people should consider going to a **major** city, like New York. Cities like New York have something for everyone. For those who prefer crowds, museums, and shopping, a trip into the center of town can be entertaining. For those who prefer something greener, a trip to Central Park is the thing to do. On a sunny day, people enjoy picnic lunches and long walks along the many trails through the Park. Central Park is very large, and features wide lawns, gardens, and a small zoo.

HIDE **PREVIOUS** **NEXT**

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 C-3. Writing Online hands-on editing tasks, screen 3, grade 8: 2002

MINUTE **NAEP WRITING ONLINE**

Editing

Next, make the following change to the bolded words and phrases in the paragraph on the right according to the directions below. You will have 1 minute.

- Insert the word "very" before "entertaining".

When planning a vacation, people should consider going to a major city, like New York. Cities like New York have something for everyone. For those who prefer crowds, museums, and shopping, a trip into the center of town can be **entertaining**. For those who prefer something greener, a trip to Central Park is the thing to do. On a sunny day, people enjoy picnic lunches and long walks along the many trails through the Park. Central Park is very large, and features wide lawns, gardens, and a small zoo. |

HIDE **PREVIOUS** **NEXT**

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 C-4. Writing Online hands-on editing tasks, screen 4, grade 8: 2002

MINUTE

NAEP WRITING ONLINE

Editing

Next, make the following change to the bolded words and phrases in the paragraph on the right according to the directions below. You will have 1 minute.

- Change the word "entertaining" to the word "enjoyable".

When planning a vacation, people should consider going to a major city, like New York. Cities like New York have something for everyone. For those who prefer crowds, museums, and shopping, a trip into the center of town can be **entertaining**. For those who prefer something greener, a trip to Central Park is the thing to do. On a sunny day, people enjoy picnic lunches and long walks along the many trails through the Park. Central Park is very large, and features wide lawns, gardens, and a small zoo.

HIDE PREVIOUS NEXT

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 C-5. Writing Online hands-on editing tasks, screen 5, grade 8: 2002

MINUTE NAEP WRITING ONLINE

Editing

Next, make the following change to the bolded words and phrases in the paragraph on the right according to the directions below. You will have 1 minute.

- Move the last sentence so that it comes right after sentence 4 (marked with an *).

When planning a vacation, people should consider going to a major city, like New York. Cities like New York have something for everyone. For those who prefer crowds, museums, and shopping, a trip into the center of town can be entertaining. For those who prefer something greener, a trip to Central Park is the thing to do.* On a sunny day, people enjoy picnic lunches and long walks along the many trails through the Park. **Central Park is very large, and features wide lawns, gardens, and a small zoo.**

HIDE PREVIOUS NEXT

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 C-6. Writing Online hands-on editing tasks, screen 6, grade 8: 2002

MINUTE

NAEP WRITING ONLINE

Editing

Next, make the following change to the bolded words and phrases in the paragraph on the right according to the directions below. You will have 1 minute.

- Correct the spelling for "vacasion".

When planning a **vacasion**, people should consider going to a major city, like New York. Cities like New York have something for everyone. For those who prefer crowds, museums, and shopping, a trip into the center of town can be entertaining. For those who prefer something greener, a trip to Central Park is the thing to do. On a sunny day, people enjoy picnic lunches and long walks along the many trails through the Park. Central Park is very large, and features wide lawns, gardens, and a small zoo.]

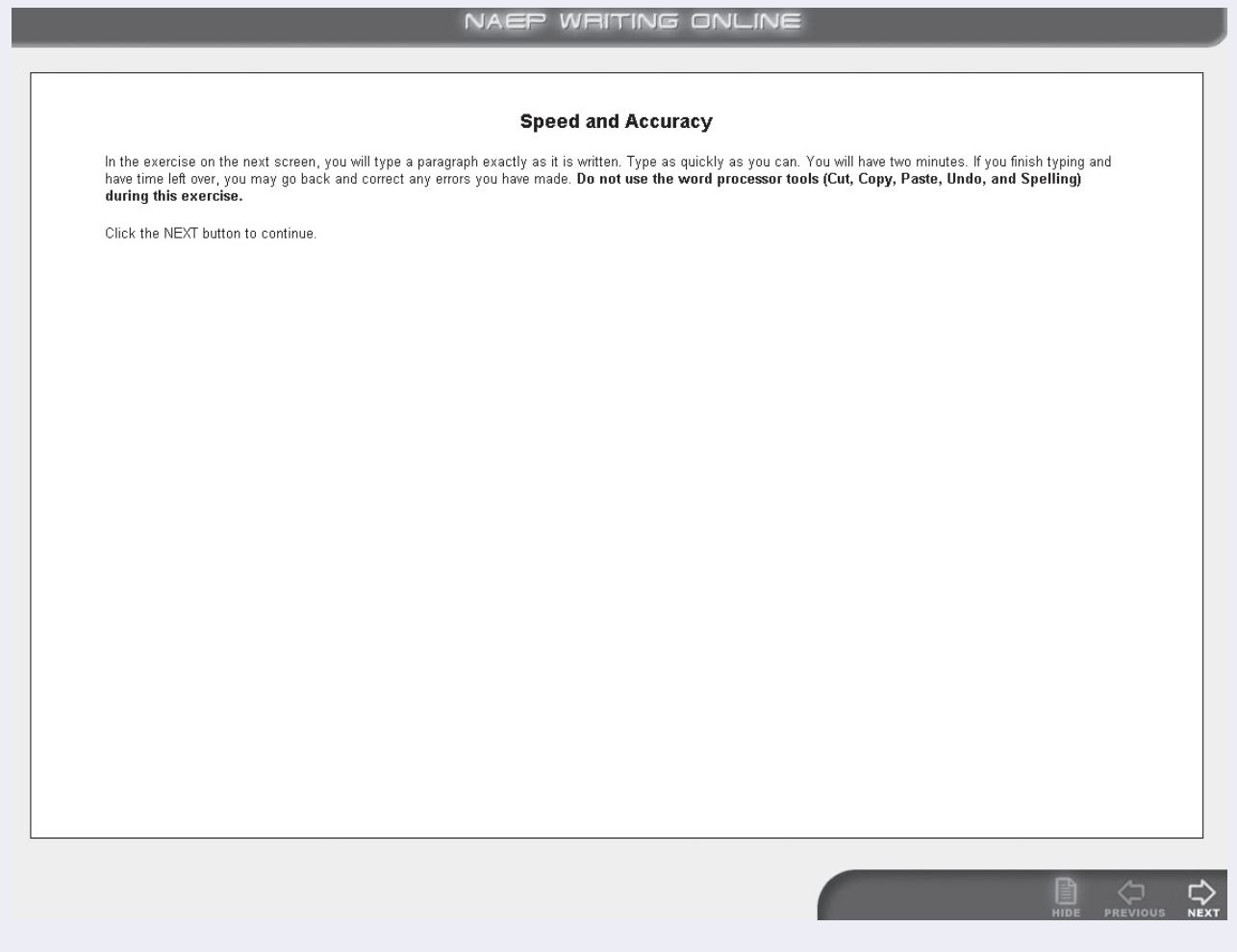
HIDE PREVIOUS NEXT

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.

Appendix D. Writing Online Speed and Accuracy Tasks

This appendix presents screen shots of the tasks used to measure students' online typing speed and accuracy.

Figure D-1. Writing Online speed and accuracy tasks, screen 1, 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 D-2. Writing Online speed and accuracy tasks, screen 2, grade 8: 2002

The screenshot shows the NAEP Writing Online interface for a speed and accuracy task. At the top, there is a header with 'MINUTES' on the left and 'NAEP WRITING ONLINE' in the center. Below the header, the interface is divided into two main sections. On the left, there is a panel titled 'Speed and Accuracy' with the instruction 'Begin typing now.' and a text box containing the following text: 'I don't know about you, but I really like new movies (the ones they make now) better than the old movies that are my parents' favorites. I think the most important difference for me is that the older movies don't seem as real. First of all, the older movies are in black and white. Color makes a big difference in a movie; it's a lot more interesting to watch things happen in color than in black and white.' On the right, there is a large writing area with a vertical scrollbar on the right side. Above the writing area is a toolbar with icons for 'Cut', 'Copy', 'Paste', 'Undo', and 'Spelling'. At the bottom of the interface, there is a navigation bar with icons for 'HIDE', 'PREVIOUS', and 'NEXT'.

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.

Appendix E. Background Questions Administered in Writing Online

Questions 1–8. 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
9. Who taught you the most about how to use a computer?
- I learned the most on my own.
 - I learned the most from my friends.
 - I learned the most from my teachers.
 - I learned the most from my family.
 - I don't really know how to use a computer.
10. How often do you use a computer at school? Include use anywhere in the school and at any time of day.
- Every day
 - Two or three times a week
 - About once a week
 - Once every few weeks
 - Never or hardly ever
11. How often do you use a computer outside of school?
- Every day
 - Two or three times a week
 - About once a week
 - Once every few weeks
 - Never or hardly ever
12. Is there a computer at home that you use?
- Yes
 - No

Questions 13–15. Please indicate the extent to which you AGREE or DISAGREE with the following statements. (Choices: Strongly agree, Agree, Disagree, Strongly disagree, I never use a computer)

13. I am more motivated to get started doing my schoolwork when I use a computer.
 14. I have more fun learning when I use a computer.
 15. I get more done when I use a computer for schoolwork.
16. Which best describes you?
- White
 - Black
 - Hispanic
 - Asian
 - Pacific Islander
 - Other

17. If you are Hispanic, what is your Hispanic background?

- I am not Hispanic
- Mexican, Mexican American, or Chicano
- Puerto Rican
- Cuban
- Other Spanish or Hispanic background

18. How far in school did your mother go?

- She did not finish high school.
- She graduated from high school.
- She had some education after high school.
- She went to college.
- I don't know.

19. How far in school did your father go?

- He did not finish high school.
- He graduated from high school.
- He had some education after high school.
- He went to college.
- I don't know.

20. About how many books are there in your home?

- Few (0-10)
- Enough to fill one shelf (11-25)
- Enough to fill one bookcase (26-100)
- Enough to fill several bookcases (more than 100)

21. Does your family get a newspaper at least four times a week?

- Yes
- No
- I don't know.

22. Does your family get any magazines regularly?

- Yes
- No
- I don't know.

23. Is there an encyclopedia in your home? It could be a set of books, or it could be on the computer.

- Yes
- No
- I don't know.

24. On a school day, about how many hours do you usually watch TV or videotapes outside of school?

- None
- 1 hour or less
- 2 or 3 hours
- 4 or 5 hours
- 6 hours or more

Questions 25–28. 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)

- 25. Brainstorm with other students to decide what to write about
- 26. Organize your paper before you write (for example, make an outline, draw a chart)
- 27. Make changes to your paper to fix mistakes and improve your paper
- 28. Work with other students in pairs or small groups to discuss and improve your paper

Questions 29–34. 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

35. How often do people in your home talk to each other in a language other than English?

- Never
- Once in a while
- About half of the time
- All or most of the time

36. When you write, how often does your teacher talk to you about what you are writing?

- Never
- Sometimes
- Always

37. When you write, how often does your teacher ask you to write more than one draft of a paper?

- Never
- Sometimes
- Always

Appendix F. NAEP Grade 8 Writing Scoring Guides

Informative Scoring Guide

6 Excellent Response

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

5 Skillful Response

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

4 Sufficient Response

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

3 Uneven Response (may be characterized by one or more of the following)

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

2 Insufficient Response (may be characterized by one or more of the following)

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

1 Unsatisfactory Response (may be characterized by one or more of the following)

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

Persuasive Scoring Guide

6 Excellent Response

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

5 Skillful Response

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

4 Sufficient Response

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

3 Uneven Response (may be characterized by one or more of the following)

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

2 Insufficient Response (may be characterized by one or more of the following)

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

1 Unsatisfactory Response (may be characterized by one or more of the following)

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

Appendix G. Statistical Procedures

Procedure for ANOVA Using WESVAR

Many of the research questions for the Writing On-line (WOL) study required repeated-measures analysis of variance (ANOVA). These analyses were complicated by the necessity of using student sampling weights with the WOL data. WESVAR, proprietary software of Westat, was used so that student sampling weights would be applied appropriately. Because WESVAR does not currently have an ANOVA option, the regression option was used to calculate ANOVA tables and tests. Contrasts, coded as categorical variables in WESVAR, were used to define the groups specified by the variables in the model. To create the contrast defining gender, for instance, male students were coded as 1 and female students were coded as 0.

Contrasts were needed for most of the independent variables to allow the WESVAR regression routines to calculate the statistics required for repeated-measures ANOVA. Any covariates included as independent variables were coded as ordinal, or continuous, variables.

Contrast coding of the type described above is necessary for any ANOVA analysis. For repeated-measures ANOVA in a regression setting, appropriate tests can be performed by creating additional variables to reflect the within- and between-group sources of variance. To do this, in a setting where scores on the two WOL essays were the outcome variables, two dependent variables were created. The between-groups variable, B , is defined as

$$B = \frac{(x_1 + x_2)}{\sqrt{2}},$$

where x_i is the score on the i^{th} essay.

The within-groups variable, W , is defined as

$$W = \frac{(x_1 - x_2)}{\sqrt{2}},$$

where x_i is the score on the i^{th} essay.

After these two dependent variables were formed, two separate regressions were run with the independent variables—one regression to estimate the between-group effects and one to estimate the within-group effects.

Correlations Used in This Report

Two types of correlations are used throughout this report. For reader reliability statistics the intraclass correlation coefficient is used. It is defined as

$$r(ICC) = \frac{MSS - MSR}{MSS + (k - 1) * MSR},$$

where MSS is the mean sum of the squares within subjects and MSR is the mean sum of the squares between subjects (i.e., within readers) obtained from a one-way ANOVA, and k is the number of readers.

For other types of correlations a standard Pearson correlation was used,

$$r(Pearson) = \frac{\text{covariance}(X, Y)}{\sqrt{\text{Var}(X) * \text{Var}(y)}}.$$

t-tests Used in This Report

The following section explains the calculation of t -tests:

Let A_i be the statistic in question (e.g., a mean for group i) and let S_{A_i} be the standard error of the statistic. The text in the reports identifies the means or proportions for groups i and j as being different if

$$\frac{|A_i - A_j|}{\sqrt{S_{A_i}^2 + S_{A_j}^2}} \geq T_{\frac{.05}{2c}}$$

where T_a is the $(1 - a)$ percentile of the t distribution with degrees of freedom, df , set to the number of replicates involved in the comparison.

Appendix H. Percentage of Writing Online Students Who Report Using a Computer for Different Specific Writing Purposes

Table H-1. Percentage of Writing Online students who report using a computer for different specific writing purposes, grade 8: 2002

Item	Always	Sometimes	Never
29. Use a computer to plan your writing (for example, by making an outline, list, chart, or other kind of plan)	15 (1.2)	48 (1.4)	37 (1.5)
30. Use a computer from the beginning to write the paper or report (for example, use a computer to write the first draft)	32 (1.7)	42 (1.6)	25 (1.4)
31. Use a computer to make changes to the paper or report (for example, spell-check, cut and paste)	57 (1.6)	32 (1.4)	10 (0.9)
32. Use a computer to type up the final copy of the paper or report that you wrote by hand	69 (1.6)	24 (1.5)	6 (0.7)
33. Look for information on the Internet to include in the paper or report	60 (1.8)	35 (1.7)	5 (0.7)
34. Use a computer to include pictures or graphs in the paper or report	37 (1.9)	48 (1.9)	14 (1.0)

NOTE: The number of students responding ranged from 1,300 to 1,304. Standard errors are in parentheses. Detail may not sum to totals 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.

Appendix I. Summary Statistics for Computer Familiarity Measures

Table I-1. Summary statistics for components of the hands-on computer skills measure, grade 8: 2002

Component	<i>n</i>	Scale range	Mean	Standard deviation
Typing speed	686	0–78	36.3	19.5
Typing accuracy	686	0–maximum number of errors made	3.1	3.8
Editing	672	0–5	3.1	1.3

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 I-2. Summary statistics for computer familiarity measures, grade 8: 2002

Measure	<i>n</i>	Scale range	Mean	Standard deviation
Extent of computer use	681	0–8.0	4.5	1.8
Computer use for writing	685	0–6.0	5.0	1.2
Hands-on computer skill	672	0–4.3	2.1	0.9

NOTE: The values for hands-on computer skill were real numbers created from a regression equation relating Writing Online (WOL) score to measures of typing speed, typing accuracy, and editing skill. The largest observed value for hands-on computer skill was just under 4.3.

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.

Appendix J. Analysis of Variance Results Relating Computer Familiarity and Gender to Writing Online Performance

Table J-1. Results of repeated-measures analysis of variance testing the effects of gender and of self-reported and hands-on computer familiarity variables on Writing Online performance, controlling for main NAEP writing performance, grade 8: 2002

Variable	F-value	Numerator df	Denominator df	p-value
Between-subjects effects				
Main NAEP writing skill	86.34	1	62	.00*
Extent of computer use	3.30	1	62	.07
Computer use for writing	1.54	1	62	.22
Hands-on computer skill	98.11	1	62	.00*
Gender	4.31	1	62	.04*
Extent of computer use x gender	0.57	1	62	.45
Computer use for writing x gender	2.96	1	62	.09
Hands-on computer skill x gender	0.22	1	62	.64
Within-subjects effects				
Main NAEP writing performance x essay	0.09	1	62	.77
Extent of computer use x essay	0.14	1	62	.71
Computer use for writing x essay	1.92	1	62	.17
Hands-on computer skill x essay	5.01	1	62	.03*
Gender x essay	0.34	1	62	.56
Extent of computer use x gender x essay	0.07	1	62	.80
Computer use for writing x gender x essay	0.23	1	62	.63
Hands-on computer skill x gender x essay	0.01	1	62	.91

* $p < .05$ for the difference of the regression coefficient from zero as calculated using an F -test.

NOTE: WOL=Writing Online. Students taking the WOL computer test were drawn from the main NAEP writing sample. The number of students responding to both essays was 660.

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.

THIS PAGE INTENTIONALLY LEFT BLANK.

THIS PAGE INTENTIONALLY LEFT BLANK.

United States
Department of Education
ED Pubs
8242-B Sandy Court
Jessup, MD 20794-1398

Official Business
Penalty for Private Use, \$300

Postage And Fees Paid
U.s. Department of Education
Permit No. G-17

