NCES
National Center for Education Statistics

The Nation's Report Card NAEP TM

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

National Center for Education Statistics

The Nation's Report Card NAEP

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

Taslima 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.

Foreword

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

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

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

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

Acknowledgments

The NAEP 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

Tables and Figures

List of Tables

List of Figures

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 multistage, 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, Lutkus, 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 interrater 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.

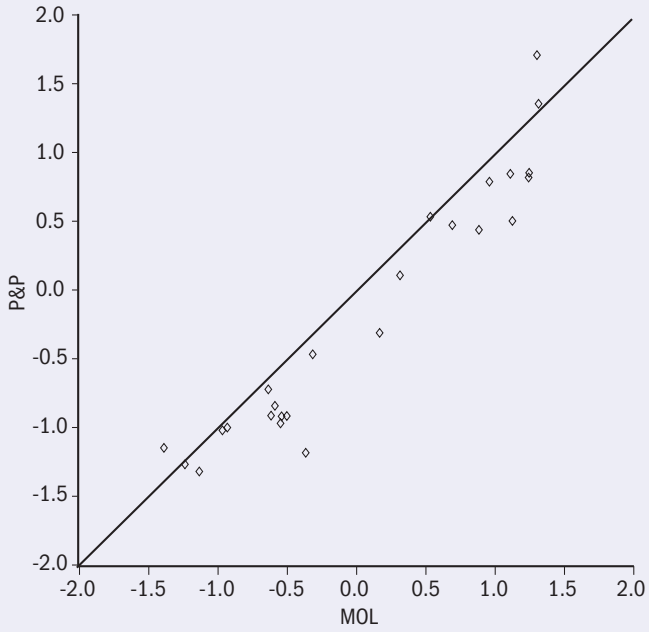
Analysis of Item Difficulty for Eighth Grade

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 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]

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.

[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 *b*		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

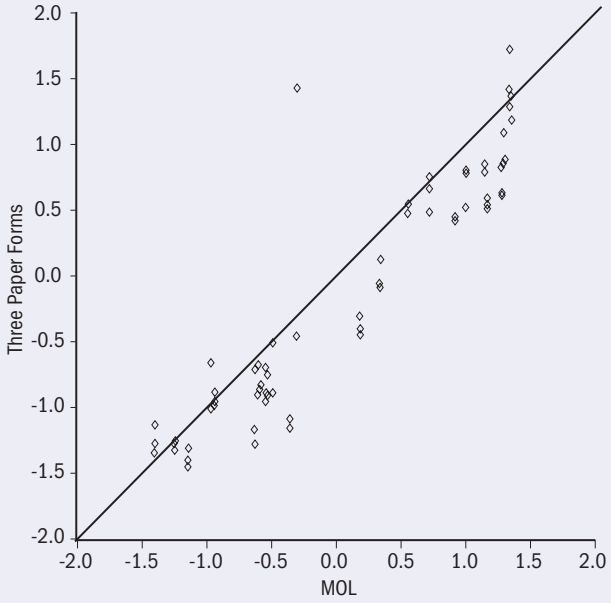
NOTE: MOL=Math Online. P&P=Paper and Pencil. Forms A and B were paper forms constructed to be parallel to P&P.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

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.

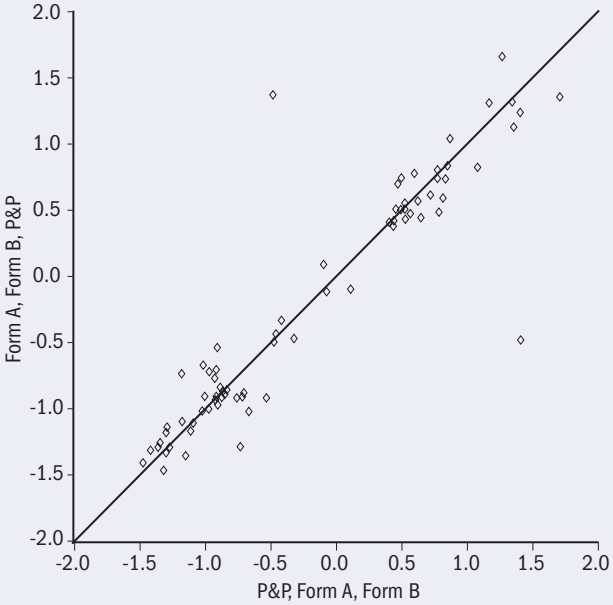
[18] This plot shows, as an outlying data point, the divergence of IRT *b* parameter estimates for the item described in the preceding footnote. The outlying data point represents the difference between the *b* parameter estimate for the item presented on Form A and the variant of that item included 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.

Table 4-1. Components of the input-skill measure, grades 4 and 8: 2001

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.

Table 4-2. Coefficient alpha values for computer familiarity measures, grades 4 and 8: 2001

	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.

Table 4-3. Sample correlations among computer familiarity measures and with mathematics performance, grades 4 and 8: 2001

	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

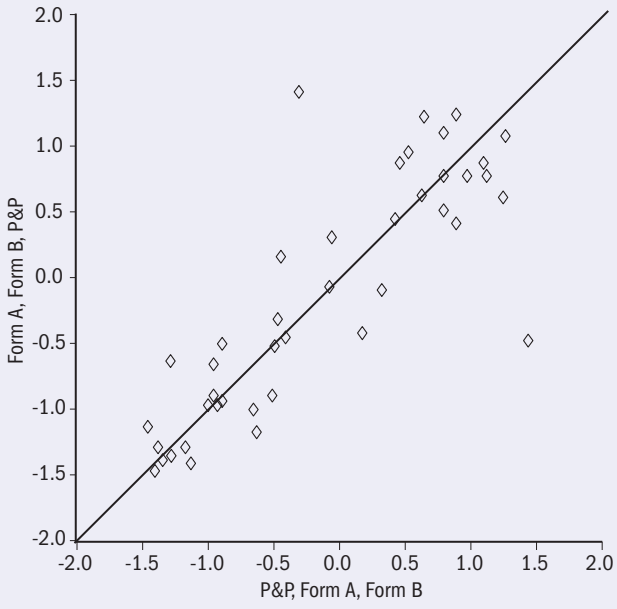
In this study, models were created for eighth-grade items using the Mathematics Test Creation Assistant (Singley and Bennett 2002). Each model resembles a test item in which elements of the stimulus, stem, and response options are treated as variables. Both linguistic and mathematical elements can be 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.)

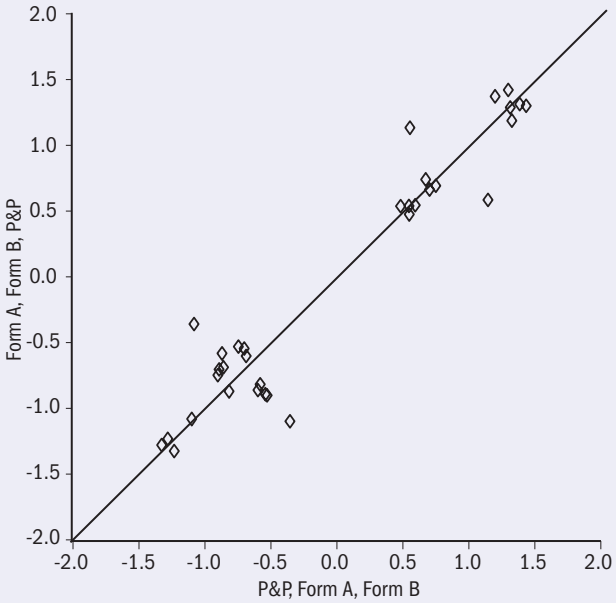
Figure 5-1 shows the IRT *b* values for each set of 14 isomorphs on the three paper forms. Each isomorph on the P&P base form appears twice on the plot, once in comparison to its sibling on Form A and once in comparison to its sibling on Form B. The parameter estimate comparisons between Forms A and 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).

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.

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 5-3. 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

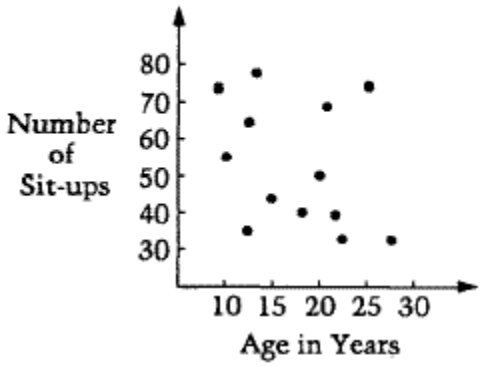
18. There are 15 girls and 11 boys in a mathematics class. If a student is selected at random to run an errand, what is the probability that a boy will be selected?

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

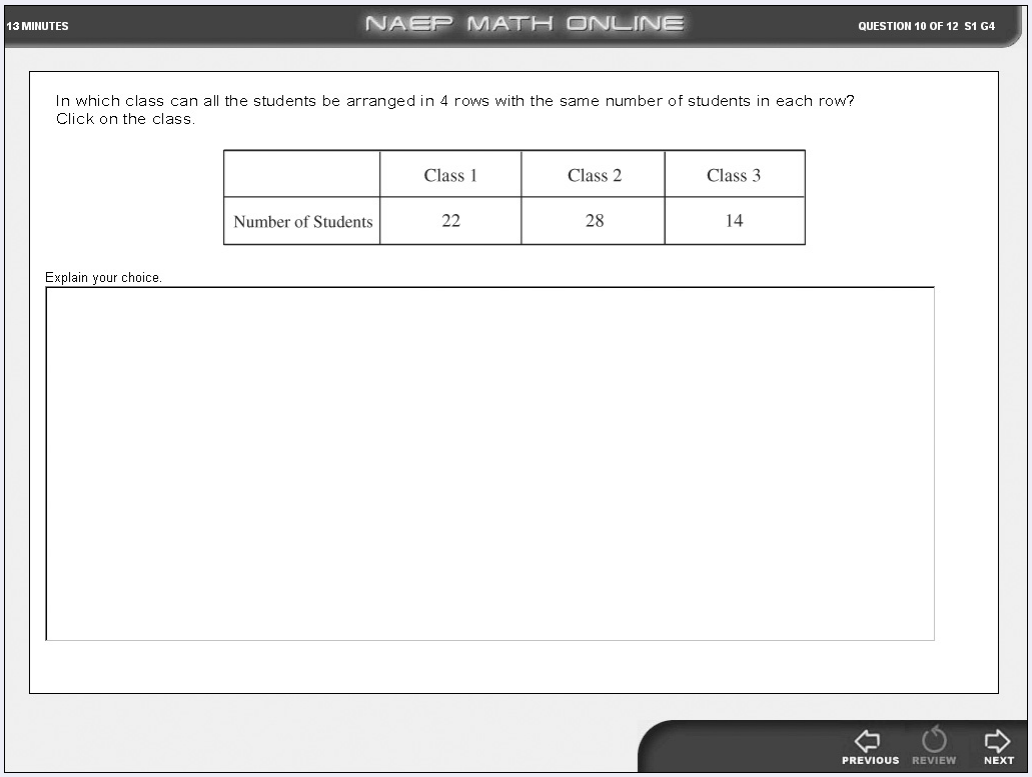
Another application of technology that could help NAEP increase efficiency is the automated scoring of constructed-response items. By reducing the need for human judges, automated scoring could potentially increase the speed with which NAEP analyses can be completed and also reduce the cost of scoring.

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).

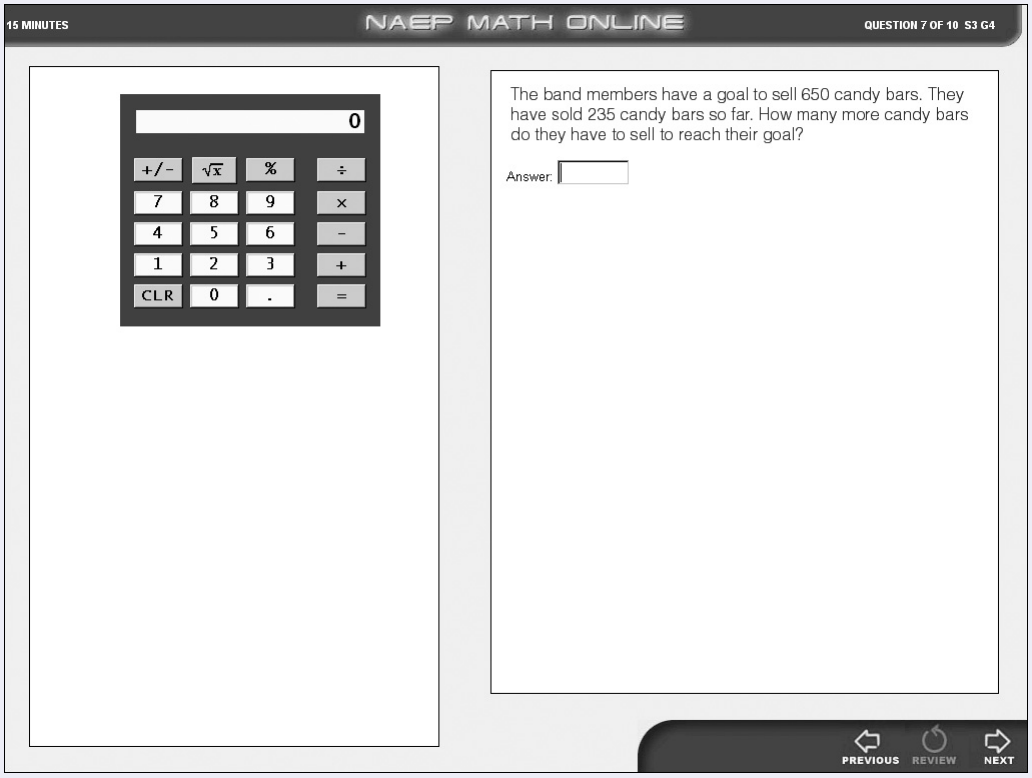
The answers students gave to items that did not require explanations were either numeric or simple text responses (e.g., "30" or "thirty"). In contrast, the answers students gave to the items requiring explanations were usually more elaborated text, consisting of phrases or sentences. These two kinds 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



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



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 distinguish 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	Percentage exact agreement: Automated grader vs. Reader 1	Percentage exact agreement: 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 "subtracking." 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:

> sutract subtact sudtract subtacted subracted substract subtracd subtrct sudtracted sbutracted subract subtarct subtracat subtrackted subtrated subtrsct subttacted

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

Pilot test

Paper delivery		Computer delivery	
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	10
• Scores sent to NAEP database	10		160
• Data sent to contractor for analysis	10		
	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	10
• Scanned responses scored on computer by NAEP raters	8		106
• Scores entered into NAEP database	10		
• Data sent to contractor for analysis	10		
	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., 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.

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

	Percent of items		
Framework content area	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

	Percent of items		
Item format	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

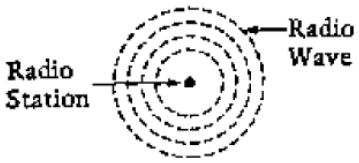
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

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

	Mean percent of students			
Test section	**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

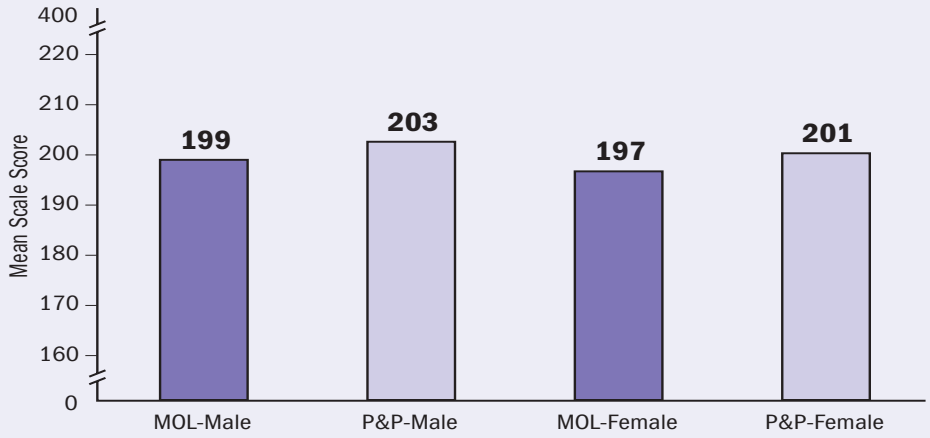
The estimate rounds to zero.
– Not available. No dichotomous CR items were included in this section.
NOTE: MOL=Math Online. CR=constructed-response. Each figure is the percentage of students omitting, not reaching, or giving an off-task response to an item, as the case may be, averaged over all items.
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

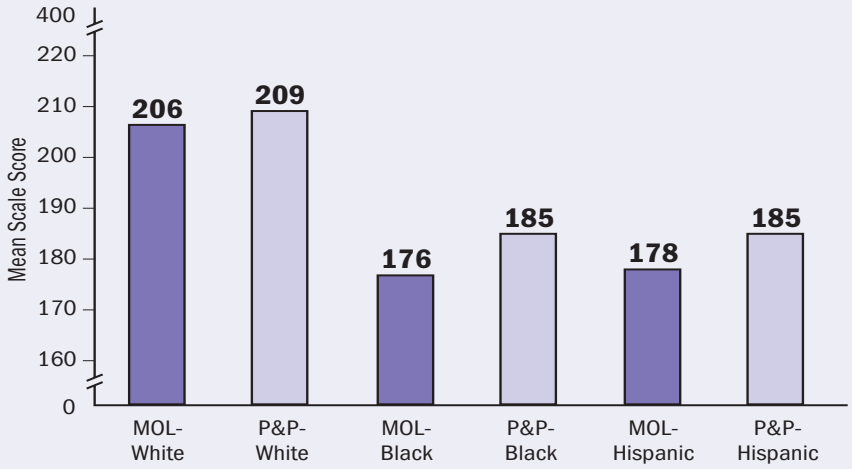
This appendix presents data on the performance of NAEP reporting groups on the 2001 mathematics online test (MOL) and on the paper-based form.

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



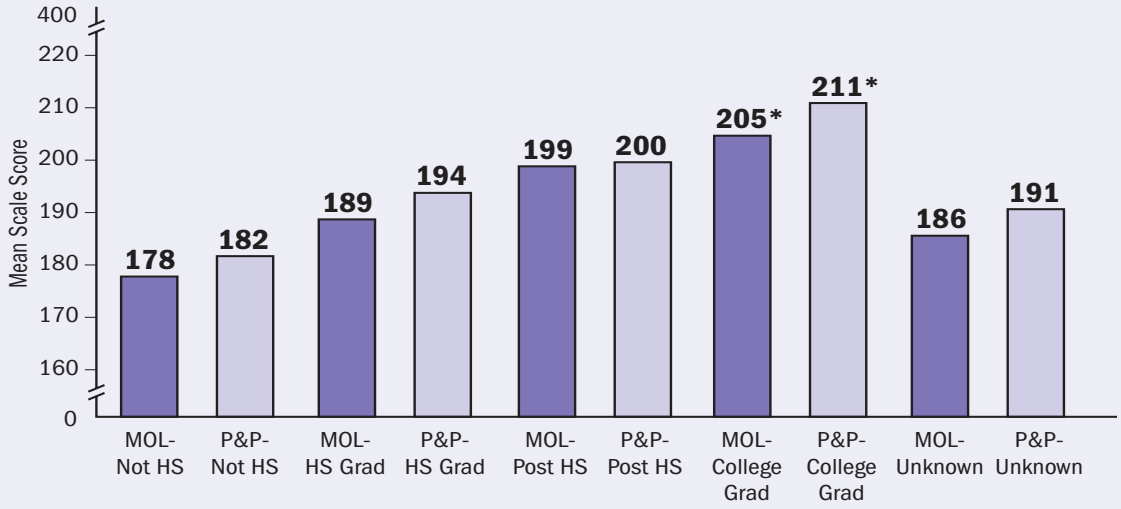
NOTE: MOL=Math Online. P&P=Paper and Pencil. Average MOL scores are reported on 0–400 scale.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

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



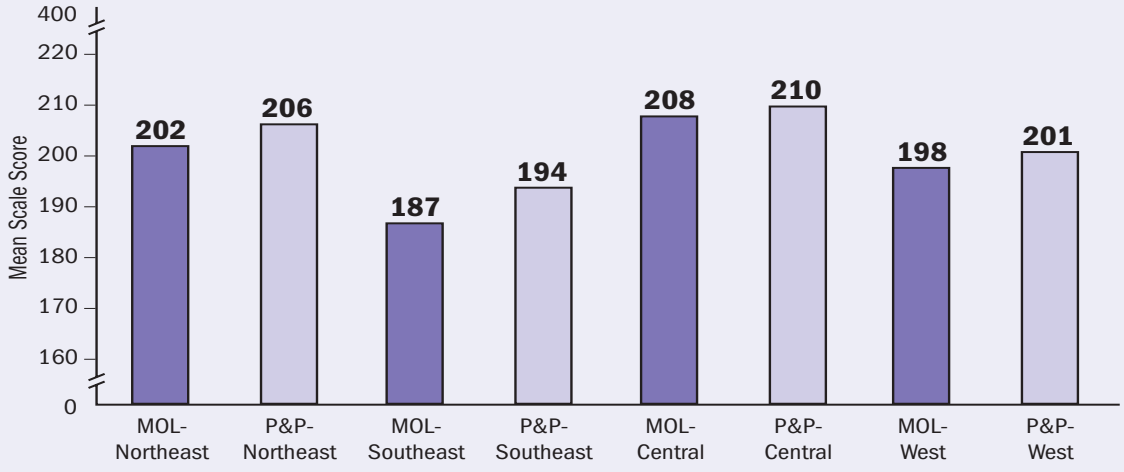
NOTE: MOL=Math Online. P&P=Paper and Pencil. Sample sizes for other racial/ethnic groups were too small to analyze statistically. Average MOL scores are reported on 0–400 scale.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment 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



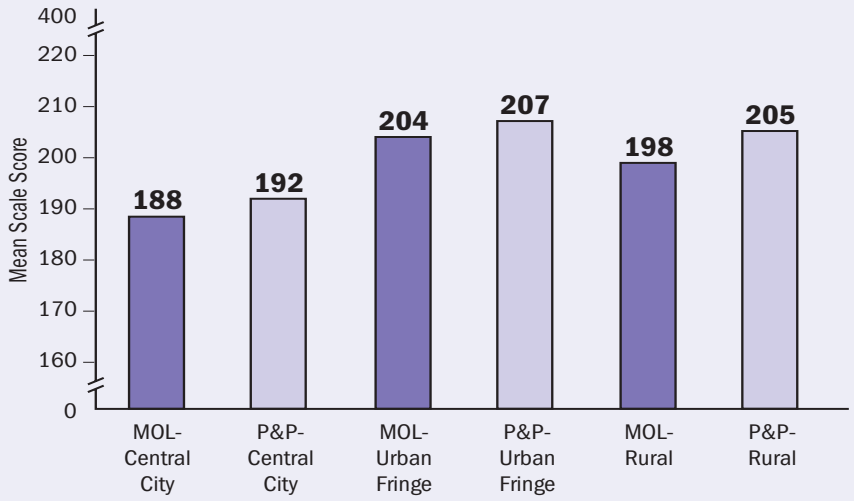
* MOL and P&P values differ significantly, $p < .05$, for students reporting parent graduated from college.
NOTE: MOL=Math Online. P&P=Paper and Pencil. Average MOL scores are reported on 0–400 scale.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education 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



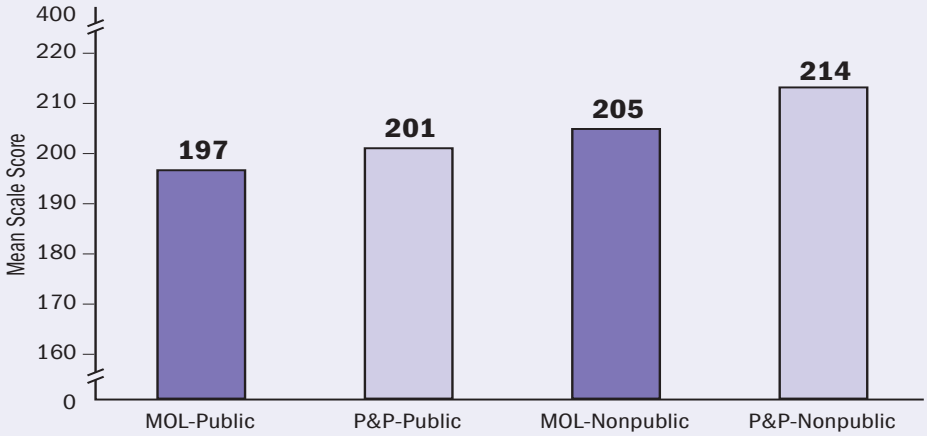
NOTE: MOL=Math Online. P&P=Paper and Pencil. Average MOL scores are reported on 0–400 scale.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2001 Math Online Study.

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.