RESEARCH REPORT

October 2001
RR-01-23

Effects of Screen Size, Screen Resolution, and Display Rate on Computer-Based Test Performance

Brent Bridgeman
Mary Louise Lennon
Altamese Jackenthal

Statistics & Research Division
Princeton, NJ 08541

Effects of Screen Size, Screen Resolution, and Display Rate on Computer-Based Test Performance

Brent Bridgeman, Mary Louise Lennon, and Altamese Jackenthal
Educational Testing Service, Princeton, New Jersey

October 2001

Abstract

Computer-based tests administered in established commercial testing centers typically have used monitors of uniform size running at a set resolution. Web-based delivery of tests promises to greatly expand access, but at the price of less standardization in equipment. The current study evaluated the effects of variations in screen size, screen resolution, and presentation delay on verbal and mathematics scores in a sample of 357 college-bound high school juniors. The students were randomly assigned to one of six experimental conditions—three screen display conditions crossed with two presentation rate conditions. The three display conditions were: a 17-inch monitor set to a resolution of 1024 x 768, a 17-inch monitor set to a resolution of 640 x 480, and a 15-inch monitor set to a resolution of 640 x 480. Items were presented either with no delay or with a five-second delay between questions (to emulate a slow Internet connection).

No significant effects on math scores were found. Verbal scores were higher, by about a quarter of a standard deviation (28 points on the SAT® scale), with the high-resolution display.

Key words: computer-based testing, display, font size, latency, presentation, resolution, screen size, test performance

Acknowledgements

Our thanks go to the following individuals for their help in conducting this study. Fred Cline conducted all the statistical analysis work. Rob Rarich authored all of the items in the test-delivery system. Janet Stumper created the graphics used in the math section of the test. Mike Wagner set up the lab server for us, provided technical assistance on a regular basis, and set up the files for data collection. Steve Szyszkiewicz helped organize the data files. Doug Forer and Pete Brittingham shared their technical expertise and helped us define the study conditions. Mary Schedl reviewed the text reference items in the reading section and shared helpful advice about item wording. Brian O'Reilly of the College Board gave us permission to use released SAT items, and Cathy Wendler and Liz Forbes helped us locate sources for those items.

Introduction

In the past, delivery of computer-based tests (CBTs) in large-scale, high-stakes testing programs (such as the Graduate Record Examinations® and the Test of English as a Foreign Language™) has used a relatively uniform standard for hardware configurations. Screen size was generally either 15 or 17 inches and resolution was set to 640 x 480. Laptops sometimes were used in temporary testing centers, but still at 640 x 480 resolution. Item presentation speed at testing centers varied only slightly because it was controlled by a local dedicated server. A new generation of CBTs is currently under development. These Web-based systems permit delivery in a diverse array of institutional settings. Rather than relying on a dedicated network of private testing centers, Web-based delivery will allow almost any university or high school computer lab to become a testing center. However, such a Web-based delivery system will have to deal with a much broader range of computer configurations. Although the intention is to set as few restrictions on the allowable configurations as possible, standards will have to be set for those factors that have a significant effect on test performance. Students and test users must be assured that no one will be disadvantaged by the particular computer configuration in the local testing center. (Alternatively, scores could be adjusted if the impact of particular configurations on test scores is known. However, it might be difficult to explain to some examinees that 10 points were deducted from their scores because they took the test on a large monitor while other students had to take the test on a smaller monitor.)

The intent of this study was to begin exploring the effect that presentation variations might have on test performance. A large number of configuration variations could conceivably have an impact on scores. Variations in visual presentation (e.g., text size, font, color, contrast), aural presentation (sound quality and volume), and timing all could potentially affect performance. An experimental design incorporating three levels of each of these features crossed with various question types yields a design with more than 1,000 cells. Thus, a truly comprehensive design was not feasible, and we instead focused on those features that previous experience and logical analysis suggested were likely to be most crucial and/or most difficult to standardize in a Web-based environment of distributed testing centers.

A few themes emerged from a review of the literature and discussions with colleagues. One clear issue of interest is the impact that variations in the amount of information displayed on a screen and the legibility of that information might have on test performance. These variations

are affected by a number of features, including screen size, the resolution at which monitors are set, and a series of font size settings.

- Screen size

  When resolution is held constant, variations in screen size do not impact the amount of information displayed on a screen but make that information appear smaller (on a smaller monitor) or larger (on a larger monitor). For example, the difference between the same information presented on a 15-inch screen and a 17-inch screen would look something like Figure 1, shown below. While the amount of information remains constant, test takers might find one easier to read than another.

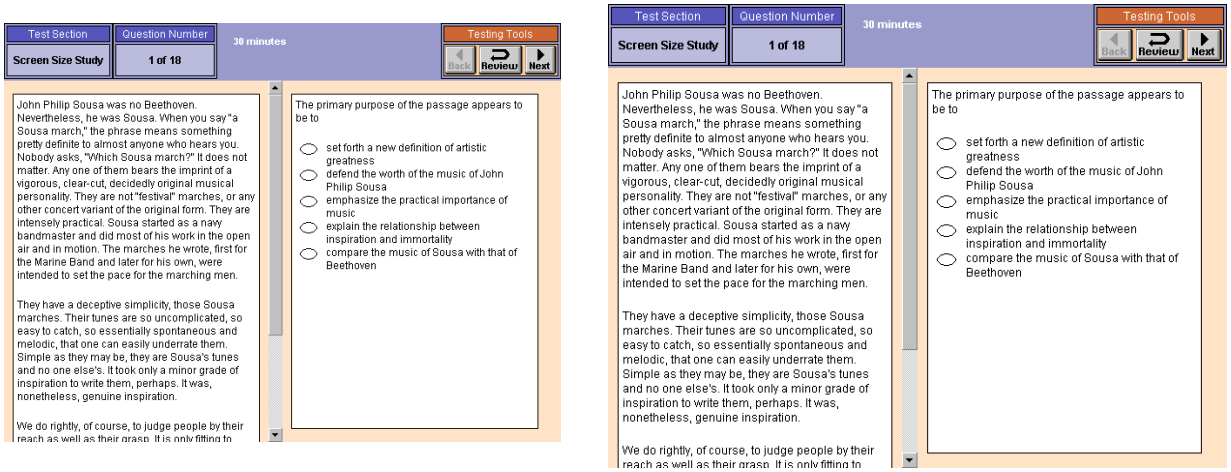


Figure 1. **Simulation of a 15-inch screen display at a resolution of 640 x 480 (on the left) and a 17-inch display at the same resolution (on the right).**

- Screen resolution

  Resolution is a critical variable because it impacts both the size of text on the screen and the actual amount of information that can be displayed. A 10-point font in a 1024 x 768 resolution will appear smaller than that same point size in a 640 x 480 resolution because the pixels are physically smaller in a 1024 x 768 display. In addition, a higher resolution allows more words per line and more lines per screen than can be displayed in a lower resolution. (See, for example, Figure 2 showing a passage that displays in its entirety in the 1024 x 768 resolution but requires scrolling in the 640 x 480 resolution.) In an

assessment context, examinees using screens set to a lower resolution who have to scroll through text to find an answer may be at a disadvantage relative to others using higher resolutions where scrolling is not required. Indeed, the cumulative effect of screen size and resolution is one aspect of display technology that has been implicated in several studies as an influence on reading speed and comprehension. Several studies (DeBruijn, de Mul, & Van Oostendorp, 1992; Dillon, Richardson, & McKnight, 1990; Haas, 1996; Haas & Hayes, 1985) found that using a larger screen at a higher resolution had a positive effect on the target performance (although it should be noted that the experimental tasks used in these studies did not resemble tasks presented in typical high-stakes tests).

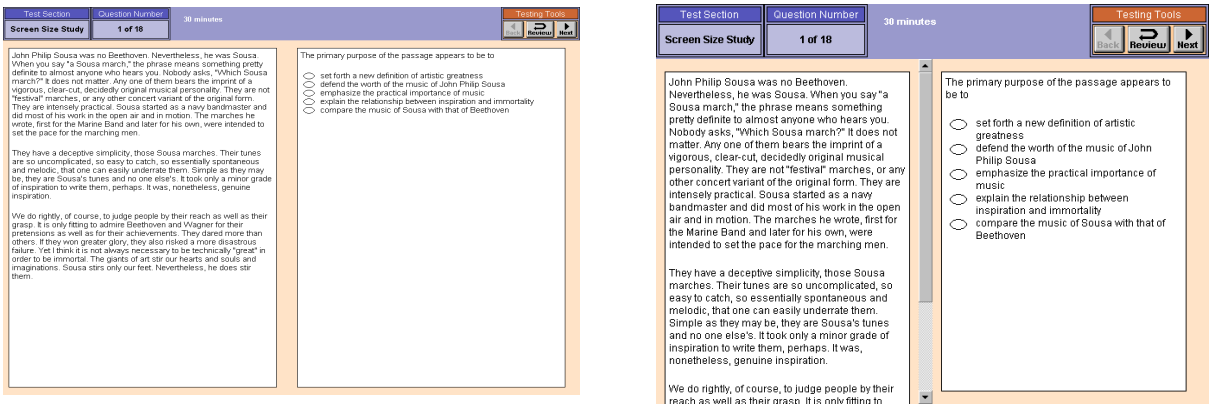


Figure 2. **Screen display on a 17-inch monitor at a 1024 x 768 resolution (on the left) and 640 x 480 resolution (on the right).**

- Font size settings

  The size of font used in rendering text to the screen directly impacts the amount of text that is visible to the test taker. Font sizes affect the number of characters that will fit on a line of text, as well as the number of lines of text that will fit in a given area on the screen. In a typical Web browser, the font size that is actually used can be affected by settings in the Web browser, in the operating system, and in the HTML coding of the Web page. As a result, variations in these settings affect the amount and appearance of text and diagrammatic information displayed on a screen. Two examples are shown in Figure 3. These illustrate two versions of an item presented on monitors of the same size

and resolution. The only variation is the font size setting in the control panel and the browser, as noted.

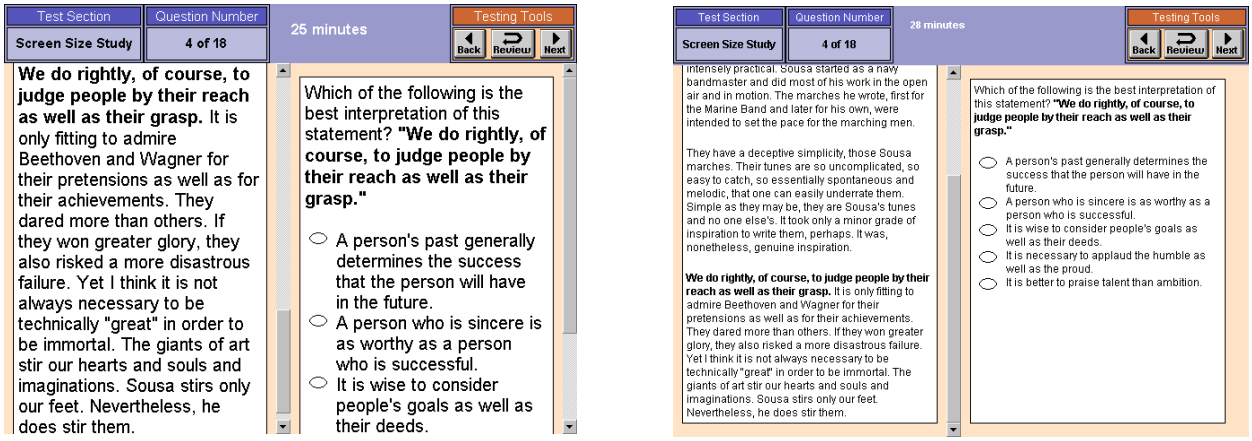


Figure 3. **On the left, font size set to "large" in control panel and "medium" in browser. Notice that the third and fourth options do not display without scrolling. On the right, font size set to "small" in control panel and "smallest" in browser.**

For the purposes of this study, variations in screen size and resolution were selected as targets of the research. Because of an interest in exploring the implications of Web-delivered assessments, question presentation latency was included as a third variable. Presentation latency is defined here as the amount of time between answering one question and the presentation of the next question. With a stand-alone computer, or a small local network with a dedicated server, presentation latency is virtually zero, but with a Web-based delivery, or even a local server on a large local network, latency could be noticeable. Although a wait of a few seconds may be trivial, it is conceivable that it could discourage examinees from going back to review previous answers on tests that allow such review.

Method

Sample

Although high-stakes admissions tests can cover a broad age range, most focus on either the end of secondary school or the end of college. We decided to focus on high school juniors because their relatively weak reading skills might be more sensitive to variations in the amount

of text on a single screen. We recruited, by mail, students who had already taken the Preliminary SAT, which is also used as the National Merit Scholarship qualifying test (PSAT/NMSQT™). This test contains the same type of questions used in the SAT I: Reasoning Test. Letters were sent to approximately 3,000 high school juniors who had taken the PSAT/NMSQT in the fall of 1999 and who lived within about 30 miles of the testing site (as estimated from postal ZIP codes). Some students were recruited by announcements at local high schools and by word of mouth, but all participants had to be high school juniors with PSAT/NMSQT scores. The incentives to participate were the opportunity to practice for the SAT I using questions from a real SAT I, and a cash payment of $25 on completion of the test.

Instruments Used

Two "test sections" were created using disclosed SAT I verbal and math items. The items were taken, with permission, from a relatively old publication (*10 SATs*, College Entrance Examination Board, 1986), which study participants were unlikely to have seen.

Section 1. The first test section consisted of four reading sets with a total of 18 accompanying items. Sets were selected based on the following criteria:

- Passage length—Two of the selected reading sets contained relatively short passages (230 and 270 words) and two had longer passages (511 and 434 words).
- Item length and number of items—The selected sets contained at least four items with stems and options of varying lengths.
- Text reference—The selected sets included items that directly referenced information in the text. These items were more likely to require participants to look back in the passage rather than depend on general recall of information to answer the questions, and it was thought these might be more sensitive to variations in presentation.
- Position in paper-and-pencil forms
- Range of item difficulty

Section 2. The second section included 15 math items, all of the quantitative comparison (QC) item type. This item type was selected for two reasons. First, many of these items contained graphically complex stimulus materials, and it was thought these materials might be most affected by variations in presentation mode. Second, templates already existed for QC

items in the system into which the items were to be authored. Using existing item templates saved both development time and cost.

Authoring. Once the paper-and-pencil SAT items were selected, they were authored using the Educational Testing Service (ETS) Research C-3 system. Because line references change with font size, 5 of the 18 reading items that referred to words on specific lines had to be revised. All revised items were reviewed to ensure that they represented realistic computer-based items. (See Appendix A for samples of each of the situations described below.)

- In three cases, the word, phrase, or sentence being referenced in the item was presented in bold text both in the item and the passage and the specific line reference was deleted.
- In one case, the line reference was changed to a paragraph reference. The original wording was changed from "mentioned in lines 1-45" to "mentioned in the first two paragraphs."
- The final change consisted of revising a stem that asked, "Which of the following is the best interpretation of lines 23-24?" The new stem included the statement made in lines 23-24. It asked, "Which of the following is the best interpretation of this statement by the author? 'We do rightly, of course, to judge people by their reach as well as their grasp.'" The statement was presented in bold text in both the item and the passage.

Questionnaire. A paper-and-pencil questionnaire was developed to gather general background data as well as information about participants' computer use and reactions to taking a computer-based test.

Equipment

Eight desktop computers were used in this study. All were Compaq DeskPro models with 500MHz Celeron processors, 128MB RAM, and 7 GB hard drives. Each machine used a 17-inch color monitor (Compaq V700) with refresh rates set to 75 Hz. Each computer was linked to an internal ETS server via a 10-Mb Ethernet connection. The server collected performance data and, for half of the participants, artificially added latencies to item downloads to control the speed of the display. The computers were set up in an ETS lab with partitions separating each workstation.

The study design called for three display conditions: a 17-inch monitor set to a resolution of 1024 x 768, a 17-inch monitor set to a resolution of 640 x 480, and a 15-inch monitor set to a

resolution of 640 x 480. Each display was presented in one of two latency conditions. Half the participants in each display condition experienced no presentation latency, while the other half had a five-second presentation delay.

The 15-inch monitor condition used only the lower resolution because it was thought that sites using older, smaller monitors would probably not be running them at a higher resolution. Because the lab was outfitted only with 17-inch monitors, the smaller screen size condition was accomplished by manually configuring the displays to a screen height and width of a typical 15-inch monitor (8 inches by 10.5 inches). The font used for all the items was Arial and the C-3 style sheet specified a font size of "small." For the 1024 x 768 resolution, the font size was set to "large" in the control panel, and for the 640 x 480 resolution the font size was set to "small." In the browser, the font size was set to "smallest" in all conditions. These settings were selected to ensure that all the directions and options were visible on a single screen in both resolutions. This avoided the problem of having some items in which all the options did not appear on screen in the 640 x 480 condition but could be seen in 1024 x 768 (as shown in Figure 3).

Procedure

Data collection took place in the Rapid Prototyping Lab (RPL) at ETS's corporate office in Princeton, NJ. The study ran over a three-month period beginning in May 2000. In order to accommodate local high school schedules, there were two late afternoon administrations (4 p.m. and 6 p.m.) available during May and June. Once schools recessed for the summer, testing was expanded to five administrations daily (9 and 11 a.m. as well as 2, 4, and 6 p.m.). There were eight seats available per administration.

A proctor was hired to run the testing sessions. The proctor was given a reference manual and underwent a week of training before being left in charge of the lab. Included in the manual was a test administration checklist that was to be used at every testing session. Utilization of the checklist was intended to help guard against the proctor missing any steps and to ensure all subjects received the same testing conditions.

Upon arrival at ETS, study participants were asked for a copy of the invitation letter, signed by a parent or guardian authorizing their participation in the study. To assure groups that were essentially random, participants were assigned sequentially to one of six study conditions: 15-inch monitor with 640 x 480 resolution, 17-inch monitor with 1024 x 768 resolution, and 17-

inch monitor with 640 x 480 resolution, with each hardware condition presented either with no latency or with a five-second time delay. Once participants were settled at their workstations, the proctor read an informational greeting. This greeting instructed participants to carefully read and complete the study's consent form. It told them that they would be taking a test composed of two sections: a verbal section with 18 items and a math section with 15 items. Thirty minutes were allowed for each section, or about 1.5 minutes for each verbal item and 2.0 minutes for each mathematics item. These time limits were intended to be generous relative to the time limits on the paper-based test, which allows about 1.2 minutes per item for verbal items related to reading passages and about 1.2 minutes per item for mathematics items. The test was linear and review of previous answers was allowed.

Before beginning the test, participants were introduced to the testing screen. The information on the title bar was reviewed, including the question number and the timer showing remaining testing time for each section. The three interface tools used in the test (Back, Next, and Review) were also explained. All the information in this introduction was printed on a sheet of paper posted at each workstation and was available to the participants throughout the testing session. After completing the two test sections, participants completed the follow-up questionnaire.

Design

The experimental conditions are outlined in Table 1.

Table 1

Experimental Conditions

Display	Delay	No Delay
17-inch screen 1024 x 768 resolution	4 ability levels and 2 sexes	4 ability levels and 2 sexes
17-inch screen 640 x 480 resolution	4 ability levels and 2 sexes	4 ability levels and 2 sexes
15-inch screen 640 x 480 resolution	4 ability levels and 2 sexes	4 ability levels and 2 sexes

Within each area (verbal and mathematics), the design was 4 (ability levels) x 3 (screen size/resolution) x 2 (presentation latency) x 2 (sex). The four ability levels were defined by scores on the matching section of the PSAT/NMSQT. The first group represented scores from 20 to 45, the second group represented scores from 46 to 55, the third group represented scores from 56 to 65, and the last group represented scores from 66 to 80. The three levels of screen size/resolution were as follows: 15-inch monitor at 640 x 480, 17-inch monitor at 640 x 480, and 17-inch monitor at 1024 x 768. Presentation latency was either virtually instantaneous or with a five-second delay between completion of one question and presentation of the next. The same delay interval was used for reviewing previous questions.

Results

Scores on the experimental test and the PSAT/NMSQT were obtained for 357 examinees (175 males and 182 females). Although a broad range of scores was represented in the study participants, examinees with relatively high scores were somewhat more likely to volunteer. For the mathematics section of the PSAT/NMSQT, the mean was 57.6 with a standard deviation of 10.9; for the verbal section, the mean was 56.4 with a standard deviation of 10.6. For the mathematics section, there were 57 students in the lowest category (scores from 20 to 45) and 97 students in the highest category (scores from 66 to 80), with 94 and 109 students, respectively, in the two middle categories. Similarly, for the verbal scores, the number of students in each category (from lowest to highest) was 56, 111, 111, and 79. The correlation of the PSAT/NMSQT verbal score with the experimental verbal score was .71, and the correlation of the corresponding math scores was .75.

Effects on Math Scores

Means and standard deviations for the mathematics test are presented in Table 2. Results of the ANOVA are summarized in Table 3. Except for the expected effect of the PSAT/NMSQT scores, none of the main effects or interactions was statistically significant. Although not statistically significant ($F = 2.82$, $p = .09$), the difference in the scores for examinees in the two latency conditions is of some interest because the pattern of the means in all three screen size/resolution groups was in the opposite direction of what was expected: means were higher with the delay than without it.

Table 2

Math Score Means by Screen Size, Resolution, and Latency

Screen Size/Resolution		Latency	
		0 seconds	5 seconds
17” 1024 x 768	N	58	62
	M	10.21	10.63
	SD	2.59	2.76
17” 640 x 480	N	62	60
	M	9.79	10.65
	SD	2.96	2.56
15” 640 x 480	N	60	55
	M	9.72	10.36
	SD	2.88	2.72

Table 3

Analysis of Variance for Math Scores

Source	df	*F*
Latency	1	2.82
Screen Size/Resolution (Screen)	2	.26
PSAT/NMSQT-Math (PSAT-M)	3	84.86**
Sex	1	.07
Latency x Screen	2	.31
Latency x PSAT-M	3	.94
Screen x PSAT-M	6	.41
Latency x Screen x PSAT-M	6	1.05
Latency x Sex	1	.12
Screen x Sex	2	.21
Latency x Screen x Sex	2	.69
PSAT-M x Sex	3	1.20
Latency x PSAT-M x Sex	3	1.04
Screen x PSAT-M x Sex	6	1.14
Latency x Screen x PSAT-M x Sex	6	.18
Error	309	(3.78)

Note. Mean square error in parentheses.

**$p < .01$.

Effects on Verbal Scores

Means and standard deviations for the verbal test are presented in Table 4. ANOVA results are in Table 5. In addition to the expected PSAT/NMSQT effect, the screen size/resolution effect also was significant. A comparison of the high-resolution group with the average of the two low-resolution groups indicated a difference of 0.91 in the means (p=.01). The standard deviation of the verbal scores, across groups, was 3.56, and the within-group standard deviation was 2.56, so the 0.91 mean difference was about a quarter of a standard deviation (or a third of the within-group standard deviation). The contrast comparing the two

screen sizes in the low-resolution groups indicated a mean difference of 0.38, which was not significant (p=.31). Thus, screen resolution, but not screen size, had a significant impact on verbal scores. The lack of significant interactions with sex or PSAT/NMSQT scores suggests that the effects were reasonably constant for males and females and for students at different ability levels.

Table 4

Verbal Score Means by Screen Size, Resolution, and Latency

Screen Size/Resolution		Latency	
		0 seconds	5 seconds
17” 1024 x 768	N	58	62
	M	10.66	11.31
	SD	3.12	3.37
17” 640 x 480	N	62	60
	M	9.92	10.82
	SD	3.34	4.28
15” 640 x 480	N	60	55
	M	10.08	9.96
	SD	3.59	3.47

Table 5

Analysis of Variance for Verbal Scores

Source	df	*F*
Latency	1	.49
Screen Size/Resolution (Screen)	2	3.60*
PSAT/NMSQT-Verbal (PSAT-V)	3	72.61**
Sex	1	1.07
Latency x Screen	2	1.15
Latency x PSAT-V	3	.16
Screen x PSAT-V	6	1.57
Latency x Screen x PSAT-V	6	1.85
Latency x Sex	1	.77
Screen x Sex	2	.24
Latency x Screen x Sex	2	.99
PSAT-V x Sex	3	1.46
Latency x PSAT-V x Sex	3	1.10
Screen x PSAT-V x Sex	6	1.35
Latency x Screen x PSAT-V x Sex	6	.79
Error	309	(6.57)

Note. Mean square error in parentheses.

*$p < .05$. **$p < .01$.

Questionnaire Results

Computer Use. The majority of the students participating in this study were regular computer users. As shown in Table 6, 97% of them reported that they use a computer on a daily or weekly basis. Over 90% regularly use e-mail and the Internet. Of the four software programs listed on the questionnaire, most students (78%) used word processing software daily or weekly and 46% used computer games. Only about 10% of the students used instructional or test preparation software regularly, with about 60% reporting that they almost never used those two types of software packages.

Table 6

Computer Use

How often do you use:	Daily/Weekly	Monthly	Almost Never
Computers	97%	2%	1%
E-mail	90%	6%	4%
Internet	91%	7%	2%
Word processing software	78%	19%	3%
Computer games	46%	28%	25%
Instructional software	11%	28%	61%
Test preparation software	10%	26%	64%

Type of Computer. Eighty-one percent of respondents reported that they had used an IBM-compatible computer, 45% said they had used a Macintosh, and 26% reported using some other type of computer. (Percentages add to more than 100% because multiple responses were allowed.) Of these, 97% reported using Windows. It should be noted that the number of students who reported using some version of Windows was greater than the number who reported using IBM-compatible computers. This discrepancy is most likely due to the fact that some of the students using an IBM-compatible computer (such as a Compaq or Dell) reported that equipment as "other."

Expertise. Most of the teenagers in this study (87%) reported that they had been using a computer for more than three years, with 32% of the group having used a computer for more than seven years. Ninety percent of the students indicated that they were comfortable or very comfortable using the computer, and most rated their ability to use a computer as good (60%) or excellent (28%). Almost 47% reported that they had taken between one and two tests on a computer.

Taking This Computer-Based Test. The questionnaire contained three questions that asked students to compare this testing experience with taking an equivalent paper-and-pencil test. Forty-four percent said they preferred taking the test on the computer, 35% preferred paper and pencil, and 20% had no preference. When asked how well they thought they would have done on a paper-and-pencil test with the same questions, most (66%) felt they would have

performed about the same as on the computer-based test. Twenty percent thought they would have done better on the paper-and-pencil test, while 14% thought they would have done better on the computer-based test. The majority of the students (59%) believed taking the test on the computer was less tiring than taking an equivalent paper-and-pencil test, 30% thought it was about the same, and 11% thought it was more tiring.

Students were asked to evaluate the extent to which the following five features interfered with taking the computer-based test: font size, use of the mouse, screen clarity, screen size, and scrolling. As shown in Table 7, the only variable rated as interfering at all for a majority of the students was scrolling, with 66% reporting that it interfered to some degree (rating it either 2, 3, or 4). Less than a quarter of the students felt that font size, screen clarity, or screen size interfered with taking the test.

Table 7

Ratings of the Extent to Which Each of the Following Interfere With Taking the Test

Feature	1 (not at all)	2	3	4 (a great deal)
Use of the mouse	61%	23%	10%	6%
Font size	75%	13%	8%	4%
Screen clarity	76%	11%	7%	6%
Screen size	79%	8%	8%	4%
Scrolling	34%	34%	21%	11%

Because we wanted to assess whether students' evaluation of these features varied by testing condition, we looked at each potential interference by screen size, resolution, and latency. For the most part, students' ratings were not influenced by these variables in any systematic way. For example, students who took the test on a 15-inch display were just about as likely to say that the screen size did not interfere with taking the test as those using a 17-inch monitor (74% versus 81%, respectively). Similarly, resolution did not influence students' ratings of the effect of font size: seventy-nine percent of students using computers set to a resolution of 1024 x 768 and 72% of those using computers set to 640 x 480 did not feel that the font size interfered with their performance.

Students' ratings of the effect scrolling had on their test-taking experience did show some relationship to the hardware used. As shown in Table 8, only 6% of the students using the 17-inch screen set to a resolution of 1024 x 768 reported that scrolling interfered a great deal.

Table 8

Ratings of the Extent to Which Scrolling Interfered With Taking the Test

	1 (not at all)	2	3	4 (a great deal)
17" 1024 x 768	36%	33%	24%	6%
17" 640 x 480	38%	31%	23%	8%
15" 640 x 480	27%	38%	24%	18%

However, 8% of the students using a 17-inch monitor set to a resolution of 640 x 480 and 18% of the students using a 15-inch monitor set to a 640 x 480 setup said scrolling interfered a great deal with taking the test. This discrepancy might be expected because in the reading section of the test, using the equipment set to the 640 x 480 resolution increased the amount of scrolling required. For example, one of the shorter passages (270 words in length) required no scrolling at 1024 x 768 but needed to be scrolled for the remaining nine lines to be seen at a resolution of 640 x 480 (see Figure 2). The two longer passages (434 and 511 words in length) required some scrolling for the remaining two and 12 lines, respectively, to be seen in the higher resolution. In the lower resolution, those same passages had to be scrolled more than once for the remaining 32 and 45 lines to be viewed. Objectively, the amount of scrolling necessary was identical for the 640 x 480 resolutions in both monitor sizes, but was apparently perceived as more interfering with the smaller monitor size. Because the 17-inch monitor was set to emulate the window of a 15-inch monitor, students may have thought that the scrolling could have been easily avoided if we had only used the whole screen.

Open-Ended Questionnaire Responses

A total of 227 examinees (64%) provided comments in the open-ended section of the questionnaire where they were asked, "Is there anything else that you particularly liked or disliked about this computer-delivered test?" Of these, 45 students (20%) had positive

comments, 57 students (25%) shared concerns, and 34 students (15%) had both positive and negative comments. Although the students' comments were varied, the majority of them addressed the following issues:

- Presentation (e.g., one question per screen, always being able to see the reading passage, font size)
- Speed of the display
- Reading from the computer screen
- Aspects related to responding on the computer (e.g., not having to fill in bubbles, not being able to mark up a passage or cross out incorrect options, using the mouse)
- Interface features of the test (e.g., having an on-screen clock, having the review screen feature, having text references in bold type rather than line reference numbers)

The number of students commenting on any single aspect of the testing experience was small. As a result, the following comments are anecdotal in nature (and are reported in raw numbers rather than percentages where the percentages are less than 1% of the 227 students responding).

Scrolling. There were more negative than positive comments about scrolling (15 versus 2). The majority of those who shared negative comments (11 participants) used computers set to the 640 x 480 resolution, where, in fact, scrolling was more necessary.

Latency. One hundred eighty-nine students experienced a five-second delay between clicking on "Next" and moving to the next screen. Of those students, 34 (18%) had negative comments about the "annoying wait." More students using the 17-inch monitors had negative comments than those using the smaller screens (26 versus 8). There was little difference between the two resolution conditions (16 students using 1024 x 768 resolution had negative comments versus 18 using 640 x 480 resolution).

Font size. A few more students had negative comments about the font size than positive comments (7 versus 2). Of those making negative comments, most used the 17-inch monitors set to 640 x 480 resolution and commented that the font was too large.

Difficulty reading. Twenty-three students (10% of those commenting) had complaints about reading from the computer screen (e.g., "It hurt my eyes." and "The glare made it hard to

see."). Those making negative comments were evenly split across the screen and resolution conditions. Six students—all but one using the larger monitor—commented that they found the text easy to read.

Additional comments. Many of the comments made by students did not address issues related to screen size, resolution, or latency. Students were very positive about not having to fill in bubbles and erase completely to change a response (100 students, or 44% of those providing comments, liked responding on the computer). However, 109 students, or 48% of those writing in comments, did not like some aspect of responding on screen. Their comments centered around the inability to write on the test, mark up a passage, cross out options they knew were incorrect, and so forth. As one of these students said, "The advantage of using paper is that you can 'interact' with the test."

Many students (70, or 31% of those responding) liked the interface features, such as the highlighted text references (used in place of line references), the on-screen clock that helped them keep track of time, and the review feature that helped them avoid inadvertently skipping questions. Most of the students who did not like the review feature were in the latency condition, where moving back and forth from the review screen to items entailed repeated five-second delays.

Discussion

The results suggest that differences in screen resolution in the range that might realistically be expected in a Web-based testing program can have an impact on verbal test scores. Although the effect size of just a quarter of a standard deviation is small by conventional standards (Cohen, 1988), it would not be considered as trivial in many testing situations. On the SAT I scale, for example, a quarter of a standard deviation is 28 points.

A major difference between the high- and low-resolution presentations was the substantially greater need to scroll through reading passages in the low-resolution condition. Furthermore, scrolling was identified by the majority of the participants as something that interfered with taking the test. In the mathematics test, little scrolling was required in any of the conditions, and no significant score differences were found. Although we cannot be certain that scrolling caused the score differences, a prudent approach to future Web delivery would be to ensure comparable scrolling regardless of the type of monitor being used at a testing site. For any

monitor configuration, the font size should meet a minimum standard, and the amount of text visible at a given time should be as constant as possible across monitors. Because font sizes can be set in a number of different places, it might be useful to have a test page so that each testing center could verify that the amount of text displayed on one screen met the specified standard.

A number of cautions should be heeded in interpreting the results and planning additional studies. First, effects could be different in a high-stakes test. Although the participants in this study seemed to be taking the test seriously, they also knew that their scores would not be used for any admissions decision. Frustrations with scrolling might reduce motivation and effort in a low-stakes test but have less of an impact in a higher-stakes test. On the other hand, difficulties with scrolling could raise anxiety levels and therefore have more of an effect on a high-stakes test. Second, effects could be different with stricter time limits. Because scrolling to search for answers takes time, larger effects of screen resolution might be found in a highly speeded test. Finally, other segments of the population could respond differently than the college-bound high school juniors in the current study. It is easy to imagine, though by no means certain, that effects might be greater for examinees with more limited computer experience or literacy skills. Although many issues remain unresolved, the current study underlines the potential importance of variations in screen displays in a Web-based testing environment.

References

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

College Entrance Examination Board. (1986). *10 SATs.* New York: Author.

deBruijn, D., de Mul, S., & Van Oostendorp, H. (1992). The influence of screen size and text layout on the study of text. *Behaviour and Information Technology*, *11* (2), 71–78.

Dillon, A., Richardson, J., & McKnight, C. (1990). The effects of display size and text splitting on reading lengthy text from screen. *Behaviour and Information Technology*, *9* (3), 215–227.

Haas, C. (1996). *Writing Technology: Studies on the materiality of literacy.* Mahwah, N.J.: Lawrence Erlbaum Associates.

Haas, C., & Hayes, J. R. (1985). *Effects of text display variables on reading tasks: Computer screen vs. hard copy*. (Carnegie-Mellon University Rep. No. CDC-TR3). Pittsburgh, Pa.: Carnegie-Mellon University.

Appendix

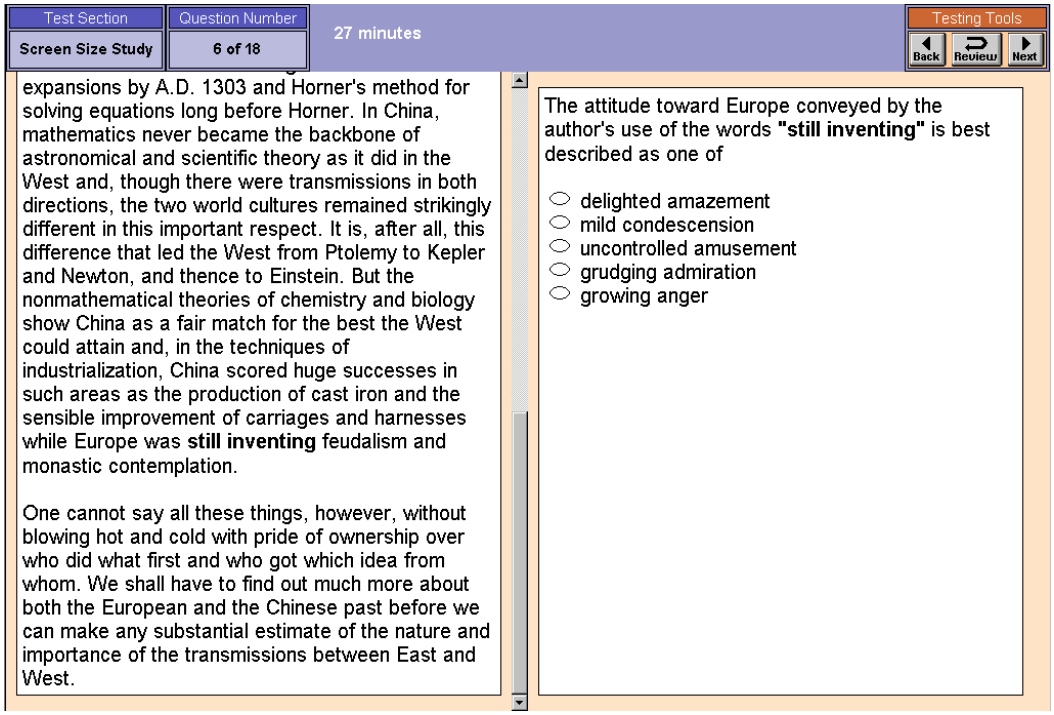
The following illustrate revisions to items containing line references.

- Example of a word, phrase, or sentence reference

  The original item wording was as follows:

  > The attitude toward Europe conveyed by the author's use of the words "still inventing" (line 45) is best described as one of . . .

  In the computer version, the line reference was removed and the phrase "still inventing" was presented in bold text in both the item and the passage, as shown below. (The bold text was more distinct on screen than it appears in this printout.)



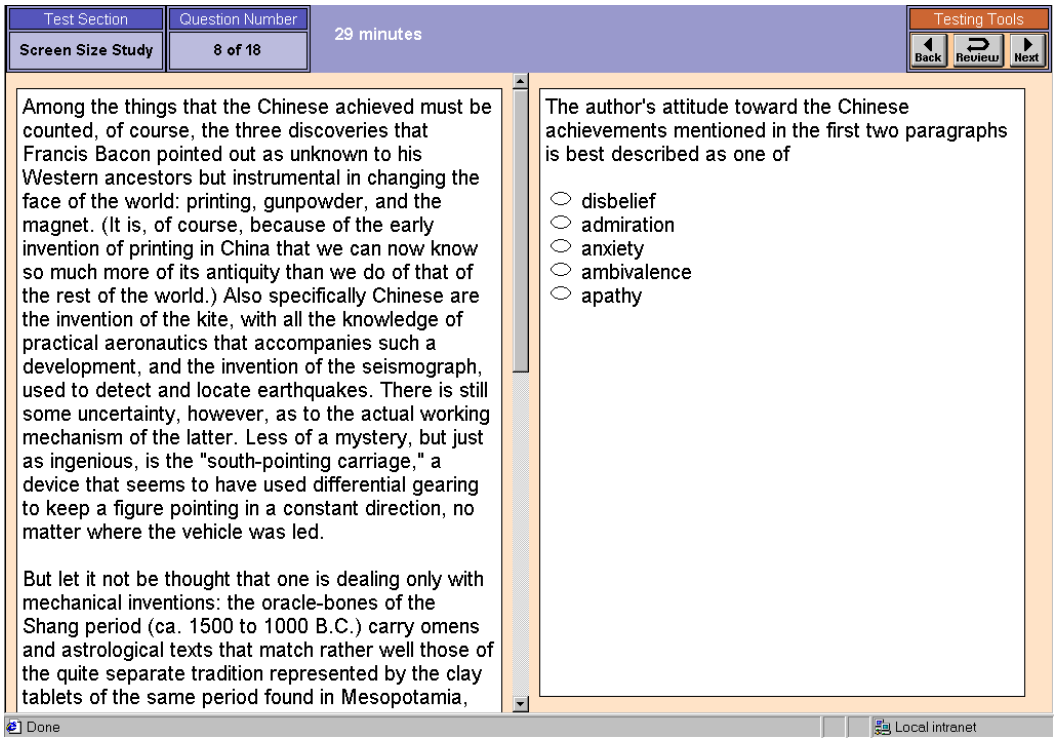
- Changing a line reference to a paragraph reference

  The original wording was as follows:

  > The author's attitude toward the Chinese achievements mentioned in lines 1-45 is best described as one of . . .

  Because lines 1-45 corresponded to the first two paragraphs of the passage, in the computer version the wording was revised to read:

  > The author's attitude toward the Chinese achievements mentioned in the first two paragraphs is best described as one of . . .



- Rewriting a stem

  The original wording was:

  > Which of the following is the best interpretation of lines 23-24?

  The new stem included the statement made in lines 23-24. It asked:

  > Which of the following is the best interpretation of this statement by the author?
  >
  > "We do rightly, of course, to judge people by their reach as well as their grasp."

  The statement was presented in bold text in both the item and the passage, as shown below.